Proceedings of
The Second Workshop on Annotation and Exploitation of Parallel Corpora

associated with
The 8th International Conference on
Recent Advances in Natural Language Processing
(RANLP 2011)

15 September, 2011
Hissar, Bulgaria
INTERNATIONAL WORKSHOP
ANNOTATION AND EXPLOITATION OF PARALLEL CORPORA

PROCEEDINGS

Hissar, Bulgaria
15 September 2011
Designed and Printed by INCOMA Ltd.
Shoumen, BULGARIA
Preface

The RANLP 2011 Workshop on *Information Extraction and Knowledge Acquisition* (IEKA 2011) took place on September 16, 2011 in Hissar, Bulgaria, following the conference on Recent Advances in Natural Language Processing (RANLP 2011).

The workshop was envisioned as a meeting place for those concerned with the fields of information extractions (IE) and knowledge acquisition (KA). Until a decade ago, these fields were mostly limited to identifying and extracting named entities, semantic and ontological relations, events, templates, and facts in relatively small text corpora using a small variety of external resources such as gazetteers, thesauri, and lexical hierarchies.

Today everything has changed. The size of corpora has grown dramatically: using Gigaword-scale data is common, and it is almost standard to use the Web, which contains quadrillions of words, or at least the Google Web 1T 5-grams. More importantly, new types of communication have emerged, such as chats, blogs and, in the last 2-3 years, Twitter, whose informal language poses many challenges to automatic IE and KA, yet they are becoming increasingly important, e.g., for learning customer opinions on various products and services. Social network analysis is another emerging topic, where data is naturally much more interconnected than in the rest of the Web.

All these recent developments have posed not only new challenges, but have also created a number of opportunities, opening new research directions, and offering new useful resources. For example, the growth of Wikipedia has given rise to DBpedia and other collaboratively-created resources such as Freebase. Today, IE and KA researchers can even create annotations and resources on demand as they need them for a very low price using crowd-sourcing tools such as Amazon’s Mechanical Turk.

We received 12 submissions, and, given our limited capacity as a one-day workshop, we were only able to accept seven full papers for oral presentation: an acceptance rate of 58%. The workshop also featured two invited talks (by Ralph Grishman and by Ralf Steinberger), and a panel discussion (involving Kevin Cohen, Kiril Simov, Petya Osenova, and Georgi Georgiev).

We would like to thank the members of the Program Committee for their timely reviews. We would also like to thank the authors for their valuable contributions.

*Preslav Nakov, Zornitsa Kozareva, Kuzman Ganchev, Jerry Hobbs
Co-Organizers*
Organizers:

Preslav Nakov, National University of Singapore
Zornitsa Kozareva, University of Southern California
Kuzman Ganchev, Google Inc.
Jerry Hobbs, University of Southern California

Program Committee:

Javier Artiles, Queens College
Kalina Bontcheva, University of Sheffield
Razvan Bunescu, Ohio State University
Hamish Cunningham, University of Sheffield
Mona Diab, Columbia University
Greg Druck, University of Massachusetts Amherst
Georgi Georgiev, Ontotext AD
Filip Ginter, University of Turku
Jennifer Gillenwater, University of Pennsylvania
Andrew Gordon, USC/ICT
João Graça, University of Pennsylvania
Chikara Hashimoto, NICT
Nitin Indurkhya, eBay
Hagen Fürstenau, Saarland University
Alex Kulesza, University of Pennsylvania
Maussam, University of Washington
Su Nam Kim, The University of Melbourne
Alexandre Klementiev, Johns Hopkins University
Ana-Maria Popescu, Yahoo!
Ryan McDonald, Google
Andrés Montoyo, University of Alicante
Rafael Muñoz, University of Alicante
Roberto Navigli, University of Rome
Manuel Palomar, University of Alicante
Siddharth Patwardhan, IBM
Marco Pennacchiotti, Yahoo!
Simone Paolo Ponzetto, University of Heidelberg
Sampo Pyysalo, University of Tokyo
Sujith Ravi, ISI, University of Southern California
Ellen Riloff, University of Utah
Alan Ritter, University of Washington
Paolo Rosso, Technical University of Valencia
Kenji Sagae, USC/ICT
Thamar Solorio, University of Alabama at Birmingham
Ralf Steinberger, JRC
Mihai Surdeanu, Stanford University
Idan Szpektor, Yahoo! Research
Valentin Tablan, University of Sheffield
Jörg Tiedemann, University of Uppsala
Ivan Titov, Saarland University
Aline Villavicencio, Federal University of Rio Grande do Sul
Alexander Yates, Temple University
Additional Reviewers:

Gabor Angeli

Invited Speakers:

Ralph Grishman, New York University
Ralf Steinberger, European Commission Joint Research Centre

Panelists:

Kevin Cohen, University of Colorado, School of Medicine
Georgi Georgiev, Ontotext AD
Petya Osenova, Bulgarian Academy of Sciences
Kiril Simov, Bulgarian Academy of Sciences
## Table of Contents

**INVITED TALK 1: The Knowledge Base Population Task: Challenges for Information Extraction**  
Ralph Grishman ................................................................. 1

**Fine-grained Entity Set Refinement with User Feedback**  
Bonan Min and Ralph Grishman ........................................... 2

**Extraction of Domain-specific Opinion Words for Similar Domains**  
Ilia Chetviorkin and Natalia Loukachevitch ............................. 7

**The Role of Predicates in Opinion Holder Extraction**  
Michael Wiegand and Dietrich Klakow .................................. 13

**Dependency-Based Text Compression for Semantic Relation Extraction**  
Marcos Garcia and Pablo Gamallo ........................................ 21

**How to Distinguish a Kidney Theft from a Death Car? Experiments in Clustering Urban-Legend Texts**  
Roman Grundkiewicz and Filip Gralinski ................................ 29

**Machine Reading Between the Lines: A Simple Evaluation Framework for Extracted Knowledge Bases**  
Avirup Sil and Alexander Yates ............................................ 37

**Temporal Expressions Extraction in SMS messages**  
Stéphanie Weiser, Louis-Amélie Cougnon and Patrick Watrin ....... 41

**INVITED TALK 2: Bringing Multilingual Information Extraction to the User**  
Ralf Steinberger .............................................................. 45
Conference Program

Friday: September 16, 2011

09:00–09:05  Welcome

Session I

09:05–10:05  Invited Talk 1
The Knowledge Base Population Task: Challenges for Information Extraction
Ralph Grishman, New York University

10:05–10:30  Fine-grained Entity Set Refinement with User Feedback
Bonan Min and Ralph Grishman

10:30–11:00  BREAK

Session II

11:00–11:25  Extraction of Domain-specific Opinion Words for Similar Domains
Ilia Chetviorkin and Natalia Loukachevitch

11:25–11:50  The Role of Predicates in Opinion Holder Extraction
Michael Wiegand and Dietrich Klakow

11:50–12:15  Dependency-Based Text Compression for Semantic Relation Extraction
Marcos Garcia and Pablo Gamallo

12:15–14:15  LUNCH

Session III

14:15–14:40  How to Distinguish a Kidney Theft from a Death Car?
Experiments in Clustering Urban-Legend Texts
Roman Grundkiewicz and Filip Graliński

14:40–15:05  Machine Reading Between the Lines:
A Simple Evaluation Framework for Extracted Knowledge Bases
Avirup Sil and Alexander Yates

15:05–15:30  Temporal Expressions Extraction in SMS messages
Stéphanie Weiser, Louis-Amélie Cougnon and Patrick Watrin

15:30–16:00  BREAK
Friday: September 16, 2011 (continued)

Session IV

16:00–17:00  **Invited Talk 2**
*Bringing Multilingual Information Extraction to the User*
Ralf Steinberger, European Commission Joint Research Centre

17:00–18:00  **Panel**
*Moderator:* Preslav Nakov
*Panelists:*
- Kevin Cohen, University of Colorado, School of Medicine
- Georgi Georgiev, Ontotext AD
- Petya Osenova, Bulgarian Academy of Sciences
- Kiril Simov, Bulgarian Academy of Sciences

18:00–18:05  **Closing Remarks**
Abstract

The Knowledge Base Population (KBP) task, being run for the past 3 years by the U.S. National Institute of Standards and Technology, is the latest in a series of multi-site evaluations of information extraction, following in the tradition of MUC and ACE. We examine the structure of KBP, emphasizing the basic shift from sentence-by-sentence and document-by-document evaluation to corpus-based extraction and the challenges it raises for cross-sentence and cross-document processing. We consider the problems raised by the limited amount and incompleteness of the training data, and how this has been (partly) addressed through such methods as semi-supervised learning and distant supervision. We describe some of the optional tasks which have been included – rapid task adaptation (last year), temporal analysis (this year), cross-lingual extraction (planned for next year) – and others which have been suggested.
Fine-grained Entity Set Refinement with User Feedback

Bonan Min
New York University
715 Broadway, 7th floor
New York, NY 10003 USA
min@cs.nyu.edu

Ralph Grishman
New York University
715 Broadway, 7th floor
New York, NY 10003 USA
grishman@cs.nyu.edu

Abstract
State of the art semi-supervised entity set expansion algorithms produce noisy results, which need to be refined manually. Sets expanded for intended fine-grained concepts are especially noisy because these concepts are not well represented by the limited number of seeds. Such sets are usually incorrectly expanded to contain elements of a more general concept. We show that fine-grained control is necessary for refining such sets and propose an algorithm which uses both positive and negative user feedback for iterative refinement. Experimental results show that it improves the quality of fine-grained sets significantly.

1 Introduction
Entity set expansion is a well-studied problem with several techniques proposed (Bunescu and Mooney 2004, Etzioni et al. 2005, Wang and Cohen 2007, Sarmento et al. 2007, Pasca 2007, Pasca 2004, Pantel et al. 2009, Pantel and Lin 2002, Vickrey et al. 2010). In practice, semi-supervised methods are preferred since they require only a handful of seeds and are more flexible for growing various types of entity sets. However, they usually produce noisy sets, which need to be refined (Vyas and Pantel, 2009). Fine-grained sets such as National Capitals are particularly noisy. Such concepts are intrinsically hard because they’re not well represented by initial seeds. Moreover, most related instances have a limited number of features, thus making it hard to retrieve them.

We examined a few sets expanded for fine-grained concepts and observed that lots of erroneous expansions are elements of a more general concept, whose sense overlaps and subsumes the intended sense. For example, the concept National Capitals is expanded to contain Major cities. In such cases, a proposed feature-pruning technique using user-tagged expansion errors to refine sets (Vyas and Pantel 2009) removes some informative features of the target concept. Moreover, since refining such sets needs more information about the target concept, it is natural to use user-tagged correct expansions as well for the refinement.

In this paper, we refer to the problem of fine-grained concepts being erroneously extended as semantic spread. We show that a rich feature representation of the target concept, coupled with appropriate weighting of features, is necessary for reducing semantic spread when refining fine-grained sets. We propose an algorithm using relevance feedback, including both positive and negative user feedback, for set refinement. By expanding the set of features and weighting them appropriately, our algorithm is able to retrieve more related instances and provide better ranking. Experimental results show that it improves the quality of fine-grained sets significantly.

2 Related work
There is a large body of research on growing named entity sets from a handful of seeds. Some are pattern-based algorithms. Sarmento et al. (2007) uses explicit patterns, e.g. “...NEa NEb and NEc...”, to find named entities of the same class. Pasca (2004) uses the pattern <[StartOfSent] X [such as|including] N [and|,|.]> (Hearst 1992) to find instances and their class labels from web logs. Some are based on distributional similarity. The distributional hypothesis states that similar terms tend to appear with similar contexts (Harris 1954). For example, Pasca (2007) extracts templates (prefixes and suffixes around seeds) from search engine query logs as features, and then ranks new instances by their similarity with the seeds in the vector space of pattern features for growing sets. Their method
outperforms methods based on handcrafted patterns (Pasca 2004) but requires extensive query logs to tolerate noisy queries. Calculating the similarity matrix between all pairs of named entities is expensive. Pantel et al. (2009) proposed a web-scale parallel implementation on the MapReduce distributed computing framework.

Observing the low quality of expanded sets, Vyas and Pantel (2009) uses negative user feedback for set refinement. They propose the Similarity Method (SIM) and Feature Modification Method (FMM), to refine entity sets by removing expansions which are similar to user-tagged errors, and removing features related to the erroneous sense from the centroid of the seed set for better ranking, respectively. Their algorithms rely on two assumptions 1) most expansion errors are caused by ambiguous seeds, and 2) entities which are similar in one sense are usually not similar in their other senses. They show average performance gain over a few sets. Vyas et al. (2009) studied the problem from the other side by selecting better seeds. They proposed three metrics and three corresponding algorithms to guide editors to choose better seeds. All three algorithms outperform the baseline.

3 Similarity modeling revisited

Given a set of candidate named entities represented by vectors of features, the goal of set refinement is to find a subset of entities which are similar to the target concept, based on a certain similarity metric (Cosine, Dice, etc). The concept is usually approximated with a set of seed instances. A previous feature pruning technique (Vyas and Pantel 2009) aims at reducing semantic drift introduced by ambiguous seeds.

We’re particularly interested in fine-grained classes since they’re intrinsically hard to expand because of the crude representation from the limited number of seeds. In practice, we observed, when expanding fine-grained classes, that semantic spread instead of semantic drift (McIntosh 2010) severely affects expansion quality. By semantic spread we mean a situation where an initial concept, represented by its member entities, changes in the course of entity set expansion into a broader concept which subsumes the original concept.

Semantic spread is usually introduced when erroneous instances, which belong to a more general concept, are incorrectly included during the set expansion process. For example, when using Google Sets (labs.google.com/sets) to expand National Capitals, we found a highly ranked error New York. By checking with our distributional thesaurus extracted from 37 years’ newspaper, we notice the following features: prep_in(embassy] *) 1, nn(*, capital), nn (*, president). These are indicators of capital cities. However, as the financial “capital” and a politically important city, New York shares lots of informative features with the National Capitals concept. Therefore, we need more sophisticated techniques for the refinement process for fine-grained concepts.

4 Refine fine-grained classes with user feedback

User feedback is a valuable resource for learning the target concept. We propose to use both positive and negative feedback to learn a rich set of features for the target concept while weighting them appropriately. Our algorithm chooses informative instances to query the user, uses positive feedback for expanding the feature set, and negative feedback for feature weight adjustment.

Relevance feedback (Harman 1992) is widely applied to improve search engine performance by modifying queries based on user feedback. Various techniques are proposed for both the vector space model and probabilistic model. Since set refinement is done in the vector space of features, we only consider techniques for the vector space model. To refine entity sets, the centroid of all vectors of seeds is used as a query for retrieving related named entities from the candidate pool. Observing that errors are usually caused by incorrect or overweighted features of seeds, we propose to incorporate user feedback for set refinement with a variant of the Rocchio algorithm (Rocchio 1971). The new centroid is calculated as follows:

$$Centroid = \frac{\sum_{I \in S \cup P} I - \sum_{C_N \in N} C_N}{|S \cup P| - \gamma \cdot \frac{|N|}{|S|}}$$

where I is an entity that is a member of seed set S or the set of user-tagged positive entities P, and C_N is a member of the set of user-tagged negative entities N. γ is the parameter penalizing features of irrelevant entities. This method does feature set expansion and iterative adjustment of feature weights for the centroid. It adds features from informative instances back into the centroid.

1 Syntactic context is used in our experiment. For the format of dependencies, please refer to the Stanford typed dependencies manual.
and penalizes inaccurate features based on user-tagged errors, thus modifying the centroid to be a better representation of the target class.

4.1 Query strategy

To be practical, we should ask the user to review as few instances as possible, while obtaining as much information as possible. Observing that 1) top-ranked instances are likely to be positive 2) random instances of a fine-grained class usually contain relatively few features with non-zero weight, thus not providing much information for approaching the target concept, our procedure selects at each iteration the \( n \) instances most similar to the centroid and presents them to the user in descending order of their number of features with non-zero weight (the user will review higher-dimension ones first). This ranking strategy prefers more representative instances with more features (Shen et al., 2004). The user is asked to pick the first positive instance.

A similar idea applies to negative instance finding. We use co-testing (Muslea et al., 2006) to construct two ranking-based classifiers on randomly split views of the feature space. Instances are ranked by their similarity to the centroid. The classifiers classify instances which ranked higher than the golden set size as correct, and classify others as incorrect. We select \( n \) contention instances – instances identified as correct expansions by one of the classifiers and incorrect by the other. These instances are more ambiguous and likely to be negative. Instances are also presented to the user in descending order of number of features with non-zero weight. Coupled with the strategy for positive instance finding, it helps to reweight a rich set of features.

Since we asked the user to review instances that are most likely to be positive and negative, and these instances are presented to the user in sequence, the user only has to review very few examples to find a positive and a negative instance in each iteration. In practice we set \( n=10 \). We observed that around 85% of the time the user only has to review 1 instance to find a correct one, and over 90% of the time has to review 3 or fewer instances to find a negative one.

5 Experiment

Corpus: we used 37 years newspaper corpus\(^2\) which is dependency parsed with the Stanford Parser\(^3\) and has all named entities tagged with Jet\(^4\) NE tagger (we didn’t use the NE tags reported by the tagger but only the fact that it is a name). We use syntactic context, which is the grammatical relation in conjunction with the words as feature, and we replace the word in the candidate NE with *. Both syntactic contexts in which the candidate entities are the heads and contexts in which the candidate entities are the dependents are used. The feature set is created from syntactic contexts of all entities tagged in the corpus. An example common feature for class National Capital is prep_in(ministry, *). We remove features in which the dependent is a stop word, and remove a limited number of less useful dependency types such as numerical modifier and determiner. We use pointwise mutual information (PMI) to weight features for entities, and cosine as the similarity measure between the centroid of the seeds and candidate instances. PMI scores are generated from the newspaper corpus statistics. Candidates are then ranked by similarity. We construct each named entity candidate pool by including similar instances with cosine score greater than 0.05 with the centroid of the corresponding golden set. This ensures that each candidate pool contains tens of thousands of elements so that it contains all similar instances with high probability.

Golden sets\(^5\): Several golden sets are prepared by hand. We start from lists from Wikipedia, and then manually refine the sets\(^6\) by removing incorrect instances and adding correct instances found as distributionally-similar instances from the corpus. The criteria for choosing the lists is 1) our corpus covers most elements of the list, 2) the list represents a fine-grained concept, 3) it contains hundreds of elements for reasons of fairness, since we don’t want the added positive examples themselves to overshadow other aspects of the evaluated algorithms. Based on these criteria, we chose three lists: National Capitals, IT companies\(^5\) and New York City (NYC) neighborhoods. All three sets have more than 200 elements. User feedback is simulated by checking membership in the golden set. Since existing

\(^2\) It contains news articles from: TDT5, NYT(94-00), APW(98-00), XINHUA(96-00), WSI(94-96), LATWP(94-97), REUFF(94-96), REUTE(94-96), and WSJSF(87-94). It contains roughly 65 million sentences and 1.3 billion tokens.

\(^3\) http://nlp.stanford.edu/software/lex-parser.shtml

\(^4\) http://cs.nyu.edu/grishman/jet/license.html

\(^5\) Golden sets are available for download at http://www.cs.nyu.edu/~min/goldset_37news.tgz

\(^6\) Manually checking indicates the golden sets are complete with high probability.

\(^7\) Set contains both software and hardware companies
golden sets such as the sets from Vyas and Pantel (2009) are not designed specifically for evaluating refinement on fine-grained concepts and they are quite small for evaluating positive feedback (with less than 70 elements after removing low frequency ones in our corpus), we decided to construct our own.

**Algorithms evaluated:** The following algorithms are applied for iteratively updating the centroid using user-tagged examples: 1) **baseline algorithm (BS),** an algorithm adding the correct example most similar to the centroid as a new seed for each iteration; this simulates using the user-tagged first positive example to assist refinement, 2) **RF-P,** relevance feedback algorithm using only positive feedback by adding one informative instance (selected using the method described in section 4.1) into seed set, 3) **FMM** (Vyas and Pantel, 2009) which uses the first user-tagged negative example for feature pruning in each iteration. 4) **RF-N,** relevance feedback algorithm using only negative feedback (selected using the method described in section 4.1), 5) Relevance feedback (**RF-all**) using both positive and negative user feedback selected using methods from Section 4.1. We use 6 seeds for all experiments, and set $\gamma=0.25$ for all RF experiments.

For each algorithm, we evaluate the results after each iteration as follows: we calculate a centroid feature vector and then rank all candidates based on their similarity to the centroid. We add sufficient top-ranked candidates to the seed and user-tagged positive items to form a set equal in size to the golden set. This set, the refined set, is then compared to the golden set. The following tables show a commonly reported metric, average R-precision$^8$ of 40 runs starting with randomly picked initial seeds (The first column shows the number of iterations):}

<table>
<thead>
<tr>
<th></th>
<th>BS</th>
<th>RF-P</th>
<th>FMM</th>
<th>RF-N</th>
<th>RF-all</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.258</td>
<td>0.298</td>
<td>0.253</td>
<td>0.246</td>
<td>0.286</td>
</tr>
<tr>
<td>4</td>
<td>0.260</td>
<td>0.317</td>
<td>0.250</td>
<td>0.251</td>
<td>0.316</td>
</tr>
<tr>
<td>6</td>
<td>0.260</td>
<td>0.323</td>
<td>0.244</td>
<td>0.255</td>
<td>0.332</td>
</tr>
<tr>
<td>8</td>
<td>0.260</td>
<td>0.325</td>
<td>0.243</td>
<td>0.255</td>
<td>0.342</td>
</tr>
<tr>
<td>10</td>
<td>0.265</td>
<td>0.325</td>
<td>0.245</td>
<td>0.256</td>
<td>0.343</td>
</tr>
</tbody>
</table>

**Table 1. Performance on class national capitals**

<table>
<thead>
<tr>
<th></th>
<th>BS</th>
<th>RF-P</th>
<th>FMM</th>
<th>RF-N</th>
<th>RF-all</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.303</td>
<td>0.340</td>
<td>0.301</td>
<td>0.303</td>
<td>0.319</td>
</tr>
<tr>
<td>4</td>
<td>0.312</td>
<td>0.403</td>
<td>0.303</td>
<td>0.311</td>
<td>0.406</td>
</tr>
</tbody>
</table>

**Table 2. Performance on class IT companies**

<table>
<thead>
<tr>
<th></th>
<th>BS</th>
<th>RF-P</th>
<th>FMM</th>
<th>RF-N</th>
<th>RF-all</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.168</td>
<td>0.190</td>
<td>0.159</td>
<td>0.161</td>
<td>0.191</td>
</tr>
<tr>
<td>4</td>
<td>0.179</td>
<td>0.217</td>
<td>0.164</td>
<td>0.163</td>
<td>0.232</td>
</tr>
<tr>
<td>6</td>
<td>0.189</td>
<td>0.235</td>
<td>0.163</td>
<td>0.166</td>
<td>0.249</td>
</tr>
<tr>
<td>8</td>
<td>0.198</td>
<td>0.243</td>
<td>0.166</td>
<td>0.176</td>
<td>0.259</td>
</tr>
<tr>
<td>10</td>
<td>0.206</td>
<td>0.248</td>
<td>0.169</td>
<td>0.181</td>
<td>0.262</td>
</tr>
</tbody>
</table>

Results show that RF-P outperforms the baseline algorithm by using positive examples with rich contexts rather than the first positive example for each iteration. The baseline algorithm shows small improvement over 10 iterations. This shows that simply adding the example which is most similar to the centroid is not very helpful. Comparing R-precision gain between RF-P and the baseline suggests that selecting informative examples is critical for refining fine-grained sets. By enriching the feature set of the centroid, RF-P is able to retrieve instances with a limited number of features overlapping the original centroid. RF-N outperforms FMM since it only reweights (penalizes some weights) but doesn’t prune out intersection features between user-tagged errors and the centroid. This flexibility avoids over-penalizing weak but informative features of the intended concept. For FMM, we observe a small performance gain with successive iterations over IT companies and NYC neighborhoods but a performance decrease for National Capitals. Inspection of results shows that FMM tends to retrieve more capital cities for small geographical regions because of removal of weak features for informative sense such as Major Cities.

Combining RF-P and RF-N, RF-all uses both positive informative examples and negative informative examples to expand feature sets of the centroid and weight them appropriately, thus achieving the most performance gain. RF-N by itself doesn’t improve performance significantly. Comparing RF-all with RF-P, using informative negative examples helps to improve performance substantially because only when both informative positive examples and informative negative examples are used can we learn a significantly large set of features and appropriate weights for them.
We also implemented a few methods combining positive feedback and FMM, and didn’t observe encouraging performance. RF-all also has the highest Average Precision (AP) for all sets, thus showing that it provides better ranking over candidates. Due to space limitations, tables of AP are not included. The quality of the top ranked elements with RF-all can be seen in the precision at rank 50 for the three sets: 84.6%, 81.6%, and 71.7%.

6 Conclusion and Future work

We propose an algorithm using both positive and negative user feedback to reduce semantic spread for fine-grained entity set refinement. Our experimental results show performance improvement over baseline and existing solutions.

Our next step is to investigate feature clustering techniques since we observe that data sparseness severely affects set refinement.

Acknowledgments

We are grateful to the anonymous reviewers for their valuable comments. We would like to also thank Prof. Satoshi Sekine and Ang Sun for their helpful discussion and comments on an early draft of this paper.

References


Extraction of Domain-specific Opinion Words for Similar Domains

Ilia Chetviorkin
Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University
ilia2010@yandex.ru

Natalia Loukachevitch
Research Computing Center Lomonosov Moscow State University
louk_nat@mail.ru

Abstract

In this paper we consider a new approach for domain-specific opinion word extraction in Russian. We suppose that some domains have similar sentiment lexicons and utilize this fact to build an opinion word vocabulary for a group of domains. We train our model in movie domain and then utilize it to book and game domains. Obtained word list quality is comparable with quality of initial domain list.

1 Introduction

The web is full of customers’ opinions on various products. Automatic extraction, processing and summarization of such opinions are very useful for future users. Opinions about products are often expressed using evaluative words and phrases that have a certain positive or negative sentiment. Therefore, important features in the qualitative classification of opinions about a particular entity are opinion words and expressions used in the domain. The problem is that it is impossible to compile a list of opinion expressions, which will be equally applicable to all domains, as some opinion phrases are used only in a specific domain while the others are context-oriented [Lu et. al., 2011]. Indeed, sentiment lexicons adapted to a particular domain or topic have been shown to improve task performance in a number of applications, including opinion retrieval [Jijkoun et. al., 2010], and expression-level sentiment classification [Choi and Cardie, 2009]. In addition there are several studies about context-dependent opinion expressions [Lu et. al., 2011].

The number of different domains is very large, and recent studies are focused on cross-domain approaches, to bridge the gap between the domains [Pan et al, 2010]. On the other side there are different subject fields that has similar sentiment lexicon. For example: «breathtaking» is an opinion word in entertainment (movies, books, games etc.) domain, but non-opinion in the politics domain. At the opposite side some words («evil», «treachery» etc.) have strong sentiment in politics domain, but are neutral in entertainment domain, these words do not express any opinion about a film, game or book.

Thus we suppose that different domains can be separated into clusters (for example: entertainment, digital goods, politics, traveling etc.) where domains of the same cluster have similar sentiment lexicons.

In this paper we focus on the problem of construction of a domain-specific sentiment lexicon in Russian, which can be utilized for various similar domains.

We present a new supervised method for domain-specific opinion word extraction. We train this method in one domain and then utilize it in two others. Then we combine extracted word lists to construct a general list of opinion words typical to this domain cluster.

Our approach is based on several text collections, which can be automatically formed for many subject areas. The set of text collections includes: a collection of product reviews with author evaluation scores, a text collection of product descriptions and a contrast corpus (for example, a general news collection). For each word in a review collection we calculate various statistical features using aforementioned collections and then apply machine learning algorithms for term classification.

To evaluate the effectiveness of the proposed method we conduct experiments on data sets in three different domains: movies, books and computer games. The results show that our approach can identify new opinion words specific to the given domain (for example “fabricated” in movie domain).

For further evaluation of the lexicon quality, we manually labeled extracted word lists, and our method is proved to be effective in construct-
ing a qualitative list of domain-dependent sentiment lexicon. The results also demonstrate the advantage of combining multiple lists of opinion words over using any single list.

The reminder of this article is organized as follows. In Section 2 we describe the state-of-the-art in the opinion words extraction sphere, Section 3 describes our approach in the movie domain, in Section 4 we utilize our approach for two other domains and combine opinion word vocabularies for all three domains.

2 Related Work

Sentiment lexicon plays an important role in most, if not all, sentiment analysis applications, including opinion retrieval, opinion question answering and summarization, opinion mining [Ding et. al., 2008]. Even though supervised machine learning techniques have been shown to be effective for sentiment classification task [Pang and Lee, 2008], authors in [Choi and Cardie, 2009] demonstrate that including features from sentiment lexicons boosts classification performance significantly.

Generally there are three main approaches to the automatic identification of opinion words in texts.

The first approach is manual labeling, which is very labor-intensive and error-prone process. In addition the coverage of this approach is usually very low.

The second approach is based on information from a dictionary or a thesaurus. In this approach a small initial set of words is usually chosen manually, and then expanded with the help of dictionaries and thesaurus entries. The basic principle of this approach is that if a word has sentiment polarity, then its synonyms and antonyms have polarity too (orientation may change). Therefore, from the initial set of words, a new, more complete set of opinion words can be constructed [Hu and Liu, 2004, Neviarouskaya et.al., 2009]. In [Esuli and Sebastiani, 2005], dictionary definitions are used for opinion words extraction. The basic idea is that words with the same orientation have “similar” glosses.

The third approach – corpus-based training. This approach is based on finding rules and patterns in the texts [Kanayama and Nasukawa, 2006]. In [Turney, 2002] word polarity is calculated by comparing the co-occurrence statistics of various words with words “excellent” and “poor”. Authors assume that words with similar semantic orientation tend to co-occur. The result-

3 Proposed method

In this section we will describe our method in respect to movie domain. We will train the model on the movie data and then try to utilize it in other domains.

3.1 Data Preparation

We collected 28773 film reviews of various genres from online recommendation service www.imhonet.ru. For each review, user’s score on a ten-point scale was extracted. We called this collection the review collection.

Example of the movie review:

Nice and light comedy. There is something to laugh - exactly over the humor, rather than over the stupidity... Allows you to relax and gives rest to your head.

We also needed a contrast collection of texts for our experiments. In this collection the concentration of opinions should be as little as possible. For this purpose, we had collected 17680 movie descriptions. This collection was named description collection.

One more contrast corpus was a collection of one million news documents. We had calculated document frequency of each word in this collection and used only this frequency list further. This list was named news corpus.

3.2 Collections with Higher Concentration of Opinions

We suggested that it was possible to extract some fragments of the reviews from review collection, which had higher concentration of opinion words. These fragments include:

- Sentences ending with a «!»;
- Sentences ending with a «…»;
- Short sentences, no more than 7 word length;
- Sentences containing the word «movie» without any other nouns.

We call this collection – small collection.
3.3 Statistical Features

Our task was to create a qualitative list of opinion words based on the calculation of various features. We used the following set of features for each word:

- Frequency of the word in the collection (i.e. number of occurrences in all documents in the collection)
- The number of documents where the word occurs
- Weirdness
- TFIDF
- Deviation from the average score
- Word score variance
- Frequency of capitalized words

We will consider some of them in more detail.

**Weirdness.** To calculate this feature two collections are required: one with high concentration of opinion words and the other – contrast one. The main idea of this feature is that opinion words will be «strange» in the contexts of the contrast collection. This feature is calculated as follows [Ahmad et. al, 1999]:

\[
\text{Weirdness} = \frac{w_s}{w_g} - \frac{t_s}{t_g}
\]

where \(w_s\) – frequency of the word in special corpus, \(t_s\) – total count of words in special corpus, \(w_g\) – frequency of the word in general corpus, \(t_g\) – total count of words in general corpus. Instead of frequency one can use the number of documents where the word occurs.

**TFIDF.** There are many varieties of this feature. We used TFIDF variant described in [Callan et. al., 1992] (based on BM25 function):

\[
\text{TFIDF} = \beta + (1-\beta) \cdot \text{idf}(l)
\]

\[
g_f(l) = \frac{\text{freq}_d(l)}{\text{freq}_d(l) + 0.5 + 1.5 \cdot \frac{\text{dl}_d}{\text{avg} \cdot \text{dl}}}
\]

\[
\text{idf}(l) = \log \left( \frac{|c| + 0.5}{g_f(l)} \right)
\]

\[
\text{freq}(l) \text{ – number of occurrences of } l \text{ in a document (collection),}
\text{dl}(l) \text{ – length measure of a document,}
\text{avg} \cdot \text{dl} \text{ – average length of a document,}
\text{df}(l) \text{ – number of documents in a collection (e.g. movie descriptions, news collection) where term } l \text{ appears,}
\beta = 0.4 \text{ by default,}
|c| \text{ – total number of documents in a collection.}
\]

**Deviation from the average score.** As we mentioned above we had collected user’s numerical score (on a ten point scale) for each review. The main idea of this feature is to calculate average score for each word (sum of review ratings where this word occurs divided into their number) in the collection and then subtract average score of all reviews in the collection from it.

\[
dev(l) = \left| \frac{\sum_{i=1}^{n} m_i k_i}{k} - \frac{\sum_{i=1}^{n} m_i}{n} \right|
\]

\[
\sum_{i=1}^{n} k_i = k
\]

where \(l\) – considered lemma, \(n\) – total count of the reviews in the collection, \(m_i\) – i-th review score, \(k_i\) – frequency of the lemma in the i-th review (may be 0).

**Word score variance.** Using review ratings we can calculate the score variance for each word. This feature can show us how often a word is used in reviews with significantly different scores. If a word has small deviation then it is used in reviews with similar scores and has high probability to be an opinion word.

\[
\text{Var}(l) = \frac{\sum_{i=1}^{n} m_i^2 k_i}{k} - \left( \frac{\sum_{i=1}^{n} m_i k_i}{k} \right)^2
\]

\[
\sum_{i=1}^{n} k_i = k
\]

where \(l\) – considered lemma, \(n\) – total count of the reviews in the collection, \(m_i\) – i-th review score, \(k_i\) – frequency of the lemma in the i-th review (may be 0).

**Frequency of words, which start with the capital letter.** The meaning of this feature is the frequency (in the review corpus) of each word starting with the capital letter and not located at the beginning of the sentence. With this feature we are trying to identify potential proper names, which are always neutral.
3.4 Feature and Collection Combinations

For our experiments we took top ten thousand words ordered by frequency from the movie review collection.

For each word from this list we had the following combinations of features and collections:

- TFIDF calculation using the pairs of collections: small-news, small-description, opinion-news, opinion-description;
- Weirdness calculation using the pairs of collections: opinion-news and opinion-description with document count and small-description, opinion-description with frequency;
- Deviation from the average score;
- Word score variance
- Word frequency in opinion and small collections;
- Total number of documents in the opinion corpus, where the word occurs;
- Frequency of capitalized words.

In addition, separately for description corpus we calculated the following features: frequency, document count, weirdness using description-news collections with document count and TFIDF using the same pair. Thus, each term had 18 features.

3.5 Algorithms and Evaluation

To train supervised machine learning algorithms we needed a set of labeled opinion words. We decided to label the full list of ten thousand words manually and then to use cross-validation. We marked up word as opinion one in case we could imagine it in any opinion context in the movie domain. All words were tagged by two authors.

As a result of our mark up we obtained the list of 3200 opinion words (1262 adjectives, 296 adverbs, 857 nouns, 785 verbs).

Our aim in this part of work was to classify words into two classes: opinion or neutral.

For this purpose Weka\footnote{http://www.cs.waikato.ac.nz/ml/weka/} data mining tool was used. We considered the following algorithms: Logistic Regression and LogitBoost. For all experiments 10 fold cross-validation was used.

Using aforementioned algorithms we obtained term lists, ordered by the predicted probability of their opinion orientation. To measure the quality of these lists we used Precision@$n$ metric. This metric is very convenient for measuring the quality of list combinations and it can be used with different thresholds. For the algorithms quality comparison in different domains we chose $n = 1000$. This level is not too large for manual labeling and demonstrates the quality in an appropriate way.

The results of classification are in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>Logistic Regression</th>
<th>LogitBoost</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision@1000</td>
<td>66.00%</td>
<td>66.80%</td>
<td>70.90%</td>
</tr>
</tbody>
</table>

Table 1. Precision@1000 of word classification

We noticed that the lists of opinion words extracted using two logistic algorithms differ significantly. So we decided to sum the weights of words in these two lists. The result of this summation can be found in the last column of the Table 1 and on the Figure 1.

As the baseline for our experiments we used lists ordered by frequency in the review collection and Deviation from the average score. Precision@1000 in these lists was 27.5% and 40.5% accordingly. Thus our algorithms gave significant improvements over baseline. All the other features can be found in Table 2.

Let us look at some examples of opinion words with the high probability value in the sum list:

- Trogatel’nyi (affective), otstoi (trash), fignia (crap), otvratitel’no (disgustingly), posredstvenniy (satisfactory), predskazuemyi (predictable), ljubimyj (love) etc.

Obtained opinion word lists can be utilized in various sentiment analysis tasks. For example...
words can be used as features for document classification by the overall sentiment.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Collection</th>
<th>Precision @1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>TFIDF</td>
<td>small – news</td>
<td>39.2%</td>
</tr>
<tr>
<td>TFIDF</td>
<td>small – descr</td>
<td>36.3%</td>
</tr>
<tr>
<td>TFIDF</td>
<td>review – news</td>
<td>33.8%</td>
</tr>
<tr>
<td>TFIDF</td>
<td>review – descr</td>
<td>30.4%</td>
</tr>
<tr>
<td>Weirdness</td>
<td>review – news (doc. count)</td>
<td>51.1%</td>
</tr>
<tr>
<td>Weirdness</td>
<td>review – descr (doc. count)</td>
<td>47.7%</td>
</tr>
<tr>
<td>Weirdness</td>
<td>small – descr (frequency)</td>
<td>49.2%</td>
</tr>
<tr>
<td>Weirdness</td>
<td>review – descr (frequency)</td>
<td>46.0%</td>
</tr>
<tr>
<td>Deviation from the average score</td>
<td>review</td>
<td>40.5%</td>
</tr>
<tr>
<td>Word score variance</td>
<td>review</td>
<td>31.7%</td>
</tr>
<tr>
<td>Frequency</td>
<td>review</td>
<td>27.5%</td>
</tr>
<tr>
<td>Frequency</td>
<td>small</td>
<td>32.1%</td>
</tr>
<tr>
<td>Document Count</td>
<td>review</td>
<td>27.9%</td>
</tr>
</tbody>
</table>

Table 2. Precision@1000 for different features

In [Chetviorkin et. al, 2011] we used opinion words in three-way review classification task and improved the quality of classification using opinion word weights.

3.6 Collection and Feature Selection

Finally, we studied the impact of each collection to the resulting quality of the opinion word classification. All collections (except review collection) were consequently excluded from constructing features. Additionally influence of the deviation from the average score, word score variance and frequency of words starting with capital letter were explored. In Table 3 results of classification with different feature sets can be found.

Thus, one can see that all collections and features improve the quality of classification. Exclusion of the description collection yields practically identical results for the sum list. Nevertheless this collection is very useful from model utilization in other domains (without it quality drops significantly).

<table>
<thead>
<tr>
<th>Feature</th>
<th>Logistic Regression</th>
<th>Logit-Boost</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>All \ small collection</td>
<td>60.7%</td>
<td>66.7%</td>
<td>66.5%</td>
</tr>
<tr>
<td>All \ descr collection</td>
<td>61.3%</td>
<td>67.2%</td>
<td>70.6%</td>
</tr>
<tr>
<td>All \ news collection</td>
<td>66.1%</td>
<td>67.1%</td>
<td>69.0%</td>
</tr>
<tr>
<td>All \ deviation from the average score</td>
<td>64.4%</td>
<td>64.1%</td>
<td>68.6%</td>
</tr>
<tr>
<td>All \ word score variance</td>
<td>62.9%</td>
<td>64.3%</td>
<td>67.6%</td>
</tr>
<tr>
<td>All \ frequency of capitalized words</td>
<td>61.1%</td>
<td>61.7%</td>
<td>64.4%</td>
</tr>
</tbody>
</table>

Table 3. Precision@1000 for different feature sets

4 Model Utilization to Similar Domains

In the previous section we constructed a new model for domain-specific opinion word extraction. We want to utilize this model in the other domains and evaluate the quality of obtained word lexicons and their combinations.

4.1 Data

We collected data on two more domains: book domain and computer games domain. The structure of the data was the same as for movie domain. Book and games review collections contained 16497 book reviews and 7928 game reviews of various genres accordingly. For each review, user’s score on a ten-point scale was extracted.

The contrast collections of texts for book domain and games domain contained 24555 book descriptions and 1853 game descriptions.

Here we used the same news corpus as for movie domain.

4.2 Model Utilization and Evaluation

For new domains we extracted ten thousand the most frequent words (or all available words with frequency more then 3) and calculated all statistical features, which were described in Section 3.3. At the next step we applied our model trained in the movie domain to the book and games word lists. To evaluate the quality of word
classification in new domains we manually labeled first thousand of words in each list. The results of classification are in Table 4.

<table>
<thead>
<tr>
<th></th>
<th>Logistic Regression</th>
<th>LogitBoost</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Books</td>
<td>69.60%</td>
<td>59.10%</td>
<td>72.20%</td>
</tr>
<tr>
<td>Games</td>
<td>49.40%</td>
<td>63.00%</td>
<td>62.90%</td>
</tr>
</tbody>
</table>

Table 4. Results of the classification in book and games domains.

At the final step we took linear combination of the words (sum of word weights) in each list from three different domains (6 lists). The Precision@1000 of the obtained opinion word list was 82.0%.

We supposed that this general opinion word lexicon could improve the quality of the best list obtained in the movie domain. We summed weights of the best combined list in movie domain and general one (from three domain lists). Weights of the latter list were normalized previously. The quality of obtained movie domain-specific word dictionary was 71.8%. So exploitation of opinion words from other similar domains improved extraction of opinion words in the initial domain (+1.26%).

5 Conclusion

In this paper, we described a method for opinion word extraction for any domain on the basis of several domain specific text collections. We utilized our algorithm in different domains and showed that it had good generalization abilities. The quality of the combined list was significantly better then the quality of each single list. Usage of the combined list improved extraction of opinion words in the initial domain.

Acknowledgements. This work is partially supported by RFBR grant N11-07-00588-a.

References


The Role of Predicates in Opinion Holder Extraction

Michael Wiegand and Dietrich Klakow
Spoken Language Systems
Saarland University
D-66123 Saarbrücken, Germany
{Michael.Wiegand|Dietrich.Klakow}@lsv.uni-saarland.de

Abstract

In this paper, we investigate the role of predicates in opinion holder extraction. We will examine the shape of these predicates, investigate what relationship they bear towards opinion holders, determine what resources are potentially useful for acquiring them, and point out limitations of an opinion holder extraction system based on these predicates. For this study, we will carry out an evaluation on a corpus annotated with opinion holders. Our insights are, in particular, important for situations in which no labeled training data are available and only rule-based methods can be applied.

1 Introduction

One of the most important tasks in sentiment analysis is opinion holder extraction in which the entities uttering an opinion, also known as opinion holders, need to be extracted from a natural language text. For example, the opinion holders in (1) and (2) are the vet and Russia, respectively.

1. The owner put down the animal, although the vet had forbidden him to do so.
2. Russia favors creation of “international instruments” to regulate emissions.

As this is an entity extraction problem it can be considered as a typical task in information extraction. Though there is much work on that subject, most work focuses on data-driven methods. Thus, to a great extent it fails to fully describe certain linguistic aspects of that task.

In this work, we will have a close look at the role of predicates involved in opinion holder extraction. Predictive predicates for this task are, for example, forbidden in (1) and favors in (2). Unlike previous work, we will examine predicates in isolation. This means we do not consider them as some feature used in a data-driven classifier but as a part of an unsupervised rule-based extraction system which almost exclusively relies on them.

Apart from carrying out a quantitative examination regarding the shape of these predicates and the relationship they bear towards opinion holders, our main contributions of this paper are the investigation of what lexical resources are potentially useful for acquiring predictive predicates and pointing out the limitations of opinion holder extraction based on these predicates.

Our insights are important for building opinion holder extraction systems, in particular, rule-based systems. In particular, we hope that our analysis will provide a realistic rule-based baseline for opinion holder extraction. We also believe that many observations from this paper carry over to languages other than English. For only few of them, there are some corpora annotated with opinion holder information available. For all the remaining languages, rule-based systems leveraging the insights of this paper could be an option for automatic analysis.

2 Related Work

There has been much research on supervised learning for opinion holder extraction. Choi et al. (2005) examine opinion holder extraction using CRFs with several manually defined linguistic features and automatically learnt surface patterns. Bethard et al. (2004) and Kim and Hovy (2006) explore the usefulness of semantic roles provided by FrameNet (Fillmore et al., 2003) for both opinion holder and opinion target extraction. The approaches of those two papers have mostly been evaluated on some artificial data sets. More recently, Wiegand and Klakow (2010) explored convolution kernels for opinion holder extraction.

Rule-based opinion holder extraction heavily relies on lexical cues. Bloom et al. (2007) use a list of manually compiled communication verbs and
identify opinion holders as noun phrases having a specific grammatical relation towards those verbs. The rule-based classifiers we evaluate in this work stem from this basic idea. However, we extend this classifier, for example, by considering a more diverse set of predicates and different grammatical relations.

Another work closely related to this study is (Ruppenhofer et al., 2008) which presents a roadmap to both opinion holder and target extraction outlining diverse linguistic phenomena involved in these tasks. In this work, we focus on the role of predicates. Moreover, we also carry out a quantitative evaluation of those related phenomena. Unlike Ruppenhofer et al. (2008), we thus try to identify the most immediate problems of this task. By also considering resources in order to solve these problems we hope to be a helpful guide for practitioners building an opinion holder extraction system from scratch.

3 Data

As a labeled (test) corpus, we use the MPQA 2.0 corpus\(^1\) which is a large text corpus containing fine-grained sentiment annotation. It (mainly) consists of news texts which can be considered as a primary domain for opinion holder extraction. Other popular domains for sentiment analysis, for example, product reviews contain much fewer opinion holders according to the pertaining data sets (Kessler et al., 2010; Toprak et al., 2010).

Opinions uttered in those texts usually express the author’s point of view. Therefore, the extraction of sources of opinions is of minor importance.

We use the definition of opinion holders as described in (Wiegand and Klakow, 2010), i.e. every source of a private state or a subjective speech event (Wiebe et al., 2003) is considered an opinion holder. This is a very strict definition and the scores produced in this work can only be put into relation to the numbers presented in (Wiegand and Klakow, 2010). The final corpus comprises approximately 11,000 sentences with more than 6,200 opinion holders.

4 Examination of Predicates

4.1 The Different Types of Predicates

Table 1 displays the distribution of the different predicate types. We divided them into three categories being: unigram predicates (\textit{verb}, \textit{noun} and \textit{adj}), multiword expressions of common syntactic structures (i.e. \textit{verb+object}, \textit{verb+prepObject}, \textit{have+object} and \textit{phrasal verb}) and a category for everything else. The table shows that the unigram predicates are most frequent. Since they cover almost 90% of the opinion holder predicate instances, we will focus on these expressions in the following experiments.

4.2 The Different Types of Grammatical Relations

Table 2 shows the distribution of the most frequent grammatical relations between opinion holder and its related predicate listed separately for each unigram predicate type. We use the Stanford parser (Klein and Manning, 2003) for obtaining all syntactic information. The table displays the percentage of that grammatical relation within the particular predicate type when it is observed as a predicate of an opinion holder in our labeled data set (\textit{Perc.})\(^2\), the property of being a fairly reliable relation for a semantic agent (\textit{Agent}), and the precision of that grammatical relation in conjunction with that opinion holder predicate type for detecting opinion holders (\textit{Precision}). As a gold-standard of opinion holder predicates we extracted all unigram predicates from our data set that co-occur at least twice with an actual opinion holder.\(^4\)

One may wonder why we did not mark the relation \textit{nsbj} for nouns as \textit{Agent} while the relation is marked as such for the other parts of speech. We found that subjects of predicate nouns can very often be found in constructions like (3). Clearly, \textit{this} is not an agent of idea.

\(^3\) Note that for verbs we display relations with a lower percentage (>1%) than for nouns or adjectives (>4%) since verb predicates occur much more often.

\(^4\) Singletons may be fairly noisy which is why we omit them.

3. This is really an unwise idea. [nsbj(This,idea)]

Table 2 shows that there are some specific grammatical relations that co-occur frequently with opinion holders. These relations are exactly those implying an agent. Moreover, these relations are also the ones with the highest precision.

This insight may suggest using semantic-role labeling (SRL) for this task. We deliberately stick to using syntactic parsing since most publicly available SRL-systems only consider verb

\(^1\)\url{www.cs.pitt.edu/mpqa/databaserelease}

\(^2\) Note that by \textit{verb}, we always only refer to full verbs, i.e. auxiliary and modal verbs are excluded.

\(^3\) Note that for verbs we display relations with a lower percentage (>1%) than for nouns or adjectives (>4%) since verb predicates occur much more often.
<table>
<thead>
<tr>
<th>Predicate Type</th>
<th>Frequency</th>
<th>Percentage</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>verb</td>
<td>4272</td>
<td>70.89</td>
<td>I believe that this is more than that.</td>
</tr>
<tr>
<td>noun</td>
<td>948</td>
<td>15.73</td>
<td>This includes a growing reluctance by foreign companies to invest in the region.</td>
</tr>
<tr>
<td>adj</td>
<td>201</td>
<td>3.34</td>
<td>Ordinary Venezuelans are even less happy with the local oligarchic elite.</td>
</tr>
<tr>
<td>verb+object</td>
<td>234</td>
<td>3.88</td>
<td>Some officials voiced concern that China could secure concessions on Taiwan.</td>
</tr>
<tr>
<td>verb+prepObject</td>
<td>58</td>
<td>0.96</td>
<td>The United States stands on the Israeli side in dealing with the Middle East situation.</td>
</tr>
<tr>
<td>have+object</td>
<td>40</td>
<td>0.66</td>
<td>The KMM had no confidence in the democratic system.</td>
</tr>
<tr>
<td>phrasal verb</td>
<td>34</td>
<td>0.56</td>
<td>Washington turned down that protocol six months ago.</td>
</tr>
<tr>
<td>else</td>
<td>239</td>
<td>3.97</td>
<td>NA</td>
</tr>
</tbody>
</table>

Table 1: Distribution of the different opinion predicates.

<table>
<thead>
<tr>
<th>Type</th>
<th>Relation</th>
<th>Perc.</th>
<th>Agent</th>
<th>Precision</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>verb</td>
<td>↓nsubj</td>
<td>80.59</td>
<td>✓</td>
<td>47.47</td>
<td>China had always firmly opposed the US Taiwan Affairs Act.</td>
</tr>
<tr>
<td>verb</td>
<td>↓byobj</td>
<td>2.69</td>
<td>✓</td>
<td>29.89</td>
<td>The agreements signed in 1960 for Cyprus were considered as nonexistent by many countries.</td>
</tr>
<tr>
<td>verb</td>
<td>↑rcmod</td>
<td>2.55</td>
<td></td>
<td>10.03</td>
<td>It was the President who banned postal voting by all Zimbabweans outside their constituencies.</td>
</tr>
<tr>
<td>verb</td>
<td>↓nsubjpass</td>
<td>1.50</td>
<td></td>
<td>8.85</td>
<td>I am shocked.</td>
</tr>
<tr>
<td>verb</td>
<td>↓dobj</td>
<td>1.24</td>
<td></td>
<td>2.38</td>
<td>Washington angered Beijing last year.</td>
</tr>
<tr>
<td>verb</td>
<td>↑partmod</td>
<td>1.08</td>
<td></td>
<td>7.13</td>
<td>Mugabe has no moral excuse for shooting people demanding a new constitution.</td>
</tr>
<tr>
<td>noun</td>
<td>↓poss</td>
<td>45.04</td>
<td>✓</td>
<td>44.56</td>
<td>President Bush’s declaration touched off questions around the globe.</td>
</tr>
<tr>
<td>noun</td>
<td>↓ofobj</td>
<td>10.75</td>
<td></td>
<td>19.06</td>
<td>Through the protests of local labor groups, foreign laborers’ working rights were protected.</td>
</tr>
<tr>
<td>noun</td>
<td>↓nsubj</td>
<td>6.12</td>
<td></td>
<td>6.42</td>
<td>Chavez is a staunch supporter of oil production cuts.</td>
</tr>
<tr>
<td>adj</td>
<td>↓nsubj</td>
<td>75.12</td>
<td>✓</td>
<td>71.63</td>
<td>We are grateful for the strong support expressed by the international community.</td>
</tr>
<tr>
<td>adj</td>
<td>↓amod</td>
<td>4.98</td>
<td></td>
<td>6.48</td>
<td>Soldiers loyal to the sacked chief of army staff exchanged gunfire with presidential guard units.</td>
</tr>
</tbody>
</table>

Table 2: Distribution of the different grammatical relations (percentage measured within predicate type).

It is interesting to note that there are also verbs occurring in argument positions that are definitely not agentive, i.e. ↓dobj and ↓nsubjpass. We inspected these cases in order to find out whether there is a set of verbs that systematically realizes opinion holders in non-agentive positions. Table 3 lists those verbs we found in our data set. 87.5% of them are also part of the so-called amuse verbs, a subset of transitive psych-verbs whose object is an experiencer and their subject is the cause of the psychological state (Levin, 1993). The subject, i.e. the cause (this does not even have to be a person), is unlikely to be the opinion holder, whereas the experiencer is often observed to denote such an entity.

4.3 The Different Resources for Opinion Holder Predicates

In this section, we want to examine in how far existing resources contain predicates that usually co-occur with opinion holders. The resources we consider are different in their design and serve diverse purposes. Only one has been specifically designed for opinion holder extraction. For the remaining resources, there may be some modification necessary, for example, by selecting a subset. As we want to examine these resources for an unsupervised (open-domain) rule-based method, these modifications should be pretty simple, fast to implement, and not require extensive knowledge about our particular data set.

4.3.1 Communication Verbs from Appraisal Lexicon (AL)

The communication verbs from Appraisal Lexicon (AL) are the only lexicon that has been designed for opinion holder extraction (Bloom et al., 2007). With 260 entries, it is the smallest resource in this paper. Little is known about the creation of this resource (e.g. whether the resource has been optimized for some domain) except that several verb classes from Levin (1993) have been considered.
Table 4: Levin’s verb classes taking opinion holders in agitative argument position (only *amuse verbs* take opinion holder in non-agitative positions).

<table>
<thead>
<tr>
<th>alienate</th>
<th>concern</th>
<th>exasperate</th>
<th>lure</th>
<th>rile</th>
</tr>
</thead>
<tbody>
<tr>
<td>anger</td>
<td>cow</td>
<td>frustrate</td>
<td>obsess</td>
<td>scare</td>
</tr>
<tr>
<td>annoy</td>
<td>disappoint</td>
<td>frustrate</td>
<td>offend</td>
<td>shock</td>
</tr>
<tr>
<td>astonish</td>
<td>discourage</td>
<td>humbleiate</td>
<td>persuade</td>
<td>stunt</td>
</tr>
<tr>
<td>baffle</td>
<td>disgust</td>
<td>infuriate</td>
<td>please</td>
<td>suit</td>
</tr>
<tr>
<td>bias</td>
<td>disturb</td>
<td>intimidate</td>
<td>rankle</td>
<td>surprise</td>
</tr>
<tr>
<td>bother</td>
<td>embarrass</td>
<td>irk</td>
<td>relieve</td>
<td>tear</td>
</tr>
<tr>
<td>captivate</td>
<td>enrage</td>
<td>irritate</td>
<td>remind</td>
<td>worry</td>
</tr>
</tbody>
</table>

Table 3: Predicates taking opinion holders in a non-agentive argument position.

A4.3.3 Levin’s Verb Classes (Levin)

Even though AL already considers verb classes from Levin (1993), we constructed a separate subset from that resource for this study. The reason for this is that we found that there are many relevant verbs (e.g. *agree, deem or disapprove*) not contained in AL but that are part of Levin’s lexicon. Our selection method was ad-hoc but we did not tune this resource for any particular data set, i.e we included every verb class in our lexicon of which the majority were verbs we would associate *a priori* with opinion holders.

Another important aspect of Levin’s work (as already mentioned in §4.2) is that it allows a distinction of verbs taking opinion holders in agentive argument positions and verbs taking them in other positions. We identified *amuse verbs* to be precisely the latter class. (Note that AL completely excludes this class.) Admittedly, other resources, such as FrameNet, also encode that distinction. Unfortunately, using FrameNet for an unsupervised classifier would be more difficult. We would need to choose from 1049 (partially overlapping) frames. In Levin’s lexicon, we only needed to choose from 193 classes. The final selection is shown in Table 4.

### 4.3.4 WordNet - Lexicographer files (WN-LF)

WordNet\(^6\) is possibly the largest and most popular general-purpose lexico-semantic ontology for the English language. Most work using this resource focuses on the relationship between the different synsets, i.e. the groups of synonymous words that represent the nodes in the ontology graph. Due to the high number of these synsets, we found it very difficult to select an appropriate subset predictive for opinion holder extraction. This is why we tried to harness another form of word grouping that this resource provides. The *lexicographer files* (WN-LF) seem to operate on a more suitable level of granularity. The entire ontology (i.e. the set of synsets) is divided in 44 of such files where each file represents a very general semantic word class (e.g. *noun.food* or *verb.motion*). We consider the files *noun.cognition*, *noun.communication*, *verb.cognition* and *verb.communication*. Due to the coarse-grained nature of the WN-LF, the resulting set of words contains 10151 words (7684 nouns and 2467 verbs).

Table 5 summarizes the properties of the different resources. Due to the high number of nouns in WN-LF, we will evaluate this lexicon both with and without nouns. For all resources only containing verbs, we also use Nomlex (Macleod et al., 1998) to find corresponding noun predicates, e.g.

\[^{5}\text{according to:}\]

http://framenet.icsi.berkeley.edu/

\[^{6}\text{wordnet.princeton.edu}\]
believe (verb) → belief (noun), as we already established in Table 1 that nouns play a significant part in the recognition of opinion holders.

### 4.4 Comparison of Resources

Table 6 displays the performance of the different resources when used in a simple rule-based opinion holder classifier. It classifies a noun phrase (NP) as an opinion holder when the NP is an agent (according to the unambiguous grammatical relations from Table 2)\(^7\) of an entry in a particular lexicon. Only for the **amuse verbs** in Levin, we consider the other grammatical relations ↓\text{nsubjpass} and ↓\text{dobj}.

The different resources produce quite different results. Surprisingly, SL is the lowest performing resource even though it has been used in previous work (Choi et al., 2005; Wiegand and Klakow, 2010). Though the recall increases by adding nouns and adjectives to verbs, the precision notably drops. For the subset \(S_{\text{strong}}\), the precision drops slightly less so that the F-Score always increases when the other parts of speech are added to the verbs. Overall, \(S_{\text{strong}}\) has a much higher precision than SL and its F-Score (considering all parts of speech) is on a par with SL even though it is a significantly smaller word list (see Table 5). SL is a resource primarily built for subjectivity and polarity classification and these results suggest that the lexical items to imply opinion holders are only partially overlapping with those clues.

Though AL and Levin are considerably smaller than SL, they perform better. Moreover, Levin is considerably better than AL. In both cases, the extension by noun predicates using Nomlex results in a marginal yet consistent improvement. Unfortunately, the usage of the **amuse verbs** does not produce a notable improvement. We mostly ascribe it to the fact that those verbs occurred only very infrequently (i.e. either once or twice in the entire data set).

**WN-LF** performs slightly better than Levin. Adding the large set of nouns is not effective. The set of verbs augmented by corresponding noun predicates obtained by Nomlex produces better results. The large F-Score of WN-LF is only due to a high recall. The precision is comparably low. For this task, another set of predicates maintaining a higher precision is clearly preferable.

### 4.5 Combination of the Resources

In this section, we combine the different resources (by that we mean taking the union of different resources). For each resource, we use the best performing configuration from the previous evaluation. Table 7 shows the performance of different combinations. As testing all combinations would be beyond the scope of this work, we mainly focus on combinations not using WN-LF.

---

\(^7\)By that we mean those relations marked with Agent.
Table 7: Performance of combined resources.

<table>
<thead>
<tr>
<th>Resource(s)</th>
<th>Prec</th>
<th>Rec</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>WN-LF (baseline)</td>
<td>32.97</td>
<td>68.73</td>
<td>44.56</td>
</tr>
<tr>
<td>SL+AL</td>
<td>37.52</td>
<td>52.80</td>
<td>43.68</td>
</tr>
<tr>
<td>SL+strong+AL</td>
<td>42.91</td>
<td>49.69</td>
<td>46.05</td>
</tr>
<tr>
<td>SL+Levin</td>
<td>34.56</td>
<td>62.95</td>
<td>44.62</td>
</tr>
<tr>
<td>SL+strong+Levin</td>
<td>40.97</td>
<td>57.43</td>
<td>47.82</td>
</tr>
<tr>
<td>AL+Levin</td>
<td>41.49</td>
<td>47.80</td>
<td>44.42</td>
</tr>
<tr>
<td>SL+AL+Levin</td>
<td>34.56</td>
<td>62.95</td>
<td>44.62</td>
</tr>
<tr>
<td>SL+strong+AL+Levin</td>
<td>40.96</td>
<td>57.43</td>
<td>47.82</td>
</tr>
<tr>
<td>SL+AL+Levin+WN-LF</td>
<td>29.47</td>
<td>78.28</td>
<td>42.82</td>
</tr>
<tr>
<td>SL+strong+AL+Levin+WN-LF</td>
<td>32.19</td>
<td>75.32</td>
<td>45.10</td>
</tr>
<tr>
<td>OraclePRED</td>
<td>46.44</td>
<td>67.83</td>
<td>55.13</td>
</tr>
<tr>
<td>OraclePRED*</td>
<td>47.04</td>
<td>68.62</td>
<td>55.82</td>
</tr>
</tbody>
</table>

We seek a classifier with a higher precision than that achieved by WN-LF. Combining WN-LF with other resources would only result in another increase in recall.

We also want to have an estimate of an upper bound of this method. OraclePRED uses all predicates that occur as a predicate of an opinion holder on our test set at least twice. We only consider predicates which have been observed in prototypical agentive positions. OraclePRED* also uses the knowledge of the predicates from the data set but is not restricted to agentive patterns. That is, we store for each predicate the grammatical relation(s) with which it has been observed (e.g. oppose+ |nsubj or anger+ |dobj); we only consider the frequent grammatical relationships from Table 2. Thus, like semantic role labeling, we should be more flexible than a classifier that exclusively considers opinion holders to be in an agentive argument position of some predicate.

Table 7 shows that a combination of resources is indeed beneficial. SL_{strong} and Levin produce a higher F-Score than WN-LF by preserving a considerably higher precision. Adding AL to this set has no effect on performance, since the few predicates of AL are already in the union of SL_{strong}+Levin. Comparing the performance of the different configurations with OraclePRED, we can conclude that the resources that are considered are not exactly modeling opinion holder predicates but a combination of them does it to a large extent. Looking at the false negatives that the best configuration produced (note that we will discuss the issue of false positives in §4.6), we could not really make out a particular group of verbs that this classifier systematically excluded. As far as Levin is concerned, however, we assume that the fact that this typology only considers 3000 verbs in total also means that many infrequent verbs, such as ratify or lobby, have simply been excluded from consideration even though their behavior would enable an assignment to one of the existing verb classes.

The performance of OraclePRED also shows that opinion holder extraction is a really difficult task as this upper bound is fairly low in absolute numbers. The oracle using the grammatical relations (OraclePRED*) improves performance only slightly. This is consistent with our experiments using amuse verbs.

4.6 Ambiguity of Predicates

In this section we evaluate individual predicates that occur very frequently and also state in which resources these expressions can be found. Table 8 shows that these predicates behave quite differently. The verb say is by far the most predictive individual predicate though this is mainly due to its high recall. Other verbs, such as want, believe or think, have a considerably lower recall but their precision is almost twice as high. In terms of coverage, WN-LF is the only resource that contains all expressions. This is consistent with its high recall that was measured in previous experiments. On the other hand, SL_{strong} only contains a subset of these expressions but the expressions are mostly those with a very high precision.

The individual examination of highly frequent predicates shows that a problem inherent in opinion holder extraction based on predicates is the lacking precision of predicates. In general, we do not think that the false positives produced by our best configuration are due to the fact that there are many predicates on the list which are wrong in general. Omitting a verb with a low precision, such as say or call, is not an option as it would always heavily degrade recall.

5 Other Clues for Opinion Holder Extraction

In this section, we want to put opinion holder predicates into relation to other clues for opinion holder extraction. We consider two types of clues that can be automatically computed. Both aim at improving precision when added to the clue based on opinion holder predicates since this clue
already provides a comparatively high recall.

The clue PERSON checks whether the candidate opinion holder is a person. For some ambiguous predicates, such as critical, this would allow a correct disambiguation, i.e., Dr. Ren in (4) would be classified as an opinion holder while the cross-strait balance of military power in (5) would not.

4. Dr. Ren was critical of the government’s decision.

5. In his view, the cross-strait balance of military power is critical to the ROC’s national security.

For this clue, we employ Stanford named-entity recognizer (Finkel et al., 2005) for detecting proper nouns and WordNet for recognizing common nouns denoting persons.

The second clue SUBJ detects subjective evidence in a sentence. The heuristics applied should filter false positives, such as (6).

6. “We do not have special devices for inspecting large automobiles and cargoes”, Nazarov said.

If an opinion holder has been found according to our standard procedure using opinion holder predicates, some additional property must hold so that the classifier predicts an opinion holder. Either the candidate opinion holder phrase contains a subjective expression (7), some subjective expression modifies the predicate (8), or the proposition that is introduced by the opinion holder predicate contains at least one subjective expression (9).

7. Angry residents looked to Tsvangirai to confront the government.

8. Thousands waited angrily to cast their votes.

9. Mr. Mugabe’s associates said it was a “bad subj decision” proposition.

The subjective expressions are again obtained by using the Subjectivity Lexicon (Wilson et al., 2005). Since in our previous experiments the subset of strong subjective expressions turned out to be effective, we examine another clue SUBJ$^{\text{strong}}$ which just focuses on this subset.

As we assume this kind of subjectivity detection to be very error prone, we also want to consider a related upper bound. This upper bound allSPEECH addresses the most frequently found reason for misclassifying an opinion holder on the basis of predicates, namely failing to distinguish between the underlying objective and subjective speech events. (We will focus on only this error source in this work, since the other error sources are much more infrequent and diverse. Their discussion would be beyond the scope of this paper.) We previously measured a fairly low precision of predicates denoting speech events, such as say or tell. This is due to the fact that these predicates may not only be involved in subjective speech events, such as (9), but may also introduce objective speech events, such as (6), that typically involve no opinion holder. Our upper bound allSPEECH undoes the distinction between different speech events in the gold standard (i.e., it always considers a source of a speech event as an opinion holder). Thus, we simulate how opinion holder extraction would work if this distinction could be perfectly automatically achieved.

Table 9 displays the results of various combinations. For the opinion holder predicates, we consider the best combination of resources from our previous experiments in §4.5 (PRED) and the upper bound of predicates (OraclePRED$^+$). The table shows that adding PERSON to PRED results in an improved F-Score. The addition of SUBJ increases precision while recall drops. allSPEECH, on the other hand, causes a boost in performance. Even though the combination of the two upper bounds OraclePRED$^+$ and allSPEECH together with the PERSON filter would largely increase performance, the total F-Score of 65% shows that it would not completely solve this task.

### 6 Discussion

If we compare our best fully automatic result, i.e. PRED+PERSON with 49.90% (Table 9) with that of data-driven methods using the same corpus and task definition, for example Wiegand and Klakow (2010), who obtain an F-Score of almost 63%, one
Table 9: Performance of opinion holder predicates in conjunction with other clues.

<table>
<thead>
<tr>
<th>Clues</th>
<th>Prec</th>
<th>Rec</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRED</td>
<td>40.97</td>
<td>57.43</td>
<td>47.82</td>
</tr>
<tr>
<td>PRED+PERSON</td>
<td>48.67</td>
<td>51.21</td>
<td>49.90</td>
</tr>
<tr>
<td>PRED+SUBJ</td>
<td>48.04</td>
<td>32.77</td>
<td>38.97</td>
</tr>
<tr>
<td>PRED+SUBJ_{strong}</td>
<td>48.89</td>
<td>23.32</td>
<td>31.58</td>
</tr>
<tr>
<td>PRED+PERSON+SUBJ</td>
<td>57.84</td>
<td>29.87</td>
<td>39.39</td>
</tr>
<tr>
<td>PRED+PERSON+SUBJ_{strong}</td>
<td>60.13</td>
<td>21.24</td>
<td>31.39</td>
</tr>
<tr>
<td>PRED+allSPEECH</td>
<td>53.79</td>
<td>58.33</td>
<td>55.97</td>
</tr>
<tr>
<td>PRED+PERSON+allSPEECH</td>
<td>64.00</td>
<td>53.27</td>
<td>58.14</td>
</tr>
<tr>
<td>OraclePRED*+allSPEECH</td>
<td>60.21</td>
<td>67.92</td>
<td>63.83</td>
</tr>
<tr>
<td>OraclePRED*+PERSON+allSPEECH</td>
<td>69.67</td>
<td>61.59</td>
<td>65.38</td>
</tr>
</tbody>
</table>

still notices a considerable gap. Of course, this particular data-driven method should be regarded as an upper bound since it uses a very large labeled training set (§3) and even incorporates some lexical resources for feature engineering we almost exclusively rely on in our rule-based classifier (i.e. AL and SL). This also shows that it is really hard to build a rule-based classifier for opinion holder extraction.

7 Conclusion

In this paper, we examined the importance of predicates from diverse resources for the extraction of opinion holders. We found that strong subjective expressions from the Subjectivity Lexicon combined with a subset of Levin’s verb classes contain very predictive words. A classifier extracting noun phrases in an unambiguous agentive position of these predicates results in an opinion holder classifier with both reasonable recall and precision but our upper bound shows that there is still room for improvement. Opinion holders in non-agentive positions are so infrequent in our test set that their consideration is less critical. The classifier based on opinion holder predicates can only be improved by restricting holder candidates to persons. Further filters ensuring subjectivity are too restrictive and thus cause a large decrease in recall. Our exploratory experiments show, however, that some additional improvement in opinion holder extraction could be achieved if subjective speech events could be better separated from objective ones.

Acknowledgements

This work was funded by the German Federal Ministry of Education and Research (Software-Cluster) under grant no. “01IC10S01” and the Cluster of Excellence for Multimodal Computing and Interaction. The authors would like to thank Josef Ruppenhofer for interesting discussions.

References


J. Kessler, M. Eckert, L. Clarke, and N. Nicolov. 2010. The ICWSM JDPA 2010 Sentiment Corpus for the Automotive Domain. In ICWSM-DCW.


Dependency-Based Text Compression for Semantic Relation Extraction

Marcos Garcia
Center for Research in
Information Technologies (CITIUS)
University of Santiago de Compostela
marcos.garcia.gonzalez@usc.es

Pablo Gamallo
Center for Research in
Information Technologies (CITIUS)
University of Santiago de Compostela
pablo.gamallo@usc.es

Abstract

The application of linguistic patterns and rules are one of the main approaches for Information Extraction as well as for high-quality ontology population. However, the lack of flexibility of the linguistic patterns often causes low coverage. This paper presents a weakly-supervised rule-based approach for Relation Extraction which performs partial dependency parsing in order to simplify the linguistic structure of a sentence. This simplification allows us to apply generic semantic extraction rules, obtained with a distant-supervision strategy which takes advantage of semi-structured resources. The rules are added to a partial dependency grammar, which is compiled into a parser capable of extracting instances of the desired relations. Experiments in different Spanish and Portuguese corpora show that this method maintains the high-precision values of rule-based approaches while improves the recall of these systems.

1 Introduction

In recent years, the interest in obtaining structured data from unstructured resources has been increased, namely due to the exponential growth of information in the Web. Regarding this objective, Relation Extraction (RE) aims to automatically identify semantic relations between entities. For instance, from the sentence “Nick Cave was born in the small town of Warracknabeal”, a RE system may identify that Warracknabeal is the birthplace of Nick Cave.

The obtained data are arranged in machine-readable formats (“Nick Cave hasBirthplace Warracknabeal”), and then incorporated into databases and ontologies, used to improve applications such as Question Answering engines or Information Retrieval systems.

RE systems usually need a set of sentences containing instances of a semantic relation (e.g., hasBirthplace). These sentences are processed in order to provide a rich linguistic space with different knowledge (tokens, lemmas, PoS-tags, syntactic dependencies, etc.). This knowledge is used to extract semantic relations by (i) training machine-learning classifiers or by (ii) applying on large corpora lexico-syntactic patterns (LSP) derived from the linguistic space.

Relation Extraction approaches rely on the assumption that lexico-syntactic regularities (e.g., LSP) may characterize the same type of knowledge, such as semantic information. However, one of the main problems of these strategies is the low coverage of LSP, which varies with small differences in punctuation, adjective or adverb modification, etc. For instance, the previous example sentence could be represented in a great variety of manners:

• “Nick Cave was born in the small town of Warracknabeal”
• “Nick Cave was born in the town of Warracknabeal”
• “Nick Cave was born in Warracknabeal”
• “Nick Cave, born in the small town of Warracknabeal”

Both machine learning and pattern-matching approaches attempt to avoid this problem by using
larger training data or by applying syntactic parsers that identify the constituents of a sentence as well as their functions. However, obtaining large collections of high-quality training data for different semantic relations is not always feasible, since a lot of manual effort is needed. Furthermore, parsers for other languages than English often perform very partial analyses, or are not freely available.

In this paper, we introduce a rule-based approach for RE that overcomes the low coverage problem by simplifying the linguistic structures: we perform a sort of sentence compression technique that uses partial dependency parsing to remove some satellite elements from the input of the extraction rules.

In order to obtain high-quality extraction rules, we use a distant-supervision strategy that takes advantage of semi-structures resources, such as Wikipedia infoboxes or Freebase:¹ First, large sets of semantically related pairs are used for automatically extracting and annotating sentences containing instances of the desired relation. Then, we transform these sentences into LSP, which are generalized through a longest common string algorithm. Finally, the generic patterns are converted into syntactic-semantic rules and added to a dependency grammar.

We performed several experiments with different semantic relations in Portuguese and Spanish, using encyclopedic and journalistic corpora. The results show that dependency-based text compression allows us to improve the recall without losing the high precision values of pattern-matching techniques.

This paper is organized as follows: Section 2 introduces some related work. Section 3 presents the motivation of our Relation Extraction method. Then, Sections 4 and 5 show the strategy for extracting patterns as well as the method for transforming them into semantic rules. In Section 6, some experiments are performed. Finally, Section 7 reports the conclusions of our work.

2 Related Work

In this section we briefly introduce some related work concerning text compression methods as well as strategies for semantic Relation Extraction.

In recent years, several approaches have been proposed for sentence compression, whose aim is to reduce the size of a text while preserving its essential information (Chandrasekar et al., 1996). There are statistical methods (with different degree of supervision) for sentence compression, which require training corpora in order to learn what constituents could be removed from the original input (Clarke and Lapata, 2006). Cohn and Lapata (2009) present Tree-to-Tree Transducer, a state-of-the-art sentence compression method which transforms a source parse tree into a compressed parse tree. We have to note that our approach differs from common sentence compression strategies in a key point: it is not centered in maintaining the grammaticality of a sentence, but just in simplifying its structure and keeping its essential information.

Regarding Relation Extraction, Hearst (1992) was the first one to experiment a pattern-based strategy for the identification of semantic relations, using a small set of initial patterns to get hyperonymy relations by means of a bootstrapping technique. In Brin (1998), a similar method is applied, but it only selects those patterns that show a good performance. Other works make use of Question-Answering pair examples to automatically extract patterns (Ravichandran and Hovy, 2002). A novelty of this system lies in the application of a suffix tree, leading to discover generalized patterns by calculating their common substrings.

In the previously cited work, the learning process starts with patterns that have high precision but low recall. So, recall is increased by automatically learning new patterns. By contrast, in Pantel and Penacchiotti (2006), the starting point are patterns with high recall and low precision. The goal is to exploit these patterns by filtering incorrect related pairs using the Web. There are also interesting works using more supervised strategies for domain-specific corpora: in Aussenac-Gilles and Jacques (2006), it is described a method and a tool to manually define new specific patterns for specialized text corpora.

Recently, distant-supervision and self-supervised approaches take advantage of large amounts of freely available structured data, in order to automatically obtain training corpora to build extraction systems (Mintz et al., 2009; Hoffman et al., 2010).

Other works perform extraction in a different way. Open Information Extraction is a new paradigm that attempts to extract a large set of relational pairs with-

¹www.wikipedia.org and www.freebase.com
out manually specifying semantic relations (Etzioni et al., 2008). WOE is an Open Information Extraction method that takes advantage of the high quality semi-structures resources of Wikipedia (Wu and Weld, 2010). Finally, Bollegala's Relational Duality (Bollegala et al., 2010) applies a sequential co-clustering algorithm in order to cluster different LSP for extracting relations.

3 Motivation

The method presented in this paper follows a common statement which suggests that some linguistic constructs reliably convey the same type of knowledge, such as semantic or ontological relations (Aussenac-Gilles and Jacques, 2006; Aguado de Cea et al., 2009). Furthermore, it is based on the following assumption:

Semantic relations can be expressed in the same simple way as syntactic dependencies

A semantic relation found in a sentence can be usually represented by a dependency link between two entities, even if there are items of extra information that can make the sentence very complex. This extra information does not express the target relation, but it may extend the meaning of the related entities or introduce knowledge not relevant for the relation. Among the most frequent patterns expressing relations, we can find variations of the same original pattern, which differ by the existence of modifiers, coordination, etc. Since these simple patterns have high precision, it is crucial to find a way of making them still more generic to increase coverage. For this purpose, we follow a two-step strategy:

1. Sentence compression: We use a partial grammar that establishes syntactic dependencies between items of extra information (modifiers, adjuncts, punctuation...). The grammar maintains only the dependency Heads and therefore allows us to obtain a sort of simplified linguistic structure.

2. Pattern extraction: We extract LSP, which are then simplified by means of a longest common string algorithm. These simplified patterns are transformed into generic semantic rules and added to our dependency grammar.

The combination of both standard dependency rules and generic semantic rules for RE allows the system to increase coverage without losing precision.

4 Partial Parsing for Sentence Compression

One of the main processes of our strategy attempts to simplify linguistic structures in order to easily extract their information. For this purpose, we use an open-source suite of multilingual syntactic analysis, DepPattern (Gamallo and González, 2011). The suite includes basic grammars for five languages as well as a compiler to build parsers from each one. The parser takes as input the output of a PoS-tagger, in our case, FreeLing (Padró et al., 2010), which also lemmatizes the sentences and performs Named Entity Recognition and Classification (NER/NEC).

The basic grammars of DepPattern contain rules for many types of linguistic phenomena, from noun modification to more complex structures such as apposition or coordination. However, for our simplification task, only some types of dependencies are required, in particular those that compress the sentences maintaining their basic meaning. Following other strategies for sentence compression (Molina et al., 2010), we modified the default grammar by making use of rules that identify the following satellites and subordinate constituents:

- Punctuation (quotation marks, commas, brackets, etc.).
- Common noun and adjective coordination.
- Noun, Adverb, and Adjectival Phrases.
- Prepositional complements, verbal periphrasis and apposition.
- Negative sentences (where the verb inherits the negative tag).

After running the parser, all the Dependents identified by these rules are removed. That is, we obtain a compressed structure without satellites, modifiers, etc. In 1 and 2 we can see two examples of our partial parsing. The elements at the tail of the arrows are the Dependents, while those at the front of the arrows are the Heads.
Nick Cave was born in the small town of Warracknabeal
SpecL AdjnL SpecL CprepR Term
(1)

Nick Cave (born in the town of Warracknabeal)
SpecR CprepR Term
(2)

Taking into account that only the Heads (that are not Dependents) are maintained, the compression process on the two initial sentences will produce an unique simplified structure (note that the Heads of location phrases (“town of NP”, “region of NP”, etc.) inherit the location information provided by the dependent proper nouns, so in the examples, “town” represents a specific location):

<Nick_Cave born in town>

Generic semantic rules are then applied on these structures. For instance:

if a personal name is the Head, a location noun is the Dependent, and the verb “to be born” is a Relator, then a hasBirthplace relation is identified.

This rule can be proposed to cover both the previous examples as well as many others. Moreover, our parsing also prevents from applying the previous extraction rule on sentences such as 3, where the Head of the first Noun Phrase is not the personal name, but a common noun (“son”).

The son of Nick Cave was born in Brazil
SpecL CprepR Term SpecL
(3)

<son born in Brazil>

This way, in this type of sentences (or in negative ones, where the verb has a negative tag), our semantic rule will not extract the incorrect pair “Nick Cave hasBirthplace Brazil” (but we will be able to know the birthplace of “the son of Nick Cave”).

The grammar formalism also allows the parser to maintain the Dependents of a rule after the rule application. Therefore, if we want to add several sets of rules for extracting various relations, the system will only need a single pass over the corpus.

In sum, the sentence compression performed by partial parsing simplifies the linguistic structures maintaining their basic information. This way, the addition of generic semantic rules (converted from LSP) at the end of a dependency grammar allows the parser to increase the coverage of the extraction.

5 Obtaining the Patterns and Rules

This section presents the distant-supervision method for extracting the lexico-syntactic patterns as well as the strategy for generating the generic rules.

5.1 Pattern Extraction

Following the assumption that most instances of a semantic relation are represented by similar LSP, we intend to obtain examples of those patterns and extract from them their original structures (without extra information), then transformed into semantic rules. In order to automate this process, we use the following strategy:

We get a large set of entity pairs of a desired relation from (semi)structured resources. For instance, for the hasBirthplace relation we get pairs from Wikipedia infoboxes (e.g., “Nick Cave - Warracknabeal”, “Josep Guardiola - Sampdor”, etc.). Note that the attributes of many relations are language-dependent (e.g., “Nick Cave hasProfession: English: “singer/songwriter”; Spanish: “cantante/cantautor”; Portuguese: “cantor/cantautor”, etc.), so the use of resources like Freebase is not always feasible. If we do not have a large amount of pairs, we manually introduce a small set of pairs regarding a particular relation.

These pairs are used to select from the unstructured text of Wikipedia sentences that contain both a named entity and an attribute of the relation. If the two terms match a known pair of the initial list, the example is annotated as positive. Otherwise, it is annotated as negative. Note that if we have a large set of pairs, the method does not require bootstrapping. However, if we only have a small set of initial pairs, a bootstrapping process is required (we use this strategy if the number of positive sentences is less than n, where n was empirically set to 200). Then, each selected sentence is tokenized, lemma-
Sentence: Nick Cave was born in the town of War-racknabeal.

Polarity: Nick Cave hasBirthplace War-racknabeal, true.

Pattern: <X be_V born_V in_PRP DA town_N of_PRP Y>

Figure 1: Example of a Sentence with the Polarity label of the related terms and its Pattern (V means verb, DT determiner, PRP preposition and N common noun).

5.2 Pattern Generalization

We use the following method for making generic patterns, then transformed into high-precision rules:

1. First, we take all the patterns of type “X[…Y]” and select the most precise ones according to their confidence value. This value is obtained as follows: we calculate the positive and negative frequencies of each pattern; then we subtract the negative frequency from the positive, and sort the patterns by this value. Finally, the top n most confident patterns are selected (where n = 20 in our experiments). The same process is made for “Y[…]X” patterns.

2. Then, we apply a generalization algorithm for extracting the longest common string (lcs) from these patterns. In order to generalize two patterns, we check first if they are similar and then all those units that they do not share are removed. The similarity, noted \( \text{Dice}_{\text{lcs}} \), between two patterns \( p_1 \) and \( p_2 \) is defined using the longest common string and Dice metric as follows:

\[
\text{Dice}_{\text{lcs}}(p_1, p_2) = \frac{2 \cdot \text{lcs}(p_1, p_2)}{\text{length}(p_1) + \text{length}(p_2)}
\]

where \( \text{lcs}(p_1, p_2) \) is the size of the longest common string between patterns \( p_1 \) and \( p_2 \), and \( \text{length}(p_i) \) represents the size of pattern \( p_i \). It means the similarity between two patterns is a function of their longest common string and their lengths.

After computing the similarity between two patterns \( p_1 \) and \( p_2 \), the \( \text{lcs} \) is extracted if and only if \( p_2 \) is the most similar pattern to \( p_1 \) and the similarity score is higher than a particular threshold (0.75 in our tests). The longest common string of two patterns is considered as the generalized pattern out of them.

3. We filter out those generalized patterns that are not in the best initial 20 patterns, so we automatically obtain a few set of very confident patterns (see Table 1 for an example).

4. All these generic patterns are added as blocks of rules into a grammar, which already has a set of dependency rules for text compression. The new semantic rules take the first entity \( X \) as the Head, and the second one \( Y \) as the Dependent of the relation. This process is made manually.

5. Finally, the grammar is compiled into a parser, which is applied on a corpus to obtain triples “X relation Y”.

Table 1 shows an example of pattern generalization, with the best extracted patterns, the generic one automatically obtained as well an extraction rule.
Extracted Patterns: <X nascer_V em_PRP Y>,
<X nascer_V em_PRP a_DA cidade_N de_PRP Y>,
<X nascer_V em_PRP NP_Fc Y>,
<X Fc_v nascer_V em_PRP a Y>,
<X nascer_V CC residir_V em_PRP Y>,

Generic Pattern: <X nascer_V em_PRP Y>

Table 1: Example of pattern generalization for the "to be born", cidade means “city” and residir, “to live”).

In sum, the application of the longest common string algorithm on the best extracted patterns allows us to obtain a small set of high-quality rules in a weakly-supervised way. These rules, added at the end of a partial dependency grammar, extract instances of pairs belonging to the initial relation.

6 Experiments

We carried out three major experiments in order to know the performance of our RE method. First, we compared the rule-based approach to two baselines in a manually revised corpus containing examples of the relation hasProfession in Spanish. We also compared the performance of the system using a large amount of initial pairs (see Section 5.1) as well as with a small set of seed pairs.

Second, we applied a parser with the obtained extraction rules for the biographical relations hasProfession and hasBirthplace on the whole Spanish and Portuguese Wikipedias.

Finally, we applied the same Portuguese parser on a journalistic corpus, in order to know the performance on the system in different text genres.

6.1 Initial Data

We first obtained about 10,000 pairs for each relation and language (Portuguese and Spanish) from the Wikipedia infoboxes. Then, we identified near 20,000 sentences containing a personal name and (i) an occupation noun (hasProfession) or (ii) a location (hasBirthplace), which were automatically classified as positive or negative using the distant-supervision strategy described in Section 5.1. Finally, we randomly selected two sets of 2,000 sentences for each relation and language as well as a small set of 200 for the relation hasProfession. The latter set was selected for evaluating the use of a small input.

For testing, we randomly selected 1,000 sentences of hasProfession (different from the previous sets), which were manually revised.²

6.2 Results

Our first experiment evaluates the performance of the rule-based method compared to two baselines (in Spanish): Baseline₁ performs a pattern-matching approach applying on the test set the whole positive sentences (except for the proper nouns, replaced by a PoS-tag) from the initial 2,000 set. Baseline₂ uses the 2,000 initial sentences to train a Support Vector Machine classifier, representing each instance with the token_TAG elements as features. For this purpose, we used the WEKA implementation of the SMO algorithm (Witten and Frank, 2005).

To evaluate the rule-based system, we performed two experiments: the first one extracted the rules from the initial 200 sentences (Rule₁, with only 2 extraction rules) while the second one used the 2,000 set of sentences (Rule₂, with 8 rules). The test only contains the 15 most frequent occupations found in the Wikipedia infoboxes, so the evaluation only takes into account the extraction containing these 15 nouns.

Table 2 shows the results of the four described methods over the test set. Precision is the number of correct positive decisions divided by the number of positive decisions (true and false positives). Recall is the number of correct positive decisions divided by the total of positive examples found in the test.

The pattern-matching baseline (Baseline₁) has a precision of 100%, but its f-score is merely 10% due to its low recall values. Baseline₂ performs better, but it produces many false positives, so its precision values do not achieve 45%.

Both rule-based methods perform clearly better than the proposed baselines. Rule₁, with only two generic rules, achieves over 55% recall, maintaining the same precision as the pattern-matching models. The use of more data allowed us to add a set of 8 generic rules, so the Rule₂ method increased its re-

²Both training and testing sets will be available at http://gramatica.usc.es/pln/
Table 2: Precision, Recall and F-score of the Baselines and the two rule-based models for the \textit{hasProfession} relation in Spanish. Test set of 1,000 sentences.

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline_1</td>
<td>100%</td>
<td>5.8%</td>
<td>10.1%</td>
</tr>
<tr>
<td>Baseline_2</td>
<td>44.51%</td>
<td>42.54%</td>
<td>43.5%</td>
</tr>
<tr>
<td>Rule_1</td>
<td>99.02%</td>
<td>55.8%</td>
<td>71.38%</td>
</tr>
<tr>
<td>Rule_2</td>
<td>99.16%</td>
<td>65.2%</td>
<td>78.7%</td>
</tr>
</tbody>
</table>

Table 3: Precision and unique extracted pairs for each relation in the whole Spanish and Portuguese Wikipedias.

<table>
<thead>
<tr>
<th>Language</th>
<th>Relation</th>
<th>Precision</th>
<th>Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish</td>
<td>hasProf.</td>
<td>85.35%</td>
<td>241,323</td>
</tr>
<tr>
<td></td>
<td>hasBirth.</td>
<td>95.56%</td>
<td>13,083</td>
</tr>
<tr>
<td>Portuguese</td>
<td>hasProf.</td>
<td>93.86%</td>
<td>17,281</td>
</tr>
<tr>
<td></td>
<td>hasBirth.</td>
<td>90.34%</td>
<td>5,762</td>
</tr>
</tbody>
</table>

call in more than 10% without losing precision.

Since the test sentences used in these experiments were filtered with a small list of frequent occupation nouns, we performed other extractions in order to know the performance of our system in real text conditions. So we used the \textit{Rule_2} method to parse the whole Spanish and Portuguese Wikipedias. For this purpose, we extracted seven \texttt{hasProfession} rules for Portuguese. Moreover, we add the \texttt{hasBirthplace} rules for each language obtained from the initial 2,000 sets of this relation (four different rules were added for each language). Semantic information obtained from the NEC was used only in those \texttt{hasBirthplace} rules that did not have verb lemmas (such as \texttt{nacer/nascer}, “to be born”).

Before evaluating the extraction in the whole corpora, we automatically remove some noise by eliminating tokens with less than 3 characters or with numbers. \texttt{hasProfession} pairs were filtered with the occupation nouns obtained from the Spanish and Portuguese Wikipedia infoboxes (about 500 and 250, respectively). To evaluate the \texttt{hasBirthplace} relation, the complete output of the extraction was used. We randomly selected and revised samples of 50 pairs from each rule, and calculate the weighted average of the extraction.

Table 3 shows the results of the two extractions over the Spanish and Portuguese Wikipedias. Only a single parsing was performed for each language (with both \texttt{hasProfession} and \texttt{hasBirthplace} extraction rules). Note that the corpora have about 3.2 and 1.8 gigabytes for Spanish and Portuguese, respectively.

In Spanish, almost 241,000 unique pairs of \texttt{hasProfession} related entities were extracted, and more than 13,000 different instances of \texttt{hasBirthplace}. Precision values for the first relation were worse than those obtained in the previous experiment (85% vs 99%). However, a deep evaluation of the errors shows that many of them were produced in previous processing steps (namely the identification of proper nouns), so the precision of these rules is likely to be better. \texttt{hasBirthplace} had better precision results (95%), but the amount of extracted pairs was noticeably lower.

In Portuguese, the system extracted about 17,000 and 5,700 \texttt{hasProfession} and \texttt{hasBirthplace} unique pairs, respectively. The differences between the Portuguese and the Spanish extractions have probably several reasons: on the one hand, the size of the Spanish corpus is almost the double. On the other hand, the number of occupation nouns used as a filter was also half in the Portuguese experiments. However, the extractions in Portuguese maintain high-precision values (90-93%).

Note that both \texttt{hasBirthplace} and \texttt{hasProfession} relations extract biographical data, so it is expected that encyclopedic resources such as Wikipedia contain many instances of these relations. Nevertheless, as we intend to perform extractions on texts of different genres, we applied the same Portuguese parser on a journalistic corpus from Público, a general-purpose Portuguese newspaper (with about 1.2 gigabytes).

In Table 4 we can see the results on the Público newspaper (evaluated in the same way as Wikipedia extractions). The first impression of these data is that the extraction doubles the number of instances with respect to the parsing of Wikipedia (which has a similar size). Precision values are between 6% and 9% lower, achieving 84% in both semantic relations. However, in a quick review of the extracted data, we also noted that many instances were incorrect due to the previous errors cited above.
### Table 4: Precision and unique extracted pairs for each relation in the Portuguese newspaper Público.

<table>
<thead>
<tr>
<th>Relation</th>
<th>Precision</th>
<th>Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>hasProfession</td>
<td>84.54%</td>
<td>41,669</td>
</tr>
<tr>
<td>hasBirthplace</td>
<td>84.67%</td>
<td>11,842</td>
</tr>
</tbody>
</table>

7 Conclusions

This paper presents a novel weakly-supervised approach for semantic Relation Extraction in different languages. We apply a sort of text compression strategy by means of partial dependency parsing which simplifies the linguistic structures, thus allowing the extraction rules to increase their coverage.

In order to (semi)automatically obtain these rules, we first extract lexico-syntactic patterns using a distant-supervision strategy. These patterns are generalized by a longest common string algorithm and finally transformed into semantic rules added at the end of a formal grammar.

Several experiments in different languages and corpora showed that this method maintains the high-precision values of pattern-matching techniques, while the recall is significantly improved.

In future work, we will carry out further experiments with other relations as well as in different corpora. Moreover, we will analyze the performance of the method with different Named Entity Classifiers, in order to avoid some noise during the extraction. Finally, we intend to take advantage of some anaphora and coreference resolution methods that might allow us to extract a large number of instances and to make a fusion process easier.

Acknowledgments

This work has been supported by the MICINN, within the project with reference FFI2010-14986.

References


How to Distinguish a Kidney Theft from a Death Car?
Experiments in Clustering Urban-Legend Texts

Roman Grundkiewicz, Filip Graliński
Adam Mickiewicz University
Faculty of Mathematics and Computer Science
Poznań, Poland
rgrundkiewicz@amu.edu.pl, filipg@amu.edu.pl

Abstract
This paper discusses a system for automatic clustering of urban-legend texts. Urban legend (UL) is a short story set in the present day, believed by its tellers to be true and spreading spontaneously from person to person. A corpus of Polish UL texts was collected from message boards and blogs. Each text was manually assigned to one story type. The aim of the presented system is to reconstruct the manual grouping of texts. It turned out that automatic clustering of UL texts is feasible but it requires techniques different from the ones used for clustering e.g. news articles.

1 Introduction

Urban legend is a short story set in the present day, believed by its tellers to be true and spreading spontaneously from person to person, often including elements of humour, moralizing or horror. Urban legends are a form of modern folklore, just as traditional folk tales were a form of traditional folklore, no wonder they are of great interest to folklorists and other social scientists (Brunvand, 1981). As urban legends (and related rumours) often convey misinformation with regard to controversial issues, they can draw the attention of general public as well1.

The traditional way of collecting urban legends was to interview informants, to tape-record their narrations and to transcribe the recordings (Brunvand, 1981). With the exponential growth of the Internet, more and more urban-legend texts turn up on message boards or blogs and in social media in general. As the web circulation of legends is much easier to tap than the oral one, it becomes feasible to envisage a system for the machine identification and collection of urban-legend texts. In this paper, we discuss the first steps into the creation of such a system. We concentrate on the task of clustering of urban legends, trying to reproduce automatically the results of manual categorisation of urban-legend texts done by folklorists.

In Sec. 2 a corpus of 697 Polish urban-legend texts is presented. The techniques used in pre-processing the corpus texts are discussed in Sec. 3, whereas the clustering process – in Sec. 4. We present the clustering experiment in Sec. 5 and the results – in Sec. 6.

2 Corpus

The corpus of \( N = 697 \) Polish urban-legend texts was manually collected from the Web, mainly from message boards and blogs2. The corpus was not gathered with the experiments of this study in mind, but rather for the purposes of a web-site dedicated to the collection and documentation of the Polish web folklore3. The following techniques were used for the extraction of web pages with urban legends:

1. querying Google search engine with formulaic expressions typical of the genre of urban legends, like \( \text{znajomy znajomego} (= \text{friend of a friend}) \), \( \text{słyszałem taką opowieść} (= \text{I heard this story}) \) (Graliński, 2009),
2. given a particular story type and its examples, querying the search engine with various combinations of the keywords specific for the story type, their synonyms and paraphrases,
3. collecting texts and links submitted by readers of the web-site mentioned above,

---

1Snopes.com, an urban legends reference web-site, is ranked #2,650 worldwide by Alexa.com (as of May 3, 2011), see http://www.alexa.com/siteinfo/snopes.com
2The corpus is available at http://amu.edu.pl/~filipg/uls.tar.gz
3See http://atrapa.net
4. analysing backlinks to the web-site mentioned above (a message board post containing an urban legend post is sometimes accompanied by a reply “it was debunked here: [link]” or similar),

5. collecting urban-legend texts occurring on the same web page as a text found using the methods (1)-(4) – e.g. a substantial number of threads like “tell an interesting story”, “which urban legends do you know?” or similar were found on various message boards.

The web pages containing urban legends were saved and organised with Firefox Scrapbook add-on\(^4\).

Admittedly, method (2) may be favourable to clustering algorithms. This method, however, accounted for about 30% of texts\(^5\) and, what’s more, the synonyms and paraphrases were prepared manually (without using any lexicons), some of them being probably difficult to track by clustering algorithms.

Urban-legend texts were manually categorised into 62 story types, mostly according to (Bruns\(\text{and}, 2002\)) with the exception of a few Polish legends unknown in the United States. The number of texts in each group varied from 1 to 37.

For the purposes of the experiments described in this paper, each urban legend text was manually delimited and marked. Usually a whole message board or blog post was marked, but sometimes (e.g. when an urban legend was just quoted in a longer post) the selection had to be narrowed down to one or two paragraphs. The problems of automatic text delimitation are disregarded in this paper.

Two sample urban-legend texts (both classified as the \textit{kidney theft} story type) translated into English are provided below. The typographical and spelling errors of the original Polish texts are preserved in translation.

Well yeah... the story maybe not too real, but that kids’ organs are being stolen it’s actually true!! More and more crimes of this kind have been reported, for example in the Lodz IKEA there was such a crime, to be more precise a girl was kidnapped from this place where parents leave their kids and go shopping (a mini kids play paradise). Yes the girl was brought back home but without a kidney... She is about 5 years old. And I know that because this girl is my mom’s colleague family... We cannot panic and hide our kids in the corners, or pass on such info via GG [Polish instant messenger] cause no this one’s going believe, but we shold talk about such stuff, just as a warning...

I don’t know if you heard about it or not, but I will tell you one story which happened recently in Koszalin. In this city a very large shopping centre- Forum was opened some time ago. And as you know there are lots of people, commotion in such places. And it so happened that a couple with a kid (a girl, I guess she was 5,6 years old I don’t know exactly) went missing. And you know they searched the shops etc themselves until in the end they called the police. And they thought was kidnapping, they say they waited for an information on a ransom when the girl was found barely alive without a kidney near Forum. Horrible...; It was a shock to me especially cause I live near Koszalin. I’m 15 myself and I have a little sister of a similar age and I don’t know what I’d do if such a thing happened to her... Horrible...\(^6\)

The two texts represents basically the same story of a kidney theft, but they differ considerably in detail and wordings.

A smaller subcorpus of 11 story types and 83 legend texts were used during the development.

Note that the corpus of urban-legend texts can be used for other purposes, e.g. as a story-level paraphrase corpus, as each time a given story is re-told by various people in their own words.

3 Document Representation

For our experiments, we used the standard Vector Space Model (VSM), in which each document

\(^4\)http://amb.vis.ne.jp/mozilla/scrapbook/

\(^5\)The exact percentage is difficult to establish, as information on which particular method led to which text was not saved.

\(^6\)The original Polish text: http://www.samosia.pl/pokaz/341160/jestem_przerazona_jak_mozna_komus_podwedzic_dziecko_i_wyciac_mu_nerke_w_ch
is represented as a vector in a multidimensional space. The selection of terms, each corresponding to a dimension, often depends on a distinctive nature of texts. In this section, we describe some text processing methods, the aim of which is to increase similarity of documents of related topics (i.e., urban legends of the same story type) and decrease similarity of documents of different topics.

3.1 Stop Words

A list of standard Polish stop words (function words, most frequent words etc.) was used during the normalization. The stop list was obtained by combining various resources available on the Internet\(^7\). The final stop list contained 642 words.

We decided to expand the stop list with some domain-specific and non-standard types of words.

3.1.1 Internet Slang Words

Most of the urban-legend texts were taken from message boards, no wonder Internet slang was used in many of them. Therefore the most popular slang abbreviations, emoticons and onomatopoeic expressions (e.g., lol, rotfl, btw, xD, hahaha) were added to the stoplist.

3.1.2 Abstract Verbs

Abstract verbs (i.e., verbs referring to abstract activities, states and concepts rather than to the manipulation of physical objects) seem to be irrelevant for the recognition of the story type of a given legend text. A list of 379 abstract verbs was created automatically using the lexicon of the Polish-English rule-based machine translation system Translatica\(^8\) and taking the verbs with subordinate clauses specified in their valency frames. This way, verbs such as mówić (= say), opowiadać (= tell), pytać (= to ask), decydować (= decide) could be added to the stop list.

3.1.3 Unwanted Adverbs

A list of Polish intensifiers, quantifiers and sentence-level adverbs was taken from the same lexicon as for the abstract verbs. 536 adverbs were added to the stop list in this manner.

3.1.4 Genre-specific Words

Some words are likely to occur in any text of the given domain, regardless of the specific topic. For example in a mathematical text one can expect words like function, theorem, equation, etc. to occur, no matter which topic, i.e. branch of mathematics (algebra, geometry or mathematical analysis), is involved.

Construction of the domain keywords list based on words frequencies in the collection of documents may be insufficient. An external, human knowledge might be used for specifying such words. We decided to add the words specific to the genre of urban legends, such as:

- words expressing family and interpersonal relations, such as znajomy (= friend), kolega (= colleague), kuzyn (= cousin) (urban legends are usually claimed to happen to a friend of a friend, a cousin of a colleague etc.),
- words naming the genre of urban legends and similar genres, e.g., legenda (= legend), anegdota (= anecdote), historia (= story),
- words expressing the notions of authenticity or inauthenticity, e.g., fakt (= fact), autentyczny (= authentic), prawdziwy (= real), as they are crucial for the definition of the genre.

3.2 Spell Checking

As very informal style of communication is common on message boards and even blogs, a large number of typographical and spelling errors were found in the collected urban-legend texts. The Hunspell spell checker was used to find misspelled words and generate lists of correction suggestions. Unfortunately, the order in which Hunspell presents its suggestions is of no significance, and consequently it is not trivial to choose the right correction. We used the observation that it is quite likely for the right correction to occur in the corpus and we simply selected the Hunspell suggestion that is the most frequent in the whole corpus. This simple method turned out to be fast and good enough for our application.

3.3 Lemmatisation and stemming

For the lemmatisation and stemming morfologik-stemming package\(^9\) was used. This tool is based on an extensive lexicon of Polish inflected forms (as Polish is a language of rather complex inflection rules, there is no simple stemming algorithm as effective as Porter’s algorithm for English (Porter, 1980).)

\(^7\)http://www.ranks.nl/stopwords/polish.html
\(^8\)http://poleng.pl/poleng/en/node/597
\(^9\)http://morfologik.blogspot.com/
3.4 Use of Thesaurus

A thesaurus of synonyms and near-synonyms might be used in order to increase the quality of the distance measure between documents. However, in case of polysemous words word-sense disambiguation would be required. As no WSD system for Polish was available we decided to adopt a naive approach of constructing a smaller thesaurus containing only unambiguous words.

As the conversion of diminutives and augmentatives to forms from which they were derived can be regarded as a rather safe normalisation, i.e. there are not many problematic diminutives or augmentatives, such derivations were taken into account during the normalisation. Note that diminutive forms can be created for many Polish words (especially for nouns) and are very common in the colloquial language.

A list of Polish diminutives and augmentatives has been created from a dump of Wiktionary pages. The whole list included above 5.5 thousand sets of words along with their diminutives and augmentatives.

4 Document Clustering

The task of document clustering consists in recognising topics in a document collection and dividing documents according to some similarity measure into $K$ clusters. Representing documents in a multi-dimensional space makes it possible to use well-known general-purpose clustering algorithms.

4.1 Clustering Algorithms

**K-Means** (KM) (Jain et al., 1999; Berkhin, 2002; Manning et al., 2009) is the most widely used flat partitioning clustering algorithm. It seeks to minimise the average squared distances between objects in the same cluster:

$$RSS(K) = \sum_{k=1}^{K} \sum_{\vec{x}_i \in C_k} ||\vec{x}_i - \vec{c}_k||^2$$  \hspace{1cm} (1)

in subsequent iterations until a convergence criterion is met. The $\vec{x}_i$ value means the vector representing the $i$th document from collection, and $\vec{c}_k$ means the centroid of the $k$th cluster. There is, however, no guarantee that the global optimum is reached – the result depends on the selection of initial cluster centres (this issue is discussed in Sec. 4.2).

In all of our tests, K-Means turned out to be less efficient than the algorithm known as **K-Medoids** (KMd). K-Medoids uses medoids (the most centrally located objects of clusters) instead of centroids. This method is more robust to noise and outliers than K-Means. The simplest implementation involves the selection of a medoid as the document closest to the centroid of a given cluster.

We examined also popular agglomerative hierarchical clustering algorithms: **Complete Linkage** (CmpL), **Average Linkage** (AvL), known as UPGMA, and **Weighted Average Linkage** (Jain et al., 1999; Berkhin, 2002; Manning et al., 2009). These algorithms differ in how the distance between clusters is determined: in Complete Linkage it is the maximum distance between two documents included in the two groups being compared, whereas in Average Linkage – the average distance, whereas in the last one, distances are weighted based on the number of documents in each of them. It is often claimed that hierarchical methods produce better partitioning than flat methods (Manning et al., 2009). Other agglomerative hierarchical algorithms with various linkage criteria that we tested (i.e. Single Linkage, Centroid Linkage, Median Linkage and Ward Linkage), were outperformed by the ones described above.

We tested also other types of known clustering algorithms. Divisive hierarchical algorithm Bisecting K-Means and fuzzy algorithms as Fuzzy K-means, Fuzzy K-medoids and K-Harmonic Means were far less satisfactory. Moreover, in the case of fuzzy methods it is difficult to determine the fuzziness coefficient.

4.2 Finding the Optimal Seeds

One of the disadvantages of K-means algorithm is that it heavily depends on the selection of initial centroids. Furthermore, the random selection makes algorithm non-deterministic, which is not always desired. Many methods have been proposed for optimal seeds selection (Peterson et al., 2010).

One of the methods which can be used in order to select good seeds and improve flat clustering algorithms is **K-Means++** (KMpp) (Arthur and Vassilvitskii, 2007). Only the first cluster centre is selected uniformly at random in this method,
each subsequent seed is chosen from among the remaining objects with probability proportional to the second power of its distance to its closest cluster centre. K-Means++ simply extends the standard K-Means algorithm with a more careful seeding schema, hence an analogous K-Medoids++ (KMdp++) algorithm can be easily created. Note that K-Means++ and K-Medoids++ are still non-deterministic.

We propose yet another approach to solving seeding problem, namely centres selection by reduction of similarities (RS). The goal of this technique is to select the $K$-element subset with the highest overall dissimilarity. The reduction of similarities consists in the following steps:

1. Specify the number of initial cluster centres ($K$).
2. Find the most similar pair of documents in the document set.
3. Out of the documents of the selected pair, remove the one with the highest sum of similarities to other documents.
4. If the number of remaining documents equals $K$ then go to step 5, else go to step 2.
5. The remaining documents will be used as initial cluster centres.

The simplest implementation can be based on the similarity matrix of documents. In our experiments reduction of similarities provided a significant improvement in the efficiency of clusters initialisation process (even though the $N - K$ steps need to be performed – for better efficiency, a random sample of data could be used). It may also cause that the outliers will be selected as seeds, so much better results are obtained with combination with the K-Medoids algorithm.

The best result for flat clustering methods in general was obtained with K-Medoids++ algorithm (see Sec. 6), it has to be, however, restarted a number of times to achieve good results and is non-deterministic.

4.3 Cluster Cardinality

Many clustering algorithms require the a priori specification of the number of clusters $K$. Several algorithms and techniques have been created to determine the optimal value of $K$ automatically (Milligan and Cooper, 1985; Likas et al., 2001; Feng and Hamerly, 2007).

For K-means, we can use a heuristic method for choosing $K$ according to the objective function. Define $RSS_{\text{min}}(K)$ as the minimal $RSS$ (see eq. 1) of all clusterings with the $K$ clusters, which can be estimated by applying reduction of similarities technique. The point at which $RSS_{\text{min}}(K)$ graph flattens may indicate the optimal value of $K$.

If we can make the assumption that $RSS_{\text{min}}$ values are obtained through the RS, we can find the flattenings very fast and quite accurate. Moreover, the deterministic feature of the introduced method favours this assumption. Starting from the calculations of the $RSS_{\text{min}}$ value for the largest $K$, we do not have to run the RS technique anew in the each next step. For $K - 1$ it is sufficient to remove only one centroid from $K$ previously selected, thus the significant increase in performance is achieved.

4.4 Evaluation Method

Purity measure is a simple external evaluation method derived from information retrieval (Manning et al., 2009). In order to compute purity, in each cluster the number of documents assigned to the most frequent class in the given cluster is calculated, and then sum of these counts is divided by the number of all documents ($N$):

$$\text{purity}(C, L) = \frac{1}{N} \sum_{k} \max_{j} | C_k \cap L_j |$$

where $L = \{L_1, \ldots, L_m\}$ is the set of the (expected) classes.

The main limitation of purity is that it gives poor results when the number of clusters is different from the real number of valid classes. Purity can give irrelevant values if the classes significantly differ in size and larger ones are divided into smaller clusters. This is because the same class may be recognized as the most numerous in the two or more clusters.

We propose a simple modification to purity that helps to avoid such situations: let each cluster be assigned to the class which is the most frequent in a given cluster and only if this class is the most numerous in this cluster among all the clusters.

Hence, each class is counted only once (for a cluster in which it occurs most frequently) and
some of the clusters can be assigned to no class. This method of evaluation will be called strict purity measure. The value of strict purity is less than or equal to the standard purity calculated on the same partition.

5 Experiment Settings

To measure the distance between two documents $d_i$ and $d_j$ we used the cosine similarity defined as the cosine of the angle between their vectors:

$$\text{sim}_\text{cos}(d_i, d_j) = \frac{d_i \cdot d_j}{\|d_i\|\|d_j\|}$$

(3)

The standard tf-idf weighting scheme was used.

Other types of distance measures and weighting models were considered as well, but preliminary tests showed that this setting is sufficient.

Table 1 presents text normalisations used in the experiment. Natural initial normalisations are $d, ch, m$, i.e.: (1) lowercasing all words (this ensures proper recognition of the words at the beginning of a sentence), (2) spell checking and (3) stemming using Morfologik package. All subsequent normalizations mentioned in this paper will be preceded by this initial sequence.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ss</td>
<td>Cut words after the sixth character</td>
</tr>
<tr>
<td>m</td>
<td>Stemming with Morfologik package</td>
</tr>
<tr>
<td>ch</td>
<td>Spell checking with Hunspell</td>
</tr>
<tr>
<td>d</td>
<td>Lowercase all words</td>
</tr>
<tr>
<td>rs</td>
<td>Remove only simple stop words</td>
</tr>
<tr>
<td>rl</td>
<td>Remove genre specific words</td>
</tr>
<tr>
<td>rc</td>
<td>Remove Polish city names</td>
</tr>
<tr>
<td>p</td>
<td>Remove stop words with abstract verbs, unwanted adverbs and Internet slang words</td>
</tr>
<tr>
<td>t</td>
<td>Use thesaurus to normalise synonyms, diminutives and augmentatives</td>
</tr>
</tbody>
</table>

Table 1: Text normalisations used.

6 Results

We compared a number of seeds selection techniques for the flat clustering algorithms using the smaller sub-corpus of 83 urban-legend texts. The results suggested that the K-Medoids++ and K-Medoids combined with centres selection by reduction of similarities perform better than other methods (see Table 2). The algorithms that are non-deterministic were run five times and the maximum values were taken. The best result for the development sub-corpus was obtained using Average Linking with the normalisation without any thesaurus.

<table>
<thead>
<tr>
<th>Alg.</th>
<th>Purity</th>
<th>$P_{\text{strict}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>KM</td>
<td>0.783</td>
<td>0.675</td>
</tr>
<tr>
<td>KMd</td>
<td>0.831</td>
<td>0.759</td>
</tr>
<tr>
<td>KMpp</td>
<td>0.916</td>
<td>0.904</td>
</tr>
<tr>
<td>KMdpp</td>
<td>0.988</td>
<td>0.976</td>
</tr>
<tr>
<td>KM$_{RS}$</td>
<td>0.807</td>
<td>0.759</td>
</tr>
<tr>
<td>KMd$_{RS}$</td>
<td>0.965</td>
<td>0.918</td>
</tr>
</tbody>
</table>

Table 2: Comparison of flat clustering algorithms using the development set.

Results obtained for the test corpus are presented in Table 3. All tests were performed for the natural number of clusters ($K = 62$). Hierarchical methods proved to be more effective than flat clustering algorithms probably because the former do not seek equal-sized clusters. The best strict purity value (0.825) was achieved for the Average Linkage algorithm with the $p, rl, t, ss$ normalisation. Average Linking was generally better than the other methods and gave results above 0.8 for the simplest text normalisation as well.

For the best result obtained, 22 clusters (35.5%) were correct and 5 clusters (8%) contained one text of an incorrect story type or did not contain one relevant text. For 10 classes (16%) two similar story types were merged into one cluster (e.g. two stories about dead pets: the dead pet in the package and “undead” dead pet). Only one story type was divided into two “pure” clusters (shorter versions of a legend were categorised into a separate group). The worst case was the semen in fast food story type, for which 31 texts were divided into 5 different clusters. A number of singleton clusters with outliers was also formed.

As far as flat algorithms are concerned, K-Medoids++ gave better results, close to the best results obtained with Average Linkage$^{11}$. The

$^{11}$The decrease in the clusters quality after adding some normalisation to K-Medoids++ algorithm, does not necessar-
value of 0.821 was, however, obtained with different normalisations including simple stemming. K-Medoids with reduction of similarities gave worse results but was about four times faster than K-Medoids++.

Both flat and hierarchical algorithms did not manage to handle the correct detection of small classes, although it seems that words unique to each of them could be identified. For example, often merged story types about student exams: a pimp (11 texts) and four students (4 texts) contain words student (= student), profesor (= professor) and egzamin (= exam), but only the former contains words alfons (= pimp), dwója (= failing mark), whereas samochód (= car), kolo (= wheel) and jutro (= tomorrow) occur only in the latter. Similarly, topics fishmac (7) and have you ever caught all the fish? (3) contain word ryba (= fish) but the first one is about McDonald’s burgers and the second one is a police joke. In addition, texts of both story types are very short.

Hierarchical methods produced more singleton clusters (including incorrect clusters), though K-Medoids can also detect true singleton classes as in a willy and dad peeing into a sink. These short legends consist mainly of a dialogue and seem to be dissimilar to others, so they have often been taken as initial centroids in K-MedoidsRS and K-Medoids++. Topics containing texts of similar length are handled better, even if they are very numerous, e.g. what is Your name? (32). But this legend is very simple and has few variants. On the other hand, the popular legend semen in fast food (31) has many variants (as semen is allegedly found in a milkshake, kebab, hamburger, salad etc.).

The results confirm the validity of the proposed text normalisation techniques: better clusters are obtained after removing the non-standard types of words and with a thesaurus including diminutives and augmentatives. Further development of the thesaurus may lead to the increase of the clusters quality.

6.1 Guessing Cluster Cardinality

Fig. 1 presents the estimated minimal average sum of squares as a function of the number of clusters $K$ for K-Medoids with centres selection by the RS

(i.e. minimal values of the $RSS(K)$ for each $K$ is approximated with this technique). The most probably natural number of clusters is 67, which is not much larger than the correct number (62), and the next ones are 76 and 69. It comes as no surprise as for $K = 62$ many classes were incorrectly merged (rather than divided into smaller ones). The most probable number of guessed cluster cardinality would not change a wider range of $K$ than one presented in Fig. 1 if were considered.

7 Conclusions

The clustering of urban-legend texts should be considered harder than e.g. clustering of news articles:

- An urban-legend text of the same story type may take very different forms, sometimes the story is summarised in just one sentence, sometimes it is a detailed retelling.

- Other legends are sometimes alluded to in a text of a given story type.

- The frequency of named entities in urban-legend texts is rather low. City names are sometimes used but taking them into account does not help much, if any (legends are rarely tied to a specific place or city, they usually “happen” where the story-teller lives). Hence it is not possible to base the clustering on named entities like in case of the news clustering (Toda and Kataoka, 2005).
<table>
<thead>
<tr>
<th>Normalization</th>
<th>Words</th>
<th>KMd</th>
<th>KMd_RS</th>
<th>KMdpp</th>
<th>CmpL</th>
<th>AvL</th>
<th>WAvL</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs</td>
<td>7630</td>
<td>0.675</td>
<td>0.732</td>
<td>0.771</td>
<td>0.747</td>
<td>0.806</td>
<td>0.798</td>
</tr>
<tr>
<td>rs,rl</td>
<td>7583</td>
<td>0.694</td>
<td>0.742</td>
<td>0.776</td>
<td>0.772</td>
<td>0.789</td>
<td>0.776</td>
</tr>
<tr>
<td>rs,t</td>
<td>6259</td>
<td>0.699</td>
<td>0.743</td>
<td>0.805</td>
<td>0.779</td>
<td>0.799</td>
<td>0.766</td>
</tr>
<tr>
<td>rs,rl,t</td>
<td>6237</td>
<td>0.688</td>
<td>0.731</td>
<td>0.775</td>
<td>0.818</td>
<td>0.785</td>
<td>0.770</td>
</tr>
<tr>
<td>p,rl</td>
<td>7175</td>
<td>0.698</td>
<td>0.743</td>
<td>0.773</td>
<td>0.77</td>
<td>0.78</td>
<td>0.798</td>
</tr>
<tr>
<td>p,rl,rc</td>
<td>7133</td>
<td>0.697</td>
<td>0.739</td>
<td>0.794</td>
<td>0.773</td>
<td>0.758</td>
<td>0.795</td>
</tr>
<tr>
<td>p,rl,ss</td>
<td>6220</td>
<td>0.684</td>
<td>0.763</td>
<td>0.758</td>
<td>0.750</td>
<td>0.808</td>
<td>0.792</td>
</tr>
<tr>
<td>p,rl,t</td>
<td>5992</td>
<td>0.699</td>
<td>0.732</td>
<td>0.786*</td>
<td>0.806</td>
<td>0.825</td>
<td>0.778</td>
</tr>
<tr>
<td>p,rl,t,rc</td>
<td>5957</td>
<td>0.618</td>
<td>0.731</td>
<td>0.777</td>
<td>0.811</td>
<td>0.824</td>
<td>0.776</td>
</tr>
<tr>
<td>p,rl,t,ss</td>
<td>5366</td>
<td>0.719</td>
<td>0.775</td>
<td>0.821*</td>
<td>0.789</td>
<td>0.813*</td>
<td>0.791</td>
</tr>
<tr>
<td>p,rl,t,rc,ss</td>
<td>5340</td>
<td>0.71</td>
<td>0.786</td>
<td>0.775</td>
<td>0.789</td>
<td>0.811</td>
<td>0.791</td>
</tr>
<tr>
<td>Mean</td>
<td>—</td>
<td>0.689</td>
<td>0.747</td>
<td>0.783</td>
<td>0.782</td>
<td>0.8</td>
<td>0.785</td>
</tr>
</tbody>
</table>

Table 3: Results of clustering urban-legend texts (strict purity) for algorithms: K-Medoids (KMd), K-Medoids++ (KMdpp), K-Medoids with seeds selection (KMd_RS), Complete Linkage (CmpL), Average Linkage (AvL) and Weighted Average Linkage (WA vL). Values with the star sign were obtained with the probabilistic document frequency instead of the idf.

- Some story types include the same motif, e.g. texts of distinct story types used the same motif of laughing paramedics dropping a trolley with a patient.
- Urban legends as texts extracted from the Internet contain a large number of typographical and spelling errors.

Similar problems will be encountered when building a system for discovering new urban-legend texts and story types.

Acknowledgement

The paper is based on research funded by the Polish Ministry of Science and Higher Education (Grant No N N516 480540).

References


Anna D. Peterson, Arka P. Ghosh and Ranjan Maitra. 2010. A systematic evaluation of different methods for initializing the K-means clustering algorithm. Transactions on Knowledge and Data Engineering.


Abstract
The traditional method to evaluate a knowledge extraction system is to measure precision and recall. But this method only partially measures the quality of a knowledge base (KB) as it cannot predict whether a KB is useful or not. One of the ways in which a KB can be useful is if it is able to deduce implicit information from text which standard information extraction techniques cannot extract. We propose a novel, simple evaluation framework called “Machine Reading between the Lines” (MRbtL) which measures the usefulness of extracted KBs by determining how much they can help improve a relation extraction system. In our experiments, we compare two KBs which both have high precision and recall according to annotators who evaluate the knowledge in the KBs independently from any application or context. But, we show that one outperforms the other in terms of MRbtL experiments, as it can accurately deduce more new facts from the output of a relation extractor more accurately. In short, one extracted KB can read between the lines to identify extra information, whereas the other one cannot.

1 Introduction
Evaluating knowledge bases (KBs), and especially extracted KBs can be difficult. Researchers typically measure the accuracy of extracted KBs by measuring precision and recall. But this only partially measures the value of a KB. Size and correctness are important intrinsic measures, but a KB that states “1 is an integer, 2 is an integer, …” contains an infinite number of correct facts, but is not very useful for most tasks. Researchers have proposed a variety of applications as testbeds for evaluating the usefulness of knowledge bases, and the Recognizing Textual Entailment Challenge (Dagan et al., 2006) has received increasing attention as an interesting testbed (Clark et al., 2007). However, evaluating a knowledge base on RTE requires implementing a functioning RTE system, which is in itself a nontrivial task. Furthermore, even if a particular kind of knowledge could be useful for RTE, it may not help improve an RTE system’s score unless all of the other knowledge required for the complex inferences in this task are already present. In short, an effective KB evaluation framework is one that:

- is easy to implement
- is able to measure a KB’s utility on a valued application such as relation extraction

In response, we propose a task called “Machine Reading between the Lines” (MRbtL). In this task, a relation extraction system first extracts a base set of facts from a corpus. An extracted KB is then used to deduce new facts from the output of the relation extractor. The KB is evaluated on the precision and amount of “new” facts that can be inferred.

We also argue that MRbtL evaluation is more rigorous than asking an annotator to evaluate the usefulness of a stand-alone piece of knowledge, because it forces the annotator to consider the application of the knowledge in a specific context. In addition, success on MRbtL provides an immediate benefit to relation extraction, an area which many NLP practitioners care about.

2 Previous Work
The MRbtL task is similar in spirit to the vision of Peñas and Hovy (2010), but they focus on a background knowledge about the names for relations between two nouns. MRbtL also provides a
quantitative evaluation framework for the implicit extractions which is missing from the Pe˜nas and Hovy work. Goyal et al. (2010) present a system for annotating stories with information about whether specific events are positive or negative for characters in the story. Viewed as an MRbtL task, they extract knowledge about whether actions and events cause people to be happy or unhappy, a very specific kind of knowledge. They then implement an inference technique, much more sophisticated than our variable substitution method, for applying this specific extracted knowledge to stories.

3 Machine Reading between the Lines

Let us assume that we have a knowledge base $KB$ and a corpus $C$ from which a Machine Reading or information extraction system has extracted a database of relational extractions ($RE$). In the MRbtL framework, a $KB$ is evaluated on the set of additional ground extractions, called implicit extractions ($IE$) which are entailed by the $KB$: $KB \land RE \models IE$. MRbtL systems are then judged on the precision, amount, and redundancy of $IE$.

By redundancy, we mean the fraction of extracted knowledge in $IE$ that overlaps with $RE$, or is obvious a priori. If $IE$ is accurate and contains many non-redundant extractions, then we judge $KB$ to be a useful knowledge base. The advantage of this setup is that a system’s score on the task depends only on the $KB$, a relation extractor, and a simple logical inference engine for performing variable substitutions and modus ponens. These last two are often freely available or quite cheap to build. Formally, our three evaluation metrics are defined as follows:

- **Accuracy**: $\frac{|\text{Correct Extractions in } IE|}{|IE|}$
- **Amount**: $|IE|$
- **Redundancy**: $\frac{|IE \cap RE|}{|IE|}$

Consider a very simple knowledge base which extracts knowledge about the President of a country. It has an axiom: $\forall pc. \text{president}_o f(p,c) \Rightarrow \text{person}(p) \land \text{country}(c)$. Using this axiom, a relation extraction system can extract $\text{president}_o f(\text{Obama}, \text{USA})$ which would then belong to $RE$. If $RE$ also contains $\text{person} (\text{Obama})$ extracted separately from the same sentence or document, then this extraction would be correct but redundant.

4 Evaluating Two STRIPS KBs with MRbtL

Common-sense knowledge about the changes in the state of the world over time is one of the most crucial forms of knowledge for an intelligent agent, since it informs an agent of the ways in which it can act upon the world. A recent survey of the common-sense knowledge involved in the recognizing textual entailment task demonstrates that knowledge about action and event semantics, in particular, constitutes a major component of the knowledge involved in understanding natural language (LoBue and Yates, 2011). In this section, we describe two example KBs of action and event semantics extracted by our previous work and also discuss an evaluation of these KBs using MRbtL.

We define actions as observable phenomena, or events, that are brought about by rational agents. One of the best-known, and still widely used, representations for action semantics is the STRIPS representation (Fikes and Nilsson, 1971). Formally, a STRIPS representation is a 5-tuple $(a, \text{args}, \text{pre}, \text{add}, \text{del})$ consisting of the action name $a$, a list $\text{args}$ of argument variables that range over the set of objects in the world, and three sets of predicates that reference the argument variables. The first, the precondition list $\text{pre}$, is a set of conditions that must be met in order for the action to be allowed to take place. For instance, in order for someone to awaken, she or he must first be asleep. The other two sets of conditions specify how the world changes when the action takes place: the add list describes the set of new conditions that must be true afterwards (e.g., after the event $\text{insert}(\text{pencil24},\text{sharpener3})$, $\text{in}(\text{pencil24},\text{sharpener3})$ holds true), and the del list specifies the conditions that were true before the action happened but are no longer true. These $\text{add}$ and $\text{del}$ conditions are sometimes collectively referred to as effects or postconditions.

Formally, the precondition, add and delete lists correspond to a set of rules describing the logical consequence of observing an event. Let $t_1$ be the time point immediately preceding an event $e$ with arguments $\text{args}$, $t_2$ the time of event $e$, and $t_3$ the time immediately following $e$. For each precondition $p$, each add effect $a$, and each delete effect $d$, the following rules hold:
∀args_{e}(args_{e}, t_2) \Rightarrow p(args_{p}, t_1)
∀args_{e}(args_{e}, t_2) \Rightarrow a(args_{a}, t_3)
∀args_{e}(args_{e}, t_2) \Rightarrow \neg d(args_{d}, t_3)

where args_{x} represents the subset of the arguments to which the predicate x applies.

4.1 Two extracted STRIPS KBs

We earlier introduced two different KBs that extract preconditions and postconditions (add and delete effects) of actions. One of the KBs (Sil et al., 2010) (henceforth, S10) uses candidate pre and postconditions which have high pointwise mutual information (PMI) with the action words. Given a corpus where each document contains an event e, S10 begins by identifying relations and arguments in a large text corpus using an open-domain semantic role labeler and OpenNLP’s noun-phrase coreference resolution system\(^1\). Taking a set of candidate predicate words, we then define different features of the labeled corpus that measure the proximity in the annotated corpus between a candidate word and the action word. Using a small sample of labeled action words with their correct preconditions and effects, we then train an RBF-kernel Support Vector Machine (SVM) to rank the candidate predicate words by their proximity to the action word.

But, S10 does not generalize adequately e.g. it extracts hammer as a precondition for the action crush. While it is true that if one has a hammer, then one can crush things, this is too strict of a precondition. Hence, we introduce another KB, HYPER (Sil and Yates, 2011), which adds generality to the extractions. HYPER uses Wordnet superclasses as additional candidates (potential pre and postconditions) of actions. Figure 1 shows sample STRIPS extractions from S10 and HYPER.

4.2 MRbtL for S10 and HYPER

We now describe how we can build a MRbtL system for the extracted STRIPS representations. We use the set of predicates and their arguments discovered by the semantic role labeler used by S10 as explicit relational extractions RE; a number of off-the-shelf extractors are available for this purpose. Next, for each occurrence of one of the action words as a predicate in the corpus, we apply the axioms (1) and the knowledge in the S10 and HYPER KBs to deduce predicates that must be true immediately before or after the occurrence of the action. For example, the semantic role labeler discovers the formula draining(a_0, a_1) \land the acid solution(a_1) from the sentence, “This is done by inverting the battery and draining the acid solution out the vent holes in the battery cover”. By applying the extracted precondition that the second argument of a draining event must be a liquid, we can infer that liquid(a_1) is true immediately before the event. Since our MRbtL setup extracts tens of thousands of implicit facts, we evaluate precision and redundancy on samples.

5 Experiments

We perform MRbtL experiments on extractions from S10 and HYPER. S10 uses a dataset of 40 actions from the lexical units in the frames that inherit from the Transitive action frame in FrameNet. The document collection has 15,088 documents downloaded from the Web for the 40 action words. We use the annotated Web corpus for HYPER with semantic role information. We measure the quality of our implicit extractions by taking a random sample of 100 and having two judges classify each extraction for accuracy and redundancy (as per the definitions in Sec 3) in the context of the sentence and document from which it was extracted. As per our earlier work, the pre-
Table 1: The knowledge base extracted by HYPER can identify more, and more accurate, implicit extrac-
tions than S10’s knowledge base, and fewer implicit ex-
tractions overlap with explicit extractions. The first two
columns record the accuracy and redundancy (averaged over
two annotators on sample of 100), and total number of
implicit extractions. \( \kappa \) indicates Cohen’s \( \kappa \) inter-annotator
agreement score, and \( p \)-values for the significance tests are
calculated using a two-sided Fisher’s exact test.

<table>
<thead>
<tr>
<th></th>
<th>S10</th>
<th>HYPER</th>
<th>( \kappa )</th>
<th>signif.</th>
</tr>
</thead>
<tbody>
<tr>
<td>accur.</td>
<td>45%</td>
<td>73%</td>
<td>0.65</td>
<td>( p &lt; 0.01 )</td>
</tr>
<tr>
<td>redund.</td>
<td>21%</td>
<td>12%</td>
<td>0.91</td>
<td>( p = .13 )</td>
</tr>
<tr>
<td>num.</td>
<td>54,300</td>
<td>67,192</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Table 1: The knowledge base extracted by HYPER can identify more, and more accurate, implicit extrac-
tions than S10’s knowledge base, and fewer implicit ex-
tractions overlap with explicit extractions. The first two
columns record the accuracy and redundancy (averaged over
two annotators on sample of 100), and total number of
implicit extractions. \( \kappa \) indicates Cohen’s \( \kappa \) inter-annotator
agreement score, and \( p \)-values for the significance tests are
calculated using a two-sided Fisher’s exact test.

Example extractions indicate that HYPER’s implicit extrac-
tions of S10 is 93% at 85% recall, whereas for HYPER the precision is 89% at 90% recall. At a
first glance, both of the systems look impressive to someone by just looking at the precision/recall
numbers.

Table 1 shows the results of our Machine Read-
ning between the Lines experiment. These extrac-
tions are based on 15,000 occurrences of the 40
action words, but as we scale the extractors to new
action words, we should increasingly be able to read
between the lines of texts. Hence, we ob-
serve that even when both S10 and HYPER re-
port similar (and high) precision and recall, they
report significantly different scores on MRbtL ex-
periments. From Table 1, we clearly see that H-
PER outperforms S10. HYPER’s implicit extrac-
tions are nearly 30% more accurate than S10’s,
and roughly half as redundant. Extrapolating from
the accuracy and redundancy scores in the evalu-
ated sample, HYPER extracts 41,659 correct, non-
redundant relationships compared with 7602 ex-
tractions for S10 from the Web corpus that does
not appear explicitly in the documents.

Example extractions indicate that HYPER’s stron-
ger performance on MRbtL is because its extracted pre and postconditions generalize
to hypernyms. From the sentence “Doctors need to heal patients.”, HYPER extracts medi-
cal.practitioner(doctors) indicating that doctors are of type medical.practitioner which is an ac-
curate and non-redundant extraction. Here, medi-
cal.practitioner is a precondition for action heal.
But S10 concludes that doctors are of type doc-
tor (a Wordnet subclass of medical.practitioner) which is a redundant extraction. Another exam-
ple: from the sentence “When a sharp object, like
a fingernail or thorn, scrupes along your skin . . . ”,
the MRbtL system extracted that the fingernail is
an object, since the instrument of a scraping action
needs to be an object. Both annotators consid-
ered this extraction correct, but redundant, since
the sentence explicitly mentions that a fingernail
is a kind of object.

6 Conclusion and Future Work

We show that the extracted knowledge base can be used to accurately identify information in a docu-
ment that is never stated explicitly. We call this
evaluation scenario “Machine Reading between
the Lines”. We demonstrate that HYPER’s ex-
tracted knowledge base outperforms the closest
comparable one though both perform extremely
well when measured under only precision and re-
call. A future direction includes comparing very
different KBs with MRbtL.

References

Peter Clark, William R. Murray, John Thompson, Phil
Harrison, Jerry Hobbs, and Christiane Fellbaum.
2007. On the role of lexical and world knowledge in
rte3. In Proceedings of the ACL-PASCAL Workshop
on Textual Entailment and Paraphrasing, RTE ’07,
pages 54–59, Morristown, NJ, USA. Association for
Computational Linguistics.

I. Dagan, O. Glickman, and B. Magnini. 2006. The
PASCAL Recognising Textual Entailment Chal-
lenge. Lecture Notes in Computer Science,
3944:177–190.

approach to the application of theorem proving to
208.

Amit Goyal, Ellen Riloff, and Hal Daumé III. 2010.
Automatically producing plot unit representations
for narrative text. In EMNLP.

Peter LoBue and Alexander Yates. 2011. Types of
common-sense knowledge needed for recognizing
textual entailment. In ACL.

Anselmo Pe nas and Eduard Hovy. 2010. Semantic
enrichment of text with background knowledge. In

Avirup Sil and Alexander Yates. 2011. Extract-
ing strips representations of actions and events. In
RANLP.

Avirup Sil, Fei Huang, and Alexander Yates. 2010.
Extracting action and event semantics from web text.
In AAAI Fall Symposium on Common-Sense Knowl-
edge (CSK).
Temporal Expressions Extraction in SMS messages

Stéphanie Weiser  Louise-Amélie Cougnon  Patrick Watrin
CENTAL - Institut Langage et Communication - UCLouvain
1348 Louvain-la-Neuve, Belgium
{stephanie.weiser, louise-amelie.cougnon, patrick.watrin}@uclouvain.be

Abstract
This paper presents a tool for extracting and normalizing temporal expressions in SMS messages in order to automatically fill in an electronic calendar. The extraction process is based on a library of finite-state transducers that identify temporal structures and annotate the components needed for the time normalization task. An initial evaluation puts recall at 0.659 and precision at 0.795.

1 Introduction
In this study, the extraction of temporal information in SMS messages is regarded as a precondition to the development of an application for the construction of a calendar. This application includes the automatic analysis of meetings and the pre-filling of calendar events. We consider the language of temporal expression in SMS messages as a sublanguage which forms a finite subset of the whole language at the syntactic and lexical levels (Harris, 1968).

Most of the recent studies (Beaufort et al., 2010; Kobus et al., 2008; Aw et al., 2006) do not process SMS messages directly. They use a heavy preprocessing step in order to standardize the SMS script. We do not deny the relevance of the transliteration process for complex applications such as SMS messages vocalisation. However, within the framework of our project, we show that, for the extraction of temporal expressions, a normalization phase is not needed, as we tend to simply identify the boundaries of precise and particular surface structures.

Before exploring the extraction task (Section 3), we briefly introduce the corpus used (Section 2). The results of the evaluation we performed are outlined in Section 4, while Section 5 shows the prospects that emerge from this preliminary work.

2 SMS Corpus
2.1 Corpus-based study
The data used for this study is a corpus of 30,000 SMS messages (Fairon et al., 2006a) that were gathered following the strict sms4science collection methodology (Fairon et al., 2006b). sms4science is an international project that promotes the study of a substantial corpus of spontaneous text messages: users are asked to send a copy of text messages that they have already sent to a real addressee in a genuine communication situation. The 30,000 SMS messages corpus that constitutes the raw material for this study was collected in 2004 in the French-speaking part of Belgium; it was semi automatically anonymized and manually normalized1 at the Université catholique de Louvain.

2.2 SMS Script Characteristics
We prefer not to talk about SMS language but about SMS script as it is not a new type of language but a new written practice through a new communication medium (Cougnon and Ledegen, 2010). This new practice shows various specificities, notably it seems to inhibit fear-related behaviour in writing — it erases traditional social, professional and academic demands. The addressee’s physical absence, in addition to the delayed character of the media, encourages SMS users to play with language and to move away from standard language2. At a syntactical level, one would identify some similarities with French oral syntax such as the recurrent lack of ne negation marker and the absence of pronouns at the beginning of sentences. We follow a more nuanced path: it appears that these characteristics

---

1“SMS normalization consists in rewriting an SMS text using a more conventional spelling, in order to make it more readable for a human or for a machine” (Yvon, 2008).

2Standard language can be understood as a graphic and syntactic demand and/or as a register standard.
are not related to the communication medium (oral/written) but to the communication situation (formal/casual) and are more related to register, in a Koch and Österreicher (1985) manner. Inspired by the theory of these authors, we consider there is a continuum between intimacy (Nähesprache) and distance (Distanzsprache) in the SMS communication model (Cougnon and Ledegen, 2010).

In addition to these variations to the norm, SMS script is also strongly influenced by social and regional features which increase the linguistic disparity (as in On va au ciné à soir/au soir/ce soir (in Canada/Belgium/France) - “We’re going to the movies tonight”). Even though these variations are versatile, they form a finite set which can be formalised within a grammar.

3 Extraction of Temporal Expressions

In natural language processing, and particularly in the field of information retrieval and extraction, temporal expressions have been widely studied. The annotation (Pustejovsky et al., 2010; Bittar, 2010) as well as the extraction process (Weiser, 2010; Kevers, 2011) have often been addressed. Indeed, both are needed if we want to compute a date representation from the textual information (Pan and Hobbs, 2006).

These studies offer a wide range of solutions to automatically process temporal information, although they limit their experiments to standard language and don’t take into account language variation. Nevertheless, some texts that could benefit from temporal analysis do not follow the norms of a standard language, notably an important part of CMC (Computer Mediated Communication) like e-mails, chats, social networks. In this study, we intend to determine if the methods used for standard language can be applied to more informal languages with specific characteristics, such as SMS script.

3.1 Typology of Temporal Expression

For an information extraction system, the typology of the data to be extracted is very important. We based this study on the typology developed by Kevers (2011) on standard language in which we selected the categories that are useful for our SMS temporal extraction purpose.

3.1.1 Existing Typology

Kevers (2011) classifies temporal information following four criteria that combine to give 16 categories: punctual or durative, absolute or relative, precise or fuzzy, unique or repetitive. For example, Le 22 octobre 2010 is an expression which is punctual, absolute, precise and unique, whereas Le 22 octobre 2010 vers 20h (that has a different granularity) is punctual, absolute, fuzzy and unique.

This typology is very rich as it includes all types of temporal expressions found in standard (French) written language like dates, durations (du 20 juin au 30 juillet - “from June the 20th to July the 30th”), relative data (le jour d’avant “the previous day”), etc. However, not everything is useful for SMS temporal extraction.

3.1.2 Temporal Expressions to be Extracted for our Application

Our aim is to build an application to identify temporal information related to meetings or events in SMS messages. We do not need to extract past information (like hier - “yesterday” or la semaine dernière - “last week” or other information like dors bien cette nuit - “sleep well tonight”). More than that, as these expressions will serve as triggers for event extraction, the recognition of irrelevant sequences could lead to the identification of “false” candidates.

The information fundamental to this research and this application concerns meetings or events that can take place in an agenda. This is the only criterion that we used to determine the temporal expressions to extract (we call it the “calendar criterion”). The temporal expressions to be extracted can be:

- a time: à 18h - “at 6:00”; de 14h à 18h - “from 2 to 6”
- a date: le 22 octobre - “on October the 22nd”
- a relative moment (day, part of day): aujourd’hui - “today”; maintenant - “now”; mardi - “Tuesday”; mardi prochain - “next Tuesday”; ce soir - “tonight”; dans 5 minutes - “in 5 minutes”
- an implicit expression: à mardi - “see you on Tuesday”; à demain - “see you tomorrow”.

According to the Kevers (2011) classification, the categories that are concerned by SMS messages events planning are PRPU (punctual, relative, precise, unique), DRPU (durative, relative, precise, unique) and PRFU (punctual, relative, fuzzy, unique). 13 categories from the original typology are not taken into account. We created a new category to deal with expressions such as à demain - “see you tomorrow” which imply that “something” will happen the next day. These expressions, which are typical of the dialogues found
in SMS messages, were not dealt with by Kevers as the corpus he studied did not contain dialogues.

3.2 Sublanguage of Temporal Expressions in SMS

The study of Temporal Expressions in SMS messages has lead us to the observation that grammars which have been created for standard language can be applied to a specific sublanguage, at least for the temporal expressions in SMS messages.

3.2.1 Comparison with SMS Script

In order to compare temporal expressions in standard language with those in SMS script, we applied the temporal grammars developed by Kevers (2011) for standard French to the normalized version of an extract of the SMS corpus (1,000 SMS messages) and compared the original SMS form and the normalized form. We found that the syntax remains the same and that only the lexicon changes. A lot of variations are introduced in SMS script, but, concerning the sublanguage of temporal expressions, they only affect the form of the words and not the word order, the syntax or the semantics.

3.2.2 Adaptation of Existing Grammars - Lexical Characteristics

As we have just mentioned, the adaptation of existing grammars to extract temporal information in SMS concerns the lexical level. As SMS messages are well known for their lexical productivity, most of the common words are subject to variation. For example demain (tomorrow) is usually invariable but can take many forms in SMS: 2m1, dem1, dm1, dmain . . . In order to solve this problem we built a specialized lexicon in which each variation (2m1) is linked to a standard lemma (demain), a POS tag (ADV for adverb) and, in some cases, some semantic features (Time): [2m1,demain,ADV+Time].

One may expect the lexicon to require constant updating, as it is intended to capture phenomena that rely on human linguistic creativity, which is potentially boundless. However, this theoretical assumption is refuted by our experiments which show that even if these forms vary consequently, they form a finite lexical set, respecting the closure property of sublanguages (Harris, 1968).

3.2.3 Resources Creation

Using the extracted expressions in normalized SMS messages, we have listed all the forms for all the words that appear in a temporal expression. This has lead to a preliminary dictionary composed of 177 forms, for 55 lemmas. This dictionary still needs to be extended but covers the main temporal expressions variants.

The grammar developed for standard French has been adapted: the invariable words have been lemmatized in order to match the variations listed in our dictionary, the sub-graphs that need to be applied have been selected and new sub-graphs have been created to cover the temporal expressions that are specific to SMS and do not appear in the original grammar (à demain - “see you tomorrow”).

4 Evaluation

We performed an evaluation for the task of temporal expression extraction. We built an evaluation corpus and manually annotated the temporal expressions. Results in terms of precision and recall are provided in Section 4.2.

4.1 Evaluation Corpus

The evaluation corpus is composed of 442 SMS messages containing temporal expressions, following the “calendar criterion”. Some SMS messages contain more than one temporal expression so the total number of temporal expressions is 666.

4.2 Results

For the task of temporal expression extraction, we obtained a recall of 0.659 and a precision of 0.795. Examples of well recognized expressions are, following the classification presented in Section 3.1.2 : N’oubliez pas: ciné Pi [ce soir,.ADV+Time+PRPU} {à 20H,.ADV+Time+PRPU} aux locaux! - “Don’t forget: movie Pi tonight at 8:00 at the office!” (PRPU), cela arrangeait pierre de venir voir asseliane [demain,.ADV+Time+PRPU] {entre 11h et midi,.ADV+Time+DRPU} - “it would suit pierre to come and see asseliane tomorrow between 11:00 and noon” (DRPU), on sera à la maison [vers cinq h trente,.ADV+Time+PRFU] - “we’ll be home around 5:30” (PRFU), à demain - “see you tomorrow” (new category). The reasons behind missing expressions or incomplete annotations are of three types. (i) The format of the expressions was not predicted and is not taken into account by the grammar, e.g. à 8.30 - 9.00; (ii) the variant of a word is missing from the dictionary,
e.g. dimanci for dimanche - “Sunday”; (iii) there is a “mistake” in the SMS, e.g. un peu près 15 minutes instead of à peu près 15 minutes - “about 15 minutes”. The results can easily be improved by working on the first two sources of errors (by extending grammars and dictionaries), while the third source of errors is more problematic, because they are really unpredictable.

5 Conclusion and Future Work

This preliminary study shows that the linguistic specificities of the SMS sublanguage of temporal expressions can be structured in order to eliminate the need for a transliteration process which can lead to errors that are difficult to deal with during the extraction process itself. This study points to numerous opportunities for future work as informal texts, such as informal texts such as SMS but also Tweets, chats, e-mails and Facebook status updates, become increasingly present and contain a lot of information that could be automatically processed.

We intend to apply this research to a calendar application that would find in an SMS all the data about events and time in order to open the calendar on the right date and help the user to fill it in. This approach suggests two complementary steps that we are currently working on:

- **Extracting the event itself**: it implies finding the subject (activity, event), the actants (in SMS, it is mostly the sender and the addressee), the time and place. On a linguistic level, we will try to find out if the properties of the sublanguage (a finite list of graphic and syntactic variations that can be formalized) can also be applied to the different items of events (place, subject, actants).

- **Importing the event in a calendar**: the important task in filling a calendar is to open it on the right date (and time). In order to do this, temporal expressions extracted from the SMS needs to be standardized and formalized in “calendar information” format.

References


Cédrick Fairon, Jean-René Klein, and Sébastien Pauvier. 2006a. Le corpus SMS pour la science. Base de données de 30.000 SMS et logiciels de consultation. CD-Rom, Louvain-la-Neuve, P.U.Louvain, Cahiers du Cental, 3.2.


Bringing Multilingual Information Extraction to the User
(invited talk)

Ralf Steinberger
European Commission Joint Research Centre
Ralf.Steinberger@jrc.ec.europa.eu

Abstract

The speaker will give an overview of how various text mining tools (information extraction, aggregation of multilingual information, document classification, trend analysis, and more) are combined in the Europe Media Monitor (EMM) family of applications to help users in their daily work. EMM was developed by the European Commission’s Joint Research Centre (JRC), whose users include EU Institutions, national EU member state organisations, international organisations such as United Nations sub-organisations, and selected international partners (e.g. in the USA, in Canada and in China). The presentation will thus have an overview character rather than going into much technical detail. EMM applications are publicly accessible at http://emm.newsbrief.eu/overview.html. For scientific details and publications, see http://langtech.jrc.ec.europa.eu/.
Author Index

Chetviorkin, Ilia, 7
Cougnon, Louis-Amélie, 41

Gamallo, Pablo, 21
Garcia, Marcos, 21
Graliński, Filip, 29
Grishman, Ralph, 1, 2
Grundkiewicz, Roman, 29

Klakow, Dietrich, 13

Loukachevitch, Natalia, 7

Min, Bonan, 2

Sil, Avirup, 37
Steinberger, Ralf, 45

Watrin, Patrick, 41
Weiser, Stéphanie, 41
Wiegand, Michael, 13

Yates, Alexander, 37
Organizers:
Kiril Simov (IICT, Bulgarian Academy of Sciences)
Petya Osenova (Sofia University “St. Kl. Ohridski” and IICT Bulgarian Academy of Sciences)
Radovan Garabik (JÚL’Š, Slovak Academy of Sciences)
Jürg Tiedemann (Uppsala University)

Program Committee:
António Branco (University of Lisbon)
Nicoletta Calzolari (Institute of Computational Linguistics of the National Research Council)
Koenraad De Smedt (University of Bergen)
Dan Flickinger (Stanford University)
Dale Gerdemann (University of Tübingen)
Voula Giouli (Institute for Language and Speech Processing)
Silvia Hansen (University of Mainz)
Erhard Hinrichs (University of Tübingen)
Valia Kordoni (University of Saarland)
Vladislav Kubon (Charles University)
Lothar Lemnitzer (Berlin-Brandenburg Academy of Sciences and Humanities)
Preslav Nakov (National University of Singapore)
Cristina Vertan (University of Hamburg)
Eline Westerhout (University of Utrecht)

Invited Speaker:
Preslav Nakov (National University of Singapore)
# Table of Contents

*Reusing Parallel Corpora between Related Languages*
  Preslav Nakov ................................................................. 1

*Discontinuous Constituents: a Problematic Case for Parallel Corpora Annotation and Querying*
  Marilisa Amoia, Kerstin Kunz and Ekaterina Lapshinova-Koltunski ................. 2

*A tagged and aligned corpus for the study of Proper Names in translation*
  Emeline Lecuit, Denis Maurel and Duško Vitas .................................. 11

*Building the multilingual TUT parallel treebank*
  Manuela Sanguinetti and Cristina Bosco ........................................... 19

*Bulgarian-English Parallel Treebank: Word and Semantic Level Alignment*
  Kiril Simov, Petya Osenova, Laska Laskova, Aleksandar Savkov and Stanislava Kancheva ...... 29

*Parallel Corpora in Aspectual Studies of Non-Aspect Languages*
  Maria Stambolieva ........................................................................... 39

*Coreference Annotator - A new annotation tool for aligned bilingual corpora*
  Mara Tsoumari and Georgios Petasis .................................................. 43

*Using Manual and Parallel Aligned Corpora for Machine Translation Services within an On-line Content Management System*
  Cristina Vertan and Monica Gavrila ................................................. 53
Workshop Program

Friday, 16 September 2011

9:20–9:30  Opening

9:30–10:30  Reusing Parallel Corpora between Related Languages
Preslav Nakov

10:30–11:00  Coffee Break

11:00–11:30  Discontinuous Constituents: a Problematic Case for Parallel Corpora Annotation and Querying
Marilisa Amoia, Kerstin Kunz and Ekaterina Lapshinova-Koltunski

11:30–12:00  Parallel Corpora in Aspectual Studies of Non-Aspect Languages
Maria Stambolieva

12:00–12:30  Coreference Annotator - A new annotation tool for aligned bilingual corpora
Mara Tsoumari and Georgios Petasis

12:30–14:00  Lunch

14:00–14:30  A tagged and aligned corpus for the study of Proper Names in translation
Emeline Lecuit, Denis Maurel and Duško Vitas

14:30–15:00  Using Manual and Parallel Aligned Corpora for Machine Translation Services within an On-line Content Management System
Cristina Vertan and Monica Gavrila

15:00–15:30  Building the multilingual TUT parallel treebank
Manuela Sanguinetti and Cristina Bosco

15:30–16:00  Coffee Break

16:00–16:30  Bulgarian-English Parallel Treebank: Word and Semantic Level Alignment
Kiril Simov, Petya Osenova, Laska Laskova, Aleksandar Savkov and Stanislava Kancheva

16:30–16:40  Closing Remarks
Reusing Parallel Corpora between Related Languages
(invited talk)

Preslav Nakov
National University of Singapore
nakov@comp.nus.edu.sg

Abstract

Recent developments in statistical machine translation (SMT), e.g., the availability of efficient implementations of integrated open-source toolkits like Moses, have made it possible to build a prototype system with decent translation quality for any language pair in a few days or even hours. This is so in theory. In practice, doing so requires having a large set of parallel sentence-aligned bilingual texts (a bi-text) for that language pair, which is often unavailable. Large high-quality bi-texts are rare; except for Arabic, Chinese, and some official languages of the European Union (EU), most of the 6,500+ world languages remain resource-poor from an SMT viewpoint. This number is even more striking if we consider language pairs instead of individual languages, e.g., while Arabic and Chinese are among the most resource-rich languages for SMT, the Arabic-Chinese language pair is quite resource-poor. Moreover, even resource-rich language pairs could be poor in bi-texts for a specific domain, e.g., biomedical text, conversational text, etc.

Due to the increasing volume of EU parliament debates and the ever-growing European legislation, the official languages of the EU are especially privileged from an SMT perspective. While this includes “classic SMT languages” such as English and French (which were already resource-rich), and some important international ones like Spanish and Portuguese, many of the rest have a limited number of speakers and were resource-poor until a few years ago. Thus, becoming an official language of the EU has turned out to be an easy recipe for getting resource-rich in bi-texts quickly.

Our aim is to tap the potential of the EU resources so that they can be used by other non-EU languages that are closely related to one or more official languages of the EU.

We propose to use bi-texts for resource-rich language pairs to build better SMT systems for resource-poor pairs by exploiting the similarity between a resource-poor language and a resource-rich one.

We are motivated by the observation that related languages tend to have (1) similar word order and syntax, and, more importantly, (2) overlapping vocabulary, e.g., casa (house) is used in both Spanish and Portuguese; they also have (3) similar spelling. This vocabulary overlap means that the resource-rich auxiliary language can be used as a source of translation options for words that cannot be translated with the resources available for the resource-poor language. In actual text, the vocabulary overlap might extend from individual words to short phrases (especially if the resource-rich languages has been transliterated to look like the resource-poor one), which means that translations of whole phrases could potentially be reused between related languages. Moreover, the vocabulary overlap and the similarity in word order can be used to improve the word alignments for the resource-poor language by biasing the word alignment process with additional sentence pairs from the resource-rich language. We take advantage of all these opportunities: (1) we improve the word alignments for the resource-poor language, (2) we further augment it with additional translation options, and (3) we take care of potential spelling differences through appropriate transliteration.

Speaker’s Bio

Dr. Preslav Nakov is a Research Fellow at the National University of Singapore. He received his PhD in Computer Science from the University of California at Berkeley in 2007. Dr. Nakov’s research interests are in the areas of Web as a corpus, lexical semantics, machine translation, information extraction, and bioinformatics.
Discontinuous Constituents: a Problematic Case for Parallel Corpora Annotation and Querying

Marilisa Amoia, Kerstin Kunz, Ekaterina Lapshinova-Koltunski
Department of Applied Linguistics, Saarland University
{m.amoia,k.kunz,e.lapshinova}@mx.uni-saarland.de

Abstract

In this paper, we discuss some linguistic phenomena that pose potential problems for multilevel linguistic annotation of parallel corpora in general and specifically for data encoding with state-of-art multilevel corpus querying tools such as CQP. We describe the strategy we use for integrating the standard hierarchical XML representation used to annotate such phenomena in our aligned bilingual corpus GECCo into a timeline-based format as used in CQP. Thus, our framework supports efficient multilevel representation as well as corpus exploitation and querying of linguistic data of arbitrary complexity.

1 Introduction

Gathering and providing a natural language corpus of good quality requires the definition of data models that mirror the complexity of natural language data from written as well as spoken discourse. In recent years, much work has been done to develop standards for annotations, annotation schemes and coding practice guidelines (c.f. (McKelvie et al., 2001), (Blache et al., 2010)) with the aim of allowing data exchange between different annotations tools and portability of corpora to other platforms as well as the integration of corpora. Yet, relative little attention has been devoted to interfacing annotation schemes with the encoding formats required by corpus query engines.

Although several efficient automatic systems for parallel corpus exploitation have been developed, these systems are generally specialized for the storage and retrieval of a very limited number of annotation levels. For instance, UNITEX (Pauvier, 2000) only allows alignment on sentence level, and although EMDROS (Petersen, 2004) is a system for storing and retrieving annotated texts that is very generic and applicable to almost any kind of linguistic annotation, it does not allow alignment.

In fact, very few corpus query tools such as CQP (Christ, 1994), ANNI2 (Zeldes et al., 2009) or MATE (McKelvie et al., 2001) exist that can handle multilevel annotated corpora. To our knowledge, ANNI2 is still in the development phase, at the moment of writing, and MATE (McKelvie et al., 2001) does not easily support alignment of parallel corpora.

In this paper, we present our experience with the multilevel query engine CQP developed within the CWB Open Corpus Workbench (Christ, 1994), a collection of open-source tools for managing and querying large text corpora (ranging from 10 million to 2 billion words).

Our focus will be on some problematic issues that have been raised by our attempt to automatically encode our multilevel-annotated bilingual parallel corpus into CQP. The GECCo corpus, which was developed in our research group for the contrastive study of cohesion in English and German combines automatic and manual annotation on different layers of linguistic knowledge ranging from pos-tagging, syntax chunking to semantic information such as linguistic chains and coreference. We noticed that certain annotations were difficult to encode employing state-of-art query tools, namely those representing discontinuous segments.

The paper is structured as follows: Section 2 gives an overview of linguistic phenomena that might lead to discontinuous segments. Section 3 describes the XML-based data format on which multi-layer annotations in the corpus are based. Section 4 deals with the strategy we adopted to encode corpus annotations into CQP and in particular describes the strategy for encoding problematic constituents such as discontinuous segments into a timeline-based data format, as the one used
by CQP, so as to allow corpus querying and exploitation. Section 5 concludes with pointers for future research.

2 Differences in Information Distribution between English and German

This section is concerned with differences in information distribution between English and German as these complicate annotation and exploitation of parallel corpora. Here, structural shifts between originals and translations have turned out to be particularly problematic in view of semi-automatic annotation and querying of translational equivalents or extraction for further processing. They cause discontinuity in cases where the translational equivalents are aligned on the basis of semantic criteria.

2.1 Contrasts between English and German

General differences between English and German such as case marking and word order (see e.g. (Hawkins, 1986), (Koenig and Gast, 2007), (Steiner, 2001) and (Teich, 2003)) are believed to have implications with respect to the positional options for the integration of information into sentences. For instance, (Doherty, 2004) but also (Teich, 2003) and (Steiner and Teich, 2004) note that the order of information is more flexible in English at the beginning of declarative main clauses where more than one constituent may occur before the verb complex. In contrast, German offers more structuring options after the finite verb (in the Mittelfeld) and an additional option behind the non-finite verb (Nachfeld). This is due to the topological peculiarity of the German verbal bracket. Fabricius-Hansen (1999) highlights the tendency of German to structure experiential meaning more vertically and metaphorically in contrast to a more horizontal and congruent distribution of information in English. She indicates "recursive compounding, repeated nominalization, heavy prenuclear and postnuclear noun phrase modification, and accumulation of adverbial adjuncts" (Fabricius-Hansen, 1999) as grammatical features that enhance hierarchical information packaging. In summary, the differences between English and German described above may provoke the following relevant shifts between originals and translation: Meaning that is expressed inside phrases in German may be expressed by a subordinate or main clause or may appear in a separate sentence in English. Meaning that is expressed in a medium sentence position in German may be shifted to the beginning or end of a sentence or be incorporated in a separate sentence in English. As a consequence of these shifts we assume that meaning may be conveyed in English by a contiguous element at one particular position, while corresponding meaning may be realized in German by separate elements in different syntactic positions. We thus expect a higher number of discontinuous segments in German, both at phrase and at clause level. In the following section of this paper some examples of discontinuous segments will be discussed.

2.2 Some examples for discontinuous segments

We now go on to examine some stretches of text from the GECCo corpus in which discontinuous segments are encoded in case of semantic alignment. In German, discontinuous segments at sentence level may be caused by a tendency to encode relevant information in the form of complex appositions in a middle position of the sentence:

(1) a. Dieser Lösung - und das ist für mich das Wunder - haben zum Schluss alle zugestimmt:

b. The miraculous thing for me was that in the end everyone agreed to this solution:

In the German original (1a), relevant and focused information is inserted as a clausal apposition into another sentence, without being related to one specific constituent. The same meaning is expressed in the English translation (1b) by the subject and predicate of the main clause turning the predicate plus arguments of the German main clause in a subordinate clause. The alignment of translational equivalents therefore requires the annotation of a discontinuous segment. Below is an example of a discontinuous segment in German that is not only annotated for alignment of translational equivalents but also for the annotation of coreference.

(2) a. Sehr erfolgreich ist - und das bestätigen mir vor Ort nicht nur sozialdemokratische Kommunalpolitiker - das Förderprogramm InnoRegio.
b. The InnoRegio funding programme has been very successful – something local politicians, and not just Social Democrats, have confirmed.

In the German original (2a), a clausal apposition is again inserted into another main clause. However, the anaphoric pronoun *das* in the apposition refers to the whole main clause. Thus, the latter has to be annotated as discontinuous segment in order to mark it as the antecedent of the pronoun. In the English translation (2b), the apposition is retained but appears after the main clause, without splitting it into two linear parts. Thus only one continuous element needs to be segmented in the translation for the annotation and alignment of the antecedent. Another cause of discontinuous segments on sentence level are prepositional phrases which, are again distributed more freely in German than in English.

(3) a. Dieser Konsens ist trotz aller möglichen Vorbehalte ein hohes politisches Gut.

b. Despite all possible reservations, this consensus is a key political asset.

A prepositional phrase functioning as an adverbial occurs after the predicate (in the Mittelfeld) in German (3a) but is moved to the beginning of the sentence in the English translation (3b). Consequently, alignment of the main clause according to semantic criteria would result in a discontinuous segment in German but not in English.


b. Despite all possible reservations, this consensus is a key political asset which the Foundation “Remembrance, Responsibility and the Future” must preserve on its Board of Trustees. The latter is chaired by Ambassador Dieter Kastrup. Board of Trustees members Michael Jansen, Hans-Otto Bräutigam, and Ambassador Avi Primor were elected the foundation’s executive officers.

In the German excerpt, a prepositional phrase functioning as an adverbial occurs in the Mittelfeld (4a) before the right verbal bracket. The meaning of this PP is realized in a separate sentence in the English translation (4b). A meaning-based alignment of the first English sentence therefore includes a discontinuous element in the German original.

Discontinuous segments on phrase level in German may be due to distinct NP pre-modification conventions (see (Koenig and Gast, 2007), (Doherty, 2004), (Fabricius-Hansen, 1999) and (Teich, 2003)). In contrast to English, merely prepositional phrases and finite relative clauses follow the head noun in German. Constructions of medium complexity are usually placed before the head noun. These contrasts may complicate coreference chaining, on the one hand, and alignment of elements of these chains in the parallel corpora, on the other hand.


b. We look across two rows of warehouses at [the motionless grey surface of the harbor basin and the tongue of land that extends between it and the river]_A2. Enclosed on three sides by water, [this area]_B2 has been a preserve of the chemical industry for as long as anyone can remember.

The German antecedent (A1) and its English translational equivalent (A2) exhibit similar NP structures. At the same time, there are some positional differences between the German anaphor (B1) and the corresponding anaphor in the English translation (B2): While the German noun phrase contains several premodifying elements, the English anaphor only consists of the demonstrative determiner and the head noun. The reason for this is that the non-finite predicate argument construction “auf drei Seiten von Wasser umgebene” inserted between the demonstrative determiner and the nominal head in German could not be realized as premodifier in English. The translator chose to separate it from the rest of the noun phrase.
and transformed it into an adverbial clause functioning as a clausal adverbial at sentence level. Hence, semantic alignment of the English subject "this area" results in the annotation of a discontinuous segment in German, consisting of "diese" and "Gelände". Flexible positioning of complex NP postmodifiers in German may also yield discontinuous segments:

(6) a. This occurred just after I took a turning and found myself on a road curving around the edge of a hill.

b. Dies geschah kurz nach einer Abzweigung, als ich mich plötzlich auf einer Straße befand, die in Kurven an einem Hang entlangführte.

The relative clause occurs after the nominal head in both the English original and its German translation. However, the heavy NP shift enables the German relative clause to be postponed after the predicate. The alignment of the corresponding relative clauses entails annotating a discontinuous element in German.

(7) a. Aber wenn die Notwendigkeit von Reformen besser verstanden wird, als die Bereitschaft verbreitet ist, diese zu unterstützen (...)

b. However, if the awareness of necessary reforms is greater than the willingness to support these reforms (...)

In the example above, the infinitive plus argument postmodifying the NP head "Bereitschaft" occurs after the predicate, while the corresponding infinitive construction appears directly after the NP head "willingness" in English. The alignment of both noun phrases requires the creation of a discontinuous segment in German.

Although we assume that the number of discontinuous segments may be higher in the German than in the English corpus, for the reasons highlighted above, note should be made of the fact that English-German contrasts may also trigger discontinuous elements in the English corpus as illustrated by the following example:

(8) a. What is now clear from the historical evidence of the last century is that in every case where a poor nation has significantly overcome its poverty, this has been achieved while engaging in production for export markets and opening itself to the influx of foreign goods, investment and technology; that is, by participating in globalization.

b. Anhand der historischen Beweise des letzten Jahrhunderts ist jetzt klar, da in jedem Fall, in dem eine arme Nation ihre Armut in beträchtlichem Maße überwunden hat, dies durch die Produktion für Exportmärkte und die eigene Öffnung für ausländische Waren, Investitionen und Technologie geschah - das heißt, durch die Beteiligung an der Globalisierung.

Pseudo-cleft constructions as employed in the example above are a rather frequent strategy in English for realizing clauses as subjects in Theme position (see (Teich, 2003)). Equivalent constructions are relatively rare in German, and indeed, the meaning of the English pseudo-cleft clause is realized as a main clause in the German translation. As a consequence, the complex prepositional phrase of the English pseudo-cleft is moved to the beginning of the sentence in German. An alignment of these two PPs therefore entails the creation of other discontinuous segments in the English original.

Other differences between English and German causing discontinuous segments especially in English may result from the greater availability of non-finite verb constructions or a more verbal realization of meaning in general.

3 Annotation of Parallel Corpora

3.1 GECCo: A Multilingual Parallel Corpus

Our multilingual parallel corpus GECCo, which is an extended version of the CroCo corpus (cf. (Neumann, 2005)), was specifically designed to support contrastive studies of English and German texts as described in the above examples. To our knowledge, it represents one of the few existing resources containing annotation of cohesive devices in parallel multilingual corpora. This type of information plays a crucial role not only in contrastive linguistics and translation studies but also in numerous NLP research areas. Most of the information encoded in the corpus was annotated manually. Further, the corpus includes manual clause alignment.
<table>
<thead>
<tr>
<th>Aligned Clauses</th>
</tr>
</thead>
<tbody>
<tr>
<td>English:</td>
</tr>
<tr>
<td>{when they put it back in} cl:53_EN</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Word Layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>English:</td>
</tr>
</tbody>
</table>
| \(<\text{token id="t310" string="when"/}\>\)
| \(<\text{token id="t311" string="they"/}\>\)
| \(<\text{token id="t312" string="put"/}\>\)
| \(<\text{token id="t313" string="it"/}\>\)
| \(<\text{token id="t314" string="back"/}\>\)
| \(<\text{token id="t315" string="in"/}\>\) |
| German: |
| \(<\text{token id="t326" string="wenn"/}\>\)
| \(<\text{token id="t327" string="sie"/}\>\)
| \(<\text{token id="t328" string="es"/}\>\)
| \(<\text{token id="t329" string="wieder"/}\>\)
| \(<\text{token id="t330" string="einsetzten"/}\>\) |

<table>
<thead>
<tr>
<th>Chunk Layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>English:</td>
</tr>
</tbody>
</table>
| \(<\text{chunk id="ch132" type="conj" gf="conj"/}\>\)
| \(<\text{tok xlink:href="t310"/}\>\)
| \(<\/chunk>\)
| \(<\text{chunk id="ch133" type="np" gf="subj"/}\>\)
| \(<\text{tok xlink:href="t311"/}\>\)
| \(<\/chunk>\)
| \(<\text{chunk id="ch134" type="vp_fin" gf="fin"/}\>\)
| \(<\text{tok xlink:href="t312"/}\>\)
| \(<\text{tok xlink:href="t315"/}\>\)
| \(<\/chunk>\)
| \(<\text{chunk id="ch135" type="np" gf="dobj"/}\>\)
| \(<\text{tok xlink:href="t313"/}\>\)
| \(<\/chunk>\)
| \(<\text{chunk id="ch136" type="advp" gf="adv_loc"/}\>\)
| \(<\text{tok xlink:href="t314"/}\>\)
| \(<\/chunk>\) |
| German: |
| \(<\text{chunk id="ch123" type="conj" gf="conj"/}\>\)
| \(<\text{tok xlink:href="t326"/}\>\)
| \(<\/chunk>\)
| \(<\text{chunk id="ch124" type="np" gf="subj"/}\>\)
| \(<\text{tok xlink:href="t327"/}\>\)
| \(<\/chunk>\)
| \(<\text{chunk id="ch125" type="np" gf="dobj"/}\>\)
| \(<\text{tok xlink:href="t328"/}\>\)
| \(<\/chunk>\)
| \(<\text{chunk id="ch126" type="advp" gf="adv_temp"/}\>\)
| \(<\text{tok xlink:href="t329"/}\>\)
| \(<\/chunk>\)
| \(<\text{chunk id="ch127" type="vp_fin" gf="fin"/}\>\)
| \(<\text{tok xlink:href="t330"/}\>\)
| \(<\/chunk>\) |

Figure 1: Example of Corpus Annotation Layers in GECCo.
For the time being, GECCo contains 10 different registers, i.e. the eight registers of written language of the CroCo corpus and two new registers (interviews and academic discourse) of spoken language (see (Kunz and Koltunski, 2011) for a more detailed description of the GECCo corpus architecture). We are currently trying to enhance the automatic annotation of the new registers by means of manual annotation. Encoding the different layers of manual annotation into CQP, we are faced with the difficulty of encoding discontinuous constituents as illustrated in section 2.

In conclusion we can say that the complexity of linguistic annotations required for studying contrasts in English-German cohesive devices necessitates both

(i) an annotation scheme capable of coping with multilevel annotations, i.e. graph structures and

(ii) a multilevel corpus query engine that can cope with the complexity of our annotation layers and data model.

3.2 Annotation data model

XML is generally considered to be a useful tool for encoding complex structured language data. Indeed, XML is a widely used standard for encoding annotations of natural language corpora. Although the base formalism cannot describe overlapping structures since it was originally designed to represent tree structures only, its extension (Isard and Thompson, 1998) with hyperlinks (href) enables the representation of crossing and overlapping structures.

In our corpus annotation framework we have adopted a modular strategy. Each annotation layer is represented as a different XML file generated by MMAX2 (Müller and Strube, 2006) that supports the manual annotation. The mapping of different representation layers (the graph structure) is guaranteed by the (href) hyperlinks between the different XML files. Figure 1 shows some example annotations from the corpus.

In order to allow further corpus query and exploitation, the linguistic information contained in the XML files needs to be merged into a format readable by a corpus query engine. As this operation is not straightforward in the case of discontinuous segments, an overview of the potential difficulties will be provided in the following section.

4 Interfacing XML Annotations of Discontinuous Segments in CQP

4.1 CQP data model

CQP is based on an XML-like corpus encoding language that is compatible with the data model we use for corpus annotation.

The primary data used in CQP are tokens. The CQP language is a rigid positional system on the token positions, i.e. the tokens are totally ordered, providing a timeline for the incremental encoding of structural attributes. CQP provides annotations of two types of attributes:

- positional attributes: describe features related to the tokens or token position such as part-of-speech, morphological features, etc.
- structural attributes: describe features related to ordered sets of tokens, such as syntactic chunks, clauses, sentences, etc.

CQP allows for incremental information merging, i.e. structural attributes can be sequentially integrated with the positional attributes so as to refine the linguistic information present in the corpus. Figure 2 displays an example of incremental annotation encoding in CQP.

Further, CQP enables the representation of overlapping structures, which is not allowed in standard XML. However, as CQP uses the positions of tokens for storage and retrieval, discontinuous segments cannot be directly represented.

In conclusion we can say that, in order to encode the GECCo corpus annotation data into CQP, the hierarchical XML representation used for encoding multi-layer annotations needs to be translated into the CQP timeline-based corpus representation on the basis of the position of tokens. The next section describes the strategy we employ for encoding discontinuous constituents into CQP.

4.2 Representing discontinuous segments in CQP

As we have seen previously, structural attributes are encoded in CQP as ordered sets of token positions. Thus, a structural attribute \(TAG\) describing an XML tag (e.g. token or chunk) can be defined as the following sequence of token positions:

\[ TAG = [t_1, t_2, \ldots, t_n], \]

with \([1, 2, \ldots, n]\) being a continuous sequence. Therefore, in a \(TAG\) attribute no gaps are allowed.
Step 1: tokens
311: they
312: put
313: it
314: back
315: in

Step 2: morphology
311: they Pro plural
312: put Verb
313: it Pro singular
314: back Adv
315: in Adv

Step 3: syntax

<table>
<thead>
<tr>
<th>np</th>
<th>311: they Pro plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>vp</td>
<td>312: put Verb</td>
</tr>
<tr>
<td></td>
<td>313: it Pro singular</td>
</tr>
<tr>
<td></td>
<td>314: back Adv</td>
</tr>
<tr>
<td></td>
<td>315: in Adv</td>
</tr>
</tbody>
</table>

Figure 2: Merging multi-layer XML annotations into CQP.

In order to describe the strategy used to encode discontinuous segments into CQP, we first give the formal definition of a discontinuous structural attribute.

Let $TAG$ be a sequence of tokens describing the structural attribute represented by an XML tag

$$TAG = [t_1, ..., t_j, ..., t_{j+n}, ..., t_k]$$

and $GAPS$ a set of integer pairs such that

$$(x_i, y_i) \in GAPS$$

 iff 

$$[x_i, x_i+1, ..., y_i]$$

 is a sequence of integer numbers without gaps

Then, the definition of a discontinuous sequence is as follows:

$TAG$ is discontinuous iff

$$| GAPS | > 1$$

As CQP does not support the representation of such discontinuous segments we adopt the following strategy: First, we split a $TAG$ containing gaps into the set of its continuous subsets ($\cup TAG_i$), i.e. sequences of tokens without gaps.

$$TAG = \cup TAG_i, \text{e.i.}$$

$$= \cup[x_i, ..., y_i], \forall (x_i, y_i) \in GAPS$$

Then, after having assigned an identical coin-
dex $gap.id$ to all the subsets of a discontinuous $TAG$, we represent each of them as a standard CQP structural attribute. At the query stage, the segments that have been split are linked together into a unique segment by a query macro that selects structural attributes with the same $gap.id$.

Summing up, the strategy we adopt consists of three steps:

- partitioning the discontinuous segment into a set of continuous subsets,
- representation of the continuous partitions of the original set as standard CQP structural attributes,
- reconstruction of the original discontinuous segment at the query stage.

An example of a structure that cannot be di-
rectly encoded in CQP was given in Figure 1. The English aligned clause contains a discontinuous $vp$ segment representing a finite verb $vp_fin$ (put in).

$$vp_fin = [t312, t315],$$

$$Gap_{vp_fin} = [(312,312),(315,315)]$$

Figure 3 shows the CQP encoding of the continuous subsets of $vp_fin$ defined by $Gap_{vp_fin}$ for this example.

After segment reconstruction, CQP will extract the expected aligned finite verb chunks from the clause-aligned German/English corpus:
Figure 4: CQP representation of alignment in GECCo.

\[
\begin{align*}
vp\_fin\_EN &= [\text{put in}] \\
vp\_fin\_GE &= [\text{einsetzen}]
\end{align*}
\]

In this paper, we have discussed problematic issues that may arise in connection with the automatic encoding of a manually annotated corpus into the multilevel corpus query engine CQP. Manual corpus annotation often produces complexly structured representations of the linguistic information displayed in the corpus that are difficult to encode using general state-of-art corpus query tools.

While much research has addressed the issue of providing annotation standards for linguistic corpora, only a few resources (e.g. ANNIS2 and MATE) exist that provide efficient interfacing of those multi-layer annotations standards with corpus query engines. However, MATE (McKelvie et al., 2001) does not support parallel corpora encoding. ANNIS 2 (Zeldes et al., 2009) for instance provides translation utilities from arbitrary XML data structures to the ANNIS format. The An- nis2 representation format allows the representation and graphs and discontinuous constituents of arbitrary complexity. However, the corpus query language provided by this system is highly complex and requires a high level of expertise on the part of the user.

In this paper we proposed a CQP-based alternative to ANNIS2. We described the strategy we implemented that allows the encoding and querying in CQP of multi-layer parallel corpora that include linguistic phenomena of arbitrary complexity.

Our framework compares well with frameworks such as the one implemented into ANNIS 2 in that it combines all the advantages of the corpus query engine CQP, e.g. efficient querying of very large text corpora, efficient querying of parallel corpora and an intuitive and user-friendly corpus query language, with a framework for encoding arbitrary complex data structures into CQP.
Acknowledgments

The authors thank the DFG (Deutsche Forschungsgemeinschaft) for supporting this project.

References


A tagged and aligned corpus for the study of Proper Names in translation

Emeline Lecuit
LLL
Université François Rabelais
France
emeline.lecuit@univ-tours.fr

Denis Maurel
LI
Université François Rabelais
France
denis.maurel@univ-tours.fr

Duško Vitas
Faculty of mathematics
University of Belgrade
Serbia
vitas@matf.bg.ac.rs

Abstract
In this paper, we propose the creation of a tagged and aligned corpus for the study of a linguistic phenomenon, the translation of proper names. We try to modify the hypothesis according to which proper names cannot be translated and should therefore appear as borrowings in a target-language. To do so, we introduce a parallel multilingual corpus made of eleven versions in ten different languages of a novel. One of these versions, the French one, which appears to be the source-text, undergoes named entity extraction so as to localize more easily the phenomenon we try to study. We focus on the tools used for the creation of our corpus and present some results refuting the idea that proper names are not translatable.

1 Introduction
The idea according to which proper names cannot be translated seems to be unfortunately widely spread. This can lead to big translation mistakes. Nevertheless it can easily be explained by a long tradition of presenting proper names as belonging to a linguistic category often defined using very reductive criteria which seem to have a very long life ahead of them. We have a different opinion and believe that proper names can be translated and are translated more often than people seem to think. We therefore introduce a multilingual corpus which will help us defend our idea. This corpus is created using several NLP tools, including cascade transducers for the extraction of named entities and an alignment tool, for the alignment of the eleven versions of the same text composing our corpus.

This study is therefore a good example both of the creation of an annotated multilingual corpus and of its usability.

In Section 2 we present the text(s) composing our corpus and the problem we try to tackle, giving details about what a proper name can be. In Section 3 we describe the extraction and annotation of the proper names in the French text of our corpus, using the Named Entity extractor CasEN. In section 4, the different steps for the creation of our multilingual corpus, using the alignment tool XAlign, are presented. Section 5 contains some preliminary answers to our translation problem and a conclusion.

2 A corpus for the observation of a linguistic phenomenon
When talking about translating proper names, a French person could argue: “Je m’appelle Paul et mon nom ne change pas si je me rends à Londres”\(^1\), which is correct. But Paul’s plane is going to land in London, and not Londres.

A lot of people believe that proper names are never translated. This idea, though widely spread and defended by many (from Moore\(^2\) to Kleiber, 1981) can be discussed.

Our hypothesis is that proper names are, just like any other linguistic unit, subject to translation processes of all sorts (from borrowing to adaptation through calque and literal translation, etc.) when transferred from a text in source-language to a text in target-language (as demonstrated by Agafonov et al.,

---

\(^1\) “I am called Paul and my name doesn’t change if I go to London.”

\(^2\) See Ballard, 2001
To defend our hypothesis we use a parallel multilingual corpus, built with different versions (i.e. in different languages) of the same novel, *Le Tour du Monde en quatre-vingts jours* (*Around the World in eighty days*), written by the famous French author Jules Verne, in 1872. The choice of this novel amongst others was motivated by two main reasons. Firstly, there exist lots of translations of this novel. Indeed, Verne’s novel was translated in many languages and is nowadays available on the Internet in almost all European languages. Secondly, there is an important number of proper names of all sorts in this novel. This may be due to the fact that the novel deals with the adventures of Phileas Fogg, a rich and enigmatic English character who, after a bet with his fellows from the Reform-Club, has to go around the world in less than 80 days and who therefore travels through many countries and also happens to meet a lot of people. The novel references proper names belonging to almost all the existing categories and sub-categories of proper names.

Proper names can refer to people (or group of people) real or fictitious (we call these proper names anthroponyms), to places (toponyms), to human productions (ergonyms), or to events (pragmonyms). Though the common idea of a proper name is a simple lexical unit (in the form of a family name, for example), proper names can be complex lexical units, composed of several proper names and/or adjectives, common names, etc.

Consider the following examples: *Passepartout* and *l’Institution royale de la Grande-Bretagne* (*the Royal Institution of Great-Britain*), both taken from the novel, though very different in structure, are proper names.

In our corpus, we gather eleven versions of the novel: starting from the original French version. We also have two English versions (by two different translators, at two very different periods and oriented towards two very different audiences3), as well as one version in German, one in Spanish, one in Italian, one in Portuguese, one in Serbian (using a Roman alphabet), one in Bulgarian, one in Polish and one in Greek. This variety of languages allows us to observe the phenomenon on languages belonging to different families. Once the different versions of the text gathered, we need to isolate the units we want to study in the French version of our text and to align the different versions to facilitate the study.

3 Comparing these two versions will show us if the phenomenon of translation of proper names can be affected when these factors vary.

### 3 Annotation of the proper names using CasEN

To have a clearer view of the items we want to study, it seems a good idea to isolate them using a named entity extractor. We decided to use the resource CasEN (Friburger and Maurel, 2004), which uses the tool CasSys, which is now available on the well-known platform Unitex (Paumier, 2006)4. The CasSys system applies a series of finite-state transducers to a text. Each transducer describes a local grammar for the recognition of some entities. The result is a text in which the objects to be studied are marked with indicative tags. The transducer cascade can only be applied to texts which have undergone a preprocessing (division of the text into sentences, tagging using dictionaries, etc.). Only after this first stage the series of transducers can be applied (one after the other, in a defined order) to the text and locate the different contexts that can indicate the presence of the object looked for. In our case, the objects looked for are all kinds of proper names. The transducers we use are extracted from a list of transducers created for the French Ester campaign.

The objects we need to extract are basically persons, organizations and places. Once localized, these objects receive the following tags:

- pers (person)
- pers.hum (human), pers.anim (animal)
- org (organization)
- org.pol (political), org.edu (educational), org.com (commercial), org.non-profit (non-commercial), org.div (media and recreation), org.gsp (administrative)
- loc (location)
- loc.geo (geographical), loc.admi (administrative), loc.line, loc.fac (facilities)
- loc.addr (address), loc.addr.post, loc.addr.tel, loc.addr.elec

4 For more information, see http://tln.li.univ-tours.fr/Tln_CasEN.html. The transducers are available for download from this website.
Let us illustrate the annotation of proper names in our corpus. When we apply the selected transducers (as explained above) to our French text; the input text:

\[
\text{En l'année 1872, la maison portant le numéro 7 de Saville-row, Burlington Gardens - maison dans laquelle Sheridan mourut en 1814 - , était habitée par Phileas Fogg, esq., l'un des membres les plus singuliers et les plus remarqués du Reform-Club de Londres, bien qu'il semblât prendre à tâche de ne rien faire qui pût attirer l'attention.}
\]

becomes the tagged text:

\[
\text{En l'année 1872, la maison portant le numéro 7 de } \text{<ENT type="loc.line">Saville-row</ENT>}, \text{<ENT type="loc.line">Burlington Gardens</ENT> -- maison dans laquelle <ENT type="pers.hum">Sheridan</ENT> mourut en 1814 --, était habitée par <ENT type="pers.hum">Phileas Fogg</ENT>, esq. } \text{<ENT type="org.div">Reform-Club de Londres</ENT>, bien qu'il semblât prendre à tâche de ne rien faire qui pût attirer l'attention.}
\]

where each recognized proper name receives a tag indicating its category and sub-category.

After applying the cascade, a checking was carried out and some corrections were manually made (some tags were expanded or reduced, i.e. the brackets were moved to adjust to the entity, and others were deleted, added, or modified).

This first phase of our work provides us with a tagged text, in which all the proper names can be easily located using simple requests.

Our French version of the text comprises 3415 proper names (519 different). These proper names represent 8.6% of all the characters in the text and 8% of all the words in the text\(^5\).

The following stage consists in aligning all the different versions of our text with the French one and all together.

\(^5\) According to Coates-Stephens (1993), this figure can reach 10% in newspaper articles, which shows the importance of these units in texts and explains our involvement in this subject.
4 Alignement of the texts using XAlign

XAlign is a text aligner developed by the LORIA (2006) and available on the Unitex Platform. It combines the performances of an alignment tool to those of a well-know corpus processing system. One of the advantages offered by XAlign is the possibility to reuse an alignment already existing. This NLP tool allows the treatment of two texts at a time, which means that to obtain our multilingual corpus, we first have to align the texts two by two. In fact, we align the French text with all the other versions individually.

Prior to the alignment each translation is transformed into a TEI format and marked at a sentence, paragraph and division level with respectively <s>, <p> and <div> tags. Id attributes are also added to the texts. All these markers will function as explicit anchor points which will help the alignment of the texts. Other potential anchor points, such as proper names, for example will also help the alignment. The alignment will extract the complete optimum path (following a pre-defined set of transitions, 1:1 equivalence, 1:2 equivalence, 2:1 equivalence, etc.). This alignment is represented as a double window, with one version of the text (in one language) on the left side and the other version of the text (in another language) on the right side. Between these two versions, red lines link the translation equivalents. The alignment is therefore visual and easy to consult (see Paumier and Dimitriu, 2008). Below is an example of alignment.

This alignment will be saved as an alignment file in the XAlign directory in Unitex. The alignment file, in XML format, lists all the “linkings” and “alignments” between the two texts. The linkings correspond to links between two (or more) segments of one of the two versions, when the alignments are of type 1:2 or 2:1, for example, meaning that two segments in one version correspond to one segment in the other version or vice versa. The linking indicated in the alignment file (see Figure 3) means that the two segments will be considered as a whole.

![Figure 2 : Extract of an alignment using XAlign](image)

![Figure 3 : XAlign Alignment file (linking)](image)

The alignments indicate, using the id codes applied during the preprocessing, equivalent segments in the first and second texts (see Figure 4).
Figure 4: XAlign alignment file (alignment)

One of the advantages offered by XAlign is that, because it is hosted by Unitex, it is quite easy to do requests on the texts, thanks to the option “XAlign Locate Pattern”. Another advantage is that the alignment can easily be modified/corrected.

Now that we have created all our bitexts (alignments of the French text with the other versions individually), proofread and corrected them when needed, we can gather all the bitexts in one big multitext (alignment of all the texts). This is easily done manually. We obtain a big table, allowing us to visualize all the equivalent segments of our texts in the different languages. The table in Annexe is an extract of our Multilanguage corpus (It shows the first sentence of the text aligned in the different versions, with the French tagged version on the left side).

This tagged and aligned corpus allows us to carry on our study of proper names in translation.

5 Results and conclusion

We have created a tagged and aligned corpus for the study of a linguistic phenomenon. Tagged corpora and aligned corpora exist. What makes our corpus interesting is the high number of languages represented and the nature of the text used. Indeed, most multilingual corpora are made of versions of law texts (see for example multilingual corpora of the European Union law texts). Vaxelaire (2006) explains that choice of non-literary texts for the study of proper names because in literary texts “tous les types de noms propres peuvent être modifiés […] ou changés par des noms qui ne peuvent être considérés comme des équivalents que dans ce contexte précis[…].”, which can be translated as follows : “ all the types of proper names can be modified […] or changed into names which cannot be considered as equivalents except on this special occasion”. We propose to study a novel. We will therefore study proper names translated by their equivalents but will also meet the case when a proper name is translated with names which cannot be considered as equivalents of translation. The novel we chose is a bit dated but this makes it available and free of use. Moreover, our corpus is extendable. Indeed, there are lots of other versions of the text not considered here which could easily be added to our corpus. We have already mentioned that our corpus is ideal for the study of proper names, since there are many of them in the text and of very various types, as can be observed in the table below (see Figure 5).

<table>
<thead>
<tr>
<th>hypertexts</th>
<th>total number of occurrences</th>
<th>number of different occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>anthroponyms</td>
<td>2079</td>
<td>162</td>
</tr>
<tr>
<td>toponyms</td>
<td>1142</td>
<td>320</td>
</tr>
<tr>
<td>ergonyms</td>
<td>186</td>
<td>31</td>
</tr>
<tr>
<td>pragmonyms</td>
<td>8</td>
<td>6</td>
</tr>
</tbody>
</table>

Figure 5: The proper names in the original version

Our study is still in progress. We only present here figures concerning 10% of the proper names of each of the types presented above, i.e. 16 anthroponyms, 32 toponyms, 3 ergonyms and 1 pragmonym. These samples are made of the most used items in each category. These 52 proper names represent 2029 occurrences in the French version, i.e. about 60% of all the proper names in the text. Figure 6 is a table containing the results for these occurrences.
What we can conclude from the study of these linguistic units is that the wide-spread hypothesis according to which proper names cannot be translated can be discussed.

Indeed, it appears that according to their type (fictitious or real proper names), according to their category (anthroponyms, toponyms, ergonyms, etc.), according to their use (as simple references, or as metaphors for example), according to their construction (simple or complex units), according to the target-language (sometimes implying different morphologic behaviors, sometimes using different alphabets, etc.), proper names can undergo a variety of translation processes.

These phenomena are easily observable thanks to our corpus. Indeed, we can use the French tagged part of our corpus to identify a segment of the text containing a proper name. Then, on the same line of our table, we can visualize the translations of this proper name in the various different languages and analyse them.

If most proper names are simple borrowings from the source-text, as can be seen in Figure 6, many are subject to various assimilation (graphic and/or phonetic), as illustrated in the following example (for complete details about translation processes, see Vinay and Darbelnet, 2004).

Our corpus also highlights the transcription processes (also accounted for in the “assimilation” column of Figure 6), which are not surprising in Bulgarian and Greek, both languages using a non Latin alphabet, but more striking in our Serbian (using a Latin alphabet) version. In this version, Passepartout, the name of the hero’s manservant becomes Paspartu, for instance. Partial or total calques mentioned in Figure 6 (see Figure 8 for an example), mainly concern proper names which Jonasson (1994) described as “mixed” and “descriptive-based” proper names, i.e. composed of “pure” proper names and/or other lexical elements, such as adjectives, common names, etc. The absences of translation, especially in the first English version and in

---

**Figure 6: Translation processes (results)**

<table>
<thead>
<tr>
<th>Target language</th>
<th>Borrowing</th>
<th>Assimilation</th>
<th>Partial or total calque</th>
<th>Absence of translation</th>
<th>Other processes</th>
</tr>
</thead>
<tbody>
<tr>
<td>English (1st version)</td>
<td>69,1%</td>
<td>11,3%</td>
<td>2,2%</td>
<td>12,1%</td>
<td>5,4%</td>
</tr>
<tr>
<td>English (2nd version)</td>
<td>74,2%</td>
<td>13,2%</td>
<td>2,0%</td>
<td>6,6%</td>
<td>4,1%</td>
</tr>
<tr>
<td>German</td>
<td>79,7%</td>
<td>10,6%</td>
<td>3,7%</td>
<td>5,2%</td>
<td>0,6%</td>
</tr>
<tr>
<td>Polish</td>
<td>31,1%</td>
<td>53,4%</td>
<td>4,5%</td>
<td>10,7%</td>
<td>0,2%</td>
</tr>
<tr>
<td>Serbian (Latin alphabet)</td>
<td>4,9%</td>
<td>89,1%</td>
<td>4,5%</td>
<td>0,3%</td>
<td>1,3%</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>0,0%</td>
<td>90,6%</td>
<td>6,3%</td>
<td>2,5%</td>
<td>0,7%</td>
</tr>
<tr>
<td>Greek</td>
<td>0,0%</td>
<td>86,8%</td>
<td>4,6%</td>
<td>3,5%</td>
<td>5,1%</td>
</tr>
<tr>
<td>Italian</td>
<td>72,0%</td>
<td>21,5%</td>
<td>2,6%</td>
<td>3,0%</td>
<td>1,0%</td>
</tr>
<tr>
<td>Portuguese</td>
<td>73,7%</td>
<td>16,0%</td>
<td>5,9%</td>
<td>4,1%</td>
<td>0,2%</td>
</tr>
<tr>
<td>Spanish</td>
<td>51,6%</td>
<td>15,5%</td>
<td>25,1%</td>
<td>7,4%</td>
<td>0,4%</td>
</tr>
</tbody>
</table>

---

**Figure 7: Borrowings with graphic and/or phonetic assimilation**

<table>
<thead>
<tr>
<th>FRA</th>
<th>ENG2</th>
<th>SPA</th>
<th>ITA</th>
</tr>
</thead>
<tbody>
<tr>
<td>{ ENT type&quot;pers.hum&quot; }Mrs. Aouda{/ENT}, ne voulant pas être vue, se rejeta en arrière.</td>
<td>Not wishing to be seen, Mrs Aouda jumped back.</td>
<td>Mistress Aouida, no queriendo ser vista, se echó para atrás.</td>
<td>Mrs Auda, non volendo esser visita, si ritrasse indietro.</td>
</tr>
</tbody>
</table>

---

6 “Mixtes” or “à base descriptives” in the original version.
the Polish version, mainly concern anthroponyms, which are replaced either by pronouns or defined descriptions. The “other processes” are various: transpositions, free translations, to name just a few. The examples below (Figure 9, Figure 10) illustrate some of these translation techniques.

<table>
<thead>
<tr>
<th>FRA</th>
<th>ENG</th>
<th>GREK</th>
<th>POR</th>
<th>POL</th>
<th>SPA</th>
</tr>
</thead>
</table>

**Figure 8: Partial Calques from the English New York (except for the French, borrowing)**

<table>
<thead>
<tr>
<th>FRA</th>
<th>ENG</th>
<th>POL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Je suis un agent de la {ENT type=&quot;org.com&quot;}Compagnie péninsulaire{/ENT}.</td>
<td>I work for P. and O.’</td>
<td>- Jestem agentem Towarzystwa Morskiego Indii Wschodnich.</td>
</tr>
<tr>
<td>Litterally, the Peninsular Company</td>
<td>For Peninsular and Oriental</td>
<td>Literally, the Society Maritime of Oriental India</td>
</tr>
</tbody>
</table>

**Figure 9: Free translation (also called adaptation)**

<table>
<thead>
<tr>
<th>FRA</th>
<th>ENG</th>
<th>BUL</th>
<th>GER</th>
<th>POL</th>
<th>SRP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canal de Suez</td>
<td>Suez Canal</td>
<td>Суецкия канал</td>
<td>Suez-Canal</td>
<td>Kanal Sueski</td>
<td>Suecki kanal</td>
</tr>
</tbody>
</table>

**Figure 10: Transposition (same semantic content but different syntactic structure)**

All these examples, which are just a few of all the examples localized thanks to our corpus, seem to prove that it is wrong to promote the systematic use of borrowings when translating proper names.

**References**


Annexe : Extract of the multitext corpus

Abstract

The paper introduces an ongoing project for the development of a parallel treebank for Italian, English and French annotated in the pure dependency format of the Turin University Treebank, i.e. Parallel–TUT. We hypothesize that the major features of this annotation format can be of some help in addressing the typical issues related to parallel corpora, e.g. alignment at various levels. Therefore, benefitting from the tools previously used for TUT, we applied the TUT format to a multilingual sample set of sentences from the JRC-Acquis Multilingual Parallel Corpus and the whole text of the Universal Declaration of Human Rights.

1 Introduction

Parallel corpora are currently considered among the crucial resources both for a variety of NLP tasks, e.g. machine translation and cross-lingual information extraction, and for research in the field of translation studies and contrastive linguistics with respect to terminology and syntax in particular.

Since the utility of parallel corpora is increased by forms of annotation which make explicit the linguistic knowledge involved in the raw data, parallel treebanks have proved to be valuable resources for a number of purposes (see e.g. (Ahrenberg et al., 2010; Grimes et al., 2010; Rios et al., 2009)). As far as translation studies are concerned, the FuSe project (Cyrus, 2006), for example, aims at studying translation shifts in an English-German corpus annotated with regard to the predicate-argument structure, while the LinEs parallel treebank for Swedish and English (Ahrenberg, 2007) focuses on this aspect by means of complete alignments of segment pairs. As for contributes to the improvement of machine translation quality (both rule-based and statistical), a few examples are provided by SMULTRON (Volk et al., 2010), with a constituency-based parallel treebank for English, German and Swedish; the Prague Czech-English Dependency Treebank (Čmejrek et al., 2004); the Copenhagen Dependency Treebank\(^1\) for Danish, English, German, Italian and Spanish; and the Swedish-Turkish Parallel Treebank (Megyesi et al., 2008).

In this paper, we introduce a new parallel treebank for Italian, English and French, henceforth Parallel–TUT. The annotation schema for this new resource is that of the Turin University Treebank (TUT), which has been applied in a dependency-based treebank used for training of parsing systems and as reference for the evaluation campaigns for Italian parsing. By featuring a rich set of grammatical relations, it shows a representation centered on the predicate-argument structure, a linguistic knowledge that is proximate to semantics and underlies syntax and morphology, essential for the efficient processing of human language. We developed our project also in order to test the hypothesis that this kind of knowledge, and thus the schema representing it, can be useful also in bridging the differences among languages, e.g. in translation.

Therefore, as far as the annotation of the Parallel–TUT corpus is concerned, our approach consists in extending and applying the same tools designed for Italian, within the TUT project, to two other languages, i.e. English and French. The result is the extension of the same format and relations for all the languages of the new parallel corpora, with the same granularity in the representation of the linguistic knowledge. On the one hand, this is motivated by the fact that, as suggested in (Paulussen and Macken, 2010), the use of

\(^1\)http://code.google.com/p/copenhagen-dependency-treebank/
different annotating tools and formats for each monolingual corpus may have a negative impact on the following exploitation and processing of corpora, such as alignment at various levels. On the other hand, the literature shows several examples of application to different languages of formats originally developed for a given language, by using the same features of the native format to address new linguistic phenomena encountered in the other languages. For instance, the format of the Prague Dependency Treebank (PDT), developed for Czech, has been afterwards applied to Arabic (Hajič and Zemánek, 2003), or the Penn Treebank format, which has been applied e.g. to Chinese\(^2\) and Arabic\(^3\). An especially relevant side effect of the application of such kind of methodology consists in increasing the portability across languages of NLP tools and in making available data useful for the comparison and study of models and strategies underlying NLP tools when applied to different languages.

The work presented here aims at going beyond the creation of a parallel treebank where Italian language is included. It aims, in fact, at extending and applying a single treebank schema to other languages, and study how the schema can be meaningfully used to address issues typically related to parallel corpora, e.g. alignment at various levels. The focus of this work is therefore the format of the treebank and the consequence of the application of this format on a parallel corpus.

The remainder of the paper is structured as follows. The next section describes the TUT annotation schema while Section 3 shows the content and size of the corpus on which the schema has been applied. Section 4 describes the annotation process for the three monolingual corpora, while Section 5 shows the alignment issues related to the effects of applying the TUT format to English and French. Finally, we discuss the current state of the project, analyze the future developments of Parallel–TUT and briefly summarize the project.

2 The Turin University Treebank: the resource and its annotation schema

TUT is a resource developed in the last ten years by the Natural Language Processing group of the University of Turin (http://www.di.unito.it/~tutreeb). It currently consists in more than 102,000 annotated tokens (around 3,500 sentences).

The treebank annotation is automatically performed by the Turin University Linguistic Environment (henceforth TULE\(^4\)) (Lesmo et al., 2002; Lesmo, 2007; Lesmo, 2009) and then semi-automatically checked in order to recover errors in the morphological and syntactic annotation. TULE is a rule-based dependency parsing system which includes also the modules needed for tokenization, PoS tagging and morphological analysis, as well as parsing. The parsing module produces a projective dependency tree for each given sentence in input. In the last evaluation campaign for Italian parsing, held in 2009 (Bosco et al., 2009b), TULE achieved the best scores currently at the state of the art (Labelled Attachment Score 88.73), which are very close to the scores known for English parsing.

The core of the treebank is a dependency-based annotation scheme (on which we will focus in this paper), but the resource has been also enriched by the converted versions of all the annotated data in a Penn-like format (Bosco, 2007), in a Combinatory Categorial Grammar format (Bos et al., 2009)\(^5\) and in other constituency-based annotations. This results both in an increased quality of the annotated material and portability of the resource. Beyond allowing the training of parsing systems, TUT has been used as a testbed for evaluation campaigns (Bosco et al., 2007; Bosco et al., 2009a; Bosco et al., 2009b) and analyses of parsing models’ performance with respect to variation in tag sets, paradigms and annotation schemes (Bosco and Lavelli, 2010).

As far as the native annotation schema is concerned, a typical TUT tree shows a pure dependency format centered upon the notion of argument structure and applies the major principles of the Word Grammar theoretical framework (Hudson, 1984). This is mirrored, for instance, in the annotation of Determiners and Prepositions, which are represented in TUT trees as complementizers of Nouns or Verbs. For instance, in figure 1 the tree for the sentence NEWS-355 from TUT, i.e. “L’accordo si è spezzato per tre motivi principali” (The agreement has been broken for three main reasons)\(^6\), shows the features of the an-

\(^2\)See http://www.cis.upenn.edu/~chinese/
\(^3\)See http://www.ircs.upenn.edu/arabic/
\(^4\)http://www.tule.di.unito.it/
\(^5\)http://www.di.unito.it/~tutreeb/CCG-TUT/
\(^6\)English translations of the Italian examples are literal
The frequency of pro–drop varies from 0.17 to 0.64 times per sentence according to the text genre included in the TUT corpora, where this phenomenon occurs an average of 0.28 times per sentence. On the one hand, advantage in using null elements in the annotation is that they permit dependency trees to be without crossing edges and projective structures also for non–projective sentences. On the other hand, by using null elements it is possible to give an explicit representation also of parts of the argument structure that can be missing, but crucial for some task. For instance, in machine translation, if the source language allows argument deletion and the target language does not, in order to make possible for the system to handle the translation, it is crucial that in the source language the dropped argument is explicitly marked. An alike situation can happen in a translation from Italian to English or French, where, on the contrary, the subject is always lexically realized in tensed clauses.

For what concerns the dependency relations that label the tree edges, TUT exploits a rich set of grammatical items designed to represent a variety of linguistic information according to three different perspectives, i.e. morphology, functional syntax and semantics. The main idea is that a single layer, the one describing the relations between words, can represent linguistic knowledge that is proximate to semantics and underlies syntax and morphology, i.e. the predicate-argument structure of events and states, which has proven essential for efficient processing of human language. Therefore, each relation label can in principle include three components, i.e. morpho-syntactic, functional-syntactic and syntactic-semantic, but can be made more or less specialized, including from only one (i.e. the functional-syntactic) to three of them (see e.g. (Bosco and Lavelli, 2010) for more details). For instance, the relation used for the annotation of the Prepositional modifiers in figure 1, i.e. PREP-RMOD-REASONCAUSE (which includes all the three components), can be reduced to PREP-RMOD (which includes only the first two components) or to RMOD (which includes only the functional-syntactic component). This variable degree of specificity is a useful means for the human annotator in that it meets his/her different degree of confidence about a given relation. Moreover, it can also be applied in particular tasks in order to increase the comparability of TUT with other existing resources, by exploiting the amount of linguistic information more adequate for the comparison, e.g. in terms of number of relations.

Last but not least, as Italian requires, the TUT format provides an extended morphological tag set including all the categories and features needed to describe morphologically rich languages. This tag set allowed therefore for an accurate description both for French, whose morphological richness resembles that of Italian, and English, which is morphologically poorer.

Observing related works, we think that the TUT schema can be a good candidate for the development of a parallel treebank for various reasons. First of all, it is oriented to the representation of the predicate-argument structure, a kind of information that can be useful as a pivot for...
the alignment in translation, but is also crucial
in tasks such as Information Extraction. As ob-
served above, both the dependency core and the
inventory of null elements introduced in the an-
notation schema of TUT contribute to a more ac-
curate representation under this respect. Second,
this schema gives the means for the development
of annotations at various degrees of specificity of
grammatical relations, thus extending the compa-
rability and compatibility with other existing re-
sources. Finally, another aspect to be taken into
account is the availability of automatic tools for
the conversion of the native TUT format in other
constituency-based representations, among which
the most known and used format in the world (i.e.
that of the Penn Treebank), and in a Combinato-
ry Categorial Grammar format too, which is a seman-
tic-oriented representation.
In the next sections we describe the parallel corpus
on which we have applied the TUT format for the
development of the Parallel-TUT.

3 The data in the Parallel–TUT

The Parallel–TUT currently comprises a small set
of sample texts, which have been annotated in
order to assess our methodology and test our hy-
pothesis. They are organized in two sub-corpora,
as outlined in Table 1.

The first sub-corpus consists of about 50 sen-
tences extracted from the JRC-Acquis multiling-
al parallel corpus\(^8\) (Steinberger et al., 2006)
for each of the three languages involved in the
Parallel–TUT. In particular, the sentences for Ital-
ian are shared by TUT and the corpus used within
the French parsing evaluation campaign Passage\(^9\),
respectively in Italian version annotated in the
TUT format, and in French version annotated in
the EASY format. The English counterpart
of the corpus was retrieved from English section
of the TUT corpus. We will refer
to these data as JRCAcquis–ITA, JRCAcquis–
FR and JRCAcquis–EN, respectively for Italian,
French and English.
The second sub-corpus, which will be referred
as UDHR–ITA, UDHR–FR and UDHR–EN, in-
cludes the entire text of the Universal Declaration
of Human Rights, as available in the official Web

\(^8\)See http://langtech.jrc.it/JRC-Acquis.
\(^9\)http://optima.jrc.it/Acquis/
\(^{10}\)See http://www.ohchr.org/EN/UDHR/Pages/
SearchByLang.aspx
\(^{11}\)Around the 30% of TUT data are extracted from legal
texts, i.e. the Codice Civile and the Costituzione Italiana.
\(^{12}\)Available from the TUT Web page at
http://www.di.unito.it/~tutreeb/ (EUDIR
Section)

page of the UN Office of the High Commissioner
of Human Rights\(^{10}\), and consists of about 76 sen-
tences for each language.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>sentences</th>
<th>tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>JRCAcquis–ITA</td>
<td>50</td>
<td>2,205</td>
</tr>
<tr>
<td>JRCAcquis–FR</td>
<td>52</td>
<td>2,297</td>
</tr>
<tr>
<td>JRCAcquis–EN</td>
<td>50</td>
<td>1,895</td>
</tr>
<tr>
<td>UDHR–ITA</td>
<td>76</td>
<td>2,387</td>
</tr>
<tr>
<td>UDHR–FR</td>
<td>77</td>
<td>2,537</td>
</tr>
<tr>
<td>UDHR–EN</td>
<td>77</td>
<td>2,293</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td><strong>382</strong></td>
<td><strong>13,614</strong></td>
</tr>
</tbody>
</table>

Table 1: Corpus overview.

For what concerns the texts of the JRCAcquis
corpus in particular, they were selected because
of their availability in two different annotation
formats developed by two independent research
groups, as mentioned above. Moreover, choosing
texts from legal documents, we benefitted from the
expertise in the field of legal language processing
acquired within the TUT project\(^{11}\). Last but not
least, the data included in our corpus are repre-
sentative of the development of raw text parallel
corpora developed in the last decades, e.g. from
the European Community. Nevertheless, we know
that analyses based on such kind of unbalanced
material may lead to misleading results if applied
in general context, as the syntax in this corpus is
typical of a quite particular kind of documents.
This will be taken into account in the further de-
velopment of our corpus.

In general, our selection of texts includes raw ma-
terials which are in translation relation to each
other, and free of Intellectual Property Rights
problems, which allows us to release treebank data
under an open license.

4 Treebank Development

Except for the Italian part of the first sub-corpus
of the Parallel–TUT, i.e. JRCAcquis–ITA (which
was already available in the annotated version\(^{12}\)
as described above), for the English and French
counterparts, as well as for the entire second

\(^{12}\)Available from the TUT Web page at
http://www.di.unito.it/~tutreeb/ (EUDIR
Section)
sub-corpus (UDHR), we processed the texts following the same strategies applied in the TUT project and using the same tools both for parsing and checking.

Being the original materials in XML format (eg. texts collected in the JRC multilingual corpus) or directly extracted from a Web page, the first step was to clean up files from noisy data (eg. markups) and to convert them to plain text files with UTF-8 encoding. In this way, texts can be exploited for our further linguistic analyses.

Despite other parallel treebanks, where monolingual corpora were processed independently with different tools (cf. (Megyesi et al., 2008)), or created from already existing monolingual treebanks (cf.(Klyueva and Mareček, 2010)), the texts of our collection were analyzed from scratch with the same tool, i.e. TULE. Although TULE supports in principle linguistic analysis in several languages (English in particular, but also French, Spanish, Catalan and Hindi), its output quality achieves satisfactory results mostly for Italian, since it has been extensively tested in the development of the Italian treebank TUT. Since TULE is a rule-based parser, the annotation phase for English and French therefore entailed alternating steps of rules insertion in TULE and automatic analysis, until an acceptable output was produced. Rule-insertion steps included mainly the enrichment of lexical knowledge, e.g. insertion of new lexical entries (including proper nouns, named entities, compounds and locutions), modifications in the suffix tables, and new disambiguation rules for linguistic phenomena previously unseen in Italian. A typical example of such phenomena is the English genitive for regular plural nouns (‘-s’). Since in Italian (and French too) the apostrophe is normally considered a graphic sign indicating an elision, during the automatic analysis, tokenization in particular, it is kept attached to the previous token. The English possessive case, however, is normally isolated and treated as a single token. Its recognition in this form by the TULE tokenizer has therefore requested the integration of a new condition in the set of disambiguation rules. Other types of intervention focused on the syntactic representation of those phenomena that distinguish the two languages from Italian. For example, the French superlatives formed by the definite article and *plus/minus* follow a word order which is quite different from that of the Italian superlatives: it was therefore necessary to modify the representation scheme already present in the TUT annotation guidelines for Italian. The treatment of the expletive subject (ie. a purely syntactic subject, not semantically realized), which is a common occurrence both in English and French, but not in Italian (where, as we said, the subject can be omitted) also required the inclusion of additional labels in the annotation schema.

The whole procedure above described had a twofold goal: to improve the output quality of TULE for English and French, and, as a result, to reduce to a feasible extent manual intervention of human annotators in future annotation work.

Because of the current small size of the corpus and the consequently limited training on English and French of our tools, we expect that a considerable amount of manual intervention (eg. enriching the knowledge base of the parsing system) will be necessary also in the next step of the development of our parallel treebank. In fact, the variety of new syntactic structures encountered so far in English and French data is quite small, and the probability that the treebank could miss some syntactic phenomena is high.

The relatively lower quality of the output of TULE for English and French with respect to Italian (as reported in Section 6) made the final stage of manual correction crucial to verify that linguistic phenomena were annotated appropriately and consistently. In this stage, the same tools used in the development of TUT were exploited. For instance, for displaying the dependency trees, the viewer TULETUT Java graphical interface was used, thus allowing the observation of the structures in a more readable graphic form.

It is known that the conversion of dependency trees into phrase structures is in itself a comparative test of the adequateness of the involved representation formats with reference to the features of the language and the quality and consistency of annotation (Musillo and Sima’an, 2002). Therefore, some preliminary experiment was also performed by applying to the English and French data the procedures for the conversion in the Penn Treebank format developed for Italian. The results are promising in particular for English, as we expected, since this is the reference language for the
Penn format. For French the conversion should be further refined by including in the Penn format the representation of particular phenomena. As far as the annotation phase of the Parallel–TUT is concerned, it can be currently considered as concluded and the corpus will be soon released and made available for research purpose. In the next section, we describe the alignment phase which is the less advanced part of the project, currently under development.

5 Aligning the Parallel–TUT

Several techniques have been developed and made available for aligning texts at various granularities. They vary from document-structure to sentence, word, phrases or dependency subtrees (see e.g. (Wu, 2010; Li et al., 2010)). Each level implies several and different issues that are currently in part unresolved also because does not exist an objective and universally shared notion of correspondence between sentence units. For instance, it is difficult to decide which words in a given target string correspond to which words in its source string (especially where idiomatic expressions are involved) and often, an alignment includes effects such as reorderings, omissions, insertions (Och and Ney, 2003). Moreover, tools implementing alignment techniques are often designed with reference to some particular kind of annotation and schema, and cannot be applied to different formats, such as TUT. This is currently the major limit of the project that should be addressed in the next future. In fact, even if in our project we are interested in the alignment at various levels, we applied until now only some preliminary form of alignment, and the most of the time devoted to this part of the Parallel–TUT project has been spent in the analysis and report of the issues raised by our data. First of all, the Parallel-TUT has been developed taking into account the issues related to the alignment at sentence and word level. Therefore, after the linguistic annotation, a further step has been the detection of lexical and structural correspondences between language pairs. As for the sentence level, the alignment was performed with Omega Aligner,13, a simple Python script used for the alignment of translation units within Computer Aided Translation (CAT) systems. The files produced conform to the Translation Memory eX-

13http://www.omegat.org/en/resources.html
6 Discussion and future work

In this section we discuss the implications of applying the TUT format to English and French for the development of Parallel–TUT. The first aspect we focused on, while evaluating our methodology and its effects, was the parser output, the type of errors produced and their investigation.

After the work phase described in Section 4, TULE, when evaluated according to its precision in building and labelling dependency trees, reached an error rate of around 9% for Italian, but 15.6% for English and 17.8% for French. Errors detected during manual correction mainly dealt with tokenization and, to a larger extent, morphological analysis and Part of Speech tagging. This is maybe due to an incorrect application of disambiguation rules by the parser or to a lack of information about the lexical items in the TULE dictionary. As a result, these errors deeply affected the parser performance, and, despite rule-insertion operations, its output quality for English and French languages is still lower if compared to Italian. This suggests that further improvements in the system are required.

In addition to these errors, two other types have been identified. For their special character, we could define them as “language-dependent” and “genre-dependent” errors. In the first case, errors have to do with the distinctive feature of each language. The most frequent phenomenon (among those encountered in our corpus) included
in the former is that of the pre-modification in English, i.e. all those cases of noun phrases where one or more units preceding the head of the phrase are syntactic modifiers of the head itself\footnote{This can be noticed, in the annotated texts, by the higher frequency of nominal modifiers (expressed by the NOUN-RMOD label) in English texts, rather than in the French and Italian sub-parts of the corpus; the occurrences of such relation are 103 in English texts, 25 in the French and 17 in the Italian ones, covering respectively 2.5%, 0.5% and 0.4% of the total amount of relational labels.} structured in a hierarchic order. Since Italian language prefers post-modification, a parser trained for such linguistic patterns, in most cases, is unable to recognize the appropriate syntactic order between the units of the pre-modification.

As for the second type of errors, defined here as “genre-dependent”, we include all those cases of errors directly attributable to the genre of the texts collected and analyzed in our small corpus. As we said, the collection comprises legal documents, where the recurrence of complex and ambiguous syntactic constructions (a feature shared by the three languages considered) is quite common. The high number of embedded prepositional phrases, subordinate clauses and parentheticals contributes to the lowering of the output quality.

As for the application of the TUT format and schema to the other two languages, distinctive features of these linguistic systems result in a lack of an appropriate structural representation, for which new relational labels were introduced, as described in Section 4. We tackled this problem with the two-fold goal of providing a coherent framework of annotation (like for Italian\footnote{See the linguistic notes of TUT at http://www.di.unito.it/tutreeb/documents/noteling-engl-15-11-08.pdf}, and taking into account the linguistic peculiarities of each language. This was made possible by a number of factors. First, the choice of a dependency (rather than constituent) structure better suits for both morphologically rich languages (such as Italian and French) and morphologically simpler ones (English). Moreover, the richness of relations provided in the TUT scheme, in addition to the use of null elements, which is another feature of the TUT format, allows a flexible annotation and the coverage of those linguistic phenomena which distinguish French and English from Italian (to name a few, the relative superlative in French, or the possessive case in English, as already mentioned in Section 4).

As said at the beginning, the Parallel–TUT is currently an ongoing project, and the aim of the present work is mainly at raising and investigating issues related to its development. Nevertheless, in this phase of our project we observed that using the same format, and the TUT format in particular, has proved useful in the detection of similarities during the alignment phase at all the levels currently taken into account. The decision to adopt the same annotation scheme and grammatical description for the three languages can also contribute to the comparison of grammatical patterns.

As for future development of this work, a number of issues must be further pursued.

First of all, by taking into account the directions collected in the alignment guidelines developed during this first phase of the Parallel–TUT project, we will address the development and the integration of suitable tools, in particular for the alignment at the predicative structure level and for displaying such kind of information.

Secondly, considering the opportunity of converting TUT into a Penn-like format, we can extend the conversion to our parallel treebank as well, in order to develop alignment procedures also for phrases and information expressed in constituency-based formats.

Thirdly, in order to address the languages involved beyond the limits of a toy domain, it is crucial to enlarge the corpus of the Parallel–TUT. On the one hand, applying to a larger corpus our methodology to a larger corpus will give us the opportunity for addressing a larger and more meaningful set of linguistic phenomena typical of French and English, though not represented in Italian. On the other hand, this will allow more detailed analyses, like e.g. in (Ahrenberg, 2010), not affected by the sparseness of data that can be currently detected using our small corpus.

Finally, we observe that currently our corpus covers a selection of texts from a specific linguistic subfield broadly corresponding to legal language; one of the main future tasks should therefore consist not only in extending the size of the annotated corpus, but also in orienting to a more balanced direction its further development, comprising different sources, e.g. technical and specialized texts, fiction, newspapers (Paulussen and Macken, 2010).
7 Conclusions

In this paper we presented preliminary results in the creation of Parallel–TUT, a multilingual parallel treebank for Italian, English and French represented in the format of the Italian resource TUT. The project mainly aims at testing the hypothesis that the annotation schema and the knowledge annotated in the TUT format can be useful also to address the issues related to parallel corpora. Therefore, the same parsing system and the tools used for the improvement of the quality of the data annotated within TUT have been extended and applied to the other two languages.

Although this attempt has produced encouraging results, the project is currently ongoing and we presented several directions for its further development, extension and improvement.

References


L. Lesmo. 2007. The rule-based parser of the NLP group of the University of Torino. Intelligenza artificiale, 2(IV).


Bulgarian-English Parallel Treebank: 
Word and Semantic Level Alignment

Kiril Simov  
LMD, IICT-BAS  
kivs@bultreebank.org

Petya Osenova  
LMD, IICT-BAS  
petya@bultreebank.org

Laska Laskova  
LMD, IICT-BAS  
laska@bultreebank.org

Aleksandar Savkov  
LMD, IICT-BAS  
savkov@bultreebank.org

Stanislava Kancheva  
LMD, IICT-BAS  
stanislava@bultreebank.org

Abstract
The paper describes the basic strategies behind 
the word and semantic level alignment in the 
Bulgarian-English treebank. The word level 
alignment has taken into consideration the ex-
perience within other NLP groups in the con-
text of the Bulgarian language specific fea-
tures. The semantic level alignment builds on 
the word level alignment and is represented in 
the framework of the Minimal Recursion Se-
namics.

1 Introduction
Manually created aligned bi- or multilingual cor-
pora have proven to be useful resources in vari-
ety of tasks, e.g. for the development of automatic 
alignement tools, but also for lexicon extrac-
tion, word sense disambiguation, machine trans-
lation, annotation transfer and others.

In this paper we describe the word level align-
ment of the Bulgarian-English Parallel HPSG Treebank (BulEngTreebank) and its connection to the semantic level alignment. The aim of con-
structing such a treebank is to use it as a source for learning of statistical transfer rules for Bul-
garian-English machine translation along the lines of (Bond et al. 2011 to appear). The transfer rules in this framework are rewriting rules over MRS (Minimal Recursion Semantics) structures. The basic format of the transfer rules is:

\[ C : [I ! F ] \rightarrow O \]

where \( I \) is the input of the rule, \( O \) is the output. \( C \) determines the context and \( F \) is the filter of the rule. \( C \) selects positive context and \( F \) selects neg-

ative context for the application of a rule. For 
more details on the transfer rules consult (Oepen 
2008). This type of rules allows for the ex-
tremely flexible transfer of factual and linguistic 
knowledge between the source and the target lan-
guages. Thus the treebank has to contain parallel 
sentences, their syntactic and semantic analyses 
and correspondences on the level of MRS.

In the development of such a parallel treebank 
we rely on the Bulgarian HPSG resource gram-
mar BURGER, and on a dependency parser (Malt Parser – Nivre et al. 2006), trained on the 
BulTreeBank data. Both parsers produce semantic representations in terms of MRS. The treebank is a parallel resource aligned first on a sentence 
level. Then the alignment is done on the level of 
MRS. This level of abstraction makes possible 
the usage of different tools for producing these 
alignments, since MRS is meant to be compatible 
with various syntactic frameworks. The chosen 
procedure is as follows: first, the Bulgarian sen-
tences are parsed with BURGER. If it succeeds, 
then the produced MRSes are used for the align-
ment. In case BURGER fails, the sentences are 
parsed with Malt Parser, and then MRSes are 
constructed on the base of the dependency ana-
lysis. The latter MRSes are created via a set of 
transfer rules (see Simov and Osenova 2011). In 
both cases we keep the syntactic analyses for the 
parallel sentences.

With respect to the MRS alignments, a very 
pragmatic approach has been adopted – namely, 
the MRS alignments originated from the word 
level alignment. This approach is based on the 
following observations and requirements:
• Both approaches for generation of MRS over the sentences are lexicalized;
• Non-experts in linguistics can do the alignments successfully on word level;
• Different rules for generation/testing are possible.

Both parsers (for Bulgarian and English), which we use for the creation of MRSes, are lexicalized in their nature. Thus, they first assign elementary predicates to the lexical elements in the sentences, and then, on the base of the syntactic analysis, these elementary predicates are composed into MRSes for the corresponding phrases, and finally of the whole sentence.

Our belief is that having alignments on word level, syntactic analyses and the rules for composition of MRS, we will be able to determine correspondences between bigger MRSes than only lexical level MRSes, using the ideas of (Tinsley et al, 2009). They first establish the mapping on word level (automatically), then for candidate phrases they calculate the rank of the correspondences on the base of the word level alignment. Thus, our idea is to score the correspondences between two MRSes on the base of involved elementary predicates as well as the syntactic structure of the parallel sentences.

As it was mentioned, the alignment on word level allows us to do more reliable alignments using annotators who are non-experts in linguistics. Currently, the inter-annotator agreement is 92 %. Also this kind of alignment does not require any initial knowledge of MRS from the annotators. Another advantage is that the result might be used for training tools for automatic word alignment, and thus automatic extension of the treebank can be performed. Additionally, the word level alignment might be done before the actual analysis of the sentences. This is especially useful in case of Bulgarian, where the BURGER grammar is underdeveloped in comparison with the English grammar.

The paper is structured as follows: the next section discusses the related works on word alignment strategies. Section 3 focuses on the basic principles behind the word alignment between Bulgarian and English. Section 4 describes the level of MRS alignments. Section 5 outlines the conclusions.

2 Previous Work on Word Level Alignment

The annotation guidelines for Bulgarian-English word alignment, presented here, gained from the tradition established by the guidelines used in similar projects, aiming at the creation of golden standards for different language pairs, such as the Blinker project for English-French alignment (Melamed 1998), the alignment task for the Prague Czech-English Dependency Treebank 1.0 (Kruiff-Korbayová et al. 2006), the Dutch parallel Corpus project (Macken 2010), among others.

As Lambert et al. (2006) point out, the alignment decisions presented in the guidelines reflect different tasks. There are projects such as ARCADE (Véronis, 2000) and PLUG (Ahrenberg et al., 2000), which aim at building a reference corpora with word, not sentence pairs, and have a different annotation strategy in contrast to those that focus on sentence level. Different linguistic theoretical backgrounds appear to be another source of divergence that affects the rules of phrase alignments as well as the specific grammatical techniques. This holds especially in correspondences between synsemantic words (like prepositions, determiners, particles, auxiliary verbs) and synsemantic and/or autosemantic words (Macken 2010). In addition, some tools for manual word alignment, e.g. HandAlign1, allow the user to link both phrases and their elements with different kind of links, which might be simulated in other tools, which are more restrictive. Finally, the use of the so-called possible (also ambiguous, fuzzy or weak) links that signal correspondence between semantically and/or structurally nonequivalent words or phrases is also a matter of dispute. While some argue that alignment with possible links should be determined by unambiguous rules, formulated with consideration of inter-annotation agreement, others (Lambert et al. 2006) allow for different decisions to be kept, which is true to the role originally ascribed to this kind of links: “P (possible) alignment which is used for alignments which might or might not exist” (Och and Ney 2000).

3 Word Level Alignment

The word level alignment was performed by the WordAligner2 – a web-based tool for word alignment, built on top of the word alignment interface developed by C. Callison-Burch. It allows the user to provide parallel input of non-aligned text through the interface or to upload file(s) with sentence level aligned texts. Editing and/or completion of alignments is also supported. Each pair

1 Available at http://www.cs.utah.edu/~hal/HandAlign/
2 http://www.bultreebank.bas.bg/aligner/index.php
of sentences is represented as a grid of squares (Fig. 1). For convenience English is considered to be the source and Bulgarian – the target language, but that has no implications for the translation direction. Correspondence between two tokens is marked by clicking on a square – once (black square) or twice (dark grey square). Originally, the two colours were introduced to allow the annotator to mark his/her degree of certainty about the alignment decision: sure link (S link, black) or possible link (P link, dark grey). It is worth noting that in an alignment there can be only one type of link between two tokens or, more precisely, there is no distinction between phrase and word levels.

Subsequently the colours were used to distinguish between strong and weak alignment (Kruijff-Korbayová et al. 2006), thus P link (dark grey) represents either weak alignment, or that the annotator is uncertain about the pairing, or both. S link (black) represents either strong alignment, or that the annotator is certain about the pairing, or both.

**General rules**

We adopt the general rules that have proven to be shared by the different annotation tasks and alignment strategies. The number of corresponding tokens to be aligned can be estimated by following these two rules (Veronis 1998, Merkel 1999, Macken 2010):

1. Mark as many tokens as necessary in the source and in the target sentence to ensure a two-way equivalence.
2. Mark as few tokens as possible in the source and in the target sentence, but preserve the two-way equivalence.

If a token or a phrase has no corresponding counterpart in the other language and bears no structural and/or semantic significance, it should be left unlinked (NULL link, square with no fill) (Melamed 1998).

Idioms and free translations present a special case. If two autosemantic words or phrases refer to the same object, but do not share the same meaning, they are aligned with a P link, e.g.:

(1) *this animal*

\[това куче ['this dog']\]

The same rule holds when there is a synsemantic – autosemantic correspondence:

(2)*Ivan’s mother called.*

**Неговата майка се обади.** ['His mother called. ']

P link: *Ivan’s ~ Неговата*

P link is used when a lexical item is paraphrased in the other language:

(3)*these non-Serbs*

\[тези лица от несръбски произход ['persons from a non-Serbian origin']\]

P link: *non-Serbs ~ лица от несръбски произход*

Idioms are linked with an S link; each token from the idiom in the source sentence is aligned with each token from the idiom in the target sentence.

(4)*She’ll marry him when pigs begin to fly.*

\[Тя ще се омъжи за него на куково лято.\]
Specific rules
These rules are primarily language specific and their subjects are predominantly function words (prepositions, determiners, auxiliary verbs and the like). We give preference to the semantic equivalence where possible.

Noun phrases
Determiners. Articles, demonstratives and possessive pronouns
a) English determiners like a(n) or the correspond either to Bulgarian determiners един [one] (always in preposition, see example (7), or bare NP (5), or to the so called full/short definite article (6). In both languages they are attached to the first modifier of the NP, if there is one, regardless of its position\(^3\).

(5) I live in a house.
Живея в къща.

S link: a house ~ къща

(6) Look at the house!
Виж къщата!

S link: the house ~ къщата

b) Usually if one of the two corresponding NPs has no modifier, the determiner and the head of the phrase are aligned together to the head of the other phrase (compare for example the rules presented in Kruijff-Korbayová 2006 or Macken 2010). Since in Bulgarian the article could be a morpheme attached to the first modifier (8), we decided to link both the article and the modifier from the English sentence to the corresponding Bulgarian modifier with an S link.

(8) the lovely old house
хубавата стара къща

S link: the lovely ~ хубавата
S link: house ~ къща

c) We follow (Kruijff-Korbayová 2006) in linking determiners from different word classes, based on the similarity in their function. Thus the correspondence between indefinite articles and indefinite pronouns is marked with an S link (9).

(9) a girl
някакво момиче

S link: a ~ някакво
S link: girl ~ момиче

d) English definite articles and Bulgarian demonstrative pronouns are also aligned with an S link (10).

(10) the man
tози човек

S link: a ~ този
S link: man ~ човек

---

\(^3\) There are some exceptions in Bulgarian, e.g. хубави един деца (‘pretty ones children’ – some pretty children). In this case един and some should be surely aligned.
S link: *the ~ този
S link: *man ~ човек
e) We use P link to align the with definite forms of full possessive pronouns (11) because the possessive.

(11) I heard the words, Чух техните думи.

P link: *the ~ техните
S link: *words ~ думи

Substitution with one(s)
Both lexical substitution and nominalization with the numeral one(s), which are typical for English, have no structural and semantic analogy in Bulgarian. They should be aligned to the Bulgarian lexical unit that correspond to the premodifi er of one (12), or, if there isn’t any, to the coreferential Bulgarian pronoun (13).

(12) the little ones малките

(13) the ones that we love онези, които обичаме

Prepositional phrases
a) Very often English noun premodifi ers are translated into prepositional phrases in Bulgarian (14). If that is the case, the preposition is aligned with a P link to the head noun, for example:

(14) Justice Minister Cemil Cicek Министърът на правосъдието Джемил Чичек

S link: *Justice ~ правосъдието
P link: *Justice ~ на

b) English possessive noun forms are translated into Bulgarian either with на prepositional phrase (John’s ~ на Иван), or with an adjective that has possessive meaning (John’s ~ Исанов). In case of PP translation, the preposition itself is aligned to the possessive ’s (for singular) or ’ (for plural) marker with an P link to reflect the fact that the two possessive markers are morphosyntactically different (15).

(15) JNA’s 1st Guards Motorised Brigade Първа гвардейска моторизирана бригада на ЮНА

S link: *JNA ~ ЮНА
P link: *’s ~ на

Verb forms
We follow the rules as they were first formulated in (Melamed 1998): link main verb to main verb and auxiliary verb(s) to auxiliary verb(s) if possible. Whenever the auxiliary form is not present or different in the source or target phrase, it should be aligned to the main verb (see for example (19), weakly or the two verb forms should be phrase aligned (21).

Expletive subject and pro-drop
a) Expletive subjects (it, there) usually have no correspondence in Bulgarian sentences, but they are obligatory for English. That is why we decided to link them with an S link to all Bul-
garian verb components, i.e. to the whole verb complex.

(16) *It is raining.*

Вали.

S link: *It ~ Вали*

(17) *there are many things* има много неща

S link: *there are ~ има*

b) Bulgarian language is a pro-drop language. If the subject is unexpressed (18, 19, 20), then the English subject should be linked with a P link to all Bulgarian verb components that express one of the agreement categories: person, gender, number, and the main verb form itself. This decision is similar to the decision described in (Lambert et al. 2006) concerning the correspondences between English and Spanish verb phrases with omitted subjects.

(18) *He knows* Знае

P link: *He ~ Знае*

(19) *She was not crying.* Не плачеше.

P link: *She ~ плачеше*

(20) *They would not dare.* Не биха посмели.

P link: *They ~ биха*
P link: *They ~ посмели*

Reflexive pronouns in a verb complex

a) Reflexive Bulgarian ce and cu particles may be part of the verb lemma (21, 22). If that is the case, they should be aligned with an S link to the non-reflexive English verb form.

(21) *had met earlier*

бяхме се срещнали по-рано

S link: *met ~ се срещнали*

b) In contrast to the rules construed for Czech-English alignments (Kruijff-Korbayová 2006), if the reflexive particle is used to form a passive voice construction, it is aligned to the English verb phrase as a whole with a P link. The difference is due to the fact that although we also align the verb forms as phrases, we try to mark separately the correspondence between the main verbs.

(22) *the house is being built* къщата се строи

S link: *is being ~ ce*
S link: *built ~ строи*

To and да particles

a) The correspondence between to and да is usually pretty straightforward.

(23) *the decision to stay*

решението да остана
In the case when to is not present in the source sentence, да should be linked with a P link to the English verb that is aligned to the Bulgarian verb following the particle. Not surprisingly this rule resembles the rule for aligning Dutch (om)…te constructions (Macken 2010) with English full infinitive or -ing forms – as an infinitival particle Bulgarian да occupies similar syntactic positions and has similar functions.

(24) they stopped yelling
tе спряха да викат

P link: yelling ~ да
S link: yelling ~ викат

(25) they may go
tе може да тръгват

P link: go ~ да
S link: go ~ тръгват

(26) You will not perish.
Ти няма да загинеш.

P link: perish ~ да
S link: perish ~ загинеш

Double negation

a) Double negation is typical for Slavic languages like Czech and Bulgarian, but not for English. In Czech the verb itself has a morphologically marked negative form that is weakly aligned with the positive form in English (Kruijff-Korbayová 2006). In Bulgarian the negative marker is not a morpheme, but a particle (не, 27) or an auxiliary verb with negative meaning (няма, нямаше 28). Often it is the case that one or more negative pronouns from the Bulgarian sentence correspond to indefinite English pronouns (27). They should be mapped with a P link.

(27) I couldn't see anything.
Не можах да виđа нищо.

S link: could'nt ~ не можах
S link: anything ~ нищо

(28) I wouldn't come.
Нямаше да дойда.

S link: would n't ~ Нямаше

If it is the English verb, that doesn’t have negative form, then we use a P link to align the Bulgarian negative particle to the English word that bares negative meaning.

(29) I felt nothing.
Нищо не почувствах.
Cardinal and ordinal multiword numerals are treated as compound nouns and thus they are aligned as a block within which one-to-one correspondences are sure aligned (see for alternative decision Graça et al. 2008).

\[
\text{(30)} \quad \text{one hundred and twenty two men} \\
\text{сто двайсет и два мъже}
\]

\section{MRS Level Alignment}

As it was mentioned above, we use the word level alignment in order to establish alignment on the level of MRS. For both languages the phrases are assigned an MRS structure which represents the semantic value of the phrase (in the case of dependency parse this MRS incorporates the semantic values of all dependent elements). The intuition behind our approach is that the lexical data of each structure in the syntactic analysis for a pair of sentences are aligned on word level. Then we assume that their MRS structures are equivalent modulo the meaning of the language specific elementary predicates. We exploit this intuition in constructing the semantic alignment in our treebank.

MRS is introduced as an underspecified semantic formalism (Copestake et al, 2005). It is used to support semantic analyses in HPSG English grammar – ERG (Copestake and Flickinger, 2000), but also in other grammar formalisms like LFG. The main idea is the formalism to rule out spurious analyses resulting from the representation of logical operators and the scope of quantifiers. Here we will present only basic definitions from (Copestake et al, 2005). For more details the cited publication should be consulted. An MRS structure is a tuple \(<GT, R, C>\), where \(GT\) is the top handle, \(R\) is a bag of EPs (elementary predicates) and \(C\) is a bag of handle constraints, such that there is no handle \(h\) that outsscopes \(GT\). Each elementary predication contains exactly four components: (1) a handle which is the label of the EP; (2) a relation; (3) a list of zero or more ordinary variable arguments of the relation; and (4) a list of zero or more handles corresponding to scopal arguments of the relation (i.e., holes). Here is an example of an MRS structure for the sentence “Every dog chases some white cat.”

\[
\text{<h0, \{h1: every(x,h2,h3), h2: dog(x), h4: chase(x, y), h5: some(y,h6,h7), h6: white(y), h6: cat(y)}, \}}
\]

The top handle is \(h0\). The two quantifiers are represented as relations \(\text{every}(x, y, z)\) and \(\text{some}(x, y, z)\) where \(x\) is the bound variable, \(y\) and \(z\) are handles determining the restriction and the body of the quantifier. The conjunction of two or more relations is represented by sharing the same handle (\(h6\) above). The outscope relation is defined as a transitive closure of the immediate outscope relation between two elementary predications – EP immediately outscopes EP’ iff one of the scopal arguments of EP is the label of EP’. In this example the set of handle constraints is empty, which means that the representation is underspecified with respect to the scope of both quantifiers. Here we finish with the brief introduction of the MRS formalism.

First we establish correspondences on lexical level. Each two lexical items in the corresponding analyses are made equivalent on the basis of word alignment. Special attention is paid to the analytical verb forms and clitics. The next step is to traverse the trees in bottom-up manner. For each phrase or head for which the components are aligned, a correspondence on the MRS level is established. It should be explicitly noted that a correspondence on a sentence level is also established. Here we present an example:

Let us consider the following pair of sentences from the English Resource Grammar datasets:

- Kucheto na Braun lae.
- Dog-the(neut) of Browne barks.

\(Browne\text{'s dog barks.}\)

The word level alignment is:

- \(\text{Kucheto = dog}\)
- \(\text{na = 's}\)
- \(\text{na Braun = Browne 's}\)
- \(\text{lae = barks}\)
- \(\text{Braun = Browne}\)

36
Here are the MRS structures assigned to both sentences by ERG and BURGER. Some details are hidden for readability:

**ERG:**

\[
\begin{align*}
&<h1, \\
&\quad \{ h3: \text{proper}_q\text{_rel}(x3,h4,h6), \\
&\quad \ h7: \text{named}_r\text{el}(x5,"Browne"), \\
&\quad \ h8: \text{def}_\text{explicit}_q\text{_rel}(x10,h9,h11), \\
&\quad \ h12: \text{poss}_\text{rel}(e13,x10,x5), \\
&\quad \ h12: \text{dog}_n\text{_1}_\text{rel}(x10), \\
&\quad \ h14: \text{bark}_v\text{_1}_\text{rel}(e2,x10) \}, \\
&\quad \{ h4 \text{ qeq } h7, \ h9 \text{ qeq } h12 \} >
\end{align*}
\]

**BURGER:**

\[
\begin{align*}
&<h1, \\
&\quad \{ h3: \text{kuche}_n\text{_1}_\text{rel}(x4), \\
&\quad \ h7: \text{named}_\text{rel}(x6,"Braun"), \\
&\quad \ h8: \text{exist}_q\text{_rel}(x6,h9,h10), \\
&\quad \ h11: \text{exist}_q\text{_rel}(x4,h12,h13), \\
&\quad \ h1: \text{laya}_v\text{_rel}(e2,x4), \\
&\quad \{ h12 \text{ qeq } h3, \ h9 \text{ qeq } h7 \} >
\end{align*}
\]

The result of correspondences between MRS on the basis of word level establishes the following mappings of elementary predicates lists:

**(m1)**

**(Braun = Browne)**

\[
\begin{align*}
&\text{h}3: \text{proper}_q\text{_rel}(x5,h4,h6), \\
&\text{h}7: \text{named}_\text{rel}(x5,"Browne")
\end{align*}
\]

to

\[
\begin{align*}
&\text{h}7: \text{named}_\text{rel}(x6,"Braun"), \\
&\text{h}8: \text{exist}_q\text{_rel}(x6,h9,h10)
\end{align*}
\]

**(m2)**

**(na = 's)**

\[
\begin{align*}
&\text{h}12: \text{poss}_\text{rel}(e13,x10,x5)
\end{align*}
\]

to

\[
\begin{align*}
&\text{h}3: \text{na}_p\text{_1}_\text{rel}(e5,x4,x6)
\end{align*}
\]

**(m3)**

**(na Braun = Browne 's)**

\[
\begin{align*}
&\text{h}3: \text{proper}_q\text{_rel}(x5,h4,h6), \\
&\text{h}7: \text{named}_\text{rel}(x5,"Browne"), \\
&\text{h}8: \text{def}_\text{explicit}_q\text{_rel}(x10,h9,h11), \\
&\text{h}12: \text{poss}_\text{rel}(e13,x10,x5)
\end{align*}
\]

to

\[
\begin{align*}
&\text{h}3: \text{na}_p\text{_1}_\text{rel}(e5,x4,x6), \\
&\text{h}7: \text{named}_\text{rel}(x6,"Braun"), \\
&\text{h}8: \text{exist}_q\text{_rel}(x6,h9,h10)
\end{align*}
\]

**(m4)**

**(Kucheto = dog)**

\[
\begin{align*}
&\text{h}12: \text{dog}_n\text{_1}_\text{rel}(x10)
\end{align*}
\]

to

\[
\begin{align*}
&\text{h}3: \text{kuche}_n\text{_1}_\text{rel}(x4), \\
&\text{h}11: \text{exist}_q\text{_rel}(x4,h12,h13)
\end{align*}
\]

**(m5)**

**(lae = barks)**

\[
\begin{align*}
&\text{h}14: \text{bark}_v\text{_1}_\text{rel}(e2,x10)
\end{align*}
\]

to

\[
\begin{align*}
&\text{h}1: \text{laya}_v\text{_rel}(e2,x4)
\end{align*}
\]

As we mentioned above, our goal is to have MRS alignment not just on word level, but also on phrase level in the sentence. Thus, using the correspondences described in the previous section and the syntactic analyses of both sentences we can infer the following mapping:

**(m6)**

**(Kucheto na Braun = Browne's dog)**

\[
\begin{align*}
&\text{h}3: \text{proper}_q\text{_rel}(x5,h4,h6), \\
&\text{h}7: \text{named}_\text{rel}(x5,"Browne"), \\
&\text{h}8: \text{def}_\text{explicit}_q\text{_rel}(x10,h9,h11), \\
&\text{h}12: \text{poss}_\text{rel}(e13,x10,x5), \\
&\text{h}12: \text{dog}_n\text{_1}_\text{rel}(x10)
\end{align*}
\]

to

\[
\begin{align*}
&\text{h}3: \text{na}_p\text{_1}_\text{rel}(e5,x4,x6), \\
&\text{h}7: \text{named}_\text{rel}(e5,"Braun"), \\
&\text{h}8: \text{exist}_q\text{_rel}(x6,h9,h10), \\
&\text{h}3: \text{kuche}_n\text{_1}_\text{rel}(x4), \\
&\text{h}11: \text{exist}_q\text{_rel}(x4,h12,h13)
\end{align*}
\]

Additionally, such correspondences might be equipped with similarity scores on the basis of word alignment types involved in the corresponding phrase, as well as the type of the phrase itself. For example, if the word alignment of two corresponding phrases involves only sure links, then the MRS alignment for these phrases also is assumed to be sure. Respectively, if on word level there are unsure links, then the MRS alignment could be assumed to be unsure. This idea could be developed further depending on the application. Also, in some cases the MRS level alignment could be assumed to be sure, although it includes some unsure links on word level. For example, in case of analytical verb forms many elements will be aligned only by possible links, but the whole forms are linked as a sure correspondence. We believe that such pairs of sentences with appropriate syntactic and semantic analyses and word alignment are a valuable source for construction of alignments on semantic level.

In our project, the mappings (explicit or inferred) are used for definition of a procedure for generating transfer rules as outlined in the introductory section.

**5 Conclusion**

In this paper we presented the alignment strategies behind the Bulgarian-English parallel treebank. The focus was on word and MRS level. On the base of each word alignment, an MRS alignment is produced together with the corresponding elementary predicates.
Although the current interannotator agreement on the word level is promising - 92 %, we will continue with the development of the guidelines in parallel to the alignment process.

The language specific features, which are likely to influence the transfer of information from Bulgarian to English, are as follows:

- Similarly to English and in contrast to other Slavic languages, Bulgarian is analytic language with a well-developed temporal system;
- Unlike English and similarly to other Slavic languages, Bulgarian has a relatively free word order and is a pro-drop language;
- Like other Slavic languages, Bulgarian verbs encode the aspect lexically;
- Being part of the Balkan Sprachbund, Bulgarian has clitics and clitic reduplication;
- Like other Slavic languages, Bulgarian has a double negation mechanism;
- Bulgarian polar questions are formed with a special question particle, which has also a focalizing role;
- Like other Slavic languages, the modification is mostly done by the adjectives (garden dog (EN) vs. gradinsko kuche (BG, ‘garden-adjective dog’)).

We hope that the MRS alignment in the treebank provides a good abstraction over the language specific features of Bulgarian as well as adequate equivalents to the English linguistic phenomena.

Acknowledgments

This work has been supported by the European project EuroMatrixPlus (IST-231720).

References


Parallel Corpora in Aspectual Studies of Non-Aspect Languages

Maria Stambolieva
Laboratory for Language Technologies, New Bulgarian University
mstambolieva@nbu.bg

Abstract

The paper presents the first results, for Bulgarian and English, of a multilingual Trans-Verba project in progress at the NBU Laboratory for Language Technologies. The project explores the possibility to use Bulgarian translation equivalents in parallel corpora and translation memories as a metalanguage in assigning aspectual values to "non-aspect" language equivalents. The resulting subcorpora of Perfective Aspect and Imperfective Aspect units are then quantitatively analysed and concordanced to obtain parameters of aspectual build-up.

1 Aims of the investigation

At the time of the appearance of the first studies on Aspect in the early 20th century, this term (a calque from the all-Slavonic "vid"), was solely used for the description of a category typologically characterising Slavonic languages, setting them apart from "non-aspect" languages. After a century of aspectual studies, the term has undergone considerable widening of meaning and forms part, in modern linguistics, of the grammatical description of languages of different groups. Thus, "aspectual classes" are set out for Romance and Germanic languages; the English opposition "non-progressive-progressive" is called "Aspect"; even the category of Correlation is often described as a "Perfect Aspect".

Far from supporting cross-language investigations, foreign language teaching and translation, the showing of different language phenomena in the same Aspect-bag is nothing but misleading and problem-raising. Bulgarian teachers of English who have tried to draw a parallel between Bulgarian and English Aspect to their pupils are well aware of the unsatisfactory results. Translators from Bulgarian to English and back, and their editors, point to Aspect as a major pitfall. Aspect is, again, the category where systems for automatic translation seem to offer the least help – Cf. the translation equivalents provided by Google Translate for a few English sentences:

1. He sang the song. – Toy izpya pesenta. (Perfective Aspect, Aorist)
2. He sang for an hour – ?Toy peeshe za edin chas. (Imperfective Aspect, Imperfect Tense)
3. They ate the sandwich. – *Te yade sandvich. (Imperfective Aspect, Aorist/Present?)

In what follows, I will try to:
• define the essence of Slavonic aspect and, in particular, aspect as expressed in the Bulgarian language – in an attempt to demonstrate why, and with respect to the expression of what semantic oppositions, Bulgarian can be used as a metalanguage in aspectual studies;
• contrast Bulgarian aspect to the aspectual system of English;
• demonstrate the possibilities of using parallel corpora and translation memories in the cross-language study of aspect and present first quantitative results of the computational analysis of the data, with parameters of English aspect construal.

2 The Slavonic Category of Aspect

Slavonic aspect is an equipollent lexico-grammatical category covering the entire verbal system and unambiguously defined in the Lexicon. The semantic basis of the opposition is the presence or absence of a bound ([+Bound] / [-Bound]) in the topological structure of a situation or, in other words, the +Event/-Event nature of the situation. Events and non-events in Slavonic languages define a small set of Situation Types, which, after lexical filling, result in a large number of 'Action Modes'.

Depending on their situation type, eventive verbs may mark one-bound or two-bound situations: zapeya ('start to sing') / izpeya ('sing from beginning to end'). One-bound verbs mark either the beginning of a situation or its end phase - compare zapeya above:

.......... [..........]

Fig. 1. One-bound situations with initial bound and dopeya ('finish singing'):
Two-bound situations can be minimal – namigna (‘wink’), padna (‘fall’), otlepya (‘unglue’):

Fig. 3. Two-bound minimal situations

or extended: procheta (‘read through’), prepluvam (swim through), pospya (‘sleep a while’):

Fig. 4. Two-bound extended situations

Non-Eventive verbs may mark simple non-bounded situations of the Action Modes Statal: hasresvam (‘like’), imam (‘have’), izglezhdam (‘seem’, ‘appear’), cherveneya (‘be red’), mladeya (‘appear young’) or Processual - ticham (‘run’), zreya (‘ripen’):

Fig. 5. Simple noun-bounded situations

or else complex non-bounded situations: preparative situations, i.e. processes preceding an event - zapyavam (‘be about to start singing’):

Fig. 6. Complex non-bounded situations: Preparatives

and iterative situations, i.e. series of similar events – kiham (‘sneeze’), izpyavam (‘repeatedly sing’):

Fig. 7. Complex non-bounded situations: Iteratives.

Preparative and iterative situations are generally expressed by verbs which are derivatively (prefixally or suffixally) formed out of perfective verbs marking momentary or extended events.

The grammaticalisation of the opposition Non-Event/Event is a typological feature of Slavonic languages which sets them apart from languages of the Germanic and Romance groups. Further, in Slavonic languages the expression of aspectual information is concentrated in the verb. Hence, the presence of a perfective or imperfective verb defines unambiguously the aspectral value of the sentence.

Bulgarian stands out among Slavonic languages in that it manifests the Perfective-Imperfective opposition to the highest degree of regularity and grammaticalisation within the language group. As Yu. Maslov points out (Maslov 1984, p.97):

'It should not be thought that the principle of the positive suffixal expression of the Imperfective Aspect and the negative, null expression of the Perfective Aspect forms an exclusive feature of the Bulgarian language area [...] However, it is precisely in the Bulgarian language area that this principle has found its fullest and most consistent development. The specifics of the Bulgarian system in this respect [...] is not in the deviation from the Slavonic language type, but in the fullest expression of the developmental tendencies built in the Slavonic grammatical system [...]'.

It is the regular, systematic character of the expression of the eventive/non-eventive nature of a situation in the verb and the richness of lexical verb types that defines the possibility to use Bulgarian as a metalanguage sui generis in aspectual studies.

3 Aspect Studies for the English Language

Even though Aspect forms part of the verbal categories claimed by English grammar, little -- if any -- of the defining features of the Slavonic category can be said to be applicable to the English data.

In harmony with the analysis of other non-aspect languages, aspectual studies of the English verb start with Verb Classes. A proliferation of classifications of these is in circulation, ranging from Aristotle's tripartition through Vendler (1957), Kenny (1963), Mourelatos (1981), Smith (1997), to name but a notable few. Surprisingly, not one of these classifications parallels the grammaticalised Slavonic opposition in distinguishing, first and foremost, events from non-events. Quite the reverse, the first line is, as a rule, drawn between states and processes. J.-P. Descles (1990) even goes as far as to claim a topological distinction between states as non-bounded situations against processes and events as bounded situations. Such verb classifications are not very helpful in event construal and cannot form the basis of cross-language parallels with Aspect languages.
Unlike other non-Aspect languages, the grammatical system of English does, in fact, incorporate an opposition of an aspectual type - the so-called "Progressive Aspect". This is a private opposition between an unmarked form and a marked form expressing non-boundedness, plus a large number of other components of meaning of a non-topological nature - such as limited duration, irritation and other emotional colouring, increasing or decreasing activity, etc. The non-progressive form in the English "aspectual" opposition is unmarked with respect to boundedness. In other words, the English non-progressive verb cannot unambiguously define a situation as eventive or not. Seeing that, on average, English non-progressive forms occur approximately 20 times oftener than progressive ones in an English narrative text, this means that English verbs are, largely, unmarked for boundedness.

In his 1972 dissertation, Henk Verkuyl tried to demonstrated that in non-Aspect languages such as English, events are construed, i.e. boundedness obtains at VP and Sentence level as a result of the combination of verbs belonging to particular verb classes with quantified or unquantified complement or subject NPs. About the same time and independently of Verkuyl, M. Ridjanovic (1969) and A. Danchev, B. Alexieva (1974) in their English-Serbo-Croatian and English-Bulgarian contrastive studies, respectively, arrived at similar results, namely: aspect markers in English occupy a large stretch of the discourse. While Ridjanovic concentrated on the articulated/non-articulated noun phrases as major markers of Aspect, Danchev/Alexieva, processing a large parallel corpus (20,000 file-cards of English Simple Past Tense sentences and their Bulgarian equivalents!) arrived at a much greater variety of contextual markers. The authors ranked these as follows: adverbial phrases, verb semantics, subject phrase semantics, object quantification.

4 Parallel Corpora in the Aspectual Study of English

In view of the abundance of English-Bulgarian or Bulgarian-English parallel texts, (mainly in the form of TRADOS or Wordfast translation memories, but also simply aligned -- whether with tools for automatic alignment such as WinAlign or computer-assisted aligners such as MIX), the idea of using translation units and the aspectual values of the Bulgarian verbs to assign aspectual values to English sentences seems to make sense. While a wider-scope study based on a set of registers from a balanced corpus is the ultimate task of this project, the data presented below are drawn from a smaller parallel corpus of fiction texts. Even this corpus, however, clearly pinpoints lines of investigation and possibilities for applications of the approach.

The Bulgarian verbs in the parallel corpus were aspect-tagged with a choice of PA (perfective aspect) or IA (imperfective aspect) values. Translation units containing one or the other tag were assigned to one of three sub-corpora: an IAcorpus, a PA corpus and a "Mixed" corpus, with sentences containing both perfective and imperfective verbal forms. Each of the subcorpus was processed with the NBU BUILD segmentation programme, yielding quantitative information. At a next stage, concordancing was performed for larger segment identification.

Setting aside some 7% verbless sentences, our corpus yielded the following quantitative information: appr. 31% of the Bulgarian sentences contained Imperfective verbs only; appr. 23% of the sentences contained perfective verbs only; appr. 29% of the sentences contained both perfective and imperfective verbs, in different patterns.

4.1 Analysing the PA subcorpus

The analysis of the PA corpus quantitative data points to the following major PA markers in the English sentences:

Adverbial modifiers of time:
- when - upon concordancing, found to present, in about all cases, an instance of the relative adverbial, introducing a time clause;
- then, now, now that, before, as (=when), eventually, finally, in+year (e.g. in 1984), at lunch, to begin with, the moment +subject+V.

Coordination:
- and - as a coordinative link between event clauses;
- commas - Cf. above.

Lexical meaning of the verbs:
- communication verbs in the simple past tense, esp. admitted, announced, insisted, lied, mumbled, prompted, said, thought (to myself), urged;
- phrasal verbs: drove away, went away, sat down, etc.
- process verbs in the simple past tense.

4.2 Analysing the IA subcorpus

The following were found to be the major IA markers in the corpus:
Adverbial modifiers:
- temporal adverbials, e.g. still, sometimes, repeatedly, when (= whenever, closely followed by would), as (= while)
- for-phrases: e.g. for a few minutes;
- do nothing but, e.g. We did nothing but quarrel.
- adverbial modifiers of time containing NPs with attributes pointing to iterative situations, e.g. every summer.

Lexical meaning of the verb:
- link verbs, e.g. was, seemed, grew;
- extended state verbs, e.g. know, hope, love, remember.

Subject phrase semantics:
- Subjects semantically characterised as [-Animate], and esp. 'Inalienable property' subjects, e.g. the symmetrical limbs, her expression, etc. are systematically present in IA clauses.

4.3 Analysing the Mixed subcorpus

The most frequent patterns were found to be: IP (appr.9%), PI (appr. 4.5%), IPP and PPI (appr. 2.5% for each subtype). Typical factors defining the "mixed" status of the sentences are: complex verbal predicates, V + complement clause groups, presence of verbs of communication (typically Perfective), presence of verbs of thinking (typically Imperfective), Frame and Event situations. Conjunctions and complementizers, as markers of coordination and subordination, appear high in the rank list of most "mixed" subgroups.

5 Conclusions

The approach not only yielded results paralleling closely those of Danchev and Alexieva's corpus-based study (op. cit.) and the Stambolieva 2008 system-based one, but also contributed interesting additional information. Thus, coordination/compounding, of which no mention has ever been made in previous work, was found in the present study to occupy an important position in the hierarchy of English contextual PA markers. On the other hand, argument NP quantification was not found to hold the high-rank position predicted by Verkuyl (op. cit. and 1993). Concordancing elements of context occurring in both corpora - such as when - allows to arrive at structures which disambiguate them as PA or IA markers. Another important advantage of the approach is the possibility to obtain reliable quantitative information defining the hierarchy of units participating in IA or PA-marked predications. Above all, the specialised corpus thus obtained can be used as valuable translation memory or teaching aid.

References


Abstract

This paper presents the main features of an annotation tool, the Coreference Annotator, which manages bilingual corpora consisting of aligned texts that can be grouped in collections and subcollections according to their topics and discourse. The tool allows the manual annotation of certain linguistic items in the source text and their translation equivalent in the target text, by entering useful information about these items based on their context.

1 Introduction

The annotation tool, Coreference Annotator, has been developed within the framework of wider research in the analysis of parallel texts from a translation point of view. More specifically, the research attempts a theoretical classification of the translation of European Union texts in the light of Relevance Theory (Tsoumari, 2008), and examines a special use monodirectional bilingual corpus consisting of aligned English (originals/source texts) and Greek (translations/target texts) versions of press releases of the European Commission.

The aim of the annotation tool is for the researcher to trace and annotate manually certain linguistic items in the source text and their translation equivalent in the target text, by entering useful information about these items based on their context. The focus for this study is on identifying discourse markers and conjunctions that express concession/contrast/adversity in the source text and then locating their translation equivalent in the target text. To the group of markers mentioned above, the conjunction ‘and’ has been added. Cases of omission of source text conjunctions or discourse markers, or addition of conjunctions or discourse markers in the target text are also marked.

2 Motivation

The scope of the research that motivated the creation of this tool combines mainly translation, parallel corpora (original-source texts and translation-target texts), semantics, pragmatics, and discourse. A parallel aligned corpus of press releases of the European Commission is examined both translationally and linguistically to reach conclusions about how certain linguistic items are translated, potentially reflecting the intention of the authors; the expectations of the readers; whether intentionality and expectations change when moving from the source text to the target text; and effects from genre, discourses depending on the topics of the documents, public sentiment or culture.

2.1 Translation in the EU

There is an intriguing matter in the translation of European Union documents into all or some of the official languages of the European Union. On the one hand, there are rules and regulations governing the operation of European countries together as a whole, as a single unity forming the European Union, and EU culture and mentality. On the other hand, the European countries-member states maintain their national cultures and mentalities. Research has shown that the culture of the
EU edifice is different from national cultures, has a culture of its own, despite the likely blurred borderlines between them (Koskinen, 2001; Koskinen, 2004). EU texts and their translation serve a primary communicative situation, since original texts are written to be translated so as to help EU (source text) authors reach different national (target) language users. Some of the characteristics of EU texts are that they are often produced and translated almost at the same time (Koutsivitis, 1994); translation may constitute the starting point to improve the ‘original’ (Koutsivitis, 2003); the writers are usually a group of people or a committee; most source texts are written in English and to a lesser degree in French and German (three procedural languages); the authors are not necessarily native speakers of the language they use for writing; source texts may not always be written in one language and have special linguistic, syntactical and stylistic characteristics called Eurojargon, Eurobabble or Eurospeak (Trosborg, 1997). Thus the translation process, strategies and methods are also affected by the particular circumstances of the production of target texts.

2.2 Press releases of the European Commission

EU press releases are one of the types of documents produced in the framework of the European Union and are distinct from non-EU press releases. The reason is that if we accept that the European Union has a culture of its own, as Koskinen (2001; Koskinen (2004) argues, then it is only normal to expect the production of EU culture-specific texts and genres. EC (European Commission) press releases are produced under the same EU-specific conditions as most EU documents are, i.e. multiple versions drafted and translated at the same time, non-native speakers drafting the documents etc. Culture has its own manner to construct and partition reality which is mirrored in its discourses, that is “modes of talking and thinking which can become ritualised” (Hatim and Mason, 1990). EU culture is no exception to that. In a corpus of aligned EC press releases an issue worth examining is whether the translation is affected by the different topics and discourses of the press releases.

2.3 Connectives: Relevance theory and Sentiment analysis

Connectives have been selected to be examined because they draw attention due to their status. According to Relevance Theory (Wilson and Sperber, 2002), the author produces his/her speech in such a way so that the reader will reach the speaker-intended interpretation with the least processing effort. The speaker, in order to achieve this, makes certain assumptions about the reader’s background knowledge and, thus, expectations, and based on these assumptions formulates his/her discourse. From a relevance-theoretic perspective (Wilson and Sperber, 1993; Blakemore, 1987), connectives are not linking items, but devices whose meaning plays a part in the interpretation of an utterance. Among the different interpretations available, the hearer will decide which the speaker-intended one is, and connectives can facilitate the elimination of some of the available interpretations in order to achieve optimal relevance (Rouchota, 1998), i.e. the best possible interpretation for the hearer in terms of processing effort and effect.

Connectives have also been discussed in sentiment analysis. There is research which uses linguistic analysis and techniques to explore the sentiment of each sentence or phrase in a document. Meena and Prabhakar (2007) addressed the effects of conjunctions and sentence constructions in extracting sentiments associated with the phrases or sentences of reviews. Conjunctions are seen as crucial constituents when determining the polarity of a sentence. They found that, usually, either they alter the sentiment orientation to the opposite direction or they enhance the sentiment of the sentence.

Agarwal et al. (2008) involved in automatic sentiment analysis at sentence level in movie, car and book reviews observed that sentence structure has a fair contribution towards sentiment determination; conjunctions play a major role in defining the sentence structure. Their basic assumption is: “Not all phrases joined by a conjunct have same level of significance in overall sentiment determination”.

3 Related tools

Parallel corpora are often used as linguistic resources in translation. Special tools have been designed to facilitate research in translation and mul-
Callisto is a multilingual, multiplatform tool providing a set of “annotation services” (Day et al., 2004). Its standard components are textual annotation view and a configurable table display. Some of the tasks performed are automatic content extraction entity and relation detection, characterization and co-reference, temporal phrase normalization, named entity tagging, event and temporal expression tagging etc.

The IAMTC Project combines already existing facilities and newly developed ones and has developed an annotation tool for text manipulation. The Project involves the creation of multilingual parallel corpora with semantic annotation to be used in natural language applications (Farwell et al., 2008). Annotation includes dependency parsing, associating semantic concepts with lexical units, and assigning theta roles.

MULTEXT (Ide and Véronis, 1994) is a project involving the development of tools on the basis of "software reusability", and multilingual parallel corpora. It combines NLP and speech, and examines the possibilities for such a combination by harmonizing tools and methods from both areas. The annotation is performed with a segmenter, a morphological analyser, a part of speech disambiguator, an aligner, a prosody tagger, and post-editing tools. Thus, the annotated data provide information about syntax, morphology, prosody and the alignment of parallel texts.

Propbank is a project where a corpus is annotated with semantic roles for verb predicates (Choi et al., 2010). Annotation is performed with the help of Jubilee by simultaneously presenting syntactic and semantic information. The process is facilitated by Cornerstone, a user-friendly xml editor, customized to allow frame authors to create and edit frameset files.

Finally, there is ParaConc (Barlow, 2002) whose main characteristics are an alignment function, concordance search, search for specific words and their possible translations, corpus frequency and collocate frequency. But the tool has no annotating function.

These tools cannot fully meet the particularities of this research for the reasons discussed next.

4 Need for a new tool

The underlying factor that can bring the above different aspects and approaches together is an annotation tool that features certain specific characteristics that are hard to find all in one annotation tool. Coreference Annotator has those characteristics. In particular, a) uploading aligned texts already processed in an efficient alignment tool so as to achieve maximum alignment performance. The tool’s ability to have as input aligned documents allows a corpus builder to use a reliable external aligner of one’s own choice and then use the annotation scheme for the manual annotation of the aligned corpus; b) depicting the aligned texts in such an arrangement that each pair of aligned texts is clearly separated from the other pairs of aligned texts; each translation unit consisting of the source text segment and the target text segment in each pair of aligned texts is clearly and easily detectable from the other translation units. At the same time, it keeps its place in the text manifesting coherence and flow of text meaning in each language; c) allowing the location of possible translation equivalents in context of the instances of the linguistic items examined, always keeping the source text item and its target text equivalent in a close, binary relationship. This unfolds the variety of equivalents an item can have that may be either context dependent or context independent, and also highlights translation procedures and strategies; d) allowing the creation of a comparable profile at sentence level of the source text entry and the target text equivalent entry by entering accompanying information based on their context (distribution of the entries, collocations etc.) in the appropriate sections and fields of attributes – the source text entry and its equivalent text entry are seen comprehensively as a whole; e) displaying all the attribute sections and fields for each source text entry and its target text equivalent with one click to provide easy access which is important due to the large amount of data; f) allowing the examination of the target text in its own right to identify the cases, if any, of linguistic items under investigation that are present in the target text without being a translation equivalent of a source text entry; the annotation tool also provides for the creation of a profile for each target text addition entry; g) allowing the correlation of discourse topics with the frequency of the linguistic items and their translation equivalents in the two languages, and also with their microenvironments, thanks to the arrangement of the aligned texts; h) allowing the correlation of discourse topics, the frequency
of the linguistic items and their translation equivalents, and the frequency of the items added in the target text; i) providing statistics based on the relationship of the source text entry and its target text equivalent where each result is fully and directly traceable in the corpus not only in terms of which pair of aligned texts it is found in but also in terms of its exact location in the pair, thus keeping track of text meaning and structure, and discourse; j) providing detailed statistics which allows the grouping of information of the profile of the entries for specialized analysis of results; k) producing tables of statistics exportable to widely commercial formats e.g. excel for further processing, e.g. SPSS. Such a sophisticated annotation tool allows multidisciplinary analysis. Finally, the tool has been implemented as a component of the Ellogon language engineering platform (Petasis et al., 2002), making extensive use of its infrastructure for the easy creation of annotation tools.

5 Corpus of Collections

This tool has been tested with a corpus of English-Greek press releases issued by the European Commission from 1/1/2007 to 1/1/2009. The corpus was drawn from the electronic text library of all EU press releases (RAPID)1. The criteria for text selection of that corpus are the availability of a Greek version and the currency of topics. The corpus consists of three thematic collections: the Environment, Agriculture, and Presidency Conclusions, which are further subdivided into thematic subcollections within each collection to make transparent the different discourses. The corpus has been aligned using the WinAlign alignment tool – an application of the SDL Trados 2007 suite. Exporting the aligned corpus in plain text format made it an appropriate input for the annotation tool which has been adapted to accommodate such input. The use of a long-standing professional alignment tool aims at achieving effective performance in the segmentation of the parallel texts at the level of equivalent sentences or text segments, i.e. translation units (SDL, 2007).

6 Annotation scheme

Annotation is conducted by associating attributes to the linguistic items. The annotation tool contains three sections of attribute fields. The first section is general and the most frequently used.

In the first section, the focus is on the source text entry (ST EN) and the target text entry (TT EL) where the latter is considered the translation equivalent of the former in that context. The ST EN fields that follow relate to accompanying information of that token based on the particular context. The same goes for the TT EL fields. The next section, TT Addition, involves the addition of the items in question in the target texts. The third section, Context, involves the context of the texts. The original concept of that section is an attempt to map the differences emerging from the translation process between the two texts. There is great flexibility in designing the annotation scheme since using xml language allows the creation of different attributes and values or sections of attributes or the change of the existing attributes and values or sections of attributes.

6.1 Toolbar

The toolbar is on the top of the screen (see Figure 1) where the collections and the filenames of the aligned documents of each collection are found. The arrow icons guide the annotator to the next or previous document of the collection. Few more icons facilitate managing the documents.

After selecting a collection and an aligned document, on the left side of the tool we can see the document in an aligned form – one column with the source text (ST) and one column with the target text (TT). The aligned document is presented in translation units, i.e. linked source and target text segments, with serial numbers for each unit for easier reference/retrieval when analysing a corpus. Also, to facilitate the visual separation of the translation units the background colour of the units alternates between white and light blue.

6.2 First section of attributes – General

On the right side of the tool, the three sections of attributes are presented. In the first section, the focus is on the source text entry (ST EN) and the target text entry (TT EL) where the latter is considered the translation equivalent of the former in that context. The ST EN and TT EL fields that follow relate to accompanying information of those tokens based on the particular context. When there is an arrow icon on the fields, there is a drop-down list of attributes to select. When an item is annotated the tool highlights it. Different annotated entries are highlighted with different colours but each ST EN entry has the same colour with its TT.

1http://europa.eu/rapid/searchAction.do
EL equivalent entry. The fields ST EN/TT EL Expression accommodate cases where the ST EN/TT EL entries are part of an expression or form a collocation with the surrounding words. Each entry is also annotated for its rhetorical relation and category in that particular context. The values in these fields have been selected in relation to the connectives and discourse markers of interest. For cases where the discourse marker or connective has another function besides the linking one, the value “0” in the ST/TT Rhetorical Relation fields and the value “Other” in the ST/TT Category fields have been provided. There is also provision if a punctuation mark is in place of a TT EL entry.

The checkbox of the ST/TT Phrase-level connection provides information about how often the ST and TT markers/connectives in question link predicates or non-predicates (noun phrases, adjectival phrases etc.) in their language respectively. Difference in the type of connection between the ST EN entry and its TT EL equivalent entry manifests different syntactic structures, and perhaps participant roles in the source and target languages. This in turn may reflect translation strategies e.g. shifts, transpositions, modulations etc. The ST/TT Position fields relate to the distribution of the tokens. When the ST EN entry and its TT EL equivalent are seen in parallel and a change in position is noted, then different thematic and rhematic structures, and focus may be reflected in the two languages. Omission of an ST EN entry in the target text is also checked. The last two fields, “ST Comment” and “TT Comment”, allow comments by the annotator of the corpus that can be used either in revising or in analysing the corpus annotation.

An example can be a token of the additive conjunction ‘and’ (see Figure 1): This entry involves the token ‘and’, highlighted with blue colour in the translation unit 20. Based on its attributes, it is a conjunction of addition (ST Rhetorical Relation = “Addition”), a coordinator in particular (ST Category = “Coordinator”), and connects phrases (non-predicates) (“ST Phrase-level Connection” box checked). The token acting as its equivalent in the target text is και (kae) ‘and’, which is also a conjunction of addition (TT Rhetorical Relation = “Addition”), a coordinator (TT Category = “Coordinator”), and connects phrases (non-predicates) (“TT Phrase-level Connection” box checked).

6.3 Second section of attributes – TT Addition

The next section, TT Addition, involves the addition of the items in question in the target texts (see Figure 2 – TT Addition). There are similar fields.
as in the first section of attributes. Because in this section of attributes the starting point is the target text, a couple of extra fields of attributes have been added: the “TT Rendering of” field which attempts to classify the category of the word/phrase in the ST, if any, that motivated the addition of the discourse marker/connective in the TT; the “TT Analysis/Rendering of Text/Expression” field where the ST word/phrase is entered. Finally, there is one more field, ST Clue for Additional TT EL. Practically, this and the previous field have a similar function. An example can be found in translation unit 5 (see Figure 2): According to the annotation, the TT EL entry και (kae) ‘and’ was added in translation unit 5, is not used as a conjunction (TT Rhetorical Relation=0) and performs a different function from coordination in the structure of the sentence (TT Category=Other).

6.4 Third section – Context

The third section involves the context of the texts (see Figure 3). The original concept of that section is an attempt to map the differences that emerge from the translation process. These differences can be grammatical e.g. a change in the tense of a verb form, semantic e.g. the choice of a slightly/a lot different semantically TT EL equivalent, pragmatic e.g. the choice of a completely different expression in the TT to render ST meaning, or lexical e.g. the addition or omission of a word/phrase in one of the two texts. The following pairs of fields have been designed: ST Verb (or verb phrase) – TT Verb (or verb phrase), ST Adjective (or adjectival phrase) – TT Adjective (or adjectival phrase), ST Adverb (or adverbial phrase) – TT Adverb (or adverbial phrase), ST Other – TT Other. The last pair involves differences that do not fall under any of the other pairs. Then the differences recorded can be evaluated compared with each other based on which of the two options – ST option or TT option – is more or less strong in meaning, more or less informative, more or less appellative, and more or less affective. Some of these differences between the two texts are mandatory driven by language restrictions, for instance, or optional driven by cultural preferences, register, politics etc. Either way, these differences create an effect to the reader. So under the ST fields there are two checkboxes ST More, ST Less and under the TT fields respectively TT More, TT Less. For each difference entered the relevant box is checked; ST entry evaluated as ST More or ST Less and TT equivalent evaluated as TT More or TT Less. There is one last checkbox in this section, Compensation, called after the translation strategy. Compensation refers to making up for the loss of meaning
or effect in some part of the sentence in another part of that sentence or in a contiguous sentence (Newmark, 1988). This box is checked when the difference in context in the two texts is due to the translation strategy of compensation.

An example can be in translation unit 7 (see Figure 4): According to the annotation, the ST phrase ‘This aims to’ in translation unit 7, entered in the ST Other field is classified as ST Less compared to its TT equivalent phrase Με τη μεταρρύθμιση επιδιωκεται (Mae ti metarythmisi epidiokeita) ‘With the reform it is aimed’. The reason is the act of referring in the English segment where the demonstrative pronoun ‘This’, a lexicalized deictic element or indexical, is clarified in the Greek segment with the nominal referent μεταρρύθμιση (metarythmisi) ‘reform’. So the TT phrase is more informative than the ST phrase. Because the foregrounded nominal in the TT phrase Με τη μεταρρύθμιση (Mae ti metarythmisi epidiokeita) ‘With the reform it is aimed’ refers to the pronominal fronted in the ST, this is another factor which enhances the effect of the referring act in relation to the transposition between active and passive voice. Thus, the referring act prevails and classifies the TT phrase as TT More.

7 Statistics

Detailed statistics tables are produced covering all possible search criteria. The findings are easily traceable in the corpus in terms of collection, aligned text and position of the translation unit where the item is found in the aligned text. In particular, three tables are generated. The first table (see Figure 4) presents all the source text tokens of interest per aligned document and collection, their frequency, their translation equivalents along with their frequency, and cases of omission of the source text connectives/discourse markers in the target text. At the end of each collection, there is the subtotal of the frequency of source text connectives/discourse markers and their translation equivalents. After all the collections have been examined the table presents the total results of the total of collections. An important element is that next to each result there are the numbered translation units where the source text connective/discourse marker and its target text equivalent are found. This last feature allows easy retrieval of the translation unit, which ensures keeping track of text meaning and structure, and flow of discourse.

The second table (see Figure 5) presents grouped data based on the first section of at-
Figure 4: Statistics Table 1.

Figure 5: Statistics Table 2.
tributes. It includes the elements of the first statistics table enriched with the accompanying attributes of both source and target text entries. The results present linearly, focusing on the ST entry – TT equivalent entry pair, the attributes which accompany the pair. Every time an attribute of the pair changes, there is a different entry in the results. Again, information on the document, collection and translation unit where the pairs with the specific attributes are found satisfies any search criteria.

The third statistics table involves results from the second section of attributes – TT Addition. It follows the rationale of statistics table 2 (Figure 5) but it focuses only on the target text items that have been added without being a translation equivalent of the source text items in question. Statistics for the third section of attributes about Context has not been designed yet because this section of attributes has not been fully tested in the corpus.

8 Conclusion

The Coreference Annotator is an annotation tool which is user friendly in its operation. It gives the researcher the advantage of selecting an external alignment tool for aligning a corpus of parallel texts according to his/her needs. It allows great flexibility in the study of various linguistic items and the translation process at the same time providing, therefore, multiple levels of analysis. Thus the researcher works with a tool that is easily adjustable to his/her varied needs in relation with the annotation of bilingual data.

References


Using Manual and Parallel Aligned Corpora for Machine Translation Services within an On-line Content Management System

Cristina Vertan  
University of Hamburg  
cristina.vertan@uni-hamburg.de

Monica Gavrila  
University of Hamburg  
gavrila@informatik.uni-hamburg.de

Abstract

Web content management systems (WCMSs) are a popular instrument for gathering, navigating and assessing information in environments such as Digital Libraries or e-Learning. Such environments are characterized not only through a critical amount of documents, but also by their domain heterogeneity, relative to format, domain or date of production, and their multilingual character. Methods from Information and Language Technology are the “plug-ins” necessary to any WCMS in order to ensure a proper functionality, given the features mentioned above. Among these “plug-ins”, machine translation (MT) is a key component, which enables translation of meta-data and content either for the user or for other components of the WCMS (i.e. cross-lingual retrieval component). However, the MT task is extremely challenging and lacks frequently the availability of adequate training data. In this paper we will present a WCMS including machine translation, explain the related MT challenges, and discuss the employment of corpora as training material, which are manually and automatically parallel aligned.

1 Introduction

During the last couple of years, the number of applications which are entirely Web-based or offer at least some Web front-ends has grown dramatically. As a response to the need of managing all this data, a new type of systems appeared: the web-content management systems. In this article we will refer to this type of systems as WCMS. Existent WCMSs focus on storage of documents in databases and provide mostly full-text search functionalities. These types of systems have limited applicability, due to reasons such as the following:

- data available on-line is often multilingual;
- documents within a content management system (CMS) are semantically related (share some common knowledge or belong to similar topics).

Shortly, currently available CMSs do not exploit modern techniques from information technology like text mining, semantic web or machine translation.

The recently launched ICT PSP EU project ATLAS (Applied Technology for Language-Aided CMS\(^1\)) aims to fill in this gap by providing three innovative Web services within a WCMS. These three Web services (i-Librarian, EUDocLib and i-Publisher) are not only thematically different, but also offer different levels of intelligent information processing.

The ATLAS WCMS makes use of state-of-the-art text technology methods in order to extract information and cluster documents according to a given hierarchy. A text summarization module and a machine translation engine, as well as a cross-lingual semantic search engine are embedded. The system is addressing for the moment seven languages (Bulgarian, Croatian, English, German, Greek, Polish and Romanian) from four different language families. However, the chosen framework allows additions of new languages at a later point.

Machine Translation is a key component of the ATLAS-WCMS and it will be embedded in all three services of the system. The development of the engine is particularly challenging as the translation should be used in different domains and on

\(^1\)http://www.atlasproject.eu.
different text-genres. Additionally, the considered language-pairs belong most of them to the lesser resourced group of languages, for which bilingual training and test material is available only in limited amount.

The availability of adequate and comparable training data for all language pairs in the ATLAS system played an important role in the architectural design of the MT-engine. The selection of training data was preceded by experiments on selected language pairs. Through these experiments we intended to investigate if small parallel corpora can be also used and with which implications on the translation quality. We investigated additionally the automatic (sentence) alignment in larger corpora in order to understand which implications alignment errors may have on the translation process.

In the following sections we report about our findings as follows: in Section 2 we present briefly the ATLAS functionality and describe the corresponding challenges for the machine translation engine. In section 3 we present the data we used for experiments and analyze it from the linguistic point of view. Section 4 deals with experiments which investigate the dependency between the amount of the training data and the translation quality. Section 5 gives an overview of future experiments and implementation steps.

2 MT-challenges in the ATLAS-System

2.1 The ATLAS-System

The core on-line service of the ATLAS platform is i-Publisher, a powerful Web-based instrument for creating, running and managing content-driven Web sites. It integrates language-based technologies to improve content navigation e.g. by interlinking documents based on extracted phrases, words and names, providing short summaries and suggested categorization concepts. Currently two different thematic content-driven Web sites are being built on top of ATLAS platform, using i-Publisher as content management layer: i-Librarian and EUDocLib. i-Librarian is intended to be a user-oriented web site which allows visitors to maintain a personal workspace for storing, sharing and publishing various types of documents and have them automatically categorized into appropriate subject categories, summarized and annotated with important words, phrases and names. EUDocLib is planned as a publicly accessible repository of EU legal documents from the EUR-Lex collection with enhanced navigation and multilingual access. All three services operate in the multilingual setting described in Section 1. To justify the need of embedded language technology tools within the ATLAS platform we detail here only the functionalities of i-Librarian.

The i-Librarian service (see Figure 2.1):

- addresses the needs of authors, students, young researchers and readers,
- gives the ability to easily create, organize and publish various types of documents,
- allows users to find similar documents in different languages, to share personal works with other people, and to locate the most essential texts from large collections of unfamiliar documents.

The facilities described above are supported through intelligent language technology components like automatic classification, named entity recognition and information extraction, automatic text summarization, machine translation and cross-lingual retrieval. These components are integrated into the system in a brick-like architecture, which means that each component is built on top of the other. The baseline brick is the language processing chains component which ensures a heterogeneous linguistic processing of all documents independent of their language (Ogrodniczuk, 2011). A processing chain for a given language includes a number of existing tools, adjusted and (or) fine-tuned to ensure their interoperability. In most respects a language processing chain does not require development of new software modules, but rather combining existing tools.

With respect to the machine translation engine the language processing tools provide the...
part-of-speech (PoS) annotation necessary for factored models and ensure named entity recognition. Other bricks of the ATLAS architecture feed information into the translation engine as follows:

1. the document categorization gives information about the domain of a particular document;
2. the automatic summarization deals with anaphora resolutions and pre-processes the document in order to simplify the translation task.

2.2 Challenges of the MT-Task

The machine translation (MT) engine is integrated in two distinct ways into the ATLAS platform:

- the MT-engine is serving as a translation aid tool for publishing multilingual content for i-Publisher. Text is submitted to the translation engine and the result is subject to the human post processing;
- for i-Librarian and EuDocLib, the MT-engine provides a translation for assimilation, which means that the user retrieving documents in different languages will use the engine in order to get a clue about the documents, and decide if he wants to store them. If the translation is considered as acceptable, it will be stored into a database.

The integration of a machine translation engine into a web based content management system, presents from the user point of view two main challenges:

- the user may retrieve documents from different domains. Domain adaptability is a major issue in machine translation, and in particular in corpus-based methods. Poor lexical coverage and false disambiguation are the main issues when translating documents out of the training domain;
- the user may retrieve documents from various time periods. As language changes over time, language technology tools developed for the modern languages do not work, or perform with higher error rate, on diachronic documents.

With the current available technology it is not possible to provide a translation system which is domain and language variation independent and works for a couple of heterogeneous language pairs. Therefore our approach envisage a system of user guidance, so that the availability and the foreseen system-performance is transparent at any time.

From the development point of view the main challenge is provided by the high number of language pairs, most of them involving languages with rich morphology and belonging to structural different language families. For most of the language pairs a limited number of parallel aligned corpora are available. Additionally, the ATLAS platform should provide a basic comparable functionality for all language pair, so we cannot train models for different language pairs on completely different corpora.

After collecting information regarding parallel corpora for all involved language pairs, we decided to focus the development of basic training models on those summarized in Figure 2.2.

It can be observed that with exception of Croatian, for all other involved languages the JRC-Acquis corpus offers a good training basis (coverage and size). In order to ensure domain portability we decided to train domain factored models as in (Niehues and Waibel, 2010). This approach allows the usage of small domain specific corpora. Small corpora have the advantage that they can be manually aligned, or at least manually corrected. In order to see how the translation engine behaves when exposed to large but automatically trained corpora and to small but manually aligned texts, we performed several analyses described in sec-

![Figure 2: Available parallel corpora for all language pairs within the ATLAS system.](http://optima.jrc.it/Acquis/)
Manually Aligned Small Corpora vs. Automatically Aligned Large Corpora

We decided to make selective experiments on corpora involving following language pairs: English, Romanian and German. Our choice is based on the availability of human evaluators speaking all three languages, but also by the fact that the languages belong to structural different families (Romania is in the Latin language family, English and German are Germanic languages). Additionally Romanian and German are highly inflected.

3.1 JRC-Acquis

The JRC-Acquis Communautaire is nowadays one of the mostly used parallel aligned corpus for training models in statistical machine translation (Koehn et al., 2009). We do not make here an extensive presentation of the SMT system but present in Table 1 and 2 just a comparative statistics on the three selected languages. From these tables we can infer that the size of the training material has large variations across different language pairs within the JRC-Acquis.

<table>
<thead>
<tr>
<th>Language pair</th>
<th>No. of documents</th>
<th>No. of links</th>
</tr>
</thead>
<tbody>
<tr>
<td>German-Romanian</td>
<td>6558 docs</td>
<td>391972 links</td>
</tr>
<tr>
<td>German-English</td>
<td>23430 docs</td>
<td>1264043 links</td>
</tr>
<tr>
<td>English-Romanian</td>
<td>6557 docs</td>
<td>391334 links</td>
</tr>
</tbody>
</table>

Table 2: JRC-Acquis alignment statistics (docs=documents).

The corpus is automatically paragraph-aligned, where a paragraph is a simple or complex sentence or a sub-sentential phrase (such as noun-phrase).

3.2 RoGER

RoGER (Romanian German English, Russian) is a parallel corpus, manually aligned at sentence level. It is domain-restricted, as the texts are from a users’ manual of an electronic device. The languages included in the development of this corpus are Romanian, English, German and Russian. The corpus was manually compiled. It is not annotated and diacritics are ignored. The corpus was manually verified: the translations and the (sentence) alignments were manually corrected.

The initial PDF-files of the manual were automatically transformed into text files (.RTF), where pictures were either left out (pictures around the text), or replaced with text (pictures inside the text). The initial text was preprocessed by replacing numbers, websites and images with “metanotions” as follows: numbers by NUM, pictures by PICT and websites by WWWSITE. In order to simplify the translation process, some abbreviations were expanded. The sentences were manually aligned, first for groups of two languages. This way we obtained two alignment files. Finally, the two alignment files obtained were merged, so that, after all, RoGER contained all four languages. The merged text files are XML encoded, as shown below:

```xml
<?xml version='1.0' encoding='UTF-8'?>
<sentences>

.......

<sentence id='1010'>
<en>Press Options and some of the following options may be available .</en>
<de>Druecken Sie Optionen . und einige der folgenden Optionen sind ggf. verfuegbar .</de>
<ro>Apasati Optiuni dupa care unele din urmatoarele optiuni pot fi disponibile .</ro>
<ru>...</ru>

</sentence>

.......

</sentences>
```

The corpus contains 2333 sentences for each language. More statistical data about the corpus is presented in Table 3. The average sentence length is eleven tokens for English, Romanian and German and nine for Russian. Punctuation signs are considered tokens. More about the RoGER corpus can found in (Gavrila and Elita, 2006).

3.3 Linguistic Analysis of the Corpora

From both corpora we randomly extracted about 100 sentences, i.e. 100 sentences from the JRC-Acquis corpus for Romanian-English and 100 sentences from the RoGER corpus and the same language pair and direction of translation. These sentences were analyzed with respect to translation divergences and translation mismatches.

Translation divergence means that the same information appears in both SL and TL, but the structure of the sentence is different. Translation
divergences are presented in the literature in (Dorr et al., 1999) and (Dorr, 1994). In the case of a translation mismatch the information that can be extracted from the SL and TL sentence is not the same. Translation mismatches have received less attention in the literature (Kameyama et al., 1991), but for corpus-based approaches they are important, as they directly influence the translation process.

Following translation challenges were observed within the JRC-Acquis:

- **Divergences**
  - Noun (NN) - adjective (Adj) inversion
  - Noun-Preposition-Noun (NN-prep-NN) translated as adjective-Noun (Adj-NN)
  - Subordinate clause translated as adjective
  - Different argument structure
  - Different type of articles
  - Voice change (for verbs)

- **Mismatches**
  - Extra information (the TL sentence is more explicit than the SL one)
  - Reformulations

- **Wrong translation (due to incorrect alignment)**

All these phenomena have a direct (negative) influence on the automatic evaluation scores. Although the corpus is domain restricted, the likelihood of at least one divergence or mismatch type occurring in a sentence is high. Only in approximately 10% of the sentences no phenomenon was encountered. As we encountered totally wrong translations in the corpus, it shows that the (paragraph-) alignments in JRC-Acquis are not always correct.

We also analyzed 100 sentences from the center of the RoGER corpus. We noticed that the diversity of the challenges is reduced, while the number of challenges is sometimes higher compared to what had been encountered in JRC-Acquis, with up to five challenges in an example (a sentence and its translation). Usually there is a one-to-one translation. Only in 12% of cases additional information appeared for one of the languages and in only 9% reformulations have been used. Two phenomena have been found most often: NN-prep-NN translated as NN-NN (or Adj-NN) and Adj-NN inversions.

### 3.4 JRC-Acquis vs. RoGER

The average number of challenges in JRC-Acquis (1.89 challenges per sentences) is lower than the average number in RoGER (2.20 challenges per sentence) for the languages analyzed. However, challenges with a more negative impact on the translation quality (such as “Wrong translation” or “Reformulations”) appear more frequently in JRC-Acquis. The phenomenon encountered more often for the language-pair analyzed is noun-adjective inversions.

### 4 Implications on the Design of the MT-Engine in ATLAS

The MT-Engine within the ATLAS System follows the hybrid approach combining a statistical based component and an example-based one. Both approaches are highly dependent from the quality
and size of the training data. The linguistic analysis above shows that both corpora present translation challenges which influence negative any further automatic processing. Therefore we argue that small domain-specific corpora should be aligned manually at sentence level, or at least the alignment has to be checked manually.

Additional experiments presented in (Gavrila and Vertan., 2011) show that using ROGER as training and test corpus, the performance of the system does not decrease dramatically. Our explanation relies on the linguistic observations in Section 3. The linguistic challenges are balanced by the manual alignment. In this way the corpus, although small has a more correct sentence alignment which triggers a more correct word alignment.

These experiments lead to the conclusion that for the ATLAS-System:

- JRC-Acquis will be used as basis training corpus, without making manual corrections. This is impossible by the size of the corpus.
- Small domain-specific corpora will be first manually aligned at sentence level and afterwards injected in domain facored models.

5 Conclusion and Further Work

In this paper, we described the integration of a machine translation engine within a WCMS system, dealing with a large number of less resourced languages. We investigated the linguistic characteristics of two parallel corpora and show how these influence the translation quality. Further work concerns a statistical relevant analysis of the linguistic phenomena presented in Section 3, involving other manually built corpora and other language-pairs.

Acknowledgments

The present work contains ideas from the ATLAS EU-Project, supported through the ICT-PSP-Programme of the EU-Commission (Topic “Multilingual Web”) (Sections 2 and 4) and from Monica Gavrila’s Ph.D research conducted at the University of Hamburg (Section 3).

References

Author Index

Amoia, Marilisa, 2
Bosco, Cristina, 19
Gavrila, Monica, 53
Kancheva, Stanislava, 29
Kunz, Kerstin, 2
Lapshinova-Koltunski, Ekaterina, 2
Laskova, Laska, 29
Lecuit, Emeline, 11
Maurel, Denis, 11
Nakov, Preslav, 1
Osenova, Petya, 29
Petasis, Georgios, 43
Sanguinetti, Manuela, 19
Savkov, Aleksandar, 29
Simov, Kiril, 29
Stambolieva, Maria, 39
Tsoumari, Mara, 43
Vertan, Cristina, 53
Vitas, Duško, 11