Same same, but different: Compositionality of Paraphrase Granularity Levels

Darina Benikova and **Torsten Zesch**
Language Technology Lab
Department of Computer Science and Applied Cognitive Science
University of Duisburg-Essen, Germany

Abstract

Paraphrases exist on different granularity levels, the most frequently used one being the sentential level. However, we argue that working on the sentential level is not optimal for both machines and humans, and that it would be easier and more efficient to work on sub-sentential levels. To prove this, we quantify and analyze the difference between paraphrases on both sentence and sub-sentence level in order to show the significance of the problem. First results on a preliminary dataset seem to confirm our hypotheses.

1 Introduction

Paraphrases are differently worded texts with approximately same content, whose automatic detection is useful in tasks such as summarization, information extraction, plagiarism detection, machine translation, question answering, and natural language generation (Bhagat and Hovy, 2013).

Most previous approaches work on the sentence level, but it is often hard to decide whether two sentences are indeed paraphrases due to only partially overlapping content as shown in the example below:

a) She gives a red apple to Snow White.
b) The witch envies the princess, so she gives Snow White a red apple.

The decision of whether (a) and (b) are paraphrases is difficult, because only a part of (b) has the same content, whereas the rest is additional information.

We thus propose to work on the *event level*, which we define as a predicate-argument structure, similarly to other researchers (Roth and Frank, 2012; Ritter et al., 2012; Li et al., 2010; Li and Ji, 2016; Shwartz et al., 2017). However, there is no study analyzing and quantifying the difference in performance on different paraphrase granularity levels of clearly distinguishable linguistic units. The contribution of this paper is the analysis and aligned annotation on different granularity levels – namely sentences, events, and event elements.[1] This analysis will show whether it is beneficial to work on the event level and whether sentence paraphrases are composed of event paraphrases.

2 Granularity Levels of Paraphrases

Madnani and Dorr (2010) discuss that paraphrases exist on several granularity levels, namely *sentences*, *phrases*, and *individual lexical items* (or *words*). To illustrate the difference between paraphrase detection on different levels, we provide a running example for each individual level.

2.1 Sentence Level

Paraphrase detection is mostly performed on the sentence level (Dolan and Brockett, 2005; Ganesan et al., 2010; Xu et al., 2014; White et al., 2015; Socher et al., 2011; Fernando and Stevenson, 2008). An exemplary sentential paraphrase would be the following:

a) The witch gives a red apple to Snow White.
b) She gives a red apple to Snow White.

However, in case of the partial overlap between a sentence pair it is more difficult to decide whether this is a paraphrase, as shown in the next sentence:

c) The witch envies the princess, so she gives Snow White a red apple.

Sentence (c) is possibly a paraphrase of (a) and (b), but contains more information which makes the decision more difficult.

[1] https://github.com/MeDarina/SameSame

Proceedings of Recent Advances in Natural Language Processing, pages 90–96,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_014

2.2 Phrase Level

Depending on the exact definition, a *phrase* can range from single words to full sentences. Hence, we only regard a specific kind of phrase, namely predicate-argument structures, as these are syntactically delimited and more suited for automatic detection. Predicate-argument structures have been previously used in paraphrasing and closely related tasks (Roth and Frank, 2012; Xu et al., 2014; Shwartz et al., 2017; Li and Ji, 2016), as they are considered to contain the most salient information in a form that is easier to process than full sentences. The structure consists of a predicate, which is most often a verb, and all its arguments.

To distinguish this specific use of phrase, we call it *event*. An *event* consists of several *event elements*, more specifically one verbal predicate and its arguments. We do not consider auxiliary verbs, if they only add temporal information. Negations and other modifications to the semantics brought e.g. by modal verbs are considered as additional information. Events, similar to previous predicate-argument structures such as e.g. propositions (Stanovsky et al., 2016), allow nesting, i.e. that an argument can also be another predicate-argument structure. For further information regarding the handling of e.g. negations and modals, see our annotation guidelines.[2]

Exemplary event structures are the following:

a) The witch gives a red apple to Snow White.
b) She gives a red apple to Snow White.

In the case of sentence (c) shown in the previous section, the sentence contains two events:

c) The witch envies the princess, so she gives Snow White a red apple.

By using a predicate-argument structure, the sentence is separated in two structures and thus it is easier to decide that the first part of the sentence does not convey the same information as examples (a) and (b), whereas the second part, does.

By regarding events we try to capture the amount of information overlap that composes a paraphrase. The example also shows that in our representation not all parts of the sentence are considered, e.g. the conjunction *so*.

[2]https://github.com/MeDarina/SameSame

2.3 Word Level

Similar to the phrase level, paraphrasing on the word level is mostly regarded in the context of the surrounding sentence. Within the sentence pair from the sentence level example, the word pair marked in bold is an example for a word paraphrase.[3]

(a) *The **[witch]** gives a red apple to Snow White.*
(b) *[**She**] gives an apple to Snow White.*

Without context, *witch* and *she* would not be considered paraphrases, while *witch* and *sorceress* probably would.

Cohn et al. (2008) annotated words and phrases in the context of sentences in order to analyze the nature of paraphrases and corresponding corpora. In our work, the lowest paraphrase level is the *event element* level which are seen in context of their sentence. *Event elements* are verbal predicates and their arguments. The predicate is always one word. In general, arguments can span from lexical items to phrases, but have clear boundaries with regard to the verb. In our work, only verb-verb and argument-argument paraphrases are considered. We do not consider words that are not arguments of a verbal predicate such as conjunctions (e.g. *so*, *because*, *if*) or interjections (e.g. *oh*, *wow*, *hello*).

2.4 Annotatability

It is likely that annotation of paraphrases on the different levels is of different difficulty. However, a comparison is challenging, as there are few studies and they are not comparable between levels.

On the SemEval 2015 Task 1 data (Xu et al., 2014), which is based on the Twitter Paraphrase Corpus (TPC) – this means the tweets are roughly equivalent to 'sentences' – the IAA measured in terms of F-measure is .82. For the phrase level, Cohn et al. (2008) report an F_1 IAA between .71 and .76. They also report IAA on the word level, which is between .74 and .79. In line with our hypothesis, IAA is higher on the word level than on the phrase level, but they did not compare their results to the sentence level. We also cannot directly compare with the results from the Twitter dataset.

In this paper, we annotate a single dataset on all three levels in order to gain insights on which level works best and possibly also how to break down the task of paraphrase detection.

[3]Note that we assume co-reference as given if it can be grammatically implied.

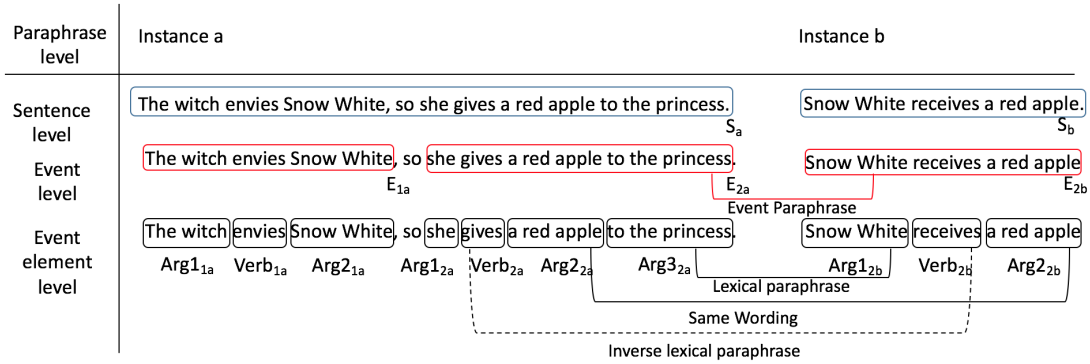


Figure 1: Paraphrase levels annotated in our model

3 Compositionality

In the trivial case of two identical sentences, they are paraphrases and so are all the events they consist of. The same holds for each of the identical events from the two sentences where each of the event elements has a perfect match on the other side. However, there certainly are sentence paraphrases, where there is no such perfect overlap. In those cases it is an open question whether the question of sentence paraphrases can be settled by only looking at the events or the question of event paraphrases by looking at the event elements.

Although many approaches in NLP are build on the assumption of semantic compositionality (Sammons et al., 2010), to our knowledge, there has been no explicit and empirical analysis of the paraphrase compositionality on different independently annotated levels. However, there have been several approaches where different granularity levels have been annotated in one corpus. Vila et al. (2015) and Cabrio and Magnini (2014) classified the paraphrases according to paraphrase classes and also classified lexically differing parts within the pairs according to the same classification. Similarly, Sammons et al. (2010) took existing textual entailment corpora that are classified according to classes including paraphrases and classify the *arguments* according to paraphrase classes. [4]

Cohn et al. (2008) performed an annotation on all three levels in parallel, by using existing sentential paraphrase corpora such as the Microsoft Paraphrase Corpus (MSPC) and adding the other two layers upon those.

Our work differs from the previous efforts in the following: we work on two sub-sentential levels: elements (individual predicates and arguments) and event elements (predicate-argument structures), as both are clearly separated and potentially compositional. Gold standard paraphrase annotations on the lower levels might thus also be helpful for higher levels. A level between the word and the sentence level, similar to the phrase level, might be the solution to issues in paraphrase detection, as it contains more semantics than a word, but presents a reduced amount of information compared to a sentence.

4 Dataset Construction

We annotate paraphrases of verb-argument structures based on existing sentential paraphrase corpora, such as the Microsoft Paraphrase Corpus (MSPC) (Dolan and Brockett, 2005) and the Twitter Paraphrase Corpus (TPC) (Xu et al., 2014). Our dataset is based on 41 sentence pairs from the MSPC and 47 tweet pairs from the TPC. We choose these corpora because, (i) they have been widely used, which makes our approach comparable to others, (ii) they contain many action verbs, which are more likely describe real-world events which fits our goal of finding similar descriptions of the same event, (iii) they also contain negative examples of paraphrases, and (iv) the Twitter corpus contains non-standard data which proves the robustness of our model.

The full set of sentence pairs was annotated by one annotator, while 75% of the corpus where also annotated by a second annotator in order to measure inter-annotator agreement. Figure 1 shows an example of the annotation on all three levels. As the first step, the annotators re-annotated the sentential paraphrases. This step is done in order to analyze the compositionality of the granularity

[4] Two sentences entailing each other are considered a paraphrase by many definitions (Rus et al., 2014; Hovy et al., 2013)

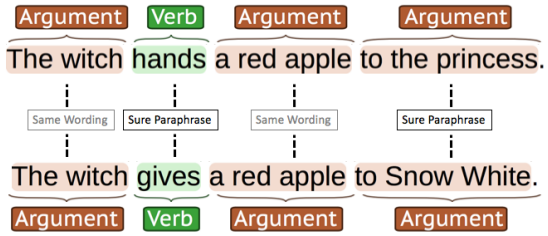


Figure 2: Exemplary event element paraphrase

levels and to compare the annotatability between them. As a second step, they annotate the event paraphrases in each sentence pair and then as a third step the event element paraphrases in each event pair. The sub-sentential tasks can be basically separated in two substasks: (i) finding the events and (ii) aligning paraphrased events and elements.

Finding Events We pre-annotate the first sentence of each pair with the Stanford Dependency Parser in the version provided by DKPro Core (Eckart de Castilho and Gurevych, 2014), using the RNN model with collapsed dependencies, as this parser worked best for our purpose. By using a dependency parser we simplify the task of finding the event with its verb-argument structure. Unfortunately, the quality of the parsing output was insufficient and had to be manually corrected for both datasets. This was mainly due to arguments with very long spans in MSPC and large amounts of non-standard language in TPC. The event is then connected with its elements by marking the span between the elements. Thus, the annotators are shown the annotation of the event and the individual elements before performing the alignment annotation.

Aligning Events and Elements The annotation of paraphrases is performed by an alignment annotation between two instances on the same granularity level in a sentence pair. The alignment annotation is performed on each level independently, in this way reducing the bias of annotating similarly on all levels on purpose. If there is no paraphrase, the annotators do not perform any alignment.

Figure 2 shows an exemplary annotation. There are special alignments for verbal antonyms, verbal negations, and modal verbs, as these change the semantics of the event. Furthermore, we distinguish between *same wording*, *sure paraphrase*, and *unsure paraphrase* on all three levels.

	Sentence	Event	Element
κ	.61	.55	.73
F	.91	.88	.93

Table 1: Inter-annotator agreement on the three granularity levels

5 Dataset Analysis

Our final corpus consists of 88 sentence pairs with 161 event pairs. We use this corpus to analyze the annotatability of each level as well as the compositionality of the levels.

5.1 Annotatability

The results in Table 1 show that in general the inter-annotator agreement is rather high for a task of that difficulty. For the sake of comparability we also report F-measure, but using chance-uncorrected measures like Cohen's Kappa κ is certainly more appropriate. F-measure is higher than in previous studies, but not directly comparable. For both measures, we do not observe the expected result that smaller units get higher agreement. While elements are clearly easier than sentences, events are even worse than sentences. As our sample size is rather small, no definitive conclusions should be drawn from these results.

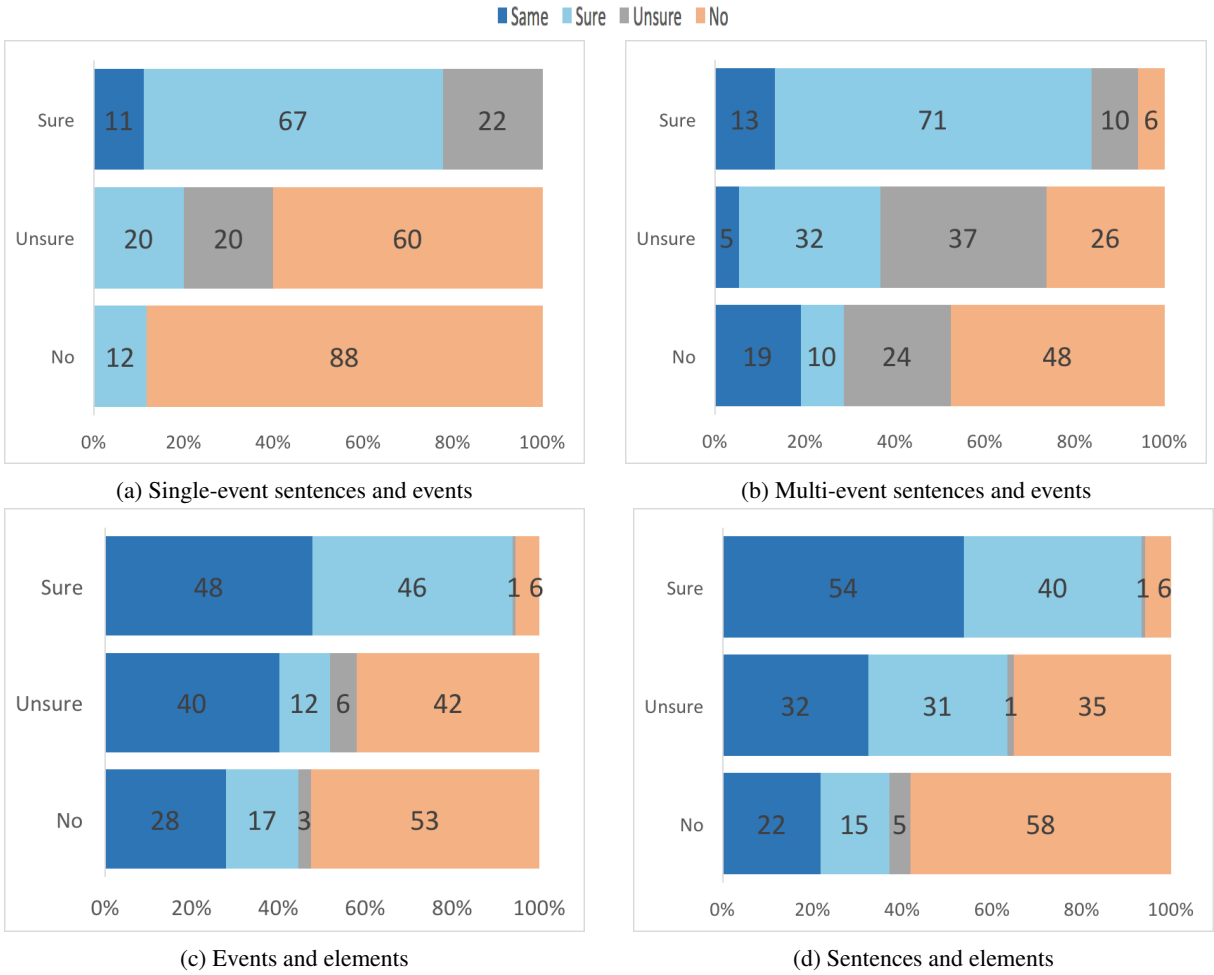
5.2 Compositionality

Using our newly created paraphrase annotations on the three granularity levels, we can now turn towards the question of compositionality. In our analysis, we differentiate between sentences with one event only –single-event sentences– and sentences with more than one event –multi-event sentences. We perform two analyses: first we check the compositionality between all three granularity pairings to empirically analyze whether paraphrases are compositional in general. Furthermore, we compare the differences of the higher classes in more detail in order to show the advantages of working on lower granularity levels.

5.2.1 All Granularity Levels

Figure 3 shows the results of the averaged percentage values between the paraphrase classes of two granularity levels.

Figures 3a and 3b show 67%-71% of *Sure Paraphrase* sentence pairs consist of *Sure Paraphrase* event pairs. Figure 3a shows that single-event sentence pairs that are not paraphrases do not con-



(a) Single-event sentences and events

(b) Multi-event sentences and events

(c) Events and elements

(d) Sentences and elements

Figure 3: Compositionality of the three granularity levels in percent

tain any *Same Wordings* or *Sure Paraphrases* on the event level. Especially when looking at the compositionality between the higher levels and elements, it is clear that there is a big lexical overlap, as *Sure Paraphrases* on the event level consist of 48% *Same Wording* element pairs. Furthermore, the figures show that event elements having the same wording are the most frequent label in each higher leveled paraphrase class. Although they are more often components of *Sure* paraphrases on the higher levels, they are also present in higher leveled instances that are not paraphrases.

This means that although *Sure Paraphrases* are composed of *Sure Paraphrases* or *Same Wording*, these two labels are also present in instances that are not paraphrases, which may be due to the highly lexically overlapping construction of the source datasets, as discussed by Rus et al. (2014). In any case, it means that only looking at the paraphrases on the lower levels is not sufficient to decide over paraphrases on the higher levels and other features need to be also considered, as pairs that are not paraphrases on the sentence and event level also contain 22% or 28% of event elements that are of the label *Same Wording*.

All figures show that both *Sure Paraphrase* and *No Paraphrase* primarily consist of the identical labels on the lower levels, or in the case of *Sure Paraphrase* of *Same Wording*, meaning that if a paraphrase is surely existent or non-existent on the upper level, its lower-leveled components have the same paraphrase label. This shows that paraphrases are compositional in most cases, especially when regarding single-event sentence pairs or pairs with a high lexical overlap.

Event element paraphrases are nearly never *Unsure*, meaning that insecurities about whether pairs are paraphrase are more frequent on the higher levels. *Unsure paraphrases* on the higher levels consist of different components, meaning that a clearer definition of paraphrases could improve the security on paraphrase annotation.

5.2.2 Sentence Level vs. Event Level

To compare the differences of paraphrases on the upper two granularity levels, we consider three different cases, namely: 1) Same paraphrase label, 2) event paraphrase only, and 3) sentence paraphrase only.

Same paraphrase label This is the case of full compositionality, meaning that the paraphrase pair of the higher level consists of paraphrase pairs on the lower level that have the same label as the higher level.

The compositionality of sentences with one or with several events differs slightly, although most sentences consist of events with the same paraphrase label as the sentence. Figure 3a shows that sentences with only one event have paraphrase labels differing from that of their event in 33% of the cases, of which 11% are *Same Wording*, which means that 78% of *Sure Paraphrases* consist of either *Sure Paraphrase* or *Same Wording* event pairs. Sentence pairs that are labeled as *No Paraphrase* in 88% of the cases consists of event pairs that are also labeled as *No Paraphrase*.

Event Paraphrase Only Additionally to the finding of single-event sentences having homogeneous labels with their events, Figure 3b, shows that sentences with multiple events also contain events with differing labels. This shift is especially prominent in the case when multi-event sentences are *No Paraphrase*, but 10% of them are *Sure Paraphrase* and 19% are of *Same Wording*, which is also the previously discussed case of partially overlapping information.

Sentence Paraphrase Only This means that the full sentences are paraphrases of each other, but the events mentioned in them are distinct. This may occur especially in cases were the information in the sentence is not expressed through verb-argument structures as considered in this work, as e.g.

- [The witch's envy is the reason for giving Snow White a red apple]
- [The witch envies Snow White] so [she gives her a red apple]

In our dataset, there is no case of a sentence pair with only one event that is a paraphrase, but its event pair is not.

6 Summary and Future Work

In this work we have examined the compositionality of paraphrases on different levels by analyzing our newly produced corpus which was manually annotated with paraphrases on three granularity levels - namely the sentence, event, and event element level. Although we could not prove that human annotation performance is better on the event level, the compositionality analysis shows that this level is the way to go when trying to find more complex paraphrases on a sub-sentential level. However, we must admit that our sample size is quite small and thus our findings may not generalize.

We plan to improve existing paraphrase methods on event paraphrases, especially by the use of clustered event representations (Benikova and Zesch, 2016). As many state-of-the-art sentence paraphrasing methods apply simple metrics such as lexical overlap or semantic word similarity, we plan to analyze their performance on event paraphrases. We also plan to explore *event embeddings*, which may be more useful in paraphrasing than word or sentence embeddings.

Additionally, we plan to experiment with textual entailment rules, similar to those proposed by Szpektor et al. (2004), Szpektor et al. (2007) and Shwartz et al. (2017) as there is a close connection between paraphrasing and entailment and the entailment rules are based upon predicate-focused structures (Szpektor et al., 2007).

We also consider to additionally annotate entailment on the three levels and investigate whether our model is also helpful for this task.

We intend to enlarge the corpus with more Twitter data, especially with more complex negative examples, which we will try to find by searching for Tweets with the same entities and differing verbs.

Acknowledgements

This work is supported by the German Research Foundation (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media".

References

Darina Benikova and Torsten Zesch. 2016. Bridging the gap between computable and expressive event

representations in Social Media. In *Workshop on Uphill Battles in Language Processing*. pages 6–10.

Rahul Bhagat and Eduard Hovy. 2013. What Is a Paraphrase? *Computational Linguistics* 39(3):463–472.

Elena Cabrio and Bernardo Magnini. 2014. Decomposing semantic inferences. *LiLT (Linguistic Issues in Language Technology)* 9.

Trevor Cohn, Chris Callison-Burch, and Mirella Lapata. 2008. Constructing corpora for the development and evaluation of paraphrase systems. *Computational Linguistics* 34(4):597–614.

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing*. pages 9–16.

Richard Eckart de Castilho and Iryna Gurevych. 2014. A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*. pages 1–11. http://www.aclweb.org/anthology/W14-5201.

Samuel Fernando and Mark Stevenson. 2008. A semantic similarity approach to paraphrase detection. In *Proceedings of the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics*. pages 45–52.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd international conference on computational linguistics*. pages 340–348.

Eduard Hovy, Andrew Philpot, and Marina Rey. 2013. Events are Not Simple : Identity , Non-Identity , and Quasi-Identity (June):21–28.

Hao Li and Heng Ji. 2016. Cross-genre Event Extraction with Knowledge Enrichment. In *The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 1158–1162.

Hao Li, Xiang Li, Heng Ji, and Yuval Marton. 2010. Domain-Independent Novel Event Discovery and Semi-Automatic Event Annotation. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*. pages 233–242.

Nitin Madnani and Bonnie J Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics* 36(3):341–387.

Alan Ritter, Oren Etzioni, and Sam Clark. 2012. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. pages 1104–1112.

Michael Roth and Anette Frank. 2012. Aligning predicate argument structures in monolingual comparable texts: A new corpus for a new task. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*. pages 218–227.

Vasile Rus, Rajendra Banjade, and Mihai C Lintean. 2014. On Paraphrase Identification Corpora. In *Proceedings of LREC*. pages 2422–2429.

Mark Sammons, V.G. Vinod Vydiswaran, and Dan Roth. 2010. Ask not what textual entailment can do for you... In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. pages 1199–1208.

Vered Shwartz, Gabriel Stanovsky, and Ido Dagan. 2017. Acquiring predicate paraphrases from news tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*. pages 155–160.

Richard Socher, Eric H Huang, Jeffrey Pennin, Christopher D Manning, and Andrew Y Ng. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in Neural Information Processing Systems*. pages 801–809.

Gabriel Stanovsky, Jessica Ficler, Ido Dagan, and Yoav Goldberg. 2016. Getting More Out Of Syntax with PropS. *arXiv preprint arXiv:1603.01648* .

Idan Szpektor, Eyal Shnarch, and Ido Dagan. 2007. Instance-based evaluation of entailment rule acquisition. In *Proceedings of Association for Computational Linguistics*. pages 456–463.

Idan Szpektor, Hristo Tanev, Ido Dagan, Bonaventura Coppola, et al. 2004. Scaling web-based acquisition of entailment relations. In *In Proceedings of EMNLP*. volume 4, pages 41–48.

Marta Vila, Manuel Bertran, M Antònia Martí, and Horacio Rodríguez. 2015. Corpus annotation with paraphrase types: new annotation scheme and inter-annotator agreement measures. *Language Resources and Evaluation* 49(1):77–105.

Lyndon White, Roberto Togneri, Wei Liu, and Mohammed Bennamoun. 2015. How well sentence embeddings capture meaning. In *Proceedings of the 20th Australasian Document Computing Symposium*. page 9.

Wei Xu, Alan Ritter, Chris Callison-Burch, William B Dolan, and Yangfeng Ji. 2014. Extracting lexically divergent paraphrases from Twitter. *Transactions of the Association for Computational Linguistics* 2:435–448.