Classification based extraction of numeric values from clinical narratives

Maximilian Zubke
Hochschule Hannover
Dept. of Information and Communication
Expo Plaza 12, 30539 Hannover
maximilian.zubke@hs-hannover.de

Abstract

The robust extraction of numeric values from clinical narratives is a well known problem in clinical data warehouses. In this paper we describe a dynamic and domain-independent approach to deliver numerical described values from clinical narratives. In contrast to alternative systems, we neither use manual defined rules nor any kind of ontologies or nomenclatures. Instead we propose a topic-based system, that tackles the information extraction as a text classification problem. Hence we use machine learning to identify the crucial context features of a topic-specific numeric value by a given set of example sentences, so that the manual effort reduces to the selection of appropriate sample sentences. We describe context features of a certain numeric value by term frequency vectors which are generated by multiple document segmentation procedures. Due to this simultaneous segmentation approaches, there can be more than one context vector for a numeric value. In those cases, we choose the context vector with the highest classification confidence and suppress the rest.

To test our approach, we used a dataset from a german hospital containing 12 743 narrative reports about laboratory results of Leukemia patients. We used Support Vector Machines (SVM) for classification and achieved an average accuracy of 96% on a manually labeled subset of 2073 documents, using 10-fold cross validation. This is a significant improvement over an alternative rule based system.

1 Introduction

Driven by the digitalization, also hospitals have begun to process their documentation more and more in a digital manner. The resulting databases establish new opportunities for efficient analysis of patient data. However, many parts of those data are described by a free text, so that concrete information first has to be extracted from text before they become available for further analysis. This paper focuses on the extraction and correct semantic interpretation of numeric values from clinical narratives. Indeed, some numeric values like in example *E:G-Verhältnis=0,4:1* can extracted by regular expression or template filling due to unambiguous formattings or keywords. But there are also numeric values, which are difficult to process on that way. Reasons for the complexity are general number descriptions, like e.g. percentage values, or a variety of keywords for the associated, semantic information. In front of many different medical areas with different informations and formulations, we assume that machine learning can be used to simplify and improve this task.

After an overview of related work in section 2, we introduce a method to assign numeric values of a given document to their semantic meanings in section 3. In contrast to rule-based systems, we use a system that is able to learn and identify descriptive context features for certain numeric values by example sentences. We consider this task as a supervised machine learning problem and examine the feasibility to replace rule based systems by a more flexible machine learning approach. In section 5 we compare a rule based system with our approach and substantiate our recommendation to use machine learning procedures for information extraction processes.

Proceedings of the Biomedical NLP Workshop associated with RANLP 2017, pages 24–31,
Varna, Bulgaria, 8 September 2017.
https://doi.org/10.26615/978-954-452-044-1_004

2 Related Work

There are various research activities in the field of clinical text mining which can be divided into research in the field of Information Retrieval and research in the field of Information Extraction. We position our work in the field of Information Extraction. In general, Information Extraction in context of medical text mining often addresses one of the following tasks:

- Named Entity Recognition (Ruch et al., 2003)
- Negation Detection (Elkin et al., 2005)
- Temporal Information (Hripcsak et al., 2005)
- Extraction of Codes (ICD,OPS) (Baud, 2003)

We noticed that most of the related studies use *regular expressions* and some kind of terminology, dictionary or ontology. Especially, a robust mapping (Sager et al., 1994) between clinical narratives and *UMLS* (Lindberg, 1990), *SnomedCT* or a self-defined coding scheme appear to be the frequent goals of research in this field. Using annotation engines like *GATE* or *UIMA* text parts are connected to the corresponding concept of the given knowledge organization system (Liu et al., 2005).

In addition, some authors define or describe a complete natural language processing tool for clinical narratives, that integrates typical text mining operations like tokenization, POS-Tagging to enhance the process of information extraction. Besides *MedLEE* (Friedman et al., 1995), *Apache cTakes* (Savova et al., 2010) is such a software solution that combines the concepts, mentioned above.

It should be noticed, that many knowledge organization systems, like e.g. *SnomedCT*, are not directly available for german. Thus Becker and Böckmann (2016) describe an approach to extract *UMLS* concepts from german clinical notes using the german version of *UMLS* and find the corresponding *SnomedCT* concept by the previously detected *UMLS* concept.

Summarizing, we observe that mapping of documents to knowledge organization systems like *UMLS* or *SnomedCT*, supported by classical text mining operations, seems to be the most common approach for information extraction from clinical narratives. One often mentioned argument against the use of *machine learning* is the high effort to generate suitable training sets.

3 Method

Instead of executing a traditional Natural Language Processing (NLP) pipeline and process each word, e.g. by associating it with an *UMLS* concept, we are only interested on numeric values specified in text documents. Hence, we introduce a method to determine the meaning of a numeric value by the surrounding words using *machine learning* algorithms. This approach represents an alternative to the explicit definition of information extraction rules or ontology based document processing.

As illustrated in figure 1, our information extraction method consists of five steps:

1. Extraction of numeric values
2. Document segmentation by . and ;
3. Generation of description candidates for each numeric value
4. Classification of candidates
5. In case of multiple positive classified candidates: Suppression of all candidates, except the one with highest score.

Furthermore we use topic-based classifiers. Each topic, like i.e. *Blasts* have to be described by positive an negative example sentences. Based on this sentence sets the topic classifier determines, if a given documents belongs to that topic or not. The mentioned processing steps are explained in detail below. The performance of this approach can be found in section 5. Further details about our implementation are described in section 4.

3.1 Initial Extraction of numeric values

Because we aim to extract numeric values from clinical narratives, we are only interested in documents of the corpus C that contain at least one numerical value. Therefore we use *regular expressions* to detect and extract numerical intervals or single values from every document. The result of this initial filtering is a subset $C_{num} \subseteq C$. After this initial processing step each document $d_i \in D$ is defined as

$$d_i := (t, N_i) \tag{1}$$

where $t \in C_{num}$ represents the original text and N_i the set of numerical values that appears in that document.

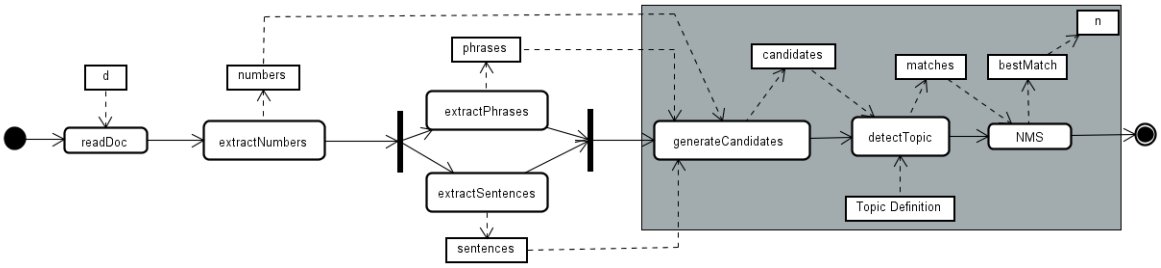

Figure 1: (1) Extraction of numeric values from document d (2) segmentation of sentences and phrases of d (3) for each n (gray area): sentence and phrases that contain n are candidates (4) topic related candidates are matches (5) Choose the match with the highest confidence

3.2 Document Segmentation

In simple clinical information systems, an unstructured text is often represented by a string. However, for advanced information extraction strings do not fit very well. Thus, the transformation of a string in a more complex data structure is the initial processing step of many text mining applications. There are several concepts to represent a document by such a complex data structure. Beside graph-based approaches (Jiang et al. (2010)), a document can also be described by bag of words or a collection of sentences.

As illustrated in Figure 2, we believe, that a numeric value is more related to certain segments like sentences or phrases and less to the whole document. Furthermore we assume, that different

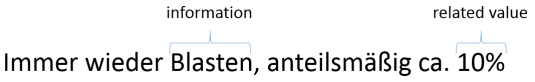

Figure 2:

generated text segments could be different expressive descriptions of the contained numeric value. Due to this assumptions we describe a text d_i both as a set of sentences D_i^s and as a set of phrases D_i^p. The elements of D_i^s are produced by a common *sentence tokenizer* which splits the document into n sentences based on the dot-sign(.) without destroying point numbers or abbreviations. The elements of D_i^p are the result of the same procedure, which separates a document by semicolon instead of a dot-sign. It should be noted, that in our context the term *phrase* means a document snippet that results from the semicolon based splitting of the document. There are two motivations for this additional segmentation: First, many clinical narratives are more written like a note and less like a formal, well structured document. Therefore, it can happen, that a document transports several informations which are separated by semicolons, but do not contain any dot-signs. In those short documents, a pure dot-sign based segmentation would fail and the whole document would be considered as the related context of a certain numeric value. Second, it is possible, that an author describes a documented quantity by a dedicated sentence, but also by the beginning of the following sentence. This related part of the following sentence is usually separated by a semicolon from the rest of the sentence. An example of such a situation can be seen in figure 3.

…, aber vollständig ausreifend bis zu den Segmentkernigen, noch einzelne Blasten vorhanden. Blastenanteil aber deutlich unter 5%; eingestreut reifzellige LOymphozyten, aber keine Lymphozytenvermehrung.

Figure 3: Underlined: Result from pure dot-sign-based segmentation; Bold: Relevant text snippet which is delivered by semicolon based segmentation.

So finally, we have extended our definition of a document 1 to:

$$d_i := (D_i^s, D_i^p, N_i) \tag{2}$$

for all $d_i \in D$. It is possible to extend this concept by a comma based document splitting. But we omitted it due to many for our use case useless segments.

3.3 Candidate Generation

After the generation of overlapping document segments, we are only interested on segments, which are related to a numeric value n_j of d_i. Due to the use of multiple segmentation procedures, there can be more than one snippet which is directly related to n_j. We call such segments *candidates*.

In our current version, a related text segment of a numerical value n_j of document d_i can only be a sentence or phrase from the same document that contains this value, so that the *candidate set* of each $n_j \in N_i$ is defined as:

$$cand(n_j) := \{c|((c \in D_i^s) \vee (c \in D_i^p)) \wedge (n_j \in c)\} \quad (3)$$

In our implementation we keep track of relations between numerical values, sentences and phrases of d_i, so that we are able to retrieve the correct candidates even if the same numerical value appears multiple times in d_i.

3.4 Topic Learning

Usually, quantities and their numerical values appear in the same sentence or text region. It is however extremely hard to define the exact construction in which the quantity and the value appear. Consider e.g. the following sentence:

(1) Immer wieder Blasten, anteilsmäßig ca. 10%
Again and again blasts, rate approx. 10 %

The quantity *Blastenanteil* (Blast rate) is expressed in two words. The second (*Anteil*) is only present as the root of a derived adjective (*anteilsmäßig*). Patterns like this are hard to capture in rules. However, when the key concept *blasts* and a numerical value appear in the same region of the text, we can almost be sure, that the number is the value for the blast rate. To recognize such a key concept or topic, our system learns the related words by a set of sample sentences.

Our system does not have any kind of knowledge from a connected ontology or terminology base like *UMLS*. Also text mining operations like *Named Entity Recognition* or *Negation detection* are not part of our processing pipeline.

Instead our system is based on a generic concept of topic definition only. In our context a topic associated with a quantity is defined as a pair of sets containing positive and negative example sentences for numeric values of that quantity. Table 1 illustrates this idea for the amount of blasts, which is mentioned in many documents of our test dataset. Based on this two sets, we train a binary topic-classifier, which determines whether a given text segment belongs to that topic or not.

$$detect_t(c) = \begin{cases} 0 & \text{if c is not about topic} \\ (1, \kappa) & \text{if c is about topic} \end{cases} \quad (4)$$

Where κ means the confidence or score of the classification.

As already explained above, c can be a sentence or a phrase, that results from the segmentation described section 3.2

We implemented 4 by *Support Vector Machines* Boser et al. (1992). The features of all candidates are *term frequencies* of a vocabulary V, so that each candidate c is described by vector $v \in \mathbb{Z}^{|V|}$ at this point. In our experiments, V contains all words from all available clinical narratives.

We assume, that c is related to topic t, if c contains a numeric value and $detect_t(c) = 1$. The definition of κ depends on the used *machine learning* algorithm. In our experiments, κ represents the distance to the hyperplane of the SVM based classifier.

3.5 Non Maxima Suppression

The trained classifier tells, whether a document segment c belongs to a certain topic t. We assume, that the numeric value mentioned in c describes the topic-related quantity, if c belongs to t. However, the classifier could find more than one candidate relevant for the given numeric value. In such cases we select the segment with the highest confidence value and assume that the value mentioned in that segment belongs to the topic. Furthermore it is possible to identify a threshold of minimum confidence to accept a candidate as an identification of a relation between a numeric value n_j and a topic t.

4 System Description

We implemented this method as a software system, which is based on Python and SQL databases. Our system should supports simple integration into a clinical data warehouse, because many clinical narratives originate from such an information system. Furthermore, adjacent data collections could be used as features of clinical narratives or vice versa in the next version of our software.

4.1 Document representation

Before the execution of any *text mining* or *machine learning* procedure, our tool first generates a database schema like shown in Figure 4. Our in section 3.2 described segmentation concept will realized by two tables, that represent D_i^s and D_i^p. This tables are filled by scripts that implement the in section 3.2 described segmentations. Further-

Positive sample sentences	Negative sample sentences
Weiterhin Monozytoide Blasten (80%) bei 300 Zellen	Ca. 80-85% kleine reife Lymphozyten, einzelne mit Granula
Es findet sich eine Verdrängung der normalen Hämatopoese durch eine monomorphe Blastenpopulation, die ca. 80% beträgt.	Granulopoese stark linksverschoben bis zu den Promyelozyten, die ca. 35% der myeloischen Zellen ausmachen
Blastenanteil 2-4%	Ausreifende granulopoese mit leichter vermehrung von eosinophilen und deutlicher vermehrung von plasmazellen mit einem anteil von 5-10%, z. t. vakuolisiert; kein signifikanter blastenanteil

Table 1: Definition of topic "Blasts" for the quantity blast rate by positive and negative example sentences; Term-related terms are underlined. The underlining is given only for illustration here and not part of the training data.

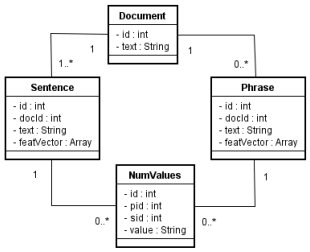

Figure 4: Documents are connected indirectly with numerical values by text segments. Each segment type is represented by a corresponding table. Currently supported segment types: Sentences and Phrases as presented in section 3.2

```
"topicName": "Blasts",
  "topicDescription": "IE of % described amount of blasts",
  "topicLanguage": "german",

      "pos":[
        "Weiterhin Monozytoide Blasten (80%) bei 300 Zellen/µL.",
        "Es findet sich eine Verdrängung der normalen Hämatopoese durch \
        eine monomorphe Blastenpopulation, die ca. 80% beträgt.",
        "Blastenanteil 2-4%."
      ],

      "neg": [
      "Ca. 80-85% kleine reife Lymphozyten, einzelne mit Granula.",
      "Vereunzelte Lymphozyten < 5%, keine pathologischen Blasten.",
      "Der Anteil an reifzelligen Lymphozyten und Plasmazellen ist deutlich \
      erhöht, stellenweise bis zu 10%."
  ]
```

Figure 5: Example of our json based topic definition format.

more we store all numerical values in a dedicated table, which is filled by the procedure, we described in section 3.1. Figure 4 also illustrates, that numerical values are directly connected with sentences and phrases, but only indirectly with the documents. We chose this structure to avoid an incorrect behavior for documents, in which exactly the same numerical values appear in multiple sentences.

4.2 Topic Definition Format

We realized our in section 3.4 presented topic concept by a json based data format. Figure 5 shows an example of this technical topic description. The example sentences can be defined via an easy to use graphical user interface, that generates the appropriate json code internally. So the topics can directly defined by doctors, that do not need knowledge about technical data description techniques for this task.

A further motivation to define such a data format was the resulting flexibility, that enables the possibility to share well defined topic definitions with other internal or external organizations.

5 Evaluation & Results

We used a collection of 12 743 clinical narratives from a german hospital to evaluate our information extraction system. The narratives consist of 1 to 29 sentences, 5 sentences on average. The collection comes from electronic health records of leukemia patients. One of the main interests of the physicians is the rate of blast cells in all reports related to one patient.

At first we defined a topic by collecting positive sentences that contain a percentage description about blast cells and negative sentences that are not related to the searched topic. Example sentences for an description of the amount of blasts are:

(2) a. Blasten (80%)
Blasts (80%)

b. Blastenanteil 2-4%
Blast percentage 2-4%

c. Die Granulopoese ist linksverschoben
mit einem Blastenanteil von $> 20\%$
der nicht erythropoetischen Zellen
The bone marrow is left-shifted
with a blast proportion of $> 20\%$
of the non erythropoietic cells.

d. Keine Markfremden Zellen,
Blastenanteil sicher unter 5%.
No marrow foreign cells,
blast percentage for sure below 5%

Then we generated a vocabulary V containing 13 400 words, based on the whole collection. A first statistic analysis shows, that the size of $|C_{num}|$ is 9 655 and only 4 162 of that documents contain known keywords about blasts and a percentage sign.

5.1 Construction of a gold standard

For the gold standard we selected a random subset of 2 073 documents, which proportion of documents is fulfilling the three conditions is the same as in the whole collection. About 75% of the documents in this selection do not contain a numerical value, or a percentage sign or a keyword related to blasts. We annotated these documents manually. Note that thus we make no difference between documents that have no information on blast rate and documents that do contain information on blast rate, but do not give a concrete value. Especially this means that we labeled all documents containing the statement *Keine Blasten* (no blasts) as documents that do not give a value for the quantity blast rate. For the remaining 435 documents, that contain keywords about blasts, a percentage sign and a numerical value, we extracted the blast percentage manually.

Our classifier is trained only on sentences containing numerical values. In our subset there are 6 805 sentences; 604 sentences contain a numerical value, 439 thereof being a blast rate, 165 not related to the amount of blasts.

5.2 Experiment setup

Each text was first split into sentences and phrases as described in section 3.2.

Next, we generated a candidate set for each numerical value that appears in the given document. As described in section 3.3, the term *candidate* means a sentence or a phrase that contains the numeric value. We processed all documents on that way.

Then we conducted two experiments: In the first experiment we examined the classification of single sentences. Beside two baselines that are described in the next section, we used a SVM based topic classifier (see section 3.4), which decides for each of the sentences, whether it is relevant for the quantity blast rate. Now we can evaluate how many sentences are classified correctly.

In the second experiment we compared methods for extracting numerical values from whole documents. We evaluated our approach in two configurations: *SVM (Sentences)* represents a variant where all elements of the candidate sets are sentences and *SVM (Sentences & Phrases)* represents the same approach using multiple text segments.

For both experiments, we consider a text as correctly processed when either (1) the correct blast rate is extracted from the text or (2) it is correctly detected that no blast rate is specified.

Our manual labeling has extracted values for each text and each sentence, obtained by splitting texts on full stops. When we make additional segments by splitting on semicolons, we can apply the classifier (trained on whole sentences) to this segments as well. However, we cannot compare the results with the manually labeled ones. On the document level, however, we can compare with the manually labeled documents.

We used ten-fold cross validation for all experiments.

5.3 Baselines

We used three baselines. Since most documents are not relevant for the quantity blast rate, we can classify most documents correctly with the majority classifier, that assumes that all documents are irrelevant.

The second baseline assumes that every percentage value is a blast rate. On the sentence level this baseline thus treats all sentences with a number and percentage sign as relevant for the blast rate and all others as irrelevant. At the document level this baseline assumes the first percentage mentioned to be the blast rate. We will refer to this baseline as the %-based approach.

As a third baseline we used an extraction method that is purely based on complex regular expressions. Motivated by the remarkable performance of the percent-based approach, a group of students developed a regular expressions based approach. Therefore they analyzed the data set and define some keywords manually. Combined with the detection of percentage values, they implemented a procedure to extract the searched informations by pattern recognition. Note that this approach processes only whole documents, which is why we could not compare this baseline with alternative approaches on sentence level described by table 2.

6 Results

Table 2 shows the result of the evaluation at sentence level We clearly observe, that the classifier treats almost all sentences correctly. With respect to precision and recall it is of course easy to beat the majority baseline, but the SVM also has an higher accuracy.

Given the good results of the %-based approach we can conclude that indeed most numerical values are related to blast rates. However, there are a number of other numerical values. Apparently, the SVM effectively distinguishes the blast rates from other numerical values.

Table 3 shows the results of the complete method on the document level. At the document level we see again very high scores. We could observe, that the additional semicolon based segmentation indeed excludes a number of mistakes. (e.g. the third negative example from Table 2) The lower precision in comparison to the pure sentence-based configuration implies, that the semicolon based approach produces a few segments which are hard to classify by the current version of our topic classifier. But *SVM(Sentences & Phrases)* also extracts significant more numeric values than *SVM(Sentences)*. As documented in table 3, the regular expression based integration of keywords improves the performance of the %-based information extraction strategy. Apparently, the rules a very precise and do almost never consider a percentage as a blast rate if that is not the case. Thus this method has the highest precision of all tested methods. However, the recall is much lower than that of the classifier based approach.

7 Conclusions and Future Work

In this paper we presented a first version of our information extraction system for medical documentations, which identifies the meaning of a numeric value by the surrounding words.

The integral difference to many similar applications is, that we had no explicit described knowledge about the content of out dataset. Instead we used *machine learning* to learn important keywords by sample sentences.

With term frequency vectors, we used a very simple kind of feature, which already works very well. In the future we want to examine, which alternative features could improve our system.

Our approach yields remarkable results. However, there are situations, that can not processed correctly by our system. We expect, that numerical values are always described by numbers. However, it is possible, that numbers are described by a words instead of number (i.e 'five' instead 5). We also observed, that especially the number zero is often replaced by a negation (i.e. 'no blasts' instead of '0% blasts'). Hence we will integrate a preprocessing step that converts textual definitions of numbers in real numbers. It should be noted, that this task is a non-trivial task, because also a quantitative value can correspond with several, very different formulation, which can be considered as an classification problem, very similar to our topic detection problem, described in section 4. Furthermore, words like 'significant' complicate or prevent a mapping to an equivalent numerical description of the information.

In general, we believe that *machine learning* could be much more efficient than rule-based concepts. Every rule engine needs someone who defines suitable rules, whereas our approach only needs sample sentences which are always available. Furthermore table 3 shows, that the *machine learning* approach is more adjustable than the more strict rule-based approach.

Acknowledgements

We would like to thank our colleagues from the Hannover Medical School to suggest the problem of extracting numerical values and making available the pseudonymized texts. Further Acknowledgements go to our students, that implemented parts of the system, along with a user interface for practical usage of the system in the Hannover Medical School hospital.

Method	Recall	Precision	Accuracy
SVM	0.987 (0.005)	0.950 (0.003)	0.996 (0)
Majority	0.0 (0)	0.0 (0)	0.935 (0)
%-based	0.893 (0)	0.727 (0)	0.971 (0)

Table 2: Results of the extraction of the percentage of blasts evaluated on **sentence** level. Results are averages of 10-fold cross-validation. Standard deviations are given in parentheses.

Method	Recall	Precision	Accuracy
SVM (Sentences & Phrases)	0.921 (0.049)	0.911 (0.044)	0.965 (0.017)
SVM (Sentences)	0.834 (0.069)	0.953 (0.037)	0.957 (0.017)
RegExp based	0.517 (0.053)	0.983 (0.021)	0.897 (0.019)
%-based	0.461 (0.082)	0.629 (0.081)	0.897 (0.023)
Majority	0.0 (0)	0.0 (0)	0.79 (0.034)

Table 3: Results of the extraction of the percentage of blasts evaluated on **document** level. Results are averages of 10-fold cross-validation. Standard deviations are given in parentheses.

References

R Baud. 2003. A natural language based search engine for icd10 diagnosis encoding. *Medicinski arhiv* 58(1 Suppl 2):79–80.

M Becker and B Böckmann. 2016. Extraction of umls® concepts using apache ctakes™ for german language. *Studies in health technology and informatics* 223:71–76.

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, pages 144–152.

Peter L Elkin, Steven H Brown, Brent A Bauer, Casey S Husser, William Carruth, Larry R Bergstrom, and Dietlind L Wahner-Roedler. 2005. A controlled trial of automated classification of negation from clinical notes. *BMC medical informatics and decision making* 5(1):13.

Carol Friedman, Stephen B Johnson, Bruce Forman, and Justin Starren. 1995. Architectural requirements for a multipurpose natural language processor in the clinical environment. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*. American Medical Informatics Association, page 347.

George Hripcsak, Li Zhou, Simon Parsons, Amar K Das, and Stephen B Johnson. 2005. Modeling electronic discharge summaries as a simple temporal constraint satisfaction problem. *Journal of the American Medical Informatics Association* 12(1):55–63.

Chuntao Jiang, Frans Coenen, Robert Sanderson, and Michele Zito. 2010. Text classification using graph mining-based feature extraction. *Knowledge-Based Systems* 23(4):302–308.

C Lindberg. 1990. The unified medical language system (umls) of the national library of medicine. *Journal (American Medical Record Association)* 61(5):40–42.

Kaihong Liu, Kevin J Mitchell, Wendy W Chapman, and Rebecca S Crowley. 2005. Automating tissue bank annotation from pathology reports–comparison to a gold standard expert annotation set. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, volume 2005, page 460.

Patrick Ruch, Robert Baud, and Antoine Geissbühler. 2003. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. *Artificial intelligence in medicine* 29(1):169–184.

Naomi Sager, Margaret Lyman, Ngo Thanh Nhan, and Leo J Tick. 1994. Automatic encoding into snomed iii: a preliminary investigation. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*. American Medical Informatics Association, page 230.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association* 17(5):507–513.