Lexical Substitution for the Medical Domain

Martin Riedl[1] **Michael R. Glass**[2] **Alfio Gliozzo**[2]
(1) FG Language Technology, CS Dept., TU Darmstadt, 64289 Darmstadt, Germany
(2) IBM T.J. Watson Research, Yorktown Heights, NY 10598, USA
riedl@cs.tu-darmstadt.de, {mrglass,gliozzo}@us.ibm.com

Abstract

In this paper we examine the lexical substitution task for the medical domain. We adapt the current best system from the open domain, which trains a single classifier for all instances using delexicalized features. We show significant improvements over a strong baseline coming from a distributional thesaurus (DT). Whereas in the open domain system, features derived from WordNet show only slight improvements, we show that its counterpart for the medical domain (UMLS) shows a significant additional benefit when used for feature generation.

1 Introduction

The task of lexical substitution (McCarthy and Navigli, 2009) deals with the substitution of a *target term* within a sentence with words having the same meaning. Thus, the task divides into two subtasks:

- Identification of *substitution candidates*, i.e. terms that are, for some contexts, substitutable for a given target term.
- Ranking the substitution candidates according to their context

Such a substitution system can help for semantic text similarity (Bär et al., 2012), textual entailment (Dagan et al., 2013) or plagiarism detection (Chong and Specia, 2011).

Datasets provided by McCarthy and Navigli (2009) and Biemann (2012) offer manually annotated substitutes for a given set of target words within a context (sentence). Contrary to these two datasets in Kremer et al. (2014) a dataset is offered where all words have are annotated with substitutes. All the datasets are suited for the open domain.

But a system performing lexical substitution is not only of interest for the open domain, but also for the medical domain. Such a system could then be applied to medical word sense disambiguation, entailment or question answering tasks. Here we introduce a new dataset and adapt the lexical substitution system, provided by Szarvas et al. (2013), to the medical domain. Additionally, we do not make use of WordNet (Miller, 1995) to provide similar terms, but rather employ a Distributional Thesaurus (DT), computed on medical texts.

2 Related Work

For the general domain, the lexical substitution task was initiated by a Semeval-2007 Task (McCarthy and Navigli, 2009). This task was won by an unsupervised method (Giuliano et al., 2007), which uses WordNet for the substitution candidate generation and then relies on the Google Web1T n-grams (Brants and Franz, 2006)[1] to rank the substitutes.

The currently best system, to our knowledge, is proposed by Szarvas et al. (2013). This is a supervised approach, where a single classifier is trained using delexicalized features for all substitutes and can thus be applied even to previously unseen substitutes. Although there have been many approaches for solving the task for the general domain, only slight effort has been done in adapting it to different domains.

3 Method

To perform lexical substitution, we follow the delexicalization framework of Szarvas et al. (2013). We automatically build Distributional Thesauri (DTs) for the medical domain and use features from the Unified Medical Language System (UMLS) ontology. The dataset for supervised lexical substitution consists of sentences, containing an annotated target word t. Considering the sentence being the context for the target word, the target word might have different meanings. Thus annotated substitute candidates $s_{g_1} \dots s_{g_n} \in s_g$, need to be provided for each context. The negative examples are substitute candidates that either are incorrect for the target word, do not fit into the context or both. We will refer to these substitutes as *false substitute candidates* $s_{f_1} \dots s_{f_m} \in s_f$ with $s_f \cap s_g = \emptyset$.

For the generation of substitute candidates we do not use WordNet, as done in previous works (Szarvas et al., 2013), but use only substitutes from a DT. To train a single classifier, features that distinguishing the meaning of words in different context need to be considered. Such features could be e.g. n-grams, features from distributional semantics or features which are extracted

[1] http://catalog.ldc.upenn.edu/LDC2006T13

relative to the target word, such as the ratio between frequencies of the substitute candidate and the target word. After training, we apply the algorithm to unseen substitute candidates and rank them according to their positive probabilities, given by the classifier. Contrary to Szarvas et al. (2013), we do not use any weighting in the training if a substitute has been supplied by many annotators, as we could not observe any improvements. Additionally, we use logistic regression (Fan et al., 2008) as classifier[2].

4 Resources

For the substitutes and for the generation of delexicalized features, we rely on DTs, the UMLS and Google Web1T.

4.1 Distributional thesauri (DTs)

We computed two different DTs using the framework proposed in Biemann and Riedl (2013)[3].

The first DT is computed based on Medline[4] abstracts. This thesaurus uses the left and the right word as context features. To include multi-word expressions, we allow the number of tokens that form a term to be up to the length of three.

The second DT is based on dependencies as context features from a English Slot Grammar (ESG) parser (McCord et al., 2012) modified to handle medical data. The ESG parser is also capable of finding multi-word expressions. As input data we use 3.3 GB of texts from medical textbooks, encyclopedias and clinical reference material as well as selected journals. This DT is also used for the generation of candidates supplied to annotators when creating the gold standard and therefore is the main resource to provide substitute candidates.

4.2 UMLS

The Unified Medical Language System (UMLS) is an ontology for the medical domain. In contrast to Szarvas et al. (2013), which uses WordNet (Miller, 1995) to generate substitute candidates and also for generating features, we use UMLS solely for feature generation.

4.3 Google Web1T

We use the Google Web1T to generate n-gram features as we expect this open domain resource to have considerable coverage for most specific domains as well. For accessing the resource, we use JWeb1T[5] (Giuliano et al., 2007).

[2]We use a Java port of LIBLINEAR (`http://www.csie.ntu.edu.tw/~cjlin/liblinear/`) available from `http://liblinear.bwaldvogel.de/`

[3]We use Lexicographer's Mutual Information (LMI) (Evert, 2005) as significance measure and consider only the top 1000 ($p = 1000$) features per term.

[4]`http://www.nlm.nih.gov/bsd/licensee/2014_stats/baseline_med_filecount.html`

[5]https://code.google.com/p/jweb1t/

5 Lexical Substitution dataset

Besides the lexical substitution data sets for the open domain (McCarthy and Navigli, 2009; Biemann, 2012; Kremer et al., 2014) there is no dataset available that can be used for the medical domain. Therefore, we constructed an annotation task for the medical domain using a medical corpus and domain experts.

In order to provide the annotators with a clear task, we presented a question, and a passage that contains the correct answer to the question. We restricted this to a subset of passages that were previously annotated as justifying the answer to the question. This is related to a textual entailment task, essentially the passage entails the question with the answer substituted for the focus of the question. We instructed the annotators to first identify the terms that were relevant for the entailment relation. For each relevant term we randomly extracted 10 terms from the ESG-based DT within the top 100 most similar terms. Using this list of distributionally similar terms, the annotators selected those terms that would preserve the entailment relation if substituted. This resulted in a dataset of 699 target terms with substitutes. On average from the 10 terms 0.846 are annotated as correct substitutes. Thus, the remaining terms can be used as false substitute candidates.

The agreement on this task by Fleiss Kappa was 0.551 indicating "moderate agreement" (Landis and Koch, 1977). On the metric of pairwise agreement, as defined in the SemEval lexical substitution task, we achieve 0.627. This number is not directly comparable to the pairwise agreement score of 0.277 for the SemEval lexical substitution task (McCarthy and Navigli, 2009) since in our task the candidates are given. However, it shows promise that subjectivity may be reduced by casting lexical substitution into a task of maintaining entailment.

6 Evaluation

For the evaluation we use a ten-fold cross validation and report P@1 (also called Average Precision (AP) at 1) and Mean Average Precision (MAP) (Buckley and Voorhees, 2004) scores. The P@1 score indicates how often the first substitute of the system matches the gold standard. The MAP score is the mean of all AP from 1 to the number of all substitutes.

- Google Web 1T:
  We use the same Google n-gram features, as used in Giuliano et al. (2007) and Szarvas et al. (2013). These are frequencies of n-grams formed by the substitute candidate s_i and the left and right words, taken from the context sentence, normalized by the frequency of the same context n-gram with the target term t. Additionally, we add the same features, normalized by the frequency sum of all n-grams of the substitute candidates. Another feature is generated using the frequencies where t and s are listed together using the words

and, *or* and *","* as separator and also add the left and right words of that phrase as context. Then we normalize this frequency by the frequency of the context occurring only with t.

- DT features:
  To characterize if t and s_i have similar words in common, and therefore are similar, we compute the percentage of words their thesauri entries share, considering the top n words in each entry with $n = 1, 5, 20, 50, 100, 200$. During the DT calculation we also calculate the significances between each word and its context features (see Section 4.1). Using this information, we compute if the words in the sentences also occur as context features for the substitute candidate. A third feature group relying on DTs is created by the overlapping context features for the top m entries of t and s_i with $m = 1, 5, 20, 50, 100, 1000$, which are ranked regarding their significance score. Whereas, the similarities between the trigram-based and the ESG-based DT are similar, the context features are different. Both feature types can be applied to the two DTs. Additionally, we extract the thesaurus entry for the target word t and generate a feature indicating whether the substitute s_i is within the top k entries with $k = 1, 5, 10, 20, 100$ entries[6].

- Part-of-speech n-grams:
  To identify the context of the word we use the POS-tag (only the first letter) of s_i and t as feature and POS-tag combinations of up to three neighboring words.

- UMLS:
  Considering UMLS we look up all concept unique identifiers (CUIs) for s_i and t. The first two features are the number of CUIs for s_i and t. The next features compute the number of CUIs that s_i and t share, starting from the minimal to the maximum number of CUIs. Additionally, we use a feature indicating that s_i and t do not share any CUI.

6.1 Substitute candidates

The candidates for the substitution are taken from the ESG based DT. For each target term we use the gold substitute candidates as correct instances and add all possible substitutes for the same target term occurring in a different context and do not have been annotated as valid in the present context as false instances.

7 Results

Running the experiment, we get the results as shown in Table 1. As baseline system we use the ranking of the ESG-based DT. As can be seen, the baseline is already quite high, which can be attributed to the fact that this resource was used to generate substitutes und thus contains all positive instances. Using the supervised approach, we can beat the baseline by 0.10 for the MAP score and by 0.176 for the P@1 score, which is a significant improvement ($p < 0.0001$, using a two tailed permutation test). To get insights of the contribution of individual feature types, we perform an ablation test. We observe that the most prominent features are coming from the two DTs as we only achieve results below the baseline, when removing DT features. We still obtain significant improvements over the baseline when removing other feature groups. The second most important feature comes from the UMLS. Features coming from the Google n-grams improve the system only slightly. The lowest improvement is derived from the part-of-speech features. This leads us to summarize that a hybrid approach for feature generation using manually created resources (UMLS) and unsupervised features (DTs) leads to the best result for lexical substitution for the medical domain.

System	MAP	P@1
Baseline	0.6408	0.5365
ALL	0.7048	0.6366
w/o DT	0.5798	0.4835
w/o UMLS	0.6618	0.5651
w/o Ngrams	0.7009	0.6252
w/o POS	0.7027	0.6323

Table 1: Results for the evaluation using substitute candidates from the DT.

8 Analysis

For a better insight into the lexical substitution we analyzed how often we outperform the baseline, get equal results or get decreased scores. According to Table 2 in around 26% of the cases we observe a decreased MAP score, which is on average 0.16 smaller then the scores achieved with the baseline. On the other hand, we see improvements in around 39% of the cases: an average improvements of 0.26, which is much higher then the loss. For the remaining 25% of cases we observe the same score.

performance	# of instances	Avg. Δ MAP
decline	180	-0.16
equal	244	0
improvements	275	0.26

Table 2: Error analysis for the task respectively to the MAP score.

Looking inside the data, the largest error class is caused by antonyms. A sub-class of this error are multi-word expressions having an adjective modifier. This problems might be solved by additional features using the UMLS resource. An example is shown in Figure 1.

[6]Whereas in Szarvas et al. (2013) only $k = 100$ is used, we gained an improvement in performance when also adding smaller values of k.

Sentence: he most common cause of thrombocytopenia during pregnancy is gestational thrombocytopenia, which is a **mild thrombocytopenia** with platelet levels remaining greater than 70,000/mL.

Gold: decreased platelet=1

Baseline: decreased platelet:17.0, severe thrombocytopenia: 16.0, macrothrombocytopenia:16.0, prolonged aptt:16.0, normal platelet count:16.0, hepatosplenomegaly:15.0, hypoxaemia:13.0, short finger:12.0, lymphadenopathy:11.0, mild symptom:11.0

System: severe thrombocytopenia:0.272, normal platelet count:0.204, macrothrombocytopenia:0.190, hepatosplenomegaly:0.174, prolonged aptt:0.168, decreased platelet:0.156, mild symptom:0.113, lymphadenopathy:0.085, hypoxaemia:0.067, short finger:0.053

Figure 1: Example sentence for the target term *mild thrombocytopenia*. The system returns a wrong ranking, as the adjective changes the meaning and turns the first ranked term into an antonym.

For feature generation, we currently lookup multi-word expressions as one term, both in the DT and the UMLS resource and do not split them into their single tokens. This error also suggests considering the single words inside the multi-word expression, especially adjectives, and looking them up in a resource (e.g. UMLS) to detect synonymy and antonymy.

Figure 2 shows the case, where the ranking is performed correctly, but the precise substitute is not annotated as a correct one. The term *nail plate* might be even more precise in the context as the manual annotated term *nail bed*. Due to the missing annotation the

Sentence: 7 Yellow nail syndrome is a rare disorder characterized by the triad of yellow and thickened **nails**, lymphedema and respiratory manifestation commonly pleural effusion and other complications like bronchiectasis and chronic sinusitis. Lymphedema in yellow nail syndrome is characteristically non-pitting and involves the lower extremities in symmetric fashion

Gold: finger nail=1, nail bed=1

Baseline: finger nail:58, nail bed:54.0, nail plate:49.0, scalp:47.0, hand:26.0, arms and legs:25.0, eye:23.0, cranial bone:19.0, palm:16.0, urinary organ:16.0

System: nail plate:0.676, finger nail:0.158, nail bed:0.144, scalp:0.106, hand:0.044, arms and legs:0.043, urinary organ:0.017, eye:0.016, palm:0.016, cranial bone:0.014

Figure 2: Example sentence for the target term *nails*. Here the ranking from the system is correct, but the first substitute from the system was not annotated as such.

baseline gets better scores then the result from the system.

9 Conclusion

In summary, we have examined the lexical substitution task for the medical domain and could show that a system for open domain text data can be applied to the medical domain. We can show that following a hybrid approach using features from UMLS and distributional semantics leads to the best results. In future work, we will work on integrating DTs using other context features, as we could see an impact of using two different DTs. Furthermore, we want to incorporate features using n-grams computed on a corpus from the domain and include co-occurrence features.

Acknowledge

We thank Adam Lally, Eric Brown, Eddie Epstein, Chris Biemann and Faisal Chowdhury for their helpful comments.

References

Daniel Bär, Chris Biemann, Iryna Gurevych, and Torsten Zesch. 2012. UKP: Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures. In *Proceedings of the 6th International Workshop on Semantic Evaluation, held in conjunction with the 1st Joint Conference on Lexical and Computational Semantics*, pages 435–440, Montreal, Canada.

Chris Biemann and Martin Riedl. 2013. Text: Now in 2D! A Framework for Lexical Expansion with Contextual Similarity. *Journal of Language Modelling*, 1(1):55–95.

Chris Biemann. 2012. Turk bootstrap word sense inventory 2.0: A large-scale resource for lexical substitution. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Thorsten Brants and Alex Franz. 2006. Web 1t 5-gram corpus version 1. Technical report, Google Research.

Chris Buckley and Ellen M. Voorhees. 2004. Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 25–32, Sheffield, United Kingdom.

Miranda Chong and Lucia Specia. 2011. Lexical generalisation for word-level matching in plagiarism detection. In *Recent Advances in Natural Language Processing*, pages 704–709, Hissar, Bulgaria.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio M. Zanzotto. 2013. Recognizing Textual Entailment: Models and Applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

Stefan Evert. 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations.* Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874.

Claudio Giuliano, Alfio Gliozzo, and Carlo Strapparava. 2007. Fbk-irst: Lexical substitution task exploiting domain and syntagmatic coherence. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 145–148, Prague, Czech Republic.

Gerhard Kremer, Katrin Erk, Sebastian Padó, and Stefan Thater. 2014. What substitutes tell us - analysis of an "all-words" lexical substitution corpus. In *EACL*, pages 540–549.

J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Diana McCarthy and Roberto Navigli. 2009. The English lexical substitution task. *Language Resources and Evaluation*, 43(2):139–159.

M. C. McCord, J. W. Murdock, and B. K. Boguraev. 2012. Deep parsing in watson. *IBM J. Res. Dev.*, 56(3):264–278.

George A. Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, 38:39–41.

György Szarvas, Chris Biemann, and Iryna Gurevych. 2013. Supervised All-Words Lexical Substitution using Delexicalized Features. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*, pages 1131–1141, Atlanta, GA, USA, June.