Applying UMLS for Distantly Supervised Relation Detection

Roland Roller and **Mark Stevenson**
University of Sheffield
Regent Court, 211 Portobello
S1 4DP Sheffield, UK
{R.Roller,M.Stevenson}@dcs.shef.ac.uk

Abstract

This paper describes first results using the Unified Medical Language System (UMLS) for distantly supervised relation extraction. UMLS is a large knowledge base which contains information about millions of medical concepts and relations between them. Our approach is evaluated using existing relation extraction data sets that contain relations that are similar to some of those in UMLS.

1 Introduction

Distant supervision has proved to be a popular approach to relation extraction (Craven and Kumlien, 1999; Mintz et al., 2009; Hoffmann et al., 2010; Nguyen and Moschitti, 2011). It has the advantage that it does not require manually annotated training data. Distant supervision avoids this by using information from a knowledge base to automatically identify instances of a relation from text and use them in order to generate training data for a relation extraction system.

Distant supervision has already been applied to the biomedical domain (Craven and Kumlien, 1999; Thomas et al., 2011). Craven and Kumlien (1999) were the first to apply distant supervision and used the Yeast Protein Database (YPD) to detect sentences containing subcellar-localization relations. Thomas et al. (2011) trained a classifier for protein-protein interactions (PPI) using the knowledge base IntAct and evaluated their approach on different PPI corpora.

There have also been recent applications of distant supervision outside the biomedical domain. The use of Freebase to train a classifier, e.g. (Mintz et al., 2009; Riedel et al., 2010), has proved popular. Other, such as Hoffmann et al. (2010), use Wikipedia info-boxes as the knowledge base.

Applications of distant supervision face several challenges. The main problem is ensuring the quality of the automatically identified training instances identified by the self-annotation. The use of instances that have been incorrectly labelled as positive can lower performance (Takamatsu et al., 2012). Another problem arises when positive examples are included in the set of negative training instances, which can occur when information is missing from the knowledge base (Min et al., 2013; Ritter et al., 2013; Xu et al., 2013).

Evaluation of relation extraction systems that use distant supervision represents a further challenge. In the ideal case an annotated evaluation set is available. Others, such as Ritter et al. (2013) and Hoffmann et al. (2011), use Freebase as knowledge base and evaluate their classifier on an annotated New York Times corpus. However, if no evaluation set is available leave-out can be used where the data identified using distant supervision used for both training and testing (Hoffmann et al., 2010).

This paper makes use of the Unified Medical Language System (UMLS) as a knowledge source for distant supervision. It is widely used for biomedical language processing and readily available. The advantage of UMLS is that it contains information about a wide range of different types of relations and therefore has the potential to generate a large number of relation classifiers. To our knowledge, it has not been used as a knowledge source to train relation extraction systems.

Evaluating such as wide range of relation classifiers is not straightforward due to the lack of gold-standard data. As an alternative approach we make use of existing annotated data sets and identify ones which contain relations that are similar to those included in UMLS.

The next section provides a short description of UMLS. We then describe how we acquire existing data sets to evaluate certain relations. In section 4 we present our first results using UMLS for distant supervision.

Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi) @ EACL 2014, pages 80–84,
Gothenburg, Sweden, April 26-30 2014. ©2014 Association for Computational Linguistics

2 Unified Medical Language System

The *Unified Medical Language System*[1] is a set of files and software which combines different biomedical vocabularies, knowledge bases and standards. The Metathesaurus is a database within UMLS which contains several million biomedical and health related names and concepts and relationships among them. All different names of a concept are unified by the Concept Unique Identifiers (CUI). MRREL is a subset of the Metathesaurus and involves different relationships between different medical concepts defined by a pair of CUIs. Many of them are child-parent relationships, express a synonymy or are vaguely defined as broader or narrower relation. Other relations are more specific, such as *has_location* or *drug_contraindicated_for*. This work focuses on more specific types of relations.

3 Acquiring Evaluation Data Sets

We examined a number of relation extraction data sets in order to identify ones that could be used to evaluate our system. The aim is to find a data set that is annotated with relations that are similar to some of those found in the UMLS. If an appropriate relation can be identified then a relation extraction system can be trained using information from the UMLS and evaluated using the data set.

To determine whether a data set is suitable we used MetaMap (Aronson and Lang, 2010) to identify the CUIs for each related item. We then compared each pair against the MRREL table to determine whether it is included as a relation. To increase coverage we also included parent and child nodes in the mapping process.

Table 1 shows the mappings obtained for two of the data sets: the DDI 2011 data set (Segura-Bedmar et al., 2011) and the data set described by Rosario and Hearst (2004).

The DDI data set contains information about drug-drug interactions and includes a single relation (DDI). The relations it contained were mapped onto 701 CUI pairs. 266 (37.9%) of these mappings could be matched to the MRREL relation *has_contraindicated_drug*. Many of the CUI pairs could also be mapped to the *isa* relationship in MRREL, but this is a very general relationship and the matches are caused by the large number of these in UMLS rather than it being a reasonable match for the DDI relation.

The data set described by Rosario and Hearst (2004) focuses on different relationships between treatments and diseases. The two most common relations TREAT_FOR_DIS (TREAT), denoting the treatment for a particular disease, and PREVENT (PREV), which indicates that a treatment can be used to prevent a disease. The MRREL *isa* relationship also matches many of these relations, again due to its prevalence in MRREL. Other MRREL relations (*may_be_prevented_by* and *may_be_treated_by*) match fewer CUI pairs but seem to be better matches for the TREAT and PREV relations.

Relation	MRREL
DDI (701)	has_contraindicated_drug (266), isa (185), may_treat (57), has_contraindication (51)
PREV (41)	isa (11), may_be_prevented_by (5)
TREAT (741)	isa (172), may_be_treated_by (118)

Table 1: Relation mapping to MRREL

It is important to note that it is not always possible to find a CUI mapping for each entity and the mapping process means that the mapping cannot be guaranteed to be correct in all cases. High coverage does not necessarily mean that a corpus is very similar to a certain MRREL relation, just that many of the CUI pairs which have been mapped to the related entities in the corpus occur often together in a certain MRREL relation. However, in the absence of any other suitable evaluation data we assume that high coverage is an indicator that the relations are strongly similar and use these two data sets for evaluation.

4 Distant Supervision using UMLS

In this section we carry out two different distant supervised experiments using UMLS. The first experiment will be evaluated on a subset of the DDI 2011 training data set using the MRREL relation *has_contraindicated_drug* and *has_contraindication*. The second experiment uses the MRREL relations *may_be_treated_by* and *may_be_prevented_by* and are evaluated on the Rosario & Hearst data set.

We use 7,500,000 Medline abstracts annotated with CUIs using MetaMap (choosing the best mapping as annotation) as a corpus for distant supervision. Our information extraction platform based on a system developed for the BioNLP

[1] https://www.nlm.nih.gov/research/umls/

Shared Task 2013 (Roller and Stevenson, 2013). In contrast to our previous work, our classification process relies on the Shallow Linguistic Kernel (Giuliano et al., 2006) in combination with LibSVM (Chang and Lin, 2011) taking the kernel as input.

4.1 Experiment 1: DDI 2011

The DDI 2011 data set was split into training and test sets for the experiments. Table 2 presents results that place the distant supervision performance in context. The *naive* classification approach predicts all candidate pairs as positive. The *supervised* approach is trained on the training set, using the same kernel method as our distant supervised experiments and evaluated on the test set. This represents the performance that can be obtained using manually labelled training data and can be considered as an upper bound for distant supervision.

Method	Prec. / Recall / F1
naive	0.098 / **1.000** / 0.178
supervised	**0.428** / 0.702 / **0.532**

Table 2: DDI 2011 baseline results

The distant supervision approach requires pairs of positive and negative CUI to be identified. These pairs are used to identify positive and negative examples of the target relation from a corpus. Pairs which occur in our target MRREL relation are used as positive CUI pairs. Negative pairs are generated by selecting pairs of CUIs that are occur in any other MRREL relation.

Sentences containing these CUI pairs are identified in the subset of the MetaMapped Medline. In the basic setup (*basic*), sentences containing a positive pair will be considered as a positive training example. There are many cases where just the occurrence of a positive MRREL pair does not express the target relation. In an effort to remove this noisy data we apply some simple heuristics. The first discards all training instances with more than five words (*5w*) between the two entities, an approach similar to one applied by Takamatsu et al. (2012). The second discards positive sentences containing a comma between the related entities (*com*). We found that commas often indicate a sentence containing a list of items (e.g. genes or diseases) and that these sentences do not form good training examples due to the multiple relations that are possible when there are several items. Finally we also apply a combination of both techniques (*5w+com*).

1000 positive examples were generated using each approach and used for training. Although it would be possible to generate more examples for some approaches, for example *basic*, applying the combination of techniques (*5w+com*) significantly reduces the number of instances available.

Method	has_contraindication *(P./R./F1)*	has_contraindicated _drug *(P./R./F1)*
basic	0.146 / 0.371 / 0.210	0.158 / **0.598** / 0.250
5w	0.109 / **0.641** / 0.187	0.207 / 0.487 / 0.290
com	**0.212** / 0.560 / **0.308**	0.177 / 0.498 / 0.261
5w+com	0.207 / 0.487 / 0.291	**0.214** / 0.471 / **0.294**

Table 3: Evaluation with DDI 2011

Table 3 presents results of the experiments. The results show that all applied techniques for both MRREL relations outperform the naive approach. The best results in terms of F1 score for the *has_contraindication* MRREL relation are obtained using the *com* selection technique. Applying just *5w* leads to worse results than using the *basic* approach. The situation for *has_contraindicated_drug* is different. The classifier provides for all techniques a better F1 score than the *basic* approach. The best results are achieved by using *5w+com*. It is interesting to see, that both MRREL relations provide similar average classification results, even if both relations are different from the target relation and cover completely different CUI pairs. It is also interesting that the MRREL relation *has_contraindication* has a lower coverage to the DDI relation than *has_contraindicated_drug*, but provides slightly better results overall. A problem with the distant supervised classification of these two MRREL relations is their low occurrence in our Medline subset. Using more training data will often lead to better results. In our case, if we apply the combined selection technique, there are fewer positive training instances than are available to the supervised approach, making it difficult to outperform the supervised approach.

4.2 Experiment 2: Rosario & Hearst

The second experiment addresses the problem of detecting the MRREL relations *may_be_prevented_by* and *may_be_treated_by*. Parts of the Rosario & Hearst data set are used to evaluate this relation. This data set differs in structure from the DDI data set. Instead of annotating the entities in the sentence according to its relation, the annotations in the data set indicate whether a certain relation occurs in the sentence. This data set does not contain any negative examples. If a sentence contains two entities, it will always describe a certain relation. A supervised classifier is created by dividing the data set into training and test sets. The test set contains 253 different sentences (221 describe a *TREAT* relation, 15 a *PREV* relation and 17 involve other relationships). Positive and negative CUI pairs are selected in a different way to the previous experiment. The two most frequent relations in the data set are *TREAT* and *PREV*. A classifier for a particular relation is trained using sentences annotated with the corresponding MRREL relation as positive instances. Negative instances are identified using the other relation. For example, the classifier for the *TREAT* relation is trained using positive examples identified using *may_be_treated_by* with negative examples generated using *may_be_treated_by*.

Table 4 shows the baseline results on the data set using a naive and a supervised approach on the two original relations *TREAT* and *PREV*. Performance of the naive approach for *TREAT* is very high since the majority of sentences in the data set are annotated with that relation.

Data Set	Method	Prec. / Recall / F1
TREAT	naive	0.874 / **1.000** / 0.933
	supervised	**0.944** / 0.923 / **0.934**
PREV	naive	0.059 / **1.000** / 0.112
	supervised	**0.909** / 0.667 / **0.769**

Table 4: Rosario & Hearst baseline results

Table 5 shows the results for the various distant supervision approaches. Again, 1000 positive training examples were used to train the classifier. Since the F-Score of the naive and the supervised approaches of *TREAT* are very high, it is difficult to compete with the *may_be_treated_by* distant supervised classifier. However, considering that just 15.9% of the *TREAT* instance pairs of the training set match the MRREL *may_be_treated_by* relation, the results are promising. Furthermore, the precision of all *may_be_treated_by* distant supervised experiments outperform the naive approach. The best results are achieved using *com* as selection technique.

The experiments using the *PREV* relation for evaluation are more interesting. Due to its low occurrence in the test set it is more difficult to detect this relation. The distant supervised classifier trained with the *may_be_prevented_by* relation easily outperforms the naive approach. The best overall F1 scoer results are achieved using the *5w* technique. As expected the distant supervised results are outperformed by the supervised approach. However, the recall for all distantly supervised approaches are at least as high as those obtained using the supervised approach.

	may_be_treated_by	*may_be_prevented_by*
	evaluated on TREAT	evaluated on PREV
Method	*(P./R./F1)*	*(P./R./F1)*
basic	0.926 / 0.733 / 0.818	0.286 / 0.667 / 0.400
5w	0.925 / 0.783 / 0.848	**0.407** / 0.733 / **0.524**
com	**0.928 / 0.819 / 0.870**	0.222 / 0.800 / 0.348
5w+com	0.924 / 0.769 / 0.840	0.361 / **0.867** / 0.510

Table 5: Evaluation with Rosario & Hearst data set

5 Conclusion and Discussion

In this paper we presented first results using UMLS to train a distant supervised relational classifier. Evaluation was carried out using existing evaluation data sets since no resources directly annotated with UMLS relations were available. We showed that using a distantly supervised classifier trained on MRREL relations similar to those found in the evaluation data set provides promising results.

Overall, our system works with some components which should be improved to achieve better results. First, we rely on a cheap and fast annotation using MetaMap, which might produce annotation errors. In addition, the use of noisy distant supervised training data decreases the classification quality. An improvement of the selection process and an improvement of the classification method, such as Chowdhury and Lavelli (2013), could lead to better classification results. In future we would also like to make further use of existing data sets with similar relations to those of interest to evaluate distant supervision approaches.

Acknowledgements

The authors are grateful to the Engineering and Physical Sciences Research Council for supporting the work described in this paper (EP/J008427/1).

References

A. Aronson and F. Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Association*, 17(3):229–236.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

Md. Faisal Mahbub Chowdhury and Alberto Lavelli. 2013. Fbk-irst : A multi-phase kernel based approach for drug-drug interaction detection and classification that exploits linguistic information. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 351–355, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Mark Craven and Johan Kumlien. 1999. Constructing biological knowledge bases by extracting information from text sources. In *In Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology (ISMB)*, pages 77–86. AAAI Press.

Claudio Giuliano, Alberto Lavelli, and Lorenza Romano. 2006. Exploiting shallow linguistic information for relation extraction from biomedical literature. In *In Proc. EACL 2006.*

Raphael Hoffmann, Congle Zhang, and Daniel S. Weld. 2010. Learning 5000 relational extractors. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 286–295, Stroudsburg, PA, USA. Association for Computational Linguistics.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, ACL '11, pages 541–550.

Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek. 2013. Distant supervision for relation extraction with an incomplete knowledge base. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 777–782, Atlanta, Georgia, June. Association for Computational Linguistics.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 1003–1011, Stroudsburg, PA, USA. Association for Computational Linguistics.

Truc-Vien T. Nguyen and Alessandro Moschitti. 2011. End-to-end relation extraction using distant supervision from external semantic repositories. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 277–282, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD '10).*

Alan Ritter, Luke Zettlemoyer, Mausam, and Oren Etzioni. 2013. Modeling missing data in distant supervision for information extraction. In *Association for Computational Linguistics Vol. 1 (TACL).*

Roland Roller and Mark Stevenson. 2013. Identification of genia events using multiple classifiers. In *Proceedings of BioNLP Shared Task 2013 Workshop*, Sofia, Bulgaria, August. Association for Computational Linguistics.

Barbara Rosario and Marti A. Hearst. 2004. Classifying semantic relations in bioscience texts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Isabel Segura-Bedmar, Paloma Martnez, and Daniel Snchez-Cisneros. 2011. The 1st ddi extraction-2011 challenge task: Extraction of drug-drug interactions from biomedical texts. In *Proceedings of DDI Extraction-2011 challenge task.*, pages 1–9.

Shingo Takamatsu, Issei Sato, and Hiroshi Nakagawa. 2012. Reducing wrong labels in distant supervision for relation extraction. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 721–729, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philippe Thomas, Illés Solt, Roman Klinger, and Ulf Leser. 2011. Learning protein protein interaction extraction using distant supervision. In *Proceedings of Robust Unsupervised and Semi-Supervised Methods in Natural Language Processing*, pages 34–41.

Wei Xu, Raphael Hoffmann, Le Zhao, and Ralph Grishman. 2013. Filling knowledge base gaps for distant supervision of relation extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 665–670, Sofia, Bulgaria, August. Association for Computational Linguistics.