ExB Medical Text Miner

Christian Hänig, Robert Remus, Vadim Demchik, and Stefan Bordag

ExB Research & Development GmbH
Seeburgstrasse 100, 04103 Leipzig, Germany
{haenig,remus,demchik,bordag}@exb.de
http://www.exb.de

Abstract. We present *ExB Medical Text Miner* – a text mining pipeline for processing biomedical documents. This application employs state-of-the-art *Named Entity Recognition*, using linguistic features and word embeddings in a fully-connected second-order Conditional Random Field model, as well as a novel two-stage *Relation Extraction* module that first detects entity-level relations using a Support Vector Classifier, then identifies document-level relations by measuring their relevance according to a document topic classification model.

Key words: Named Entity Recognition (NER), Named Entity Normalization, Relation Extraction (RE), Biocuration

1 Introduction

A major motivator for research into text mining in the biomedical domain is to provide a cost-effective way of working with information present in natural language texts about chemicals, diseases, genes and their interactions. The heavy use of domain-specific terms in biomedical documents, and the complexity of the relations between them, makes this a challenging field for research.

This year's *BioCreative V Task 3* focuses on these challenges with two subtasks: *Disease NER and Normalization* (DNER), and *Chemical-induced Diseases RE* (CID). In this paper we present ExB's NLP processing and text mining system as adapted to the biomedical domain. First, we describe our preprocessing pipeline (Section 2), and our approach to NER (Section 3) and RE (Section 4). Finally, we present the results our current system achieves on the official Task 3 test data set (Section 5).

2 Corpora and Preprocessing

To train the *ExB Medical Text Miner*, both for NER and RE, we used the official training (*CDR Train*) and development set (*CDR Devel*) of *BioCreative V: Track 3 - CDR*[1]. For the detection of chemical names, we additionally used the *CHEMDNER Corpus* from *BioCreative IV: Track 2 - CHEMDNER*[2]. The

[1] http://www.biocreative.org/tasks/biocreative-v/track-3-cdr/
[2] http://www.biocreative.org/tasks/biocreative-iv/chemdner/

ExB Medical Text Miner also uses features like unsupervised POS tags and word embeddings produced using models trained with the *PubMed Corpus*[3].

We preprocess documents using an NLP pipeline that comprises:

- Tokenization, with rules adapted from the ones described for *Model 1* in [10].
- Stop-word detection, using in-house tools.
- Sentence boundary detection, using in-house tools.
- Unsupervised POS tagging, based on SVD2 [9].
- Supervised POS tagging, using the Stanford Maximum Entropy tagger[4] [14].
- Lemmatization, using Stanford CoreNLP[5] [11].
- Measurement detection (e.g. *55.8 g/mol* or *10 mg*), using an in-house tool.
- Negation detection with scope, using NegEx[6] [3].
- Syntactic structures, using the Stanford Parser[7] [4].

3 Named Entity Recognition and Normalization

3.1 Feature Extraction

We performed NER using ExB's existing NER ensemble framework [8], but extended it with new features specifically targeting biomedical applications. We drew heavily on [10] for new features oriented towards biomedical tasks.

Feature extraction for a given token uses information about the token itself and a context window of ±3 tokens. Features extracted include plain token strings, supervised and unsupervised POS tags, semantic clusters as described in [7], word shape features as described in [1] and [6], word embeddings (at 150 dimensions) computed by word2vec[8][12] and word embeddings of syntactically dependent words. In addition, we extract binary features identifying chemical formulas, Roman numerals, Greek characters, as well as amino acids and nucleobases from a fixed list. For a given token, we also extract character n-grams up to length 5, including delimiters for the start and end of each token.

3.2 Classification

Recent work in biomedical NER [10] suggests that using second-order Conditional Random Fields (CRFs) achieves superior results in biomedical domains, so we decided to integrate two new CRF libraries into our existing NER ensemble framework:

- CRFSuite, a fast library for first-order CRFs [13].
- *ExBCRF*, our own CRF implementation inspired by CRFSuite, but adding support for second-order CRFs, transitions from all previous and all subsequent labels to the current label (symmetric, fully connected), and automatic restarts using different random seeds.

[3] `http://www.ncbi.nlm.nih.gov/pubmed/`
[4] `http://nlp.stanford.edu/software/tagger.shtml`
[5] `http://nlp.stanford.edu/software/corenlp.shtml`
[6] `http://code.google.com/p/negex/`
[7] `http://nlp.stanford.edu/software/lex-parser.shtml`
[8] `https://code.google.com/p/word2vec/`

In addition to CRF components, we evaluated word lists of chemicals and diseases as a part of the NER ensemble. These word lists were extracted from the MeSH thesaurus[9], then applied to the *CDR Train* data, and words which were over-proposed (i.e. words used more frequently in the data with meanings other than as diseases and chemicals) were dropped. We discovered that integrating the resulting disease name list into our ensemble architecture improves our results, but the chemical name list led to a drop in performance.

3.3 Normalization

Named Entity (NE) normalization maps tokens to MeSH thesaurus ID numbers.

Where a term in the training data is unambiguously associated with just one ID, we store the term in a dictionary. If the term contains a conjunction (i.e. *and*), we store both parts of the conjunction if they are unambiguously associated with a single ID.

We also use the MeSH thesaurus to learn common variations on terms, for example that *antibody* and *antibodies* refer to the same MeSH ID code, therefore if we find *surgeries* we should identify it with the same ID as *surgery*. We train this system by checking the frequencies of transformed tokens in the *CDR Train* and *CDR Devel* corpora, and keep the most significant transformations.

During the normalization process, we look up tokens in the dictionary, getting a MeSH ID directly from there if possible. If not, we test various transformations of the token, acquiring for each one a set of possible MeSH IDs. This is then filtered based on the frequency of terms corresponding to each MeSH ID in the *CDR Train* and *CDR Devel* corpora. We also assess the specificity of each term based on its place in the MeSH thesaurus subject ontology, and select more specific terms over less specific ones whenever possible.

4 Chemical-induced Diseases RE

Chemical-induced Diseases (CID) relations are binary associations between normalized NEs for diseases and for chemicals. We view their extraction as a two part task: identifying relations between diseases and chemicals in texts, and classifying them by their relevance to the documents they are in.

4.1 Entity-level RE

The official training data for this task does not provide entity-level relations for its texts, therefore we used two strategies to obtain disease-chemical entity pairs from texts:

First, we used a simple unsupervised strategy in which every NE identified as a disease in a given document is associated with every NE identified as a chemical. Those CID pairs that match gold standard document-level relations are treated as positive examples of entity-level relations for training, and the others as negative.

[9] US National Library of Medicine Subject Headings: `http://www.nlm.nih.gov/mesh/`

Secondly, we manually annotated the training data for disease and chemical NEs that participate in gold standard document-level CID pairs.

We then trained a Support Vector Classifier using a linear feature kernel to recognize these entity-level relation pairs. We note that others have had success in similar tasks using different kernel types and feature sets derived from syntactic parses (e.g. Bio Event Extraction [2], Drug-Drug-Interaction [5]). However, in our experiments with different learning strategies, we had the most success with a linear kernel. Furthermore, syntactic feature sets do not extend to relations that cross sentence boundaries. Our approach has the benefit of using the same procedures to find intra- and cross-sentence relation pairs.

We extracted the features used for training from each NE participating in an entity-level candidate pair, as well as from a context window that includes all the words between the two NEs and ± 2 tokens around each NE. We included standard features like plain token strings, lemmas, supervised and unsupervised POS tags for the tokens in each NE and in their immediate context. We also used information about the number of NEs of various types (like disease NEs, or chemical NEs) and the number of measurement phrases (e.g. *55.8 g/mol* or *10 mg*) observed in the context of an entity-level pair.

A binary feature indicates whether the entity-level candidate pair falls in whole or part within the scope of a negation. We also extract syntactic features, such as embeddings of dependency-linked tokens, dependency paths between entities and the length of dependency paths between participating NEs. For cross-sentence CID pairs, we automatically extract special cross-sentence patterns (pairs of token sequences from both entities' context), because dependency paths between both entities are not available.

The performance of supervised and unsupervised strategies are compared in Table 1. *Subtask CID, Run 1* corresponds to the supervised candidate entity-level relation selection condition, while *Run 2* corresponds to the unsupervised candidate selection.

4.2 Document-level RE

In general, when a candidate entity-level relation pair is semantically unrelated to the main topic of its document, it does not match a gold standard CID relation in the training data. Consequently, we can quickly remove many candidates by filtering out the topically irrelevant ones.

We trained a document topic classification model to identify both chemicals and diseases that are topically important to a text. For each chemical or disease found this way, we extract a number of features to predict whether or not it will occur in a document-level relation.

Features are extracted at each NE in the document corresponding to that chemical or disease, including plain token strings and sequences of tokens within a ± 4 token window, bags of words within a ± 10 token window (up to sentence boundaries), what part of the document the NE appears in (e.g. title, first sentence of abstract, last sentence of abstract), whether it is first NE of its type in

the text, whether measurement phrases are near it, and the frequency of NEs matching that chemical or disease.

Finally, we identify entity-level CID relations as document-level relations if:

- the entity-level CID relation is the only candidate related to the document's main topics,
- *or* it appears at least twice in the document,
- *or* both NEs participating in the CID relation are related to the document's main topics,
- *or* no other entity-level relation satisfies any of the above conditions.[10]

5 Results

For our participation in *BioCreative V Track 3*, we set up three "runs" for each of the subtasks, measuring the relative success of each variation in our procedure. All models were trained on both the *CDR Train* and *CDR Devel* corpora.

Subtask	Run	Description	P (%)	R (%)	F (%)
DNER	1	*CRFSuite*, without dependency embeddings	90.37	80.28	85.03
	2	*ExBCRF*, with dependency embeddings	90.53	**80.78**	**85.38**
	3	*ExBCRF*, without dependency embeddings	**90.92**	80.13	85.19
		Official baseline	42.71	67.46	52.30
CID	1	Entity-level relations, no dependency features	**50.65**	47.47	**49.01**
	2	Document-level relations, no dependency features	46.73	**48.97**	47.82
	3	Entity-level relations, dependency features	48.61	47.47	48.03
		Official baseline	16.43	76.45	27.05

Table 1. Run definitions and results on official test data for both subtasks

First, we note that for the DNER task (*Subtask DNER* in Table 1), our in-house CRF implementation *ExBCRF* slightly outperforms *CRFSuite* using identical feature sets (*Runs 1 & 3*). This difference is even greater when comparing the number of correct NE appearances identified rather than the disambiguated MeSH ID codes: an improvement of +1.2% on *CDR Devel* when trained using *CDR Train*. Adding *dependency embedding* features increases the F-score performance by another 0.19% (*Run 2*).

For the *CID* task (*Subtask CID* in Table 1), training on entity-level relations (*Run 1*) significantly increases the predictive power of our model (+1.19% in F-score when compared to *Run 2*). Surprisingly, our model with dependency features (*Run 3*) performs worse than without dependency features (−0.98%). We suspect that this is a consequence of parser errors due to atypical constructions common to biomedical scientific literature (e.g. terms like *chemical-induced disease*) that the Stanford parser does not handle correctly.

[10] This last condition makes sense only because we know in advance that there is *at least one* document-level CID relation for every document. Therefore, any relation found in the text is a better guess than no answer at all.

6 Conclusion

We have successfully adapted ExB's NLP processing and text mining system to the biomedical domain and shown its effectiveness on the BioCreative CID task. *ExB Medical Text Miner* achieves state-of-the-art results and placed 4th out of 16 teams in subtask DNER. Document-level RE is a relatively new area in NLP, and with F-scores near 50%, we finished 6th out of 18 teams. Despite not using additional knowledge bases the state-of-the-art results show that our approach generalizes well in resource-scarce situations.

References

1. Bikel, D.M., Schwartz, R., Weischedel, R.M.: An Algorithm that Learns What's in a Name. Machine Learning 34(1-3), 211–231 (1999)
2. Björne, J., Heimonen, J., Ginter, F., Airola, A., Pahikkala, T., Salakoski, T.: Extracting Complex Biological Events with Rich Graph-Based Feature Sets. In: Proceedings of the Workshop on BioNLP: Shared Task. pp. 10–18. ACL (2009)
3. Chapman, W.W., Hilert, D., Velupillai, S., Kvist, M., Skeppstedt, M., Chapman, B.E., Conway, M., Tharp, M., Mowery, D.L., Deleger, L.: Extending the NegEx Lexicon for Multiple Languages. Studies in Health Technology and Informatics 192, 677–681 (2013)
4. Chen, D., Manning, C.D.: A Fast and Accurate Dependency Parser using Neural Networks. In: Proceedings of EMNLP. ACL (2014)
5. Chowdhury, M.F.M., Lavelli, A.: 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 SemEval 2013. pp. 351–355. ACL (2013)
6. Finkel, J.R., Grenager, T., Manning, C.: Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In: Proceedings of the 43rd Annual Meeting of the ACL. pp. 363–370. ACL (2005)
7. Gamallo, P., Bordag, S.: Is singular value decomposition useful for word similarity extraction? Language Resources and Evaluation 45(2), 95–119 (2011)
8. Hänig, C., Bordag, S., Thomas, S.: Modular Classifier Ensemble Architecture for Named Entity Recognition on Low Resource Systems. In: Workshop Proceedings of the 12th Edition of the KONVENS Conference. pp. 113–116 (2014)
9. Lamar, M., Maron, Y., Johnson, M., Bienenstock, E.: SVD and Clustering for Unsupervised POS Tagging. In: Proceedings of ACL 2010. pp. 215–219. ACL (2010)
10. Leaman, R., Wei, C.H., Lu, Z.: tmChem: a high performance approach for chemical named entity recognition and normalization. Journal of Cheminformatics 7(Suppl 1), S3 (2015)
11. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP Natural Language Processing Toolkit. In: Proceedings of 52nd Annual Meeting of the ACL: System Demonstrations. pp. 55–60. ACL (2014)
12. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. In: Proceedings of Workshop at ICLR. pp. 1–12 (2013)
13. Okazaki, N.: CRFsuite: a fast implementation of Conditional Random Fields (CRFs) (2007), `http://www.chokkan.org/software/crfsuite/`
14. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the NAACL 2003. pp. 173–180. ACL (2003)