Efficient and Accurate Entity Recognition for Biomedical Text

Fabio Rinaldi[1], Lenz Furrer[1], Marco Basaldella[2]
1 University of Zurich, 2 Università degli Studi di Udine

Abstract- **This short paper briefly presents an efficient implementation of a named entity recognition system for biomedical entities, which is also available as a web service. The approach is based on a dictionary-based entity recognizer combined with a machine-learning classifier which acts as a filter. We evaluated the efficiency of the ap-proach through participation in the TIPS challenge (BioCreative V.5), where it obtained the best re-sults among participating systems. We separately evaluated the quality of entity recognition and link-ing, using a manually annotated corpus as a reference (CRAFT), where we obtained state-of-the-art results.**

Keywords- named entity recognition; text mining; machine learning; natural language processing.

I. INTRODUCTION

Named entity recognition is most often tackled with knowledge-based approaches (using dictionaries) or example-based approaches (machine learning). Currently the best results are obtained using supervised machine-learning based systems. For extracting chemical names, [9] describes how two CRF classifiers are trained on a corpus of journal abstracts, using different features and model parameters. The approach in [10] also tackles chemical name extraction with CRF, partly using the same software basis as the previous one. For tagging gene names, [14] describes another supervised sequence-labeling approach, using a CRF classifier.

There is growing interest in hybrid systems combining machine learning and dictionary approaches such as the one described in [1], which obtains interesting performance on chemical entity recognition in patent texts.

In the field of entity linking, dictionary-based methods are predominant, since the prediction of arbitrary identi-fiers cannot be modeled in a generalized way. In [6], the authors explore ways to improve established information retrieval techniques for matching protein names and other biochemical entities against ontological resources. The TaggerOne system [8] uses a joint model for tackling NER and linking at the same time – yet another example of a hybrid system that combines machine learning and dictionaries.

II. DATA

The Colorado Richly Annotated Full Text (CRAFT) cor-pus [2, 16] has been built specifically for evaluating these kinds of systems. It consists of 67 full-text articles that have been manually annotated with respect to chemicals, genes, proteins, cell types, cellular components, biological processes, molecular functions, organisms, and biological sequences. In total, the available articles are annotated with over 100,000 concepts.

For our experiments, we used all terminology resources that were distributed with the corpus (which means all annotated entities, except those grounded using Entrez Gene). We regarded species and higher taxonomic ranks (genus, or-der, phylum etc.) from both cellular organisms and viruses as a common entity type "organism". Also, we combined the two non-physical entity types (biological processes and molecular functions) into a single class.

III. METHODS

The OntoGene group has developed an approach for biomedical entity recognition based on dictionary lookup and flexible matching. Their approach has been used in several competitive evaluations of biomedical text mining technologies, often obtaining top-ranked results [12, 13, 11]. Recently, the core parts of the pipeline have been imple-mented in a more efficient framework using Python [4] and are now developed under the name OGER (OntoGene's En-tity Recognizer). These improvements showed to be effec-tive in the BioCreative V.5 shared task [7]: in the techni-cal interoperability and performance of annotation servers (TIPS) task, our system achieved best results in four out of six evaluation metrics [5].

OGER offers a flexible web API for performing dictionary-based NER. It accepts a range of input formats and provides the annotated terms along with identifiers in various output formats. We run an instance of OGER as a permanent web service which is accessible through an API and a web user interface.[1]

For the experiments with the CRAFT corpus, we used the ontologies on which the original annotation was based. We use those resources to compile a non-hierarchical dictionary with 1.26 million terms pointing to 864,000 concept identifiers.
The input documents were tokenized with a simple method based on character class, which collapsed spelling variants such as "SRC 1", "SRC-1", and "SRC1" to a common form.

[1] `https://pub.cl.uzh.ch/projects/ontogene/oger/`

TABLE 1: Performance of our system in entity recognition (top) and entity linking (a.k.a. concept recognition, bottom), compared to the best results reported in [15].

System	Precision	Recall	F1
OGER	0.59	**0.66**	0.62
OGER+NN	**0.86**	0.60	**0.70**
OGER	0.32	**0.52**	0.40
OGER+NN	**0.51**	0.49	**0.50**
cTakes Dict. Lookup	**0.51**	0.43	0.47

All tokens were then converted to low-ercase and stemmed, except for acronyms that collide with a word from general language (e.g. "WAS"). As a further normalization step, Greek letters were expanded to their letter name in Latin spelling, e.g. "α" → "alpha".

In order to improve the system's accuracy, we added a machine-learning filter to remove spurious matches. We used an approach based on neural networks (NN), as they were the best performing algorithm in our previous exper-iments described in [3]. Training is performed using 10-fold cross validation on 47 articles; the evaluation is thus performed on 20 documents only. The features used are mostly shape-based (character count, capitalization), but some include linguistic information (POS, stem) or domain knowledge (frequent pre-/suffixes).

IV. RESULTS

We examined our system in two separate evaluations. We first considered the performance of NER proper, i.e. we re-garded only offset spans and the (coarse) entity type of each annotation produced by each system, ignoring con-cept identifiers. We then evaluated the correctness of the selected concept identifiers. To this end, we augmented the ML-based output with concept identifiers taken from the dictionary-based pre-annotations, which enabled us to draw a fair comparison to previous work in entity linking on the CRAFT corpus.

A. Named Entity Recognition We have compiled very de-tailed results for different configurations and for each en-tity category, however the brevity of this paper allows us to present only aggregated results. The OGER pipeline alone (without filtering) delivers an overall 66% recall score with a precision of 59% over all the entity types consid-ered. Adding the NN-based filtering module, recall drops to 60%, with an increase in precision to 86%, leading to a very competitive F-score of 70%.

B. Concept Recognition We chose a simple strategy to reintroduce the concept identifiers provided by OGER into the output of the ML systems, based on the intersection of the original annotations from OGER's output (which in-clude identifiers) and the annotations left after applying by the NN-based filter. We did not resolve ambiguous an-notations; instead, multiple identifiers could be returned for the same span. While having no disambiguation at all is arguably a deficiency for an entity linking system, it is not imperative that each and every ambiguity is reduced to a single choice. This is particularly true when evaluating against CRAFT, which contains a number of reference an-notations with multiple concept identifiers. For example, in PMID: 16504143, PMCID: 1420314, the term "fish" (oc-curring in the last paragraph of the Discussion section) is assigned six different taxonomic ranks.

This simple strategy allows the system to reach a pre-cision of 51% with a recall of 49% in concept recognition. Compared with the results of several previous systems re-ported in [15], who carried out a series of experiments using the same dataset, our results are already state-of-the-art.

Please note that the results reported by [15] are not per-fectly comparable to the ones we obtained, since the former were tested on the whole CRAFT corpus, while our approach was evaluated on 20 documents only (since we used the remaining documents to train our system) Still, the comparison shows that even a relatively simple approach is sufficient to transform our NER pipeline into an entity linking system with reasonable quality. This is particularly true for the OGER-NN configuration, where both precision and recall are as good as or better than the figures for all the reported systems.

V. CONCLUSION AND FUTURE WORK

In this paper, we presented an efficient, high-quality system for biomedical entity recognition and linking (OGER). We evaluated both processing speed and annotation quality in a series of in-domain experiments using the CRAFT corpus. OGER's scalability and efficiency was also demonstrated in the recently held TIPS task of the BioCreative V.5 chal-lenge. For the NER performance, we used a NN classifier, which acted as a postfilter of the dictionary annotations. The combined system achieved competitive results in en-tity recognition and state-of-the-art results in entity linking over the selected evaluation corpus (CRAFT).

Currently we expose via web API only the OGER service (entity recognition and linking) but without disambigua-tion. As a next step in this research activity, we intend to make available a second web service including disam-biguation. At the same time we are performing additional experiments aimed at improving the quality of the disam-biguation step.

REFERENCES

[1] Saber A Akhondi, Ewoud Pons, Zubair Afzal, Herman van Haagen, Benedikt FH Becker, Kristina M Hettne, Erik M van Mulligen, and Jan A Kors. Chemical entity recognition in patents by combining dictionary-based and statistical approaches. *Database*, 2016:baw061, 2016.

[2] Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A Baumgartner, K Bretonnel Cohen, Karin Verspoor, Judith A Blake, et al. Concept annotation in the CRAFT corpus. *BMC bioinformatics*, 13(1):1, 2012.

[3] Marco Basaldella, Lenz Furrer, Nico Colic, Tilia R Ellendorff, Carlo Tasso, and Fabio Rinaldi. Using a hybrid approach for entity recognition in the biomedical domain. In *Proceedings of the 7th International Symposium on Semantic Mining in Biomedicine (SMBM 2016)*, 2016.

[4] Nicola Colic. Dependency parsing for relation extraction in biomedical literature. Master's thesis, University of Zurich, Switzerland, 2016.

[5] Lenz Furrer and Fabio Rinaldi. OGER: OntoGene's entity recogniser in the BeCalm TIPS task. In *Proceedings of the BioCreative V.5 Challenge Evaluation Workshop*, pages 175–182, 2017.

[6] Tudor Groza and Karin Verspoor. Assessing the impact of case sensitivity and term information gain on biomedical concept recognition. *PloS one*, 10(3):e0119091, 2015.

[7] Martin Krallinger, Martin Pérez-Pérez, Gael Pérez-Rodríguez, Aitor Blanco-Míguez, Florentino Fdez-Riverola, Salvador Cappella-Gutierrez, Anália Lourenço, and Alfonso Valencia. The BioCreative V.5/BeCalm evaluation workshop: tasks, organization, sessions and topics. In *Proceedings of the BioCreative V.5 Challenge Evaluation Workshop*, pages 8–10, 2017.

[8] Robert Leaman and Zhiyong Lu. TaggerOne: joint named entity recognition and normalization with semi-Markov Models. *Bioinformatics*, 32(18):2839, 2016.

[9] Robert Leaman, Chih-Hsuan Wei, and Zhiyong Lu. tmChem: a high performance approach for chemical named entity recognition and normalization. *J. Cheminformatics*, 7(S-1):S3, 2015.

[10] Tsendsuren Munkhdalai, Meijing Li, Khuyagbaatar Batsuren, Hyeon Ah Park, Nak Hyeon Choi, and Keun Ho Ryu. Incorporating domain knowledge in chemical and biomedical named entity recognition with word representations. *Journal of Cheminformatics*, 7(1):S9, 2015.

[11] Fabio Rinaldi, Simon Clematide, Hernani Marques, Tilia Ellendorff, Raul Rodriguez-Esteban, and Martin Romacker. OntoGene web services for biomedical text mining. *BMC Bioinformatics*, 15(14), 2014.

[12] Fabio Rinaldi, Thomas Kappeler, Kaarel Kaljurand, Gerold Schneider, Manfred Klenner, Simon Clematide, Michael Hess, Jean-Marc von Allmen, Pierre Parisot, Martin Romacker, and Therese Vachon. OntoGene in BioCreative II. *Genome Biology*, 9(Suppl 2):S13, 2008.

[13] Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Simon Clematide, Therese Vachon, and Martin Romacker. OntoGene in BioCreative II.5. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(3):472–480, 2010.

[14] Golnar Sheikhshab, Elizabeth Starks, Aly Karsan, Anoop Sarkar, and Inanc Birol. Graph-based semi-supervised gene mention tagging. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 27–35, Berlin, Germany, 2016. Association for Computational Linguistics.

[15] Eugene Tseytlin, Kevin Mitchell, Elizabeth Legowski, Julia Corrigan, Girish Chavan, and Rebecca S Jacobson. NOBLE – Flexible concept recognition for large-scale biomedical natural language processing. *BMC bioinformatics*, 17(1):1, 2016.

[16] Karin Verspoor, Kevin Bretonnel Cohen, Arrick Lanfranchi, Colin Warner, Helen L. Johnson, Christophe Roeder, Jinho D. Choi, Christopher Funk, Yuriy Malenkiy, Miriam Eckert, Nianwen Xue, William A. Baumgartner, Michael Bada, Martha Palmer, and Lawrence E. Hunter. A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. *BMC Bioinformatics*, 13(1):207, 2012.