Syntactic analyses and named entity recognition for PubMed and PubMed Central — up-to-the-minute

Kai Hakala[1,2*], **Suwisa Kaewphan**[1,2,3*], **Tapio Salakoski**[1,3] **and Filip Ginter**[1]
1. Dept. of Information Technology, University of Turku, Finland
2. The University of Turku Graduate School (UTUGS), University of Turku, Finland
3. Turku Centre for Computer Science (TUCS), Finland
kahaka@utu.fi, sukaew@utu.fi,
tapio.salakoski@utu.fi, ginter@cs.utu.fi

Abstract

Although advanced text mining methods specifically adapted to the biomedical domain are continuously being developed, their applications on large scale have been scarce. One of the main reasons for this is the lack of computational resources and workforce required for processing large text corpora.

In this paper we present a publicly available resource distributing preprocessed biomedical literature including sentence splitting, tokenization, part-of-speech tagging, syntactic parses and named entity recognition. The aim of this work is to support the future development of large-scale text mining resources by eliminating the time consuming but necessary preprocessing steps.

This resource covers the whole of PubMed and PubMed Central Open Access section, currently containing 26M abstracts and 1.4M full articles, constituting over 388M analyzed sentences. The resource is based on a fully automated pipeline, guaranteeing that the distributed data is always up-to-date. The resource is available at https://turkunlp.github.io/pubmed_parses/.

1 Introduction

Due to the rapid growth of biomedical literature, the maintenance of manually curated databases, usually updated following new discoveries published in articles, has become unfeasible. This has led to a significant interest in developing automated text mining methods specifically for the biomedical domain.

Various community efforts, mainly in the form of shared tasks, have resulted in steady improvement in biomedical text mining methods (Kim et al., 2009; Segura Bedmar et al., 2013). For instance the GENIA shared tasks focusing on extracting biological events, such as gene regulations, have consistently gathered wide interest and have led to the development of several text mining tools (Miwa et al., 2012; Björne and Salakoski, 2013). These methods have been also successfully applied on a large scale and several biomedical text mining databases are publicly available (Van Landeghem et al., 2013a; Franceschini et al., 2013; Müller et al., 2004). Although these resources exist, their number does not reflect the vast amount of fundamental research invested in the underlying methods, mainly due to the non-trivial amount of manual labor and computational resources required to process large quantities of textual data. Another issue arising from the challenging text preprocessing is the lack of maintenance of the existing databases which in effect nullifies the purpose of text mining as these resources tend to be almost as much out-of-date as their manually curated counterparts. According to MEDLINE statistics[1] 806,326 new articles were indexed during 2015 and thus a text mining resource will miss on average 67 thousand articles each month it hasn't been updated.

In this paper we present a resource aiming to support the development and maintenance of large-scale biomedical text mining. The resource includes all PubMed abstracts as well as full articles from the open access section of PubMed Central (PMCOA), with the fundamental language technology building blocks, such as part-of-speech (POS) tagging and syntactic parses, readily available. In addition, recognition of several bio-

*These authors contributed equally.

[1] https://www.nlm.nih.gov/bsd/bsd_key.html

Proceedings of the 15th Workshop on Biomedical Natural Language Processing, pages 102–107,
Berlin, Germany, August 12, 2016.

logically relevant named entities, such as proteins and chemicals is included. Hence we hope that this resource eliminates the need of the tedious preprocessing involved in utilizing the PubMed data and allows swifter development of new information extraction databases.

The resource is constructed with an automated pipeline which provides weekly updates with the latest articles indexed in PubMed and PubMed Central, ensuring the timeliness of the distributed data. All the data is downloadable in an easily handleable XML format, also used by the widely adapted event extraction system TEES (Björne and Salakoski, 2015). A detailed description of this format is available on the website.

2 Data

We use all publicly available literature from PubMed and PubMed Central Open Access subset, which cover most of the relevant literature and are commonly used as the prime source of data in biomedical text mining knowledge bases.

PubMed provides titles and abstracts in XML format in a collection of baseline release and subsequent updates. The former is available at the end of each year whereas the latter is updated daily. As this project was started during 2015, we have first processed the baseline release from the end of 2014 and this data has then been extended with the new publications from the end of 2015 baseline release. The rest of the data up to date has been collected from the daily updates.

The full articles in PMC Open Access subset (PMCOA) are retrieved via the PMC FTP service. Multiple types of data format are provided in PMCOA, including NXML and TXT formats which are suitable for text processing. We use the provided NXML format as it is compatible with our processing pipeline. This service does not provide distinct incremental updates, but a list of all indexed articles updated weekly.

3 Processing Pipeline

In this section, we discuss our processing pipeline as shown in Figure 1. Firstly, both PubMed and PMCOA documents are downloaded from NCBI FTP services. For the periodical updates of our resource this is done weekly — the same interval the official PMCOA dataset is updated. From the PubMed incremental updates we only include newly added documents and ignore other updates. As the PMCOA does not provide incremental updates, we use the index file and compare it to the previous file list to select new articles for processing.

Even though the PubMed and PMCOA documents are provided in slightly different XML formats, they can be processed in similar fashion. As a result, the rest of the pipeline discussed in this section is applied to both document types.

Both PubMed XML articles and PMCOA NXML full texts are preprocessed using publicly available tools[2] (Pyysalo et al., 2013). These tools convert XML documents to plain text and change character encoding from UTF-8 to ASCII as many of the legacy language processing tools are incapable of handling non-ASCII characters. Additionally, all excess meta data is removed, leaving titles, abstracts and full-text contents for further processing. These documents are subsequently split into sentences using GENIA sentence splitter (Sætre et al., 2007) as most linguistic analyses are done on the sentence level. GENIA sentence splitter is trained on biomedical text (GENIA corpus) and has state-of-the-art performance on this domain.

The whole data is parsed with the BLLIP constituent parser (Charniak and Johnson, 2005), using a model adapted for the biomedical domain (McClosky, 2010), as provided in the TEES processing pipeline. The distributed tokenization and POS tagging are also produced with the parser pipeline. We chose to use this tool as the performance of the TEES software has been previously evaluated on a large-scale together with this parsing pipeline (Van Landeghem et al., 2013b) and it should be a reliable choice for biomedical relation extraction. Since dependency parsing has become the prevalent approach in modeling syntactic relations, we also provide conversions to the collapsed Stanford dependency scheme (De Marneffe et al., 2006).

The pipeline is run in parallel on a cluster computer with the input data divided into smaller batches. The size of these batches is altered along the pipeline to adapt to the varying computational requirements of the different tools.

3.1 Named Entity Recognition

Named entity recognition (NER) is one of the fundamental tasks in BioNLP as most of the cru-

[2] https://github.com/spyysalo/nxml2txt

Entity type	Our system Precision/Recall/F-score	State-of-the-art system Precision/Recall/F-score	References
Cell line	89.88 / 84.36 / 87.03	91.67 / 85.47 / 88.46	(Kaewphan et al., 2016)
Chemical	85.27 / 82.92 / 84.08	89.09 / 85.75 / 87.39	(Leaman et al., 2015)
Disease*	86.32 / 80.83 / 83.49	82.80 / 81.90 / 80.90	(Leaman et al., 2013)
GGP**	74.27 / 72.99 / 73.62	90.22 / 84.82 / 87.17	(Campos et al., 2013)
Organism	77.15 / 80.15 / 78.63	83.90 / 72.60 / 77.80	(Pafilis et al., 2013)

Table 1: Evaluation of the named entity recognition for each entity type on the test sets, measured with strict entity level metrics. Reported results for corresponding state-of-the-art approaches are shown for comparison.
* The evaluation of the best performing system for disease mentions is the combination of named entity recognition and normalization.
** The official BioCreative II evaluation for our GGP model results in 84.67, 84.54 and 84.60 for precision, recall and F-score respectively. These numbers are comparable to the listed state-of-the-art method.

cial biological information is expressed as relations among entities such as genes and proteins. To support further development on this dataset, we provide named entity tagging for five entity types, namely diseases, genes and gene products (GGPs), organisms, chemicals, and cell line names. Although several tools with state-of-the-art performance are available for these entity types (Leaman et al., 2015; Leaman and Gonzalez, 2008), we have decided to use a single tool, NERsuite[3], for all types. NERsuite is based on conditional random field classifiers as implemented in the CRFsuite software (Okazaki, 2007). Having a single tool for this processing step instead of using the various state-of-the-art tools is critical for the maintainability of the processing pipeline. NERsuite was selected as several biological models are readily available for this software (Kaewphan et al., 2016; Pyysalo and Ananiadou, 2014) and as it supports label weighting (Minkov et al., 2006) unlike many other NER tools.

For cell line names we use a publicly available state-of-the-art model (Kaewphan et al., 2016), whereas for the other entity types we train our own models with manually annotated data from GENETAG (Tanabe et al., 2005), CHEMDNER (Krallinger et al., 2015), SPECIES (Pafilis et al., 2013) and NCBI disease (Doğan et al., 2014) corpora for GGPs, chemicals, organisms and diseases, respectively. All these corpora are comprised of biomedical articles and should thus reflect well the text types seen in PubMed.

All used corpora provide the data divided to training, development and test sets in advance, the SPECIES corpus being an exception. For this corpus we do our own data division with random sampling on document level, for each taxonomy category separately. For each entity type, the C2 value, as well as the label weights are selected to optimize the F-score on the development set. For the training of the final models used in the resource, we use the whole corpora, i.e. the combination of training, development and test sets.

Detailed performance evaluations for all entity types are shown in Table 1. We evaluate NERsuite in terms of precision, recall and F-score against the test data using "strict matching" criteria, i.e. only consider the tagged entities correct if they are perfectly matched with the gold standard data. These results may not be directly comparable to the results reported in other studies as relaxed evaluation methods are sometimes used. However, we can conclude that our system is on par with the methods published elsewhere and the limitation of using a single tool does not have a significant negative impact on the overall performance.

4 Data Statistics

During the time of writing this paper the dataset included 25,512,320 abstracts from PubMed and 1,350,119 full articles from PMCOA, resulting in 155,356,970 and 232,838,618 sentences respectively. These numbers are not identical to the ones reported by NCBI for couple of reasons. Firstly, at the moment, we do not process the deletion updates nor do we remove the old versions of PMCOA articles if they are revised, i.e. our dataset may include articles, which have been retracted and an article may be included multiple times if

[3] http://nersuite.nlplab.org/

Entity type	Occurrences	Most common entity spans
Cell line	6,967,903	HeLa, MCF-7, A549, HepG2, MDA-MB-231
Chemical	153,285,486	glucose, N, oxygen, Ca2+, calcium
Disease	105,416,758	tumor, cancer, HIV, breast cancer, tumors
GGP	190,543,270	insulin, GFP, p53, TNF-alpha, IL-6
Organism	69,962,111	human, mice, mouse, HIV, humans

Table 2: Occurrence counts and the most frequent entity spans for all entity types in the whole data set.

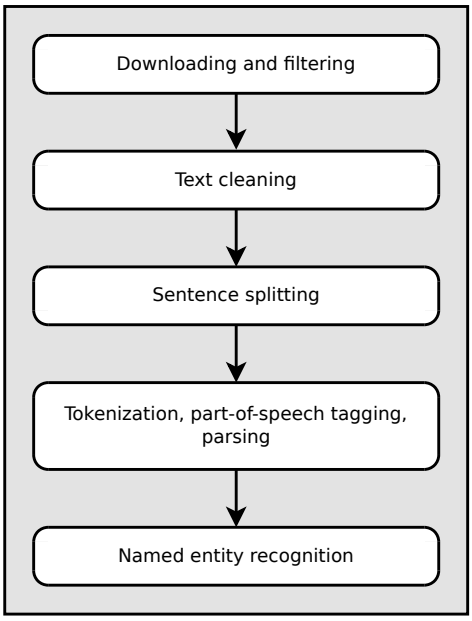

Figure 1: The main processing steps of the pipeline. First, the articles are downloaded from the source and filtered to prevent reprocessing old documents. The documents are then converted to plain text format. This text data is split to independent sentences, tokenized and tagged with POS labels and syntactic dependencies. In addition, named entity recognition for several entity types is carried out.

the content has been modified. We plan to take the deletions into account in near future. Secondly, the external tools in our pipeline may occasionally fail, in which case some of the articles are not processed. Since the pipeline processes the input data in batches, a critical error may lead to a whole batch not being processed. We are currently improving the pipeline to automatically reprocess the failed batches with the problematic articles excluded to minimize the loss of data.

Running the parsing pipeline, including tokenization, POS tagging and conversion to the collapsed Stanford scheme, is the most time consuming part of the whole pipeline. Execution of this step has taken 84,552 CPU hours (9.6 CPU years) for the currently available data.

Unfortunately we do not have exact processing time statistics for named entity recognition and thus estimate its computational requirements by extrapolating from a smaller test run. Based on this experiment NER has demanded 4,100 CPU hours thus far. The text preprocessing and sentence splitting steps are negligible and thus the overall processing time required is approximately 10 CPU years.

In total, our processing pipeline has detected 526,175,528 named entities. GGPs are the most common entities, covering 36.2% of all entity mentions, whereas the cell lines are the most infrequent, forming only 1.3% of the data. The entity type specific statistics along with the most common entity spans are listed in Table 2.

5 Future Work

Our future efforts will focus on expanding the coverage of supported entity types to mutations and anatomical entities (Wei et al., 2013; Pyysalo and Ananiadou, 2014), deepening the captured information of biological processes and bringing text mining one step closer to extracting a realistic view of biological knowledge.

As many of the NER training corpora include only abstracts and are limited to specific domains, the generalizability of the trained NER models to full articles and to the wide spectrum of topics covered in PubMed is not clear. Thus we wish to assess how well these models perform on large-scale datasets and analyze how their performance could be improved on out-of-domain documents.

We plan to also include entity normalization for all supported types, but as we wish to minimize the number of individual tools in the processing pipeline, we are developing a generic approach suitable for most entity types.

6 Conclusions

We have introduced a new resource which provides the basic linguistic analyses, essential in the development of text mining knowledge bases, for the whole of PubMed and PubMed Central Open Access section, thus drastically reducing the amount of required preprocessing efforts.

In addition, we provide named entity tagging for several biologically relevant entity types and show that the models we have used are comparable to the state-of-the-art approaches, although our focus has been on retaining the processing pipeline as simple as possible for easier maintenance.

The resource is periodically updated with an automated pipeline, and currently includes over 26M documents fully parsed with 526M named entity mentions detected. The data is available for download in XML format.

Acknowledgments

Computational resources were provided by CSC - IT Center For Science Ltd., Espoo, Finland. This work was supported by ATT Tieto käyttöön grant.

References

Jari Björne and Tapio Salakoski. 2013. TEES 2.1: Automated annotation scheme learning in the BioNLP 2013 Shared Task. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 16–25.

Jari Björne and Tapio Salakoski. 2015. TEES 2.2: biomedical event extraction for diverse corpora. *BMC Bioinformatics*, 16(16):1–20.

David Campos, Sérgio Matos, and José Luís Oliveira. 2013. Gimli: open source and high-performance biomedical name recognition. *BMC bioinformatics*, 14(1):54.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine N-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL'05, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marie-Catherine De Marneffe, Bill MacCartney, and Christopher D Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1 – 10.

Andrea Franceschini, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer Bork, Christian von Mering, and Lars J. Jensen. 2013. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*, 41(D1):D808–D815.

Suwisa Kaewphan, Sofie Van Landeghem, Tomoko Ohta, Yves Van de Peer, Filip Ginter, and Sampo Pyysalo. 2016. Cell line name recognition in support of the identification of synthetic lethality in cancer from text. *Bioinformatics*, 32(2):276–282.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 Shared Task on event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pages 1–9. Association for Computational Linguistics.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M. Lowe, Roger A. Sayle, Riza Theresa Batista-Navarro, Rafal Rak, Torsten Huber, Tim Rocktäschel, Sérgio Matos, David Campos, Buzhou Tang, Hua Xu, Tsendsuren Munkhdalai, Keun Ho Ryu, SV Ramanan, Senthil Nathan, Slavko Žitnik, Marko Bajec, Lutz Weber, Matthias Irmer, Saber A. Akhondi, Jan A. Kors, Shuo Xu, Xin An, Utpal Kumar Sikdar, Asif Ekbal, Masaharu Yoshioka, Thaer M. Dieb, Miji Choi, Karin Verspoor, Madian Khabsa, C. Lee Giles, Hongfang Liu, Komandur Elayavilli Ravikumar, Andre Lamurias, Francisco M. Couto, Hong-Jie Dai, Richard Tzong-Han Tsai, Caglar Ata, Tolga Can, Anabel Usié, Rui Alves, Isabel Segura-Bedmar, Paloma Martínez, Julen Oyarzabal, and Alfonso Valencia. 2015. The CHEMDNER corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, 7(1):1–17.

Robert Leaman and Graciela Gonzalez. 2008. BANNER: an executable survey of advances in biomedical named entity recognition. In *Pacific Symposium on Biocomputing*, volume 13, pages 652–663. Citeseer.

Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. 2013. DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917.

Robert Leaman, Chih-Hsuan Wei, and Zhiyong Lu. 2015. tmChem: a high performance approach for chemical named entity recognition and normalization. *Journal of Cheminformatics*, 7(1):1–10.

David McClosky. 2010. *Any domain parsing: automatic domain adaptation for natural language parsing*. Ph.D. thesis, Providence, RI, USA.

Einat Minkov, Richard Wang, Anthony Tomasic, and William Cohen. 2006. NER systems that suit user's preferences: adjusting the recall-precision trade-off for entity extraction. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 93–96, New York City, USA, June. Association for Computational Linguistics.

Makoto Miwa, Paul Thompson, John McNaught, Douglas B Kell, and Sophia Ananiadou. 2012. Extracting semantically enriched events from biomedical literature. *BMC bioinformatics*, 13(1):1.

Hans-Michael Müller, Eimear E Kenny, and Paul W Sternberg. 2004. Textpresso: An ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*, 2(11), 09.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of Conditional Random Fields (CRFs).

Evangelos Pafilis, Sune P. Frankild, Lucia Fanini, Sarah Faulwetter, Christina Pavloudi, Aikaterini Vasileiadou, Christos Arvanitidis, and Lars Juhl Jensen. 2013. The SPECIES and ORGANISMS resources for fast and accurate identification of taxonomic names in text. *PLoS ONE*, 8(6):1–6, 06.

Sampo Pyysalo and Sophia Ananiadou. 2014. Anatomical entity mention recognition at literature scale. *Bioinformatics*, 30(6):868–875.

Sampo Pyysalo, Filip Ginter, Hans Moen, Tapio Salakoski, and Sophia Ananiadou. 2013. Distributional semantic resources for biomedical text mining. In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine*, pages 39–44.

Rune Sætre, Kazuhiro Yoshida, Akane Yakushiji, Yusuke Miyao, Yuichiro Matsubayashi, and Tomoko Ohta. 2007. AKANE system: protein-protein interaction pairs in BioCreAtIvE2 challenge, PPI-IPS subtask. In *Proceedings of the BioCreative II*, pages 209–212.

Isabel Segura Bedmar, Paloma Martínez, and María Herrero Zazo. 2013. SemEval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (DDIextraction 2013). In *In Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval*. Association for Computational Linguistics.

Lorraine Tanabe, Natalie Xie, Lynne H Thom, Wayne Matten, and W John Wilbur. 2005. GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC bioinformatics*, 6(1):1.

Sofie Van Landeghem, Jari Björne, Chih-Hsuan Wei, Kai Hakala, Sampo Pyysalo, Sophia Ananiadou, Hung-Yu Kao, Zhiyong Lu, Tapio Salakoski, Yves Van de Peer, and Filip Ginter. 2013a. Large-scale event extraction from literature with multi-level gene normalization. *PLoS ONE*, 8(4):1–12, 04.

Sofie Van Landeghem, Suwisa Kaewphan, Filip Ginter, and Yves Van de Peer. 2013b. Evaluating large-scale text mining applications beyond the traditional numeric performance measures. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*, pages 63–71, Sofia, Bulgaria, August. Association for Computational Linguistics.

Chih-Hsuan Wei, Bethany R Harris, Hung-Yu Kao, and Zhiyong Lu. 2013. tmVar: a text mining approach for extracting sequence variants in biomedical literature. *Bioinformatics*, 29(11):1433–1439.