Distributional Semantics Resources for Biomedical Text Processing

Sampo Pyysalo[1] Filip Ginter[2] Hans Moen[3] Tapio Salakoski[2] Sophia Ananiadou[1]
1. National Centre for Text Mining and School of Computer Science
University of Manchester, UK
2. Department of Information Technology
University of Turku, Finland
3. Department of Computer and Information Science
Norwegian University of Science and Technology, Norway
sampo@pyysalo.net ginter@cs.utu.fi hans.moen@idi.ntnu.no
tapio.salakoski@utu.fi sophia.ananiadou@manchester.ac.uk

Abstract

The openly available biomedical literature contains over 5 billion words in publication abstracts and full texts. Recent advances in unsupervised language processing methods have made it possible to make use of such large unannotated corpora for building statistical language models and inducing high quality vector space representations, which are, in turn, of utility in many tasks such as text classification, named entity recognition and query expansion. In this study, we introduce the first set of such language resources created from analysis of the entire available biomedical literature, including a dataset of all 1- to 5-grams and their probabilities in these texts and new models of word semantics. We discuss the opportunities created by these resources and demonstrate their application. All resources introduced in this study are available under open licenses at http://bio.nlplab.org.

1 Introduction

Despite efforts to create annotated resources for various biomedical natural language processing (NLP) tasks, the number of unannotated domain documents dwarfs that of annotated documents by many orders of magnitude. The PubMed literature database provides access to over 23 million citations, of which nearly 14 million include an abstract. The biomedical sciences are also at the forefront of the shift toward open-access (OA) publication (Laakso and Björk, 2012), with the PubMed Central (PMC) OA subset containing nearly 700,000 full-text articles in an XML format.[1] Together, these two resources constitute an unannotated corpus of 5.5 billion tokens, effectively covering the entire available biomedical scientific literature and forming a representative corpus of the domain (Verspoor et al., 2009).

The many opportunities created by the availability of large unannotated corpora for various NLP methods are well established (see e.g. Ratinov and Roth (2009)), and models induced from unannotated texts have been considered also in a number of recent biomedical NLP studies (Stenetorp et al., 2012; Henriksson et al., 2012). A particular focus of recent research interest are models of meaning induced from unannotated text, with numerous methods introduced for capturing both the semantics of words as well as those of phrases or whole sentences (Mnih and Hinton, 2008; Collobert and Weston, 2008; Turian et al., 2010; Huang et al., 2012; Socher et al., 2012). Although such approaches generally produce better models with more data, their computational complexity has largely limited their application to corpus sizes far below that of the biomedical literature. Recently, a number of efforts have introduced new language resources derived from very large corpora and demonstrated approaches that allow word representations to be induced from corpora of billions of words (Lin et al., 2010; Mikolov et al., 2013). However, despite the relevance of such approaches to biomedical language processing, there have to the best of our knowledge been no attempts to apply them specifically to the biomedical literature.

Corpora containing billions of words can represent challenges even for fully automatic processing, and most domain efforts consequently focus

[1] In this study, we do not consider PDF supplementary materials (see e.g. Yepes and Verspoor (2013)).

	Subset		
	PubMed	PMC OA	Total
Documents	22,120,269	672,589	22,792,858
Sentences	124,615,674	105,194,341	229,810,015
Tokens	2,896,348,481	2,591,137,744	5,487,486,225

Table 1: PubMed and the PMC OA statistics, representing the entire openly available biomedical literature. Note that PubMed statistics omit documents found also in PMC OA, and that only approximately 14 million of PubMed documents include an abstract.

n	#
1	24,181,640
2	230,948,599
3	1,033,760,199
4	2,313,675,095
5	3,375,741,685

Table 2: Counts of unique n-grams.

AFUB_038070
epicardin/capsulin/Pod-1-mediated
22-methoxydocosan-1-ol
mmHg/101.50+/-12.86
5.26@1000
40.87degrees
electromyocinesigraphic
(1-5)-KDO
overpressurizing
rootsanel

Table 3: A random sample of 10 tokens appearing exactly once in the openly available literature.

only on small subsets of the literature at a time. To avoid duplication of efforts, it is therefore desirable to build and distribute standard datasets that can be utilized by the community. In this work, we introduce and evaluate new language resources derived from the entire openly available biomedical scientific literature, releasing these resources to the community under open licenses to encourage further exploration and applications of literature-scale resources for biomedical text processing.

2 Materials and methods

2.1 Text sources

Article titles and abstracts were drawn from the PubMed distribution as of the end of September 2013, constituting in total 22,723,471 records. Full-text articles were, in turn, sourced from the PubMed Central Open Access (PMC OA) section, again as of the end of September 2013, and constitute 672,589 articles. PubMed abstracts for articles that are also present in PMC OA were discarded, so as to avoid the duplication of the abstract, which is also part of the PMC full text.

2.2 Text preprocessing

We first extracted document titles and abstracts from the PubMed XML and extracted all text content of the PMC OA articles using the full-text article extraction pipeline[2] introduced for the BioNLP Shared Task 2011 (Stenetorp et al., 2011). Since the pipeline extracts all text content, also including sections not desired for the current resource such as author affiliations and lists of references, we used a custom script to post-process the output and preserve only text from the title, abstract, and main body of the articles. We further removed inline formulae. Both for the abstracts and the full-text articles, Unicode characters were mapped to ASCII using the replacement table also used in the BioNLP Shared Task pipeline. This step is motivated by the number of commonly used NLP tools which do not handle Unicode-encoded text correctly, as well as the normalization gained from mapping, for example, the character β to the ASCII string *beta* — both of which are common in the input text. The extracted text was then segmented into sentences using the GENIA sentence splitter[3] and tokenized using a custom tokenization script replicating the tokenizer used in the GENIA Tagger (Tsuruoka et al., 2005). The resulting corpus consists in total of 5.5B tokens in 230M sentences. Detailed statistics are shown in Table 1.

2.3 N-grams

All 1- to 5-grams from the data were collected using the KenLM Language Model Toolkit (Heafield et al., 2013) and a custom tool[4] based on HAT-tries (Askitis and Sinha, 2007). The counts of unique

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

[3] https://github.com/ninjin/geniass

[4] https://github.com/spyysalo/ngramcount

Word2vec				Random Indexing			
Input: cysteine		Input: methylation		Input: cysteine		Input: methylation	
Word	Distance	Word	Distance	Word	Distance	Word	Distance
cystein	0.865653	hypermethylation	0.815192	lysine	0.975116	hypermethylation	0.968435
serine	0.804936	hypomethylation	0.810420	proline	0.968552	acetylation	0.967535
Cys	0.798540	demethylation	0.780071	threonine	0.963178	fragmentation	0.961802
histidine	0.782239	methylated	0.749713	arginine	0.963163	plasticity	0.960208
proline	0.771344	Methylation	0.749538	histidine	0.962816	hypomethylation	0.959995
Cysteine	0.769645	methylations	0.745969	glycine	0.960027	replication	0.959925
aspartic	0.750118	acetylation	0.740044	tryptophan	0.959929	deletions	0.956500
active-site	0.745223	DNA-methylation	0.739505	methionine	0.959649	disturbance	0.955987
asparagine	0.735614	island1	0.738123	serine	0.958578	pathology	0.954187
cysteines	0.725626	hyper-methylation	0.730208	Cys	0.953123	asymmetry	0.953079

Table 4: Nearest words for selected inputs in the two models.

n-grams are shown in Table 2. Of the 24M unique tokens, a full 14M are singleton occurrences. To illustrate the long tail, ten randomly selected singleton tokens are shown in Table 3.

Having precomputed all n-grams enables an efficient way of building word vectors, utilizing the fact that the list of n-grams includes all unique windows focused on each word in the corpus together with their count (or, correspondingly, probability). This makes the n-gram model a compressed representation of the corpus with all salient information needed to build a distributional similarity model. As opposed to the standard technique of sliding a window across the corpus, one can instead aggregate the information directly from the n-grams.

2.4 Word vectors from n-grams with Random Indexing

Random indexing (Kanerva et al., 2000) is a method for building a semantic word vector model in an incremental fashion. First, every word is assigned an *index vector* with all elements equal to zero, except for a small number of randomly distributed +1 and -1 values. The vector space representation of a given word is then obtained by summing up the index vectors of all words in all its context windows in the corpus.

We used an existing implementation of random indexing[5] that we modified to consider each 3-gram as the left half window of the rightmost word, as well as the right half window of the leftmost word. The index vectors are weighted by their corresponding probability. For the training we used vector dimensionality of 400, 4 non-zeros in the index vectors, and shifted index vectors in the same way as was done for *direction vectors* by Sahlgren et al. (2008). We also weighted the index vectors by their distance to the target word according to the following equation: $weight_i = 2^{1-dist_{it}}$ where $dist_{it}$ is the distance to the target term. The run took approximately 7.7 hours on a 16-core system and the compressed model occupies 3.6GB on disk. See Table 4 for an illustration of the similarities captured by the word vectors.

2.5 `word2vec` word vectors

We also applied the `word2vec`[6] implementation of the method proposed by Mikolov et al. (2013) to compute additional vector representations and to induce word clusters. The algorithm is based on neural networks and has been shown to outperform more traditional techniques both in terms of the quality of the resulting representations as well as in terms of computational efficiency. A primary strength of the class of models introduced by Mikolov et al. in comparison to conventional neural network models is that they use a single linear projection layer, thus omitting a number of costly calculations commonly associated with neural networks and making application to much larger data sets than previously proposed methods feasible. We specifically induce 200-dimensional vectors applying the *skip-gram* model with a window size of 5. The model works by predicting the context words within the window focused on each word (see Mikolov et al. for details). Once the vector representation of each word is computed, the words are further clustered with the k-means clustering algorithm with $k = 1000$.

We applied `word2vec` to create three sets of word vectors: one from all PubMed texts, one from all PMC OA texts, and one from the combination of all PubMed and PMC OA texts. For the PubMed and PMC OA subsets, the processing required approx. 12 hours on a 12-core system and

[5] `http://www.nada.kth.se/~xmartin/java/`

[6] `https://code.google.com/p/word2vec/`

Method	Corpus AnEM	BC2GM	NCBID
NERsuite	69.31 / 50.16 / 58.20	74.39 / 75.21 / 74.80	84.41 / 81.69 / 83.02
+ Word clusters	66.43 / 53.11 / 59.03	78.14 / 73.96 / **75.99**	86.91 / 80.12 / **83.38**
Stenetorp et al.	72.90 / 55.89 / **63.27**	74.71 / 66.78 / 70.52	83.86 / 77.84 / 80.73

Table 5: Effect of features derived from `word2vec` word clusters on entity mention tagging (precision/recall/F-score). The best results achieved in a previous evaluation using multiple word representations (Stenetorp et al., 2012) are given for reference.

consumed at peak approx. 4.5GB of memory. The combination of the two took 24 hours and 7.5GB of memory. The resulting vector representations for the three sets are 2-3GB in size. Table 4 shows the nearest words (cosine distance) to selected input words.

3 Extrinsic evaluation

To assess the quality of the word vectors and the clusters created from these vectors, we performed a set of entity mention tagging experiments using three biomedical domain corpora representing various tagging tasks: the BioCreative II Gene Mention task corpus (Smith et al., 2008) (gene and protein names), the Anatomical Entity Mention (AnEM) corpus (Ohta et al., 2012) (anatomical entity mentions) and the NCBI Disease (NCBID) corpus (Doğan and Lu, 2012) (disease names). We compare the results with those of Stenetorp et al. (2012), who previously applied these three corpora in a similar setting to evaluate multiple word representations induced from smaller corpora.

To perform the evaluation, we applied AnatomyTagger (Pyysalo and Ananiadou, 2013), an entity mention tagger using the NERsuite[7] toolkit built on the CRFsuite (Okazaki, 2007) implementation of Conditional Random Fields. For each corpus, we trained one model with default features, and another that augmented the feature set with the cluster ID of each word. We selected hyperparameters ($c2$ and label bias) separately for each corpus and feature set using a grid search with evaluation on the corpus development set. We then trained a final model on the combination of training and development sets, and evaluated it on the test set. We measure performance using exact matching, requiring both tagged mention types and their spans to be precisely correct.[8]

Table 5 shows the extrinsic evaluation results. We find that the word representations are beneficial for tagging performance for all three corpora, improving the performance of a state-of-the-art tagger and surpassing the previously reported results in two out of three cases.

4 Conclusion

We have introduced several resources of general interest to the BioNLP community. First, we assembled a pipeline which fully automatically produces a reference conversion from the complex PubMed and PubMed Central document XML formats into ASCII text suitable for standard text processing tools. Second, we induced 1- to 5-gram models from the entire corpus of over 5 billion tokens. Third, we induced vector space representations using the word2vec and random indexing methods, producing the first word representations induced from the entire available biomedical literature. These can serve as drop-in solutions for BioNLP studies that can benefit from precomputed vector space representations and language models.

In addition to building the resources and making them available, we also illustrated the use of these resources for various named entity recognition tasks. Finally, we have demonstrated the potential of calculating semantic vectors from an existing n-gram based language model using random indexing. All tools and resources introduced in this study are available under open licenses at `http://bio.nlplab.org`.

Acknowledgments

We thank the Chikayama-Tsuruoka lab of the University of Tokyo and the CSC — IT Center for Science of Finland for computational resources and Pontus Stenetorp for input regarding word representations.

[7] `http://nersuite.nlplab.org`

[8] Note that this criterion is stricter than used in some previous studies on these corpora.

References

Nikolas Askitis and Ranjan Sinha. 2007. Hat-trie: a cache-conscious trie-based data structure for strings. In *Proceedings of the thirtieth Australasian conference on Computer science-Volume 62*, pages 97–105.

R. Collobert and J. Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of ICML 2008*, pages 160–167.

Rezarta Islamaj Doğan and Zhiyong Lu. 2012. An improved corpus of disease mentions in pubmed citations. In *Proceedings of BioNLP 2012*, pages 91–99.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of ACL 2013*.

Aron Henriksson, Hans Moen, Maria Skeppstedt, Ann-Marie Eklund, Vidas Daudaravicius, and Martin Hassel. 2012. Synonym extraction of medical terms from clinical text using combinations of word space models. In *Proceedings of SMBM 2012*.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of ACL 2012*, pages 873–882.

Pentti Kanerva, Jan Kristoferson, and Anders Holst. 2000. Random indexing of text samples for latent semantic analysis. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, page 1036. Erlbaum.

Mikael Laakso and Bo-Christer Björk. 2012. Anatomy of open access publishing: a study of longitudinal development and internal structure. *BMC medicine*, 10(1):124.

Dekang Lin, Kenneth Ward Church, Heng Ji, Satoshi Sekine, David Yarowsky, Shane Bergsma, Kailash Patil, Emily Pitler, Rachel Lathbury, Vikram Rao, et al. 2010. New tools for web-scale n-grams. In *LREC*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Andriy Mnih and Geoffrey E Hinton. 2008. A scalable hierarchical distributed language model. In *Advances in neural information processing systems*, pages 1081–1088.

Tomoko Ohta, Sampo Pyysalo, Jun'ichi Tsujii, and Sophia Ananiadou. 2012. Open-domain anatomical entity mention detection. In *Proceedings of DSSD 2012*, pages 27–36.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).

Sampo Pyysalo and Sophia Ananiadou. 2013. Anatomical entity mention recognition at literature scale. *Bioinformatics*.

L. Ratinov and D. Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of CoNLL 2009*, pages 147–155.

Magnus Sahlgren, Anders Holst, and Pentti Kanerva. 2008. Permutations as a means to encode order in word space. In *Proceedings of the 30th Conference of the Cognitive Science Society*, pages 1300–1305.

Larry Smith, Lorraine K Tanabe, Rie J Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, et al. 2008. Overview of BioCreative II gene mention recognition. *Genome Biology*, 9(Suppl 2):S2.

Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of EMNLP-CoNLL 2012*, pages 1201–1211.

Pontus Stenetorp, Goran Topić, Sampo Pyysalo, Tomoko Ohta, Jin-Dong Kim, and Jun'ichi Tsujii. 2011. Bionlp Shared Task 2011: Supporting resources. In *Proceedings of BioNLP 2011*, pages 112–120.

Pontus Stenetorp, Hubert Soyer, Sampo Pyysalo, Sophia Ananiadou, and Takashi Chikayama. 2012. Size (and domain) matters: Evaluating semantic word space representations for biomedical text. In *Proceedings of SMBM 2012*.

Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Junichi Tsujii. 2005. Developing a robust part-of-speech tagger for biomedical text. In *Advances in informatics*, pages 382–392. Springer.

J. Turian, L. Ratinov, and Y. Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of ACL 2010*, pages 384–394.

Karin Verspoor, K Bretonnel Cohen, and Lawrence Hunter. 2009. The textual characteristics of traditional and open access scientific journals are similar. *BMC Bioinformatics*, 10(1):183.

Antonio Jimeno Yepes and Karin Verspoor. 2013. Towards automatic large-scale curation of genomic variation: improving coverage based on supplementary material. In *BioLINK SIG 2013*, pages 39–43.