Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Applications Note

OXFORD

Genome analysis

Supplementary Information: A Neural Network Multi-Task Learning Approach to Biomedical Named Entity Recognition

Abstract

This document contains supplmentary information for the paper *A Neural Network Multi-Task Learning Approach to Biomedical Named Entity Recognition*.

1 Datasets

We used 20 biomedical corpora representing 5 NER tasks and one biomedical Part Of Speech Tagging (PoS Tagging) task. We chose datasets which were publicly available and represented the most important named entities in bioinformatics: Anatomy, Chemical, Disease, Gene/Protein and Species. The names of the datasets and their corresponding tasks are listed in Table **??**. Details of their creation and prior use follows.

1.1 AnatEM corpus

The extended Anatomical Entity Mention corpus (Pyysalo and Ananiadou, 2013) is the result of combining and extending the Anatomical Entity Mention (AnEM) corpus (Ohta *et al.*, 2012) and the Multi-level Event Extraction corpus (MLEE) (Pyysalo *et al.*, 2012a). AnEM consists of 500 randomly selected PubMed abstracts and full-text extracts annotated for anatomical entity mentions. MLEE consists of 262 PubMed abstracts on the molecular mechanisms of cancer, specifically relating to angiogenesis. MLEE is also annotated for anatomical entities specified in AnEM.

AnatEM was created by combining the anatomical entity annotations of the AnEM and MLEE corpora, then manual annotation was done on an additional 100 documents following the selection criteria of AnEM and 350 documents following those of MLEE, for a selection of topics related to cancer. The resulting corpus thus consists of 1212 documents, 600 of which are drawn randomly from abstracts and full texts as in AnEM, and the remaining 612 are a targeted selection of PubMed abstracts relating to the molecular mechanisms of cancer.

1.2 BC2GM corpus

The BioCreative II Gene Mention (BC2GM) task corpus consists of 20,000 sentences from biomedical publication abstracts and is annotated for mentions of the names of genes, proteins and related entities using the single NE class Gene Smith *et al.* (2008). It has become the major NER benchmark for gene/proteins names and has been used to train and evaluate many available NER systems such as BANNER (Leaman and Gonzalez, 2008) and Gimli Campos *et al.* (2013).

1.3 BC4CHEMD corpus

The BioCreative IV Chemical and Drug (BC4CHEMD) named entity recognition task corpus consists of 10,000 annotated for mentions of chemical and drug names using a single class, Chemical (Krallinger *et al.*, 2015).

1.4 BC5CDR-Chem corpus

The BioCreative V Chemical Disease Relation (CDR) corpus was created for the BioCreative V Chemical Disease Relation (CDR) Task (Wei *et al.*, 2015) and consists of human annotations of all chemicals, diseases and their interactions in 1,500 PubMed articles. 1,400 of these articles were selected from an existing 150,000 chemical-disease interactions which were annotated by CTD-Pfizer. The CTD biocurators followed CTDâĽ™s rigorous curation process and curated interactions from mostly just the abstract, but referenced the full text when it was necessary to resolve relevant issues mentioned in the abstract. The remaining 100 articles were completely new.

1.5 BioNLP09 corpus

The BioNLP'09 shared task on event extraction (Kim *et al.*, 2009) targeted semantically rich event extraction, involving the extraction of several different classes of information. To focus on these novel aspects of the event extraction task, it was assumed that NER has already been performed and the task began with a given set of gold protein annotations. The named entities in the BioNLP task data were prepared based on the GENIA event corpus. Part of the data were derived from the publicly available event corpus (Kim *et al.*, 2008), and the remainder from an unpublished portion of the corpus.

1.6 BioNLP11 corpus

Similar to the BioNLP'09 task, the BioNLP Shared Task 2011 (Pyysalo *et al.*, 2012b) was focused on semantically rich tasks: Infectious Diseases

(ID) and Epigenetics and Post-translational Modifications (EPI). The ID task was concerned with the molecular mechanisms of infection, virulence and resistance while the EPI task focused on the extraction of statements regarding chemical modifications of DNA and proteins. Though both tasks used manual annotations created specifically for the shared task, the named entities were automatically tagged.

The texts for the EPI task corpus were drawn from PubMed abstracts annotated with the MeSH term corresponding to the target event (e.g. Acetylation). Protein/Gene entity mentions in the selected abstracts were automatically tagged using the BANNER (Leaman and Gonzalez, 2008) named entity tagger trained on the GENETAG (Tanabe *et al.*, 2005) corpus. Abstracts where fewer than five entities are found were removed and documents not relevant to the targeted topic were also manually removed.

The data for the ID corpus were drawn from the primary text content of full-text PMC open access documents deemed by infectious diseases domain experts to be representative publications on two-component regulatory systems. The annotation of the Protein entity, which is what we utilized form this corpus, was performed automatically using NeMine (Sasaki *et al.*, 2008) trained on the JNLPBA data (Kim *et al.*, 2004) with threshold 0.05, filtered to only GENE and Protein types.

1.7 BioNLP13 corpus

The BioNLP 2013 Shared Task focused on knowledge-based construction. There were six tasks in this Shared Task, of which four datasets were used for our work: Genia Event Extraction (GE), Cancer Genetics (CG), Pathway Curation (PC) and Gene Regulation Ontology (GRO).

The GE corpus consists of 20 full paper articles sourced from PubMed Central Open Access subset (PMCOA) with 7721 spans manually annotated as protein names (Bio, 2013). The CG task corpus consists of 600 PubMed abstracts annotated for over 17,000 events and was prepared as an extension of the MLEE (Pyysalo *et al.*, 2012a) corpus of 250 abstracts (c.f. Section 1.1). We use both the Chemicals and Gene/Protein named entities from this corpus. The PC task corpus consists of 525 PubMed abstracts, chosen for the relevance to specific pathway reactions selected from SBML models registered in BioModels and PANTHER DB repositories (Mi and Thomas, 2009). The corpus was manually annotated for over 12,000 events on top of close to 16,000 entities. The GRO corpus consists of 300 MEDLINE abstracts, prepared as an extension of Kim *et al.* (2011).

1.8 Colorado Richly Annotated Full Text (CRAFT) corpus

The CRAFT corpus (Bada *et al.*, 2012; Verspoor *et al.*, 2012) consists of 97 full-text articles, over 790,000 Tokens, over 21,000 Sentences and approximately 140,000 concept annotations. It manually annotates all mentions of nearly all concepts from nine prominent biomedical ontologies and terminologies: Cell Type Ontology, Chemical Entities of Biological Interest ontology, NCBI Taxonomy, Protein Ontology, Sequence Ontology, Entrez Gene database entries, and the three sub-ontologies of the Gene Ontology. There was emphasis on journal articles that comprise the corpus being drawn from diverse biomedical disciplines and on them being completely annotated. We use the Chemicals, Gene and Gene Product and Species named entities from this corpus.

1.9 Ex-PTM corpus

The Exhaustive Post-Translational Modifications corpus (Pyysalo *et al.*, 2011) was part of the BioNLP Shared Task 2011 and employed a similar creation methodology to that of the BioNLP11 EPI task corpus (c.f. Subsection 1.6). It annotated 360 PubMed abstracts containing 76,806 words of which 4,698 were annotated as proteins. Though the more semantically complex PTM identification task used manual annotations, the Protein/Gene entity mentions were automatically tagged using the BANNER (Leaman and Gonzalez, 2008) named entity tagger trained on the GENETAG (Tanabe *et al.*, 2005) corpus. Abstracts containing fewer than five entities were removed and a randomly chosen subset of the remaining documents were annotated.

1.10 JNLPBA corpus

The Joint workshop on NLP in Biomedicine and its Applications corpus consists of 2,404 publication abstracts (approx. 22,400 sentences) and is annotated for mentions of five entity types: cell line, cell type, DNA, RNA, and protein Kim *et al.* (2004). The corpus was derived from GENIA corpus entity annotations. It is now a standard point of reference for evaluating multi-class biomedical entity taggers and has served as training material for tools such as ABNER (Settles, 2005) and the GENIA Tagger.

1.11 Linnaeus corpus

The LINNAEUS corpus (Gerner *et al.*, 2010) consists of 100 full-text documents from the PMCOA document set which were randomly selected. All mentions of species terms were manually annotated and normalized to the NCBI taxonomy IDs of the intended species.

1.12 NCBI-Disease corpus

The NCBI-Disease corpus (Doğan *et al.*, 2014) consists of 793 PubMed abstracts fully annotated at the mention and concept level for disease mentions. The public release of the NCBI disease corpus contains 6,892 disease mentions, which are mapped to 790 unique disease concepts. Of these, 88% link to a MeSH identifier, while the rest contain an OMIM identifier. 91% of the mentions were linked to a single disease concept, while the rest are described as a combination of concepts.

1.13 GENIA corpus

The GENIA corpus is one of the most widely used resources for biomedical NLP and has a rich set of annotations including parts of speech, phrase structure syntax, entity mentions, and events Ohta *et al.* (2002). For this work we use the GENIA POS annotations, which cover 2000 PubMed abstracts (approx. 20,000 sentences).

References

(2013). Bionlp13-ge corpus. Accessed: 2016-8-1.

Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Baumgartner, W. A., Cohen, K. B., Verspoor, K., Blake, J. A., *et al.* (2012). Concept annotation in the craft corpus. *BMC bioinformatics*, **13**(1), 1.

Campos, D., Matos, S., and Oliveira, J. L. (2013). Gimli: open source and high-performance biomedical name recognition. *BMC bioinformatics*, **14**(1), 54.

Doğan, R. I., Leaman, R., and Lu, Z. (2014). Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, **47**, 1–10.

Gerner, M., Nenadic, G., and Bergman, C. M. (2010). Linnaeus: a species name identification system for biomedical literature. *BMC bioinformatics*, **11**(1), 1.

Kim, J.-D., Ohta, T., Tsuruoka, Y., Tateisi, Y., and Collier, N. (2004). Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of JNLPBA*, pages 70–75.

Kim, J.-D., Ohta, T., and Tsujii, J. (2008). Corpus annotation for mining biomedical events from literature. *BMC bioinformatics*, **9**(1), 1.

Kim, J.-D., Ohta, T., Pyysalo, S., Kano, Y., and Tsujii, J. (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.

Kim, J.-J., Han, X., and Chua, W. W. K. (2011). Annotation of biomedical text with gene regulation ontology: Towards semantic web for biomedical literature. *Proceedings of LBM*, pages 63–70.

Krallinger, M., Leitner, F., Rabal, O., Vazquez, M., Oyarzabal, J., and Valencia, A. (2015). CHEMDNER: The drugs and chemical names extraction challenge. *J. Cheminformatics*, **7**(S-1), S1.

Leaman, R. and Gonzalez, G. (2008). BANNER: an executable survey of advances in biomedical named entity recognition. In *Proceedings of PSB*, volume 13, pages 652–663.

Mi, H. and Thomas, P. (2009). Panther pathway: an ontology-based pathway database coupled with data analysis tools. *Protein Networks and Pathway Analysis*, pages 123–140.

Ohta, T., Tateisi, Y., and Kim, J.-D. (2002). The GENIA corpus: An annotated research abstract corpus in molecular biology domain. In *Proceedings of HTL*, pages 82–86.

Ohta, T., Pyysalo, S., and Ananiadou, S. (2012). Open-domain anatomical entity mention detection.

Pyysalo, S. and Ananiadou, S. (2013). Anatomical entity mention recognition at literature scale. *Bioinformatics*.

Pyysalo, S., Ohta, T., Miwa, M., and Tsujii, J. (2011). Towards exhaustive protein modification event extraction. In *Proceedings of BioNLP 2011 Workshop*, pages 114–123. Association for Computational Linguistics.

Pyysalo, S., Ohta, T., Miwa, M., Cho, H.-C., Tsujii, J., and Ananiadou, S. (2012a). Event extraction across multiple levels of biological organization. *Bioinformatics*, **28**(18), i575–i581.

Pyysalo, S., Ohta, T., Rak, R., Sullivan, D., Mao, C., Wang, C., Sobral, B., Tsujii, J., and Ananiadou, S. (2012b). Overview of the id, epi and rel tasks of bionlp shared task 2011. *BMC bioinformatics*, **13**(11), 1.

Sasaki, Y., Tsuruoka, Y., McNaught, J., and Ananiadou, S. (2008). How to make the most of ne dictionaries in statistical ner. *BMC bioinformatics*, **9**(11), 1.

Settles, B. (2005). ABNER: An open source tool for automatically tagging genes, proteins, and other entity names in text. *Bioinformatics*, **21**(14), 3191–3192.

Smith, L., Tanabe, L. K., nee Ando, R. J., Kuo, C.-J., Chung, I.-F., Hsu, C.-N., Lin, Y.-S., Klinger, R., Friedrich, C. M., Ganchev, K., *et al.* (2008). Overview of BioCreative II gene mention recognition. *Genome biology*, **9**(Suppl 2), 1–19.

Tanabe, L., Xie, N., Thom, L. H., Matten, W., and Wilbur, W. J. (2005). Genetag: a tagged corpus for gene/protein named entity recognition. *BMC bioinformatics*, **6**(1), 1.

Verspoor, K., Cohen, K. B., Lanfranchi, A., Warner, C., Johnson, H. L., Roeder, C., Choi, J. D., Funk, C., Malenkiy, Y., Eckert, M., *et al.* (2012). 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), 1.

Wei, C.-H., Peng, Y., Leaman, R., Davis, A. P., Mattingly, C. J., Li, J., Wiegers, T. C., and Lu, Z. (2015). Overview of the biocreative v chemical disease relation (cdr) task. In *Proceedings of the fifth BioCreative challenge evaluation workshop*, pages 154–166.