Biomedical Named Entity Recognition with Multilingual BERT

Kai Hakala, Sampo Pyysalo
Turku NLP Group, University of Turku, Finland
{first.last}@utu.fi

Abstract

We present the approach of the Turku NLP group to the PharmaCoNER task on Spanish biomedical named entity recognition. We apply a CRF-based baseline approach and multilingual BERT to the task, achieving an F-score of 88% on the development data and 87% on the test set with BERT. Our approach reflects a straightforward application of a state-of-the-art multilingual model that is not specifically tailored to either the language nor the application domain. The source code is available at: https://github.com/chaanim/pharmaconer

1 Introduction

Named entity recognition (NER) is a fundamental task in information extraction, and the ability to detect mentions of domain-relevant entities such as chemicals and proteins is required for the analysis of texts in specialized domains such as biomedicine. Although a wealth of manually annotated corpora and dedicated NER methods have been introduced for the analysis of English biomedical and clinical texts (e.g. (Leaman and Lu, 2016; Crichton et al., 2017; Weber et al., 2019)), there has been comparatively little work on these basic resources for other languages, including Spanish.

The PharmaCoNER task focuses on pharmacological compound mentions in Spanish clinical texts, promoting the development of biomedical text mining tools for non-English data (Gonzalez-Agirre et al., 2019). Track 1 involves the recognition and classification of entity mentions into upper-level ontological categories (chemical, protein, etc.), and Track 2 the normalization (grounding) of these mentions to identifiers in external resources. We participate in Track 1.

We participate in the PharmaCoNER task using a collection of tools developed for English as well as out-of-domain multilingual models. In particular, we use a freely available NER toolkit, NERsuite, tailored for English biomedical literature and a multilingual neural model, BERT, pretrained on general domain Wikipedia articles. Thus, the emphasis of this work is on analyzing how well such tools can be adapted to new languages and domains with minimal effort. We cast the task as sequence labeling using a conventional in-out-begin (IOB) representation of the data for learning and prediction. The used tools are described in detail in Section 3.

Item	Train	Devel
Documents	500	250
Tokens	177 022	85 148
Annotations	3 822	1 926
Protein	1 405	745
Chemical(+)	2 304	1 121
Chemical(-)	24	16
Other	89	44

Table 1: Data statistics.

2 Data

The annotation involves four types of entities, labeled in the data as PROTEINAS (proteins, genes, and related entities), NORMALIZABLES (chemicals that can be normalized to external resources), NO_NORMALIZABLES (chemicals that cannot), and UNCLEAR (miscellaneous related entities). In the following, we refer to these respectively as Protein, Chemical(+), Chemical(-) and Other. Table 1 briefly summarizes data statistics. We note that compared to English language biomedical NER resources, the number of annotations is somewhat limited; for example, the JNLPBA shared task (Kim et al., 2004) data contains over 50,000 training examples of similar

Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, pages 56–61
Hong Kong, China, November 4, 2019. ©2019 Association for Computational Linguistics

```
.txt
… La proteína C reactiva y la VSG eran
normales. La interleucina 6 fue normal.
```

```
.ann
T18   PROTEINAS 979 998    proteína C reactiva
T17   PROTEINAS 1026 1040  interleucina 6
```

```
.nersuite
O             976     978     La
B-PROTEINAS   979     987     proteína
I-PROTEINAS   988     989     C
I-PROTEINAS   990     998     reactiva
…
O             1023    1025    La
B-PROTEINAS   1026    1038    interleucina
I-PROTEINAS   1039    1040    6
```

Figure 1: Illustration of data formats. Left: task data in separate .txt and .ann files. Right: NERsuite format.

types, the BioCreative II GM data (Smith et al., 2008) over 18,000 gene mentions, and the BioCreative CHEMDNER (Krallinger et al., 2015) data over 80,000 chemical mentions. We thus expect that methods addressing the PharmaCoNER task to benefit from pretraining or other similar methods of incorporating information from outside of just the task data.

The task data is distributed in the simple standoff format first introduced for the BioNLP Shared Task 2009 (Kim et al., 2009). To convert this data into a version the column-based IOB format popularized by the CoNLL NER tasks and used by many NER tools, we apply a simple conversion script provided with the BRAT annotation tool[1] (Stenetorp et al., 2012). We note that conversion between the standoff and the token-based IOB representations is lossless if and only if there are no overlapping annotations in the source data and the boundaries of the annotations match token boundaries. Based on an experiment on the training data, we estimate that the conversion preserves the original annotations exactly over 99% of the time. Figure 1 illustrates the two formats.

We note that one training file[2] failed conversion due to an off-by-one offset error. We excluded this file in all of our experiments.

3 Methods

3.1 NERsuite

Conditional Random Fields (CRF) (Lafferty et al., 2001) are a popular and effective model for sequence labeling and thus a relevant baseline in NER work. We perform experiments with NERsuite[3], an NER toolkit that is based on the CRFsuite (Okazaki, 2007) CRF implementation and includes rich features optimized for English biomedical text. In particular, NERsuite incorporates features derived from analysis by the GENIA tagger (Tsuruoka et al., 2005), which performs part-of-speech tagging, chunking and lemmatization and has been trained on English text. When applied on Spanish input, the tags and lemmas will necessarily very frequently be incorrect. We nevertheless opted to apply the system as an off-the-shelf baseline as its rich feature set also includes many language-independent features. We leave the NERsuite parameters at their defaults.

3.2 BERT

In our second experiment we utilize BERT (Devlin et al., 2018), a transformer (Vaswani et al., 2017) based attentive neural architecture. Whereas pretrained BERT models have shown strong performance for English NER tasks (Peng et al., 2019), to our knowledge no pretrained Spanish BERT models are readily available[4]. Thus we conduct our experiments with the multilingual BERT model (Pires et al., 2019) trained on a Wikipedia corpus, covering 104 languages. Whereas Spanish is one of the pretraining languages used for the model, the used Wikipedia corpus is not specific to clinical or biomedical content. We use the cased variant of the model, which preserves the case and accents of the characters. BERT relies on subword units, shared between all the used languages, leading to subword embeddings which can benefit from the commonalities of similar languages, yet are a compromise across different uses in different languages and domains. For fine-tuning the model, we use the Keras BERT Python library [5].

When fine-tuning the model for the NER task at hand, we replace the original pretraining output layers with a CRF layer and allow the optimizer to adjust all layers of the network. The model is optimized with Adam (Kingma and Ba, 2014) with a

[1] http://brat.nlplab.org

[2] S0211-69952015000200015-1

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

[4] Shallow word embeddings for Spanish are studied e.g. by Soares et al. (2019)

[5] https://github.com/CyberZHG/keras-bert

batch size of 16 and a learning rate of 2e-5 warmed up from 2e-7 over the first training epoch.

We train the model for 50 epochs and evaluate the model after every epoch on the development set using entity-level F-scores on subword tokens. The best performing checkpoint is used as the final prediction model, i.e. we use early stopping with a decreasing patience. In addition to early stopping, the model is regularized with dropout (Srivastava et al., 2014) within each transformer block and weight decay (Loshchilov and Hutter, 2017). The dropout is set to 0.1, but the weight decay is selected in a grid search, being the only hyperparameter optimized in our experiments.

As the input for the BERT model we use the CoNLL formatted data identical to the CRF experiments (see Section 2), which is split into sentences and tokenized on word-level. As the BERT model utilizes subword units, we further retokenize every word independently. Due to computational reasons we use a maximum length of 128 subword units for the input. This limit permits running the model on low memory consumer-grade GPUs instead of requiring data center hardware. Sentences longer than the limit are split into separate input sequences for the network. Note that this may occasionally split entities into separate example sequences leading to sequences starting with I tags. When converting the predictions back to the word-level CoNLL format, we assign the predicted entity label of the first subword unit for the entire token.

4 Results and discussion

The official PharmaCoNER evaluation criteria measure performance on the level of entity mentions (rather than e.g. tokens) and require exact identification of the offset where each mention occurs and the type of the mentioned entity. We note that this common but fairly stringent criterion penalizes many small divergences from the reference annotation twice: if a predicted entity is otherwise correct but e.g. differs in its ending offset from a gold standard entity, the predicted mention is considered a false positive, and the corresponding gold standard entity a false negative. Performance is evaluated in terms of precision, recall, and balanced F-score over all entity types (microaverage). To provide a more fine-grained look into the performance of our approach, we perform additional analyses breaking down performance by type as well as considering approximate matching criteria, namely *left* boundary matching where only the start offsets of mentions is required to match, *right* boundary matching where only end is required, and *overlap* matching, where any overlapping spans are considered a match. We require entity types to match for all criteria.

The detailed evaluation for the NERsuite and BERT models on the development set are listed in Tables 2 and 3, where *exact* matching criterion corresponds to the official evaluation. The NERsuite model achieves and overall F-score of 82% showing surprisingly strong performance considering the fact that it relies on English part-of-speech tagging, chunking and lemmatization models. The BERT model surpasses this baseline by +6.5pp with an overall F-score of 88%. We used the BERT model as our official submission to the shared task resulting in an F-score of 87.38% on the test set according to the organizers. For both of these models the *overlap* evaluation shows an improvement of 3–4pp, suggesting that the models are in effect better at detecting the entities, but suffer from slightly inaccurate boundary detection. For the BERT model the difference between *exact* and *overlap* results is slightly larger, which might be caused by the additional retokenization to subword units and detokenization back to the original CoNLL format. As the overall performance of the BERT model is notably better than NERsuite's, we focus on the former in all further analyses.

To measure the BERT model's ability to generalize to unseen entity mentions, we analyze how many of the development data entity spans are not present in the training data and how well the model performs on these entities in comparison to entity spans which the model has seen during training. We observe that 55% of the unique entity spans, covering 36% of all occurrences, in the development set are not present in the training data. This suggests that strong generalization abilities are required from the model to perform well in the task.

To obtain a rough understanding of how well the model performs on the entities unseen during training, we measure the recall of the model separately for entity spans seen and not seen during training (Table 4). As can be expected the model has an extremely high recall of 96% for spans present in the training data, but also relatively strong performance with recall of 72% for previously unseen spans. This suggests that the

Criterion	Protein	Chemical(+)	Chemical(-)	Other	Overall
Exact	88.89/74.09/80.82	93.41/75.76/83.66	0.00/0.00/0.00	79.31/52.27/63.01	91.35/73.95/81.73
Left	92.43/77.05/84.04	95.16/77.18/85.24	0.00/0.00/0.00	82.76/54.55/65.75	93.85/75.97/83.97
Right	91.63/76.38/83.31	94.73/76.83/84.84	0.00/0.00/0.00	79.31/52.27/63.01	93.21/75.45/83.40
Overlap	95.01/79.19/86.38	96.15/78.16/86.23	0.00/0.00/0.00	82.76/54.55/65.75	95.45/77.37/85.47

Table 2: NERsuite development set results for various boundary matching criteria (precision/recall/F-score).

Criterion	Protein	Chemical(+)	Chemical(-)	Other	Overall
Exact	84.87/88.86/86.82	92.99/87.51/90.17	40.00/12.50/19.05	76.47/88.64/82.11	89.05/87.44/88.24
Left	89.36/93.56/91.41	95.73/90.10/92.83	40.00/12.50/19.05	78.43/90.91/84.21	92.49/90.81/91.64
Right	87.44/91.54/89.44	93.65/88.14/90.81	40.00/12.50/19.05	76.47/88.64/82.11	90.48/88.84/89.65
Overlap	91.92/95.30/93.58	96.02/90.72/93.30	40.00/12.50/19.05	78.43/90.91/84.21	93.71/91.85/92.77

Table 3: BERT development set results for various boundary matching criteria (precision/recall/F-score).

Entities	Pretraining	No pretraining
All	87.44	54.00
Seen	96.13	70.16
Unseen	71.72	24.78

Table 4: Recall of the BERT model on development set with and without pretraining on all entities, entity spans which are also present in the training data (seen) and entity spans which do not appear in the training data (unseen).

Pretraining	Precision	Recall	F-score
Yes	89.05	87.44	88.24
No	57.62	54.00	55.75

Table 5: Development set results for BERT model with and without pretraining.

model has either learned suitable subword representations during the pretraining for detecting pharmacological entities or is able to effectively utilize the context in which they appear.

As the model is pretrained on multilingual out-of-domain data, we are also interested in the benefits of such pretraining. To this end we train an identical model with randomly initialized weights as the starting point. The same subword unit vocabulary is used. This model results in far inferior performance with an F-score of 56% (see Table 5). Moreover the recall of unseen entity spans is mere 25%, whereas for previously seen spans the recall is 70%. Thus the pretraining, even with multilingual Wikipedia data, seems to offer drastic improvements to the model, particularly for detecting entity spans not seen during training. However, using the same vocabulary makes this comparison slightly unfair as subword embeddings are left in their random initial state if not present in the training data. In the development set this impacts around 12% of the unique subword units, which however constitute only 2% of all subword occurrences.

We also note that although we have used a CRF layer as the output of the BERT model, in our preliminary experiments we observed similar results with a fully connected output layer. This suggests that the transformer architecture has the capability of implicitly modelling sequential dependencies of the output labels, unlike earlier neural models such as bidirectional LSTM networks, which still substantially benefit from the added CRF output layer (Ma and Hovy, 2016; Lample et al., 2016).

5 Conclusions and future work

In this study we have demonstrated that strong results for Spanish clinical NER can be achieved with straightforward adaptation of multilingual or English text mining tools. In particular the multilingual BERT model pretrained on general domain Wikipedia articles shows competitive performance with an F-score of 87% in the official PharmaCoNER evaluation.

As prior studies have shown that the multilingual BERT model can also be utilized in zero-shot settings (Pires et al., 2019), as a future work, we will look into optimal ways of incorporating English NER datasets in this task. This can be either achieved in zero-shot setting, training the model purely on English NER datasets and applying on Spanish texts or by combining both English and Spanish training data in a multitask setting.

In addition to studying the BERT model, we have demonstrated that a strong baseline system for this task can also be achieved with the NERsuite toolkit, even though it relies on feature representations built upon POS tagging and chunking models trained on English data, warranting the use of such freely available tools even in cross-lingual settings.

Acknowledgments

We are grateful to CSC – IT Center for Science for computational resources used to train our models.

References

Gamal Crichton, Sampo Pyysalo, Billy Chiu, and Anna Korhonen. 2017. A neural network multi-task learning approach to biomedical named entity recognition. *BMC bioinformatics*, 18(1):368.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Aitor Gonzalez-Agirre, Montserrat Marimon, Ander Intxaurrondo, Obdulia Rabal, Marta Villegas, and Martin Krallinger. 2019. PharmaCoNER: Pharmacological substances, compounds and proteins named entity recognition track. In *Proceedings of the BioNLP Open Shared Tasks (BioNLP-OST)*, Hong Kong, China. Association for Computational Linguistics.

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 BioNLP Shared Task*, pages 1–9.

Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of JNLPBA*, pages 70–75.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. 2015. The CHEMDNER corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(1):S2.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the ICML*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Robert Leaman and Zhiyong Lu. 2016. TaggerOne: joint named entity recognition and normalization with semi-markov models. *Bioinformatics*, 32(18):2839–2846.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in Adam. *arXiv preprint arXiv:1711.05101*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

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

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? *arXiv preprint arXiv:1906.01502*.

Larry Smith, Lorraine K Tanabe, Rie Johnson nee 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(2):S2.

Felipe Soares, Marta Villegas, Aitor Gonzalez-Agirre, Martin Krallinger, and Jordi Armengol-Estapé. 2019. Medical word embeddings for Spanish: Development and evaluation. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 124–133, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of EACL demonstrations*, pages 102–107.

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 *Panhellenic Conference on Informatics*, pages 382–392. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Leon Weber, Jannes Münchmeyer, Tim Rocktäschel, Maryam Habibi, and Ulf Leser. 2019. HUNER: Improving biomedical NER with pretraining. *Bioinformatics*.