Applying existing Named Entity taggers at BARR IBEREVAL 2017 task

Rodrigo Agerri and German Rigau
rodrigo.agerri@ehu.eus

IXA NLP Group, University of the Basque Country UPV/EHU

Abstract. We present our experiments applying, off-the-shelf, two existing Named Entity Recognition (NER) taggers for the Biomedical Abbreviation Recognition and Resolution (BARR) task at IberEval 2017. The first system is a Perceptron tagger based on sparse, shallow features whereas the second is a bidirectional Long-Short Term Memory neural network with a sequential conditional random layer above it (LSTM-CRF) and initialized with ngram word embeddings. Due to time constraints, we only managed to submit a run from the Perceptron tagger, although in this paper we will also report results with the LSTM-CRF tagger evaluated on the development data. Results show that both Perceptron and LSTM-CRF perform reasonably well for SHORT abbreviations whereas the Perceptron model fails to generalize properly for the LONG entity class.

1 Introduction

The literature on biomedical Named Entity Recognition and Classification is rather extensive, see for example the CHEMDNER tasks [8], but for this first edition of BARR we were particularly interested in evaluating current, general purpose, existing Named Entity Recognition (NER) taggers off-the-shelf. The idea was to establish how well can we perform *for free* in this task, namely by training and applying existing tools without feature or hyperparameter tuning. For this exercise we chose two taggers from different conceptual frameworks, ixa-pipe-nerc [2], and a LSTM-CRF NER tagger [10] in order to annotate biomedical entities and their abbreviations, as illustrated by the following example:

> (1) "Se describe la relación entre diferentes factores de riesgo cardiovasculares (FRCV) y la obesidad a partir de una muestra representativa de la población adulta de Madrid"

The BARR task consists of detecting and classifying both LONG and SHORT mentions of biomedical entities and to establish that the SHORT mention is an abbreviation of the LONG entity mention. Thus, in this example "factores de riesgo cardiovasculares" corresponds to a mention of the LONG entity class whereas "FRCV" would be a mention of the SHORT class. Furthermore, the SHORT mention is an abbreviation of the LONG one. Our interest relies purely on entity recognition and, more specifically, on evaluating current existing taggers and models on the BARR evaluation setting, so we participated only in the entity recognition subtask, and not on the relation recognition task.

2 Methodology

For our experiments we trained on the train1 subset and evaluated on the train2 set. Tokenization and pre-processing was performed using the IXA pipes tokenizer [1] without any fine-tuning for the medical domain. The BARR-E background set distributed in the task [7] is leveraged in order to induce clusters and word embeddings to train both ixa-pipe-nerc and LSTM-CRF systems. In the following we briefly describe the architecture of both systems.

2.1 ixa-pipe-nerc

The design of *ixa-pipe-nerc* aims to establish a simple and shallow feature set, avoiding any linguistic motivated features, with the objective of removing any reliance on costly extra gold annotations (POS tags, lemmas, syntax, semantics) and/or cascading errors if automatic language processors are used. The underlying motivation is to obtain robust models to facilitate the development of NER systems for other languages and datasets/domains while obtaining state of the art results. Our system consists of: (i) Local, shallow features based mostly on orthographic, word shape and n-gram features plus their context; (ii) three types of simple clustering features, based on unigram matching; (iii) publicly available gazetteers. *ixa-pipe-nerc* learns supervised models via the Perceptron algorithm as described by [5]. To avoid duplication of efforts, *ixa-pipe-nerc* uses the Apache OpenNLP project implementation of the Perceptron algorithm[1] customized with its own features. Specifically, *ixa-pipe-nerc* implements, on top of the local features, a combination of word representation features: (i) Brown [3] clusters, taking the 4th, 8th, 12th and 20th node in the path; (ii) Clark [4] clusters and, (iii) Word2vec [12] clusters, based on K-means applied over the extracted word vectors using the skip-gram algorithm. The implementation of the clustering features looks for the cluster class of the incoming token in one or more of the clustering lexicons induced following the three methods listed above. If found, then we add the class as a feature. The Brown clusters only apply to the token related features, which are duplicated.

The ixa-pipe-nerc tagger includes a simple method to *combine* and *stack* various types of clustering features induced over different data sources or corpora, with state of the art results in newswire Named Entity Recognition [2] and Opinion Target Extraction [14], both in out-of-domain and in-domain evaluations.

2.2 LSTM-CRF

Long Short-Term Memory Networks were design to address the bias of Recurrent neural networks (RNNs) to learn their most recent input so that long-range dependencies can be more effectively captured. LSTMs are a family of neural networks that operate on sequential data. They take as input a sequence of vectors $(x_1, x_2, ..., x_n)$ and return another sequence $(h_1, h_2, ..., h_n)$ that represents some information about the sequence at every step in the input. LSTMs try to learn long-range dependencies by using several gates that control the proportion of the input to give to the memory cell, and the proportion from the previous state to forget.

[1] http://opennlp.apache.org/

Class	LONG			SHORT			OVERALL
Features (bilou)	Precision	Recall	F1	Precision	Recall	F1	F1
Local (L)	72.58	43.90	54.71	91.58	81.31	86.14	70.64
L + set02-clark300	70.20	51.71	59.55	**93.72**	80.14	86.40	72.31
L + set02-clark300-char25	71.01	47.80	57.14	91.58	83.88	**87.56**	72.31

Table 1. ixa-pipe-nerc results training on train1 and testing on train2.

Lample et al. [10] present a hybrid tagging architecture consisting of LSTMs whose output is then modeled by Conditional Random Fields to tag decisions jointly instead of independently [9]. This architecture is similar to the one presented by Collobert [6]. Furthermore, the LSTM-CRF lookup table is initialized using pretrained word embeddings. They use a variation of word2vec [12] embeddings that accounts for word order [11]. Recurrent models such as RNNs and LSTMs are interesting for the BARR task because they are theoretically capable of encoding very long sequences. This system obtained similar results to ixa-pipe-nerc in newswire NER evaluated on CoNLL 2002 and 2003 data.

3 Experiments

We train both ixa-pipe-nerc and LSTM-CRF systems with the default parameters and features as described in Agerri and Rigau [2] and Lample et al. [10], using both BILOU and BIO annotation schemes. The BARR-E background is leveraged to train Brown, Clark and Word2vec clusters for ixa-pipe-nerc, and ngram word embeddings for LSTM-CRF. Tables 1 and 2 summarize the results of training and evaluating both systems on the train1 and test2 set, respectively. The reported results for both systems are evaluated using the CoNLL script for Named Entity Recognition [15].

We also submitted the runs in Table 1 to the Markyt evaluation platform [13]. Although it seems that the micro F1 score is computed following the CoNLL evaluation methodology, there were small disagreements between the results obtained using the CoNLL script and those reported by the Markyt platform. For example, the *set02-clark300* run scored 72.31 F1 using the CoNLL script whereas the Markyt scorer computed 72.24 F1. By looking at the details of the Markyt evaluation, we found a couple of possible reasons to explain this slight misalignment:

1. Our runs only contained LONG, SHORT and MULTIPLE classes, yet the Markyt system reported that we annotated some NESTED mentions.
2. The Markyt system reported results for LONG, SHORT, NESTED and UNKNOWN. However, the shared task description asked for LONG and SHORT classes only. Moreover, the UNKNOWN class we could not find in any of the annotated data provided.

Although we still do not know the official results of the other participants in the task, it looks like the results for the SHORT class are reasonably good, especially if we consider that we are applying the systems off-the-shelf. For both classes, ixa-pipe-nerc

Class	LONG			SHORT			OVERALL
Features	Precision	Recall	F1	Precision	Recall	F1	F1
BIO	65.85	69.77	**67.75**	88.08	85.88	**86.97**	76.43
BILOU	60.49	74.70	66.85	89.25	87.21	**88.22**	**76.97**

Table 2. LSTM-CRF results training on train1 and testing on train2.

is particularly strong in terms of precision, failing the clustering features to increase the recall in order to obtain more competitive F1 results. In particular, the results show that ixa-pipe-nerc fails to generalize on the LONG class. In this sense, the LSTM-CRF system models much better the LONG class, surpassing ixa-pipe-nerc results for this entity type by around 7 points in F1 score.

In terms of official results, we only managed to submit the *set02-clark300* due to the time and formatting constraints of the shared task. In this sense, it would have been interesting to send the LSTM-CRF runs which seem to be more competitive in this evaluation setting. The submitted run obtained *65.29* F1 (micro averaged) for entity recognition (71.78 precision and 59.87 recall). At the time of writing, no further information was available about the official run (e.g., false positives, evaluation per class, etc.).

4 Concluding Remarks

In this paper we present the results of directly applying, off-the-shelf, two existing NER taggers for the BARR entity recognition subtask. In this way, we could have an idea of actual performance of existing systems when applied to other discourse genre and/or domain. The results show that both systems perform reasonably well considering the lack of feature and/or parameter tuning performed in this exercise. Both systems scores are similar for the SHORT class whereas the LSTM-CRF system models better the long-range dependencies required to perform well for the LONG entity class.

We believe a number of issues made this task unnecessarily difficult. Firstly, it seems customary now for every new task to create their own dataset formats and evaluation scripts, even if the task is as old as Named Entity Recognition for which a long tradition of shared tasks exists [15]. In this case, both the documents and entities sets of the corpus were formatted differently for training and testing, that is, four different formats needed to be parsed [7].

Secondly, the test set required to be processed for every run was unusually large (26 MB of text distributed in 19K documents, and more than 20 times larger than the training set) containing, once tokenized using IXA pipes, more than 170K sentences. This made it quite cumbersome to obtain multiple test runs based on the results reported in Tables 1 and 2. For example, ixa-pipe-nerc required around 45 minutes to tag the test for each run whereas the LSTM-CRF tagger required around three hours to do so. With respect to the training process, ixa-pipe-nerc trained a model in less than a minute whereas the LSTM-CRF tagger needed around 20 hours to complete 100 epochs on the 950 documents of the training set.

Finally, there were a number of annotation issues which, in our opinion, did not help participant systems:

1. Wrong tokenization in entity annotation: expressions such as min, ml, i.e, cm and EE.UU instead of min., ml., i.e., cm. and EE.UU. were labeled as entities by the task annotators.
2. Fractions and dashes: hundreds of fraction expressions such as 23/mg and others containing a forward slash or dash such as radio/tv, angio-TC and h/da were splitted in the annotation (mistakenly in our opinion). For example, instead of *23/mg* being an entity mention, only *mg* was annotated. In the same way, instead of *angio-TC* being annotated as an entity, only *TC* was tagged.
3. Entities inside tokens: many entities were annotated inside a token. The most frequent were mmHg (splitted into mm and Hg entities instead of considering the token mmHg an entity mention) and H1N1, from which the GLOBAL H and N entities were annotated, instead of labeling H1N1 as an entity by itself. This decision means that to be successful in this task entity taggers should work at character level, because otherwise all these entities inside token words would not be modeled.

Acknowledgements

This work has been supported by Spanish Ministry of Economy and Competitiveness (MINECO/FEDER, UE), under the projects TUNER (TIN2015-65308-C5-1-R) and CROSSTEXT (TIN2015-72646-EXP).

References

1. Agerri, R., Bermudez, J., Rigau, G.: IXA pipeline: Efficient and ready to use multilingual NLP tools. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 3823–3828. Reykjavik, Iceland (May 2014)
2. Agerri, R., Rigau, G.: Robust multilingual named entity recognition with shallow semi-supervised features. Artificial Intelligence 238, 63–82 (2016)
3. Brown, P.F., Desouza, P.V., Mercer, R.L., Pietra, V.J.D., Lai, J.C.: Class-based n-gram models of natural language. Computational linguistics 18(4), 467–479 (1992)
4. Clark, A.: Combining distributional and morphological information for part of speech induction. In: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1. pp. 59–66. Association for Computational Linguistics (2003)
5. Collins, M.: Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. pp. 1–8 (2002)
6. Collobert, R., Weston, J.: A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th international conference on Machine learning. pp. 160–167. ACM (2008)
7. Krallinger, M., Intxaurrondo, A., Lopez-Martin, J., de la Pea, S., Prez-Prez, M., Prez-Rodrguez, G., Santamara, J., Villegas, M., Akhondi, S., Loureno, A., Valencia, A.: Resources for the extraction of abbreviations and terms in spanish from medical abstracts: the barr corpus, lexical resources and document collection. In: SEPLN (2017)

8. Krallinger, M., Leitner, F., Rabal, O., Vazquez, M., Oyarzabal, J., Valencia, A.: Overview of the chemical compound and drug name recognition (chemdner) task. In: BioCreative challenge evaluation workshop. vol. 2, p. 2 (2013)
9. Lafferty, J.: Conditional random fields : Probabilistic models for segmenting and labeling sequence data. Computer (1999)
10. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: 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 (NAACL-2016). pp. 260–270 (2016)
11. Ling, W., Chu-Cheng, L., Tsvetkov, Y., Amir, S., Astudillo, R.F., Dyer, C., Black, A.W., Trancoso, I.: Not all contexts are created equal: Better word representations with variable attention. In: Proceedings of Empirical Methods in Natural Language Processing (EMNLP) (2015)
12. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems. pp. 3111–3119 (2013)
13. Pérez-Pérez, M., Pérez-Rodríguez, G., Rabal, O., Vazquez, M., Oyarzabal, J., Fdez-Riverola, F., Valencia, A., Krallinger, M., Lourenço, A.: The markyt visualisation, prediction and benchmark platform for chemical and gene entity recognition at biocreative/chemdner challenge. Database 2016 (2016)
14. San Vicente, I.n., Saralegi, X., Agerri, R.: Elixa: A modular and flexible absa platform. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). pp. 748–752. Association for Computational Linguistics, Denver, Colorado (June 2015)
15. Tjong Kim Sang, E.F.: Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In: Proceedings of CoNLL-2002. pp. 155–158. Taipei, Taiwan (2002)