Simple Language Independent Sequence Labelling for the Annotation of Disabilities in Medical Texts

Rodrigo Agerri and German Rigau
{rodrigo.agerri,german.rigau}@ehu.eus

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

Abstract. In this paper we describe our participation to the DIANN 2018 shared task at IberEval 2018 for the annotation of disabilities in the medical domain. We use IXA pipes to model the annotation of disabilities as a sequence labelling task. Our system consists of a combination of clustering features implemented on top of a simple set of shallow local features. We show how leveraging distributional features obtained from large in-domain unlabelled data helps to easily develop a robust system for the medical domain. The system and models generated in this work are available for public use and to facilitate reproducibility of results.

1 Introduction

The DIANN task [8] proposes an evaluation competition focused on the annotation of disabilities in English and Spanish. Disabilities affect a large part of the world's population. Currently, there are some tools for the annotation of medical concepts, especially in English, such as Metamap [3]. However, they do not consider disabilities specifically. Therefore, currently there is not exist the possibility of automatically classifying and distinguishing disabilities from other signs associated to diseases. The following example illustrates the kind of annotations provided by the DIANN dataset:

(1) In the patients <scp><neg>without<neg> <dis>dementia</dis></scp>, significant differences were obtained in terms of functional and cognitive status (Barthel index of 52.3438 and Pfeiffer test with an average score of 1.48 3.2 (P<.001)).

In Example (1) it can be seen that, apart from disabilities such as 'dementia', both negations and their scope is also annotated.

Our submission is placed in the context of the TUNER project[1] on "Multi-faceted Domain Adaptation for Advanced Textual Semantic Processing". In Natural Language Processing (NLP) technology is crucial to extract accurate, complete, relevant, interoperable and timely structured knowledge from large amounts of unstructured multilingual text to make informed decisions. The aim of TUNER is to address these needs through the research and development of domain adaptation techniques to apply them to the NLP technology developed within the project. Summarizing, the project aims to develop domain-oriented cross-lingual systems that will provide deep semantic capabilities to process multilingual data.

[1] http://ixa2.si.ehu.es/tuner/

In this setting, we take an existing sequence labelling system developed for Named Entity Recognition (NER) on newswire text [2] and we experiment on the domain adaptation of such system by including, via semi-supervision, semantic information automatically obtained from large amounts of unlabelled domain-specific data.

2 Methodology

The DIANN task is modelled as a sequence labelling task. In order to do so, we convert an annotated sentence such as the one in Example (1) into the BIO scheme for learning sequence labelling models [12]. Example (2) shows the review in BIO format. Tokens in the review are tagged depending on whether they are at the beginning (B-target), inside (I-target) or outside (O) of the disability and negation expressions:

(2) In/O the/O patients/O **without/B-NEG dementia/B-DIS** ,/O significant/O differences/O were/O obtained/O in/O terms/O of/O functional/O and/O cognitive/O status/O . . .

Our system learns language independent models which consist of a set of local, shallow features complemented with semantic distributional features based on clusters obtained from a variety of domain-specific data sources. We show that our approach, despite the lack of hand-engineered, language-specific features, obtains competitive results in the present task.

As it can be seen in Example (2), we do not treat the scope of negation, but in a post-processing step, we proceed as follows: if at least a negation and a disability are annotated in a sentence, then the negation is considered to have a wide scope. In other words, if more than one disability appears in the same sentence as a negation, then it is interpreted that the negation affects to every disability present in that sentence.

For our experiments we trained on the training data provide for each language and choose the best settings via 5-fold cross validation. The chosen models were then used to annotated the test data. The datasets were tokenized using the IXA pipes tokenizer [1] without any fine-tuning for the medical domain. The BARR-E background set distributed in the task [9] is leveraged in order to induce clusters for Spanish. For English, we used PubMed Annual Baseline from 2017[2].

2.1 ixa-pipe-nerc

Our sequence labeling system is *ixa-pipe-nerc*, which 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 was to obtain robust models to facilitate the development of NER systems for other languages and datasets/domains while obtaining state of the art results.

[2] ftp://ftp.ncbi.nlm.nih.gov/pubmed/baseline

The 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 [6]. To avoid duplication of efforts, *ixa-pipe-nerc* uses the Apache OpenNLP project implementation of the Perceptron algorithm[3] customized with its own features. Specifically, *ixa-pipe-nerc* implements, on top of the local features, a combination of word representation features: (i) Brown [4] clusters, taking the 4th, 8th, 12th and 20th node in the path; (ii) Clark [5] clusters and, (iii) Word2vec [10] 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 [11], both in out-of-domain and in-domain evaluations.

3 Experiments

We train ixa-pipe-nerc with the default parameters and features as described in Agerri and Rigau [2] using both BILOU and BIO annotation schemes. The BARR-E background and Pubmed Annual Baseline are used to train Brown, Clark and Word2vec clusters. Tables 1 and 2 show our official DIANN results for English and Spanish, respectively.

The shared task provides partial and exact matching results. For example, for a disability such as "severe cognitive impairment", exact match would be "severe cognitive impairment" whereas a partial matching may be matching one of the terms included in the sequence, such as "cognitive impairment". However, as the released evaluator[4] only provides exact results, we will limit our discussion to the exact evaluation. The metrics used to evaluate the systems will be precision, recall and their harmonic mean F-measure. For each system, the organizers provided three types of results:

1. DIS: Identifying the sequences in the text denoting disabilities.
2. ALL: The annotation of negated disabilities, namely, the set of annotations {disability, negation trigger and scope of the negation}.
3. JOINT: The joint annotation of disabilities and negation. This evaluation is correct when both negation and disability are correct.

Best results for every setting are obtained using the BILOU encoding scheme[5]. As we expected, using clustering features substantially improves the results, especially in

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

[4] https://github.com/diannibereval2018/evaluation

[5] The BIO scheme suggests to learn models that identify the Beginning, the Inside and the Outside of sequences. The BILOU scheme proposes to learn models Beginning, the Inside and the Last tokens of multi-token chunks as well as Unit-length chunks.

	DIS			ALL			JOINT		
Features	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
(run 2) Local	70.6	49.4	58.1	75.0	39.1	51.4	68.5	45.7	54.8
(run 1) Local + CPUB300	70.1	53.1	60.4	66.7	43.5	52.6	67.2	49.0	56.7

Table 1. Official DIANN 2018 Exact English Results. CPUB300 (run 1): Clark PubMed 300 classes. Local (run 2).

	DIS			ALL			JOINT		
Features	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
(run 2) Local	64.1	61.6	62.8	92.9	59.1	72.2	63.3	59.4	61.3
(run 1) Local + BBE2000	65.0	64.2	64.6	1.0	54.5	70.6	64.4	61.6	62.9
(run 3) Local + BBEC2000	63.6	65.5	64.5	92.3	54.5	68.6	62.6	62.9	62.7

Table 2. Official DIANN 2018 Exact Spanish Results. BBE2000 (run 1): Brown BARR-E 2000 classes; BBEC2000 (run 3): Brown BARR-E corrected 2000 classes; Local (run 2).

terms of recall. However, unlike previous works using the same system [2], the combination of clusters from different data sources does not improve performance. Still, and considering the lack of hand-engineered features and resources employed for the task, the official exact results are quite competitive, especially for Spanish.

3.1 Tokenizing the Test Set

Together with the official results, the task organizers also released the Gold standard test set, which allowed us to perform an extra experiment using exactly the same models trained for the official runs and showed in Tables 1 and 2. We tokenized and segmented the test set and fed it to ixa-pipe-nerc for testing. Given that ixa-pipe-nerc works at token level, we suspected that our official runs were hindered by the fact that the Raw test set was not tokenized and we had to annotate the Raw test set without proper segmentation and tokenization.

	DIS			ALL			JOINT		
Features	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
(run 2) Local	82.0	63.7	71.7	68.7	47.8	56.4	79.2	59.6	68.1
(run 1) Local + CPUB300	85.5	65.8	74.4	68.7	47.8	56.4	82.9	61.7	70.7
IxaMed	78.6	86.0	82.1	47.6	43.5	45.5	74.6	81.1	77.7

Table 3. DIANN 2018 Exact English Results on tokenized test data. CPUB300 (run 1): Clark PubMed 300 classes. Local (run 2).

Tables 3 and 4 show the results of applying the models used for the official runs to the tokenized and segmented test set.

As it can be seen, the improvements are substantial, to the point that our system now obtains the second best results across languages and evaluations, although still lower

	DIS			ALL			JOINT		
Features	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
(run 2) Local	71.0	71.6	71.3	1.0	54.5	70.6	70.7	68.5	69.2
(run 1) Local + BBE2000	75.0	72.0	73.5	1.0	59.1	74.2	74.5	68.9	71.6
(run 3) Local + BBEC2000	71.5	74.6	73.0	1.0	59.1	74.2	70.5	71.1	70.9
IxaMed	75.7	81.7	78.6	88.9	72.7	80.0	74.6	79.5	77.0

Table 4. DIANN 2018 Exact Spanish Results on tokenized test data. BBE2000 (run 1): Brown BARR-E 2000 classes; BBEC2000 (run 3): Brown BARR-E corrected 2000 classes; Local (run 2).

than the results obtained on most evaluations by IxaMed [7], a system developed for the medical domain by our colleagues at IXA Group[6]. Nonetheless, we believe that our results are particularly interesting considering the lack of specific manual adaptation to the domain.

4 Concluding Remarks

In this paper we have presented some experiments showing how to quickly and easily adapt a general purpose sequence labeller to the medical domain. The tagger uses shallow local features enriched with semi-supervised features based on distributional semantics. The result is a robust, language independent tagger suitable to be easily applied across languages and domains. The official results obtained, while fairly competitive themselves, are substantially improved by providing the sequence labeller with a tokenized and segmented input. If we compare our results with IxaMed, the best system in most of the evaluation settings, we obtain better results in terms of precision but recall needs to be further improved. We believe that this could be achieved by training clusters using more closely related data sources to the DIANN datasets, instead of using readily available data, as we have done for these experiments. The models and tools are freely available under Apache License 2.0[7].

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)

[6] http://ixa.eus/

[7] http://ixa2.si.ehu.es/ixa-pipes/

2. Agerri, R., Rigau, G.: Robust multilingual named entity recognition with shallow semi-supervised features. Artificial Intelligence 238, 63–82 (2016)
3. Aronson, A.R.: Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In: Proceedings of the AMIA Symposium. p. 17. American Medical Informatics Association (2001)
4. 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)
5. 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)
6. 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)
7. Gojenola, K., Oronoz, M., Pérez, A., Casillas, A.: Ixamed: Applying freeling and a perceptron sequential tagger at the shared task on analyzing clinical texts. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). pp. 361–365 (2014)
8. Hermenegildo Fabregat, Juan Martinez-Romo, L.A.: Overview of the diann task: Disability annotation task at ibereval 2018. In: Proceedings of the Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) (2018)
9. 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)
10. 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)
11. 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)
12. 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)