The impact of simple feature engineering in multilingual medical NER

R. Weegar†, A. Casillas‡, A. Diaz de Ilarraza‡, M. Oronoz‡, A. Pérez‡, K. Gojenola‡
†Clinical Text Mining Group, DSV, Stockholm University
‡IXA NLP Group, University of the Basque Country (UPV-EHU)
koldo.gojenola@ehu.es

Abstract

The goal of this paper is to examine the impact of simple feature engineering mechanisms before applying more sophisticated techniques to the task of medical NER. Sometimes papers using scientifically sound techniques present raw baselines that could be improved adding simple and cheap features. This work focuses on entity recognition for the clinical domain for three languages: English, Swedish and Spanish. The task is tackled using simple features, starting from the window size, capitalization, prefixes, and moving to POS and semantic tags. This work demonstrates that a simple initial step of feature engineering can improve the baseline results significantly. Hence, the contributions of this paper are: first, a short list of guidelines well supported with experimental results on three languages and, second, a detailed description of the relevance of these features for medical NER.

1 Introduction

Named Entity Recognition (NER), such as the recognition of person names, organizations, locations or medical entities, has become a crucial task in any Natural Language Processing (NLP) application, as a first step to other types of processing as, for example, Relation Extraction (Oronoz et al., 2015). Several tools have been developed for this task, such as CRF++ (Kudo, 2013), SVM (Kudo and Matsumoto, 2001) or Perceptron (Collins, 2002). Using these tools, and training them with a set of annotated data, many people can obtain a NER system easily and apply it to the respective domain. In this paper the experiments will be performed with clinical texts, on the recognition of Medical entities such as disorder or drug brand names. The basic NER models make use of a sequence of *(word form, features, tag)* elements for training. For inference, the system will give the tag sequence with the highest score given a new text. Each model is defined by a set of features, taken from the surroundings of each word to be tagged, usually by means of a sequential tagging approach.

Many techniques have been developed in order to improve the NER results, such as the incorporation of additional information, in the form of lemmatization, POS tagging, dictionaries and ontologies (IHTSDO, 2016), or the inclusion of knowledge acquired by unsupervised techniques like Brown clusters (Brown et al., 1992; Clark, 2003), word2vec neural models (Agerri and Rigau, 2016) or deep neural network architectures (dos Santos and Guimarães, 2015) that yielded significant improvements.

However, this availability of tools and techniques has led to using only a limited set of predefined or standard models that were successful for a prototypical NER task, without any kind of time-consuming adjusting (Pradhan et al., 2014). Moreover, as most published papers center on novel techniques (Ratinov and Roth, 2009; Turian et al., 2010), sometimes less effort is devoted to data analysis or to filtering and tuning the models. Researchers rarely give the full details of feature engineering and they often present their best configurations, or otherwise they only study the impact of one or two specific types of feature. However, the benefits of sophisticated techniques would be better highlighted taking a stronger baseline as departure. In this sense, this paper may be useful to researchers that are new to the field of medical NER, showing the impact of simple feature engineering on medical texts in three languages.

As an example, looking at the systems presented at the Semeval 2014 Shared Task 7 on English Medical texts (Pradhan et al., 2014), we see that most of the system descriptions do not give a precise

Proceedings of the Clinical Natural Language Processing Workshop,
pages 1–6, Osaka, Japan, December 11-17 2016.

overview of the contribution of the simplest feature types (Ramanan et al., 2014; Leal et al., 2014; Kate, 2014; Attardi et al., 2014) and they give at most a list of the used features, but without a detailed account of each's performance. For example, while Attardi et al. (2014) describe word shape features, they do not describe the window of words used, while Parikh et al. (2014) use a window of three words ([-2,0]).

There exist several available systems for English, as cTAKES (Savova et al., 2010), which was used by some of the participants at Semeval 2014 or cLiner (Boag et al., 2010). However, for other types of languages, there is a scarcity of resources and information about the usefulness of the available features.

We will experiment the effect of using simple features on medical NER, giving a measure of the improvements that can be achieved without resorting to more sophisticated types of information. Although most of these techniques have been previously applied in many works (Pradhan et al., 2014), we think that their effectiveness has not always been clearly evaluated, and they are briefly described as a preprocessing step before applying other, more complex, techniques. The main contribution of this paper will be a thorough examination of simple features for the recognition of entities in the medical domain. To give a better account of the generalization across different languages, we will perform our experiments on English, Spanish and Swedish, hoping that these results will be useful for many researchers and will help them to follow the principle of doing the easy things first, before resorting to more complex models.

2 Experimental Setup

We will perform a set of experiments using different types of features, starting from the most basic type of information, the word form itself and its derivatives, and continuing with basic language processing tools as lemmatization, POS tagging and medical dictionaries and ontologies: **Phase 1:** using only word forms (plus lower-casing); **Phase 2:** using prefixes and suffixes of different length. For example, the four letter suffix *-itis* indicates an inflamatory disease, as in meningitis or bronchitis; **Phase 3:** using different patterns of capitalization of word forms (word starts with a letter, all letters are capitalized, or different types of numbers); **Phase 4:** using lemmas; **Phase 5:** using POS tags. **Phase 6:** using Snomed-CT tags.

With the objective of establishing measures of the contribution of several features corresponding to simple types of information to medical NER, we will examine three languages:

- **English (EN)** We will use data from the SemEval-2014 Task 7 Analysis of Clinical Text Shared Task ShARe[1]. This corpus comprises annotations of disease entities (9,694 instances) over de-identified clinical reports from a US intensive care department (version 2.5 of the MIMIC II database)
- **Spanish (SP)** The Spanish EHRs consist of patient records collected during 2008-2012 at the Galdakao-Usansolo Hospital leading to 141,800 documents, 52 million word-forms (Oronoz et al., 2015). The entire corpus was provided after anonymization, signing confidentiality agreements and passing the corresponding ethical committees. From this set of raw clinical text, a subset of 121 texts was randomly selected for manual annotation (3,362 instances of diseases and 1,406 drugs).
- **Swedish (SW)** The Swedish clinical text[2] origins from patient records from over 500 different clinical units at Karolinska University Hospital. The texts were collected during 2009-2010 and are stored in HEALTH BANK (Dalianis et al., 2015). For this study, a supervised corpora was created, annotated with medical entities (4,000 entities corresponding to body parts, disorders and findings).

Regarding the English corpus, we only had access to the train and development sets, because the test set was not public. This is not a problem, because from our experiments on the Semeval Shared Task datasets, the results on the test set increased by about 2 percent points (Pradhan et al., 2014), as using the train and development sets for training compensates the effect of evaluating on the unseen test set. For that reason, we will use the train set for training and will evaluate on the development set.

For the experiments, we will use our own implementation of the averaged structured perceptron (Freund and Schapire, 1999; Collins, 2002), a state of the art tagging model that relies on Viterbi decoding of training examples combined with simple additive updates. The algorithm is competitive to maximum-entropy taggers or CRFs (Collins, 2002). For each experiment, we trained 25 iterations on different feature templates. Although not reported in this work, similar experiments have also been performed

[1] http://share.healthnlp.org

[2] This research was approved by the Regional Ethical Review Board in Stockholm, permission number 2014/1882-31/5

Perceptron. Phase 1 (word forms)				
Features	**Model**	**EN**	**SP**	**SW**
Window size (word unigrams)	wf(-2, +2)	**51.20**	51.07	55.94
	wf(-2, +1)	49.20	52.46	56.09
	wf(-2, 0)	46.80	50.70	55.61
	wf(-1, +1)	47.70	**52.49**	57.16
	wf(-1, 0)	40.10	50.18	**58.47**
Window size (lowercase word unigrams)	wf(-2, +2)	53.30	57.20	57.98
	wf(-2, +1)	**54.30**	**57.74**	57.84
	wf(-2, 0)	52.20	54.81	58.71
	wf(-1, +1)	49.80	56.37	58.98
	wf(-1, 0)	47.70	55.01	**60.44**

Table 1: Results changing the window size and capitalization of words (wf(i, j) = unigram features of words in a window from i to j).

Perceptron. Phase 2 (prefixes and affixes)				
Features	**Model**	**EN**	**SP**	**SW**
Adding prefixes/ suffixes	p2 + s2	59.40	60.74	62.27
	p3 + s3	60.40	61.75	64.23
	p4 + s4	59.90	60.73	64.50
	p23 + s23	60.40	**63.27**	63.74
	p34 + s34	60.40	61.82	65.34
	p234 + s234	**60.50**	62.10	**66.36**
Adding prefixes/ suffixes (lowercase)	p2 + s2	59.10	62.13	62.59
	p3 + s3	60.50	63.43	64.69
	p4 + s4	60.30	63.68	64.78
	p23 + s23	60.30	64.94	64.45
	p34 + s34	59.60	64.09	64.51
	p234 + s234	**61.00**	**65.23**	66.07

Table 2: Results adding prefixes and suffixes of word forms, using the best model of phase 1 as baseline. (p$N_1N_2...N_k$ = prefix of size N_1, N_2, ... N_k for the current word).

Perceptron. Phase 3 (capitalization and numbers)			
Model	**EN**	**SP**	**SW**
(1) all capital letters	60.30	65.99	**66.66**
(2) starts with capital letter	61.00	65.88	65.74
(3) number types	61.00	65.85	66.02
(4) mixed letters and numbers	60.10	64.59	65.68
(5) = (1) + (2)	**61.80**	66.04	65.72
(6) = (1) + (3)	60.20	**66.22**	66.02
(7) = (2) + (3)	60.90	65.34	66.06
(8) = (1) + (2) + (3)	61.20	65.86	65.45

Table 3: Results adding capitalization and numbers to the best model of phase 2.

Perceptron. Phase 4 (lemmas)				
Features	**Model**	**EN**	**SP**	**SW**
Window size	lem(0)	61.40	65.82	66.12
	lem(-1, +1)	**62.10**	66.13	65.31
	lem(-2, +2)	60.00	65.67	65.74

Table 4: Results adding features based on lemmas (on the best model of phase 3) (lem(i, j) = unigram features of lemmas in a window from i to j).

with SVM and CRFs, obtaining results comparable but slightly lower than with the Perceptron. For English and Spanish, we used FreelingMed for lemmatization, POS tagging, and annotating Snomed CT concepts (Oronoz et al., 2013). For Swedish, we used Stagger (Östling, 2013). The experiments were tested following a greedy approach, taking at each phase the best model in the previous phase as a baseline. This approach can be debatable, as it could happen that the knowledge used in phase $x+1$ could not be useful when applied with the best model in phase x, but perhaps it produced improvements at phases earlier than x. We have also experimented the effect of applying each set of features independently, but our aim is to get an account of the benefits obtained by applying a simple yet coherent approach (from the simplest to more elaborated experiments), and we leave out of the scope of this work the development of more time-consuming tests, such as grid search.

3 Results and Discussion

Table 1 shows the results (F-measure) with different values of the window size (WS). There is no use on trying a single WS for all the languages as it has different impacts on different languages. Note that lowercasing improved the results considerably for all three languages, specially for Spanish. We hypothesize that this can be due to the informal writing used in the Spanish medical reports, characterized by big differences in writing style and non-consistent use of casing (either lowercase, uppercase or mixed). The use of prefix/suffixes in Phase 2 (see Table 2) helps significantly for all the languages with respect to the best results from Phase 1 (above 5 absolute points in all cases). Lower casing does not seem useful for English and Swedish (0.5 improvement for English over the best result without lower casing, and no improvement for Swedish), but it gives an increase of 2 points on Spanish.

Table 3 presents the effect of adding features to represent capitalization patterns (words formed only by capital letters and words that start with a capital letter) and number types[3]. The improvements are modest for Swedish and slightly better for English (adding 0.8 points) and Spanish (almost one point).

[3] 'number types' differentiates numbers according to four types: only digits (1234), digits with hyphen (23-35), digits with '/' (2/2012), and measure (200_mg)).

Perceptron. Phase 5 (POS)				
Features	**Model**	**EN**	**SP**	**SW**
Window size	pos(0)	61.90	**70.01**	65.55
	pos(-1, +1)	**63.80**	69.95	65.50
	pos(-2, +2)	63.10	68.94	66.21

Table 5: Results adding features based on POS tags on the best model of the previous phase.

Perceptron. Phase 6 (Snomed CT, ...)				
Features	**Model**	**EN**	**SP**	**SW**
Window size	snomed(0)	66.20	68.22	**68.41**
	snomed(-2, +2)	**66.40**	67.84	68.27
	snomed(-2, +2)	65.60	67.67	68.31

Table 6: Results adding features based on Snomed tags.

Using lemmatization, Table 4 shows that we get an improvement on English (+0.3) and a decrease for Spanish and Swedish. This seems surprising, as in principle lemmatization could be useful to normalize terms (e.g. singular/plural and feminine/masculine in Spanish). Note that, as we are performing a greedy approach, the number of features used grows from one phase to the next one, and this is the main reason why we limited the number of feature templates, because the gains are decreasing for each phase. It should be clear that our experiments do not conclude that lemmatization is not useful but, rather, they show that it is not useful after applying other features. Table 5 shows the results using POS features, helpful for English and Spanish, but not for Swedish. We hypothesize that it could be due to the poorer quality of the Swedish POS tagger (Dalianis et al., 2015). Finally, Table 6 presents the results using specialized medical dictionaries, giving the best results for English and Swedish, but no improvement for Spanish. This aspect deserves further work, because the Spanish Snomed has similar coverage to the English version regarding concepts (around 300,000), but less terms (660,000 compared to 480,000).

4 Conclusion

Standard and well-known features together with model tuning are frequently being left aside by researchers in favor of novel approaches, as though they were low-level or insignificant mechanisms. By contrast, we have showed that these simple techniques lead us to achieve significant improvements at really low computation expenses. As an example, looking at the Semeval 2014 Shared Task, we can say from our results that a simple system using only word forms and POS would outperform more than half of the presented systems[4]. It is not our aim to imply that other systems were poorly designed, as most of them had other objectives in mind, such as experimenting new approaches but, rather, our objective is to delve into the details of the simplest approaches, that are specially interesting for implemented systems, but are often neglected in scientific papers[5]. The results for our best performing systems for Swedish and Spanish are near to those obtained by more elaborated techniques like word embeddings, although they are still far from the best performing system on the Semeval English test.

To summarize, we experimented the NER task related to the biomedical domain in three languages: Semeval task in English, and EHRs in both Swedish and Spanish. The techniques presented tend to be of much benefit, particularly for domains that lack of big amounts of data, as it is the case of biomedicine:

- It is recommendable to re-case the text and well-worthy trying different window-sizes on each language (not simply using the default parameters adopted from other languages).
- While prefixes and suffixes have a different impact on each language, it seems as though taking all prefixes and suffixes of lengths 3 and 4 is a generally recommendable configuration. These techniques can be specially useful when analyzing non-formal text, as in the Spanish medical records.
- Regarding other types of information (capitalization, numbers, lemmas and POS) we have seen that, although the features can be effective, they should be carefully tested on each language and corpus.
- Overall, we see that there are important differences on the impact of different features with respect to each language. This fact opens an interesting research area for analyzing the effect of language and corpus types on the effectiveness of each feature.

For future work we will take these results as a stronger baseline and delve into state-of-the art techniques e.g. word embeddings (Bengio et al., 2006; Mikolov et al., 2013) and recursive neural networks (Lample et al., 2016) to gain an insight on their impact on medical NER for these three languages.

[4]Looking at Tables 5 and 6, and taking into account that the results on the test set bumped by 2 points (Pradhan et al., 2014).

[5]We think that, in fact, this low level tuning was performed for the Semeval 2014 best performing systems, although their system description papers do not address this issue in detail.

Acknowledgements

The authors would like to thank the Nordic Center of Excellence in Health-Related e-Sciences (NIASC) and the personnel of Pharmacy and Pharmacovigilance services of the Galdakao-Usansolo Hospital and the Pharmacy service of the Basurto Hospital. This work was partially funded by the Spanish Ministry of Science and Innovation (EXTRECM: TIN2013-46616-C2-1-R, TADEEP: TIN2015-70214-P) and the Basque Government (DETEAMI: Department of Health 2014111003).

References

Rodrigo Agerri and German Rigau. 2016. Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence*, 238:63–82.

Giuseppe Attardi, Vittoria Cozza, and Daniele Sartiano. 2014. Unipi: Recognition of mentions of disorders in clinical text. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 754–760, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Yoshua Bengio, Holger Schwenk, Jean-Sébastien Senécal, Fréderic Morin, and Jean-Luc Gauvain. 2006. Neural probabilistic language models. In *Innovations in Machine Learning*, pages 137–186. Springer.

W. Boag, K. Wacome, T. Naumann, and A. Rumshisky. 2010. Cliner: A lightweight tool for clinical named entity recognition. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Alexander Clark. 2003. 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*, pages 59–66. Association for Computational Linguistics.

Michael Collins. 2002. Discriminative training meth- ods for hidden markov models: Theory and ex- periments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 1–8. Association for Computational Linguistics, July.

Hercules Dalianis, Aron Henriksson, Maria Kvist, Sumithra Velupillai, and Rebecka Weegar. 2015. Health bank–a workbench for data science applications in healthcare. In *Proceedings of the CAiSE-2015 Industry Track co-located with 27th Conference on Advanced Information Systems Engineering (CAiSE 2015)*, volume 1381, pages 1–18. CEUR, urn:nbn:de:0074-1381-0.

Cicero dos Santos and Victor Guimarães. 2015. Boosting named entity recognition with neural character embeddings. In *Proceedings of the Fifth Named Entity Workshop*, pages 25–33, Beijing, China, July. Association for Computational Linguistics.

Yoav Freund and Robert E. Schapire. 1999. Large margin classification using the perceptron algorithm. *Machine Learning*, 37(3):277–296.

IHTSDO. 2016. SNOMED-CT, Systematized Nomenclature of Medicine-Clinical Terms, http://www.ihtsdo.org/snomed-ct/. Accessed 2014-04-09.

Rohit Kate. 2014. Uwm: Applying an existing trainable semantic parser to parse robotic spatial commands. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 823–827, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Taku Kudo and Yuji Matsumoto. 2001. Chunking with support vector machines. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics, 2001.

Taku Kudo. 2013. CRF++: yet another CRF toolkit. `https://taku910.github.io/crfpp//`.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

André Leal, Diogo Gonçalves, Bruno Martins, and Francisco M Couto. 2014. Ulisboa: Identification and classification of medical concepts. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, pages 711–715.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Maite Oronoz, Arantza Casillas, Koldo Gojenola, and Alicia Perez. 2013. Automatic Annotation of Medical Records in Spanish with Disease, Drug and Substance Names. In *Lecture Notes in Computer Science, 8259. Progress in Pattern Recognition, ImageAnalysis, ComputerVision, and Applications 18th Iberoamerican Congress, CIARP 2013, Havana, Cuba*, November 20-23.

Maite Oronoz, Koldo Gojenola, Alicia Prez, Arantza Daz de Ilarraza, and Arantza Casillas. 2015. On the creation of a clinical gold standard corpus in spanish: Mining adverse drug reactions. *Journal of Biomedical Informatics*, 56:318 – 332.

Robert Östling. 2013. Stagger: An open-source part of speech tagger for Swedish. *Northern European Journal of Language Technology*, 3:1–18.

Ankur Parikh, Avinesh PVS, Joy Mustafi, Lalit Agarwalla, and Ashish Mungi. 2014. Thinkminers: Disorder recognition using conditional random fields and distributional semantics. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 652–656, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Sameer Pradhan, Noémie Elhadad, Wendy Chapman, Suresh Manandhar, and Guergana Savova. 2014. Semeval-2014 task 7: Analysis of clinical text. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 54–62, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Sv Ramanan, Chennai Adyar, and Senthil Nathan. 2014. Relagent: Entity detection and normalization for diseases in clinical records: A linguistically driven approach. *SemEval 2014*, page 477.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394, Uppsala, Sweden, July. Association for Computational Linguistics.