Procesamiento del Lenguaje Natural, Revista nº 60, marzo de 2018, pp. 13-20 recibido 03-11-2017 revisado 05-01-2018 aceptado 09-02-2018

Ensembles for clinical entity extraction

Agrupaciones para la extracción de entidades clínicas

Rebecka Weegar[1], **Alicia Pérez**[2], **Hercules Dalianis**[1]
, **Koldo Gojenola**[2], **Arantza Casillas**[2], **Maite Oronoz**[2]
[1]Clinical Text Mining group; DSV; Stockholm University
[2]IXA (http://ixa.eus); Euskal Herriko Unibertsitatea (UPV-EHU)
Corresponding author: rebeckaw@dsv.su.se

Abstract: Health records are a valuable source of clinical knowledge and Natural Language Processing techniques have previously been applied to the text in health records for a number of applications. Often, a first step in clinical text processing is clinical entity recognition; identifying, for example, drugs, disorders, and body parts in clinical text. However, most of this work has focused on records in English. Therefore, this work aims to improve clinical entity recognition for languages other than English by comparing the same methods on two different languages, specifically by employing ensemble methods. Models were created for Spanish and Swedish health records using SVM, Perceptron, and CRF and four different feature sets, including unsupervised features. Finally, the models were combined in ensembles. Weighted voting was applied according to the models individual F-scores. In conclusion, the ensembles improved the overall performance for Spanish and the precision for Swedish.

Keywords: Clinical entity recognition, ensembles, Swedish, Spanish

Resumen: Los informes médicos son una valiosa fuente de conocimiento clínico. Las técnicas de Procesamiento del Lenguaje Natural han sido aplicadas al procesamiento de informes médicos para diversas aplicaciones. Generalmente un primer paso es la detección de entidades médicas: identificar medicamentos, enfermedades y partes del cuerpo. Sin embargo, la mayoría de los trabajos se han desarrollado para informes en Inglés. El objetivo de este trabajo es mejorar el reconocimiento de entidades médicas para otras lenguas diferentes a Inglés, comparando los mismos métodos en dos lenguas y utilizando agrupaciones de modelos. Los modelos han sido creados para informes médicos en Español y Sueco utilizando SVM, Perceptron, CRF y cuatro conjuntos diferentes de atributos, incluyendo atributos no supervisados. Para el modelo combinado se ha aplicado votación ponderada teniendo en cuenta la F-measure individual. En conclusión, el modelo combinado mejora el rendimiento general y para posibles mejoras debemos investigar métodos más sofisticados de agrupación.

Palabras clave: Reconocimiento de entidades médicas, agrupaciones, sueco, castellano

ISSN 1135-5948. DOI 10.26342/2018-60-1

1 Introduction

Natural language processing has been applied to health records for tasks as diverse as detecting adverse drug reactions (Henriksson et al., 2015), surveillance of nosocomial infections (Haas et al., 2005) and for assigning ICD codes to health records (Crammer et al., 2007). To many of the tasks utilizing natural language processing on health records, a well-functioning named entity recognition module is central (Demner-Fushman, Chapman, and McDonald, 2009).

There are European and also national projects that focus on the automatic extraction of valuable information from patient records. Three on-going projects are: firstly, CrowdHEALTH, a European project that attempts at gathering and processing multi-modal data from member states, conform ethical regulations, and exchange important information; secondly, the Spanish Ministry has involved a multi-disciplinary team to tackle natural language processing in the clinical domain among others in the so called "Plan de impulso de las tecnologías del lenguaje"; a third example is the Nordic Center of Excellence in Health-Related e-Sciences (NIASC) which is funded by NORDFORSK, the Nordic council of ministers, with one aim to detect early symptoms of cancer in patient records. Being so different from one another, the aforementioned three projects include, to different extents, the detection of key entities. While CrowdHEALTH shall incorporate languages from European states, English is still the dominating language in research articles in the clinical domain. Moreover, patient records is a type of data seldom explored due to confidentiality issues.

Motivated by this gap and shared interest, the Clinical Text Mining group at Stockholm University and the IXA research group at the University of the Basque Country cooperate with the aim to extract information from patient records and build robust methods for languages other than English.

The goal of this work is to extract medical entities from Electronic Health Records (EHRs) focusing on patient records in Swedish and Spanish, from Karolinska University Hospital and Galdakao-Usansolo Hospital respectively.

Swedish is a Germanic language with about 10 million speakers. A challenge for processing Swedish, as well as other Germanic languages, is that compounds are very common. For Swedish, a rich variety of noun compounds are possible, an example is the word *huvudvärkstablett* (*huvud*-head, *värk*-ache, *s*-, *tablett*-tablet). Spanish is a Romance language and about 360 million people has Spanish as their first language. Regarding the object of this paper, clinical entities, some examples of specific features of the Spanish language are given in (Reynoso et al., 2000). For instance, medical terms in English expressed by gerunds tend to take the form of subordinate clauses or prepositional phrases in Spanish. Some examples from SNOMED CT are as follows: *Conditions causing complications in pregnancy* that takes the form of a subordinate *"condiciones que causan [that cause] complicaciones en el embarazo"*; *dispatching and receiving clerk* takes the form of the prepositional phrase *"empleado de despacho y recepción de mercadería"*.

Text in patient records tend to pose characteristics that are not shared with other kind of texts (such as journal abstracts, social media etc.) which make them challenging to process. These characteristics include a rich vocabulary with many possible forms for the same concept and domain specific terminology, many abbreviations and acronyms which may be ambiguous, and few complete sentences. Besides, it has been found that up to 10% of tokens in health records are misspelled (Ruch, Baud, and Geissbühler, 2003; Lai et al., 2015; Ehrentraut et al., 2012).

Conditional Random Fields (CRFs) is a probabilistic model for labelling sequences of data (Lafferty, McCallum, and Pereira, 2001), which makes it suitable for named entity recognition. CRFs were previously applied in the clinical domain with good results (Skeppstedt et al., 2014). As with CRFs, Support Vector Machines (SVMs) have proven useful for entity recognition.

The works explored so far that made use of CRFs or SVMs to detect entities have the drawback of relying on vast discrete feature-spaces built up on the basis of n-grams of words. Named entity recognition has been recently shifted from symbolic representations (words, lemmas, POS, etc.) to dense representations.

In Tang et al. (2014) biomedical entity recognition was carried out on Biocreative II GM corpus making use of CRFs. With regard to the features, they used basic features (stemmed words and POS), Brown clusters, distributional word representations and word embeddings extracted with word2vec.

Regarding entity recognition for Spanish, word representations (Turian, Ratinov, and Bengio, 2010) have been incorporated as external features to infer a CRF (Zea et al., 2016; Agerri and Rigau, 2016). To this end, the entity recognition system was inferred from in-domain annotated data, however, large out-domain unannotated data were used to infer continuous word-representations. This strategy can yield results comparable to those obtained with approaches based on deep learning strategies (Zea et al., 2016), possibly due to the semantic relatedness associated to continuous spaces that lead to generalization (Faruqui and Padó, 2010).

In addition to more robust feature representations, ensembles of classifiers have previously been shown capable of improving Entity Recognition. Florian et al. (2003) applied Named Entity Recognition to English and German texts using an ensemble of four different classifiers achieving improved results. Saha and Ekbal (2013) created an ensemble of seven base-learners, including CRF and SVM, and performed Named Entity Recognition on Hindi, Bengali and Telugu. The performance of the ensemble of classifiers using weighted voting was better than that of any of the individual classifier and the weights were determined using genetic algorithms. Ensembles of classifiers have also been used on clinical texts, Kang et al. (2012) combined seven existing system for clinical entity recognition for English texts. A threshold – the number of systems needed to agree on an entity to include it – was decided by evaluating the systems on the training set. An ensemble of systems was found to give a higher performance than any of the individual systems

Exploring patient records is a challenging task. Moreover, given that this is a joint-project on Swedish and Spanish, the aim is to use robust cross-lingual techniques. The contribution of this work is the exploration of the use of ensemble techniques in a comparable task on both languages. We mean *ensemble* in two ways: on the one hand, we explored a simple combination of three base-learners (a perceptron, a CRF and an SVM); on the other hand, each base-learner was trained on ensembles of semantic spaces. The system rests on classical supervised classification techniques combined taking advantage of features derived from dense representations.

The focus of this paper is clinical entity recognition following the criteria in Pérez et al. (2017). That is, first, the decision space is set by means of semi-supervised representations that include ensembles of features derived from distributional semantics and also from classical symbolic representations (section 2 is devoted to the representation). The characterisation relied upon a big unannotated data-set, next, with an annotated set of much smaller size supervised classifiers can be inferred to decide whether a phrase is a clinical entity or not.

2 Ensembles of features for clinical entity representation

Classical entity recognition systems rested mainly upon word-forms (W) as a surface representation and lemmas with POS as a representation with linguistic (L) connotations. The linguistic features conveys helpful information, but to generate such features an analyser adapted to the medical domain is required, which is not available for all languages. Here, Freeling-Med (Oronoz et al., 2013) was used for Spanish and Stagger (Östling, 2013) paired with terminology matching following Skeppstedt et al. (2014) were used for Swedish.

In this work, the linguistic features were complemented with unsupervised (U) features. Current trends in language processing are shifting from symbolic representations based on words to distributional semantics. The benefits are multiple: while word-based representations tend to be scattered, continuous representations embed semantic information in a vector space. Classical symbolic representations (e.g. bag of words) entail a big number of components, and close vectors are rarely related. By contrast, distributional semantics keeps the dimension of the space manageable and permits a quantitative interpretation of word-relatedness. Word representations are obtained from big corpora by means of unsupervised techniques based on co-occurrences of words. The representation achieved depends not only on the corpus but also on a set of hyper-parameters influencing the training of the model. With a given corpus and different hyper-parameters, different spaces are obtained. Yet, currently there is no conclusive fine-tuning technique to decide on the parameter setting. It has also been shown that when combining different spaces, rather than being redundant, the ensembles of semantic spaces enhanced the word-representation and improved information extraction techniques (e.g. entity recognition) (Henriksson, 2015).

In the clinical domain and, particularly, working with EHRs, available data tend to be scarce. The question arising is if distributional semantics can cope, in a robust and reliable way, with data sparseness. A typical method of dealing with sparsity of data given a continuous variable is clustering. Clustering regards as equivalent close values of a given variable as if we zoomed out our variable and could not make distinguishable close values.

On this account, the semantic spaces were clustered using k-means clustering. Again, k is a key parameter that changes the representation.

All in all, two semantic spaces were built from a given unannotated corpus, each of which with different hyper-parameters. The semantic spaces were clustered using two different numbers of clusters (k) in an attempt to combine fine-grained and coarse-grained clustering.

As an additional effort to handle data sparsity, features were also generated using Brown clustering (Brown et al., 1992). In this case, the information conveyed and the approach to get it is notably different. Brown clustering is a hierarchical clustering which arranges words found in a corpus into a tree with the words at the leaf nodes and where clusters corresponds to sub trees (Liang, 2005).

3 Ensemble classifier for entity recognition

This work started from the hypothesis that a simple ensemble learner would beat the individual base-learners, the contribution of three state of the art supervised classifiers was explored and next they were combined in a simple way. All the classifiers were trained using the ensemble representation-space features described in section 2.

3.1 Base-learners

Three approaches for supervised learning of medical entities were selected. The selected learners are all discriminative classifiers that perform sequential tagging, with different characteristics:

Perceptron This algorithm performs Viterbi decoding of the training examples combined with simple additive updates, trying to find the sequence of tags with the maximum score. The algorithm is competitive to other options such as CRFs (Carreras, Márquez, and Padró, 2003).

Support Vector Machine SVMs make use of kernel functions, which provide a similarity metric between two instances and, hence, a way to get a model suitable for discriminative tasks.

Conditional Random Fields CRF is a machine learning algorithm that makes use of feature functions representing the relationships between the features and the output. To assign the current output, it takes into account both earlier and later parts of the input and, also, the previous output tag.

3.2 Ensembles

The rationale of ensemble or committee models is quite intuitive: if many estimates are averaged together the variance of the estimate is reduced (Murphy, 2012). Regarding the ensembles, there are two key-issues:

1. The **diversity** of the base-models to be combined, since there is no point in combining models that make similar decisions. In this case, two kind of combinations were explored:
   (a) Combining models obtained with the same learning approach but different input representations or parameters, in this case with four different feature sets.
   (b) Combining models obtained with different learning approaches, namely the CRFs, Perceptrons and SVMs.
2. The **combination strategy**. There are a wide variety of combination strategies: linear opinion pools (or simple voting); weighted voting; stacking or stacked generalization learns a classifier from the

predictions of the base-learners; and others. Weighted voting was selected for its simplicity, and the weights were set to the average F-score of each base-model.

4 *Experimental layout*

4.1 Task and corpus

Two data sets were used for each language, a smaller annotated set and a larger set used for the unsupervised features. For Spanish, diseases (4,296 instances) and drugs (1,862 instances) were annotated, and for Swedish the annotated entities were body parts (2,082 instances), disorders (981 instances) and findings (3,759 instances) from HEALTH BANK[1] (Dalianis et al., 2015). The annotated data was divided into training sets containing about 60% of the annotations, development sets with 20% of the data and a test set with the remaining 20%. The unannotated data sets were of similar size for both languages, 52×10^6 tokens for Spanish and 51×10^6 for Swedish. More details about the data sets can be found in (Pérez et al., 2017).

4.2 Results

Table 1 shows the results for the ensemble tagger. The ensemble model was built up of 3 base-learners (CRF, Perceptron and SVM) each of which was trained in 4 alternative spaces using different sets of features:

1. **W:** Word-forms.
2. **WL:** Word-forms and Linguistic information (lemmas and part of speech)
3. **WLU:** the previous WL and Unsupervised features (ensembles of semantic spaces clusterized and Brown clusters).
4. **WU:** just word-forms and unsupervised features.

The composition of the 12 base-models consisted of a simple weighted voting strategy where the weights associated with each base-learner were set according to their individual F-scores on the development set. To be precise, the votes were weighted by the average F-score over all the the classes (the different entities in each set), giving 12 votes in total. This means that a model that proved more successful on the development data was given a stronger influence over the final tagging.

Other strategies are possible, for example using only the three base-models trained on the feature set that was most successful (WLU), and relying only on 3 votes for the ensemble, nevertheless, these results were slightly lower than the ensemble of all the models provided in Table 1.

Spanish				
Set	**Entity**	**P**	**R**	**F**
Dev	Disease	69.98	60.61	64.96
	Drug	94.95	82.76	88.43
	Average	78.55	67.46	72.58
Test	Disease	69.92	55.82	62.08
	Drug	94.38	84.46	89.15
	Average	78.68	65.22	**71.32**

Swedish				
Set	**Entity**	**P**	**R**	**F**
Dev	Body part	88.03	76.27	81.73
	Disorder	70.81	57.00	63.16
	Finding	63.89	58.33	60.98
	Average	72.68	63.98	68.02
Test	Body part	86.14	81.45	83.73
	Disorder	70.47	55.51	62.10
	Finding	68.28	65.35	66.78
	Average	74.39	69.24	**71.65**

Table 1: Results of the ensemble tagger comprising 12 base-models for each language. Evaluation metrics: Precision (P), Recall (R) and F-score (F)

4.3 Discussion

Not all the entities are equally easy to recognize for the system. Finding drugs or body parts is by far simpler than recognizing diseases, disorders or findings. Drugs, substances and brand-names tend to follow similar patterns and the same applies to body parts. By contrasts, in EHRs diseases, disorders and findings are expressed in a variety of ways that hardly ever follow their corresponding standard term in clinical dictionaries (e.g. ICD) or ontologies (e.g. SNOMED-CT). In medical records the same disease could be described in diverse and very different ways: either formal, or colloquial, either in a specific way or in a general way. In addition, there are variations in the way of expressing numbers (e.g. *"diabetes mellitus type II"*, *"diabetes mellitus type 2"*) and abbreviations are used fre-

[1] This research has been approved by the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2014/1882-31/5.

quently (e.g. *"DM2"*). For example, Pérez et al. (2015) showed the case of the *"Malignant neoplasm of prostate"* disease that appeared in the EHRs with the variants *"Adenocarcinoma of the prostate"*, *"prostate adenocarcinoma"*, *"prostate Ca."* and *"PROSTATE CANCER"*.

The models were trained on the training set, fine-tuned on the development set and, finally, re-trained on a joint training and development set to assess the system on the test set. The results achieved in both development and test sets are comparable.

With regard to the difference of the performance across languages, while the results in the development set were better for Spanish than for Swedish, it was the other way around in the test set. Our intuition is that the differences in the performance on the development and test sets stand on the way the split was carried out. The sets were split at document level and given that the documents are much longer in the Spanish set, it might have made the inference tougher.

Previous work on medical entity recognition in this task showed that the aforementioned base-learners (SVM, CRF and Perceptron) were useful for the clinical domain. The best results for an individual model were achieved by the Perceptron using the WLU feature set. For Swedish, the average F-score in the test set for this configuration was 71.72 and for Spanish, the average F-score was 70.30 (Pérez et al., 2017). This work investigated the capability of ensemble techniques and explored diverse sets of features. The results of each of the 12 base-models involved were combined following a weighted voting strategy. We found that the ensemble approach was robust and that the overall trend, for both languages, and on both the development sets and the tests, was an improvement of the precision scores. On the test set this improvement was 1.54 points for Swedish and 4.3 points for Spanish. However, in most cases, the ensemble approach decreased the average recall. The recall was only improved on the Spanish development set. Altogether, the average F-scores were improved for both development and test data for Spanish, but only on the development data for Swedish.

The p-value given by the McNemar's test (McNemar, 1947) on the improvements achieved with respect to the best base-model on the development set (i.e. the Perceptron on WLU space) show statistical significance (p-value $\ll 0.01$) for Swedish, however, not for Spanish (p-value < 0.08). By contrast, for the best performing model in the test set, the difference with respect to the ensemble model is statistically significant with p-value $\ll 0.01$ for both languages.

It is debatable whether this increment in precision is worth the combination of 12 models. However, within the clinical domain precision tends to be crucial. Given the improvements achieved by this simple combination technique, the plans forfuture work include to test other methods for combining the learners, for example, stacked generalization could prove more efficient than the weighted voting approach.

5 Concluding remarks

Text in health records is challenging to process, and one of the biggest challenges for further work has to do with the variability associated to the spontaneous expressions found in medical records, particularly when it comes to express multi-word entities regarding the diseases, disorders and findings. A strength of this work stands on the comparable framework achieved which allows for evaluations of clinical entity recognition on clinical texts in two different languages A step ahead is made with respect to previous works combining three base-learners (CRF, Perceptron, SVM) inferred on four alternative spaces. Influenced by previous works these spaces, including unsupervised features such as clusterized ensembles of semantic spaces, brown clusters and also linguistically motivated features (word-forms, POS and lemmas), were built. All together, we constructed an ensemble that combined 12 base-models using a weighted voting paradigm where the weights were set as the average F-score of each model. The ensemble model achieved an average F-score above 71%. The combination increased the performance in terms of precision for both languages. It seems as if the upper threshold was not achieved yet and that there is room for improvement, specifically for recall. Therefore, ensemble techniques other than weighted voting should be explored for future work.

Acknowledgements

This work has been partially funded by the Spanish ministry (PROSAMED: TIN2016-

77820-C3-1-R, TADEEP: TIN2015-70214-P), the Basque Government (DETEAMI: 2014111003), the University of the Basque Country UPV-EHU (MOV17/14) and the Nordic Center of Excellence in Health-Related e-Sciences (NIASC).

References

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

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

Carreras, X., L. Márquez, and L. Padró. 2003. Learning a perceptron-based named entity chunker via online recognition feedback. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 156–159. Association for Computational Linguistics.

Crammer, K., M. Dredze, K. Ganchev, P. P. Talukdar, and S. Carroll. 2007. Automatic code assignment to medical text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136. Association for Computational Linguistics.

Dalianis, H., A. Henriksson, M. Kvist, S. Velupillai, and R. 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), J. Krogstie, G. Juel-Skielse and V. Kabilan, (Eds.), Stockholm, Sweden, June 11, 2015, CEUR, Vol-1381*, pages 1–18.

Demner-Fushman, D., W. W. Chapman, and C. J. McDonald. 2009. What can natural language processing do for clinical decision support? *Journal of biomedical informatics*, 42(5):760–772.

Ehrentraut, C., H. Tanushi, H. Dalianis, and J. Tiedemann. 2012. Detection of Hospital Acquired Infections in sparse and noisy Swedish patient records. In *Proceedings of the Sixth Workshop on Analytics for Noisy Unstructured Text Data.*

Faruqui, M. and S. Padó. 2010. Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. In *KONVENS*, pages 129–133.

Florian, R., A. Ittycheriah, H. Jing, and T. Zhang. 2003. Named entity recognition through classifier combination. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 168–171. Association for Computational Linguistics.

Haas, J. P., E. A. Mendonça, B. Ross, C. Friedman, and E. Larson. 2005. Use of computerized surveillance to detect nosocomial pneumonia in neonatal intensive care unit patients. *American journal of infection control*, 33(8):439–443.

Henriksson, A. 2015. *Ensembles of semantic spaces: On combining models of distributional semantics with applications in healthcare.* Ph.D. thesis, Department of Computer and Systems Sciences, Stockholm University.

Henriksson, A., M. Kvist, H. Dalianis, and M. Duneld. 2015. Identifying adverse drug event information in clinical notes with distributional semantic representations of context. *Journal of Biomedical Informatics*, 57:333–349.

Kang, N., Z. Afzal, B. Singh, E. M. Van Mulligen, and J. A. Kors. 2012. Using an ensemble system to improve concept extraction from clinical records. *Journal of biomedical informatics*, 45(3):423–428.

Lafferty, J. D., A. McCallum, and F. C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lai, K. H., M. Topaz, F. R. Goss, and L. Zhou. 2015. Automated misspelling detection and correction in clinical free-text records. *Journal of Biomedical Informatics*, 55(Supplement C):188–195.

Liang, P. 2005. *Semi-Supervised Learning for Natural Language*. Ph.D. thesis, Massachusetts Institute of Technology.

McNemar, Q. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Murphy, K. P. 2012. *Machine Learning: A Probabilistic Perspective*. MIT Press.

Oronoz, M., A. Casillas, K. Gojenola, and A. Perez. 2013. Automatic annotation of medical records in Spanish with disease, drug and substance names. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Springer, pages 536–543.

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

Pérez, A., K. Gojenola, A. Casillas, M. Oronoz, and A. D. a. de Ilarraza. 2015. Computer aided classification of diagnostic terms in Spanish. *Expert Systems with Applications*, 42:2949–2958.

Pérez, A., R. Weegar, A. Casillas, K. Gojenola, M. Oronoz, and H. Dalianis. 2017. Semi-supervised medical entity recognition: A study on Spanish and Swedish clinical corpora. *Journal of Biomedical Informatics*, 71:16–30.

Reynoso, G. A., A. D. March, C. M. Berra, R. P. Strobietto, M. Barani, M. Iubatti, M. P. Chiaradio, D. Serebrisky, A. Kahn, O. A. Vaccarezza, J. L. Leguiza, M. Ceitlin, D. A. Luna, F. G. B. de Quirós, M. I. Otegui, M. C. Puga, and M. Vallejos. 2000. Development of the Spanish Version of the Systematized Nomenclature of Medicine: Methodology and Main Issues. In *Proceedings of the AMIA Symposium*, pages 694–698.

Ruch, P., R. Baud, and A. Geissbühler. 2003. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. *Artificial intelligence in medicine*, 29(1):169–184.

Saha, S. and A. Ekbal. 2013. Combining multiple classifiers using vote based classifier ensemble technique for named entity recognition. *Data & Knowledge Engineering*, 85:15–39.

Skeppstedt, M., M. Kvist, G. Nilsson, and H. Dalianis. 2014. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. *Journal of Biomedical Informatics, 49*, pages 148–158.

Tang, B., H. Cao, X. Wang, Q. Chen, and H. Xu. 2014. Evaluating word representation features in biomedical named entity recognition tasks. *BioMed research international*, 2014.

Turian, J., L. Ratinov, and Y. 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. Association for Computational Linguistics.

Zea, J. L. C., J. E. O. Luna, C. Thorne, and G. Glavaš. 2016. Spanish NER with Word Representations and Conditional Random Fields. In *Proceedings of the Sixth Named Entity Workshop*, pages 34–40.