A Domain and Language Independent Named Entity Classification Approach Based on Profiles and Local Information

Isabel Moreno
Department of Software and Computing Systems, University of Alicante, Alicante, Spain
imoreno@dlsi.ua.es

María Teresa Romá-Ferri
Department of Nursing, University of Alicante, Alicante, Spain
mtr.ferri@ua.es

Paloma Moreda
Department of Software and Computing Systems, University of Alicante, Alicante, Spain
moreda@dlsi.ua.es

Abstract

This paper presents a Named Entity Classification system, which employs machine learning. Our methodology employs local entity information and profiles as feature set. All features are generated in an unsupervised manner. It is tested on two different data sets: (i) DrugSemantics Spanish corpus (Overall F1 = 74.92), whose results are in-line with the state of the art without employing external domain-specific resources. And, (ii) English CoNLL2003 dataset (Overall F1 = 81.40), although our results are slightly lower than previous work, these are reached without external knowledge or complex linguistic analysis. Last, using the same configuration for the two corpora, the difference of overall F1 is only 6.48 points (DrugSemantics = 74.92 versus CoNLL2003 = 81.40). Thus, this result supports our hypothesis that our approach is language and domain independent and does not require any external knowledge or complex linguistic analysis.

1 Introduction

The goal of Named Entity Recognition and Classification (NERC) is to recognize the occurrences of names in text, which is known as the recognition phase (NER), and assign them a category, which is referred as the classification phase (NEC). Both steps can be performed jointly or separately (Feldman and Sanger, 2007)[pp. 96–97].

NERC systems are fundamental to many text-processing applications. NERC not only is a prerequisite for many tasks, such as general language generation (Vicente and Lloret, 2016) or question answering (Marrero et al., 2013), but also a positive effect in performance has been reported when a NERC is included, as in the case of automatic text summarization (Alcón and Lloret, 2015).

Despite its proven usefulness, their usage is not always direct. Most NERC systems are typically focused on a specific domain, which has diverse requirements and, thus, different types of entities. As a result, these systems are designed ad hoc for a reduced set of predefined categories. If a NERC tool needs to be adapted to a new domain, with different constraints and a new set of entities, considerable effort is required (Marrero et al., 2013).

Furthermore, NERC not only is often domain conditioned, but also language dependent. Most systems are built for a specific corpus and, consequently, there is a dependence on such corpus. The adaptation of a NERC system to a new language is not always possible mainly due to three reasons. First, these systems often rely on linguistic analysis tools, which are not always available for all languages (Indurkhya, 2014). Second, these tools usually need resources which vary between languages (Marrero et al., 2013), if they exists. Lastly, each language poses distinct challenges that may affect the performance of NERC (Tjong Kim Sang, 2002; Sang and De Meulder, 2003).

Towards the advance of such issues, our final objective is to develop a general-purpose NERC system that consists of two separate modules for NER and NEC. To that end, this paper will focus on the development of our NEC module, assuming the output of a "perfect NER" so as to avoid any bias. The implemented NEC module is based on context information using profiles (Lopes and Vieira, 2015) and local information. This NEC approach can be used for different languages and domains.

Aiming to confirm such independence, this work is evaluated on two different corpora: a general-purpose English corpus, CoNLL2003 (Sang and De Meulder, 2003),

Proceedings of Recent Advances in Natural Language Processing, pages 510–518,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_067

and a pharmacotherapeutic Spanish corpus, DrugSemantics (Moreno et al., 2017a,b). The former represents general information needs, whereas the latter is related to specific information needs during pharmacotherapeutic day-to-day care. Both corpora are highly representative in terms of linguistic features as well as available NEs. Hence, these datasets allows us to define a language and domain independent evaluation scenario.

The rest of the paper is structured as follows. Section 2 reviews previous NERCs. Then, Section 3 defines our approach. Latter, the experiments set-up is described in Section 4. The evaluation is provided in Section 5, and Section 6 discusses our results. Last, Section 7 concludes the paper and outlines future work.

2 Background

For more than 20 years, several Natural Language Processing forums have promoted shared tasks to evaluate NERC systems. In these cases, research is focalized either in one domain and several languages (Tjong Kim Sang, 2002; Sang and De Meulder, 2003) or on one language and a narrow domain (Segura-Bedmar et al., 2013).

Concerning an example of the former, CoNLL conference hold two shared tasks (Tjong Kim Sang, 2002; Sang and De Meulder, 2003) to deal with NERC in news stories and several languages, namely English, Deutsch, Spanish and German. In both editions, systems obtained different results in each language, thus these NERCs are arguably fully language independent.

Carreras et al. (2002, 2003) obtained the best results in CoNLL 2002 (Spanish F1=81.39 and Dutch F1=77.05) but was ranked 5th on 2003 (NEglish F1=85 and German F1=69.15). This system performs NER and NEC sequentially with separate modules. Both components use Machine Learning (ML), specifically a binary AdaBoost classifier to ensemble small decision trees. Regarding its NEC module, it considers as features lexical (word forms, lemmas, their position and NE length) and orthographic (e.g. capitalization or affixes) information from context and the NE being classified, linguistic tags (such as POS and syntactic chunks) and external gazetteers.

The first place on CoNLL 2003 edition was for Florian et al. (2003) (English F1=88.76 and German F1=72.41). In this case, NER and NEC are addressed as one single task by means of a voting scheme. Specifically, diverse ML algorithms were combined: robust linear classifier, maximum entropy, transformation-based learning, and hidden Markov models. These algorithms take advantage of features of different nature: lexical information (word form in a window), orthographic information (such as affixes), together with linguistic features (e.g. POS tags or lemmas) and gazetteers.

More recently, Konkol et al. (2015) proposed a NERC in one step based on Conditional Random Fields (CRF) algorithm which employs unsupervised features from clusters of semantic spaces (COALS - Correlated Occurrence Analogue to Lexical Semantic - and HAL - Hyperspace Analogue to Language) as well as Latent Dirichlet allocation. This system obtained different results for each language (English F1=89.18, Spanish F1=82.74, Dutch F1=83.01 and Czech F1=74.08). Later, Agerri and Rigau (2016) built ixa-pipes, which tackles NER and NEC jointly, using CoNLL corpora. This tool learns Perceptron models from OpenNLP ML framework[1]. Their inferred model is based on local information (e.g. token, shape, n-grams, prefix and suffix), clusters (i.e. brown, word2vec and clark) and external knowledge (gazetteers). These systems also achieved different results for each language (Spanish F1=84.16, Dutch F1=85.04, English F1=91.36 and German F1=76.42).

Regarding an example of community challenge centered on one narrow domain and one language, the SemEval Workshop organized the DDIExtraction 2013 challenge (Segura-Bedmar et al., 2013). One of its main goals was NERC of drug names from English BioMedical Texts from two textual genres (DrugBank and MedLine abstracts). Most participants employed ML algorithms, specifically three proposals obtained the best results. One of them was the approach of Rocktäschel et al. (2013) who chose CRF algorithm. This strategy not only considers domain independent features (e.g. affixes, capitalization or tokens in a window), but also domain dependent ones such as domain-specific knowledge bases (ChEBI) or tools (ChempSpot). This system achieved the best results. However, the same configuration yield to different results in each genre (F1 STRICT, whole dataset=71.5; MedLine=58.1; DrugBank=87.8).

[1] https://opennlp.apache.org/ (last accessed: May 16th, 2017)

Other proposal was made by Grego and Couto (2013), who used five CRF models that had a domain independent feature set (stem, affixes and whether the token is a number or not). But this system also requires a domain-specific knowledge base (ChEBI) to perform lexical similarity, as well a set of post-processing rules, to obtained good but different results across genres (F1 STRICT, whole dataset=65.6; MedLine=56.7; DrugBank=77.1). Last, the proposal of Björne et al. (2013) selected TEES, which is based on Support Vector Machines (SVM) algorithm. Their features incorporated domain independent information (e.g. affixes), complex linguistic analysis (e.g. dependency chains) as well as information from a domain-specific resource (DrugBank) and a domain-specific tool (MetaMap) to reach adequate results which varied over genres (F1 STRICT, whole dataset=64.8; MedLine=52.2; DrugBank=78.1).

Outside that shared task, on the contrary, few NERC studies are specifically designed to be applied in several domains. Tkachenko and Simanovsky (2012) experimented on different textual genres from OntoNotes corpus and employed the CRF algorithm. CRF was feed with different features: lexical (e.g. tokens and bigrams), orthographic (e.g. hyphenation, shape or affixes), linguistic features (i.e. PoS tags), word clusters (i.e. Brown, Clark, Phrasal) and gazetteers from external sources. In the best case, this method achieved a F1 greater than 70%; while at worst, F1 is less than 50%. DINERS was proposed by Kitoogo and Baryamureeba (2008), who chose two corpora from journalism (CoNLL2003) and law domains. A Maximum entropy classifier was optimized via a genetic algorithm making use of: (i) gazetteers from external resources but also from the training data; (ii) orthographic information (e.g. prefixes, capitalization, presence of hyphens or digits or dollar sign); (iii) lexical information (word form, unigrams, bigrams); and (iv) linguistic information (PoS tags). Their proposal obtained a difference in terms of F1 of more than 20 points between overall law results (F1 = 92.04%) and global journalism performance (F1=70.27%).

In summary, all NERC systems presented here employed different levels of linguistic analysis (ranging from lexical to syntactical). Besides, all of them include certain semantic features to recognise and classify an entity. Such information is gathered from gazetteers, which are derived from external sources (Tkachenko and Simanovsky, 2012; Carreras et al., 2002, 2003; Florian et al., 2003; Kitoogo and Baryamureeba, 2008) or from training data (Kitoogo and Baryamureeba, 2008), from word clusters (Tkachenko and Simanovsky, 2012; Agerri and Rigau, 2016) or from distributional semantics methods (Konkol et al., 2015) or from domain-specific resources and tools (Rocktäschel et al., 2013; Grego and Couto, 2013; Björne et al., 2013). As a consequence of the usage of gazetteers, complex linguistic analysis and domain-specific resources, all these systems have shown that there is a performance gap between different languages and domains or textual genres. Furthermore, none of them analyzed the behavior of their systems changing simultaneously both language and domain or genre. Therefore, our hypothesis is that profile-based entity classification is effective using minimal linguistic information (lemmatizer and PoStagger) and out of external knowledge resources in any domain and language. To this end, our proposed approach is evaluated on two domains (general and pharmacotherapeutic) in different languages (Spanish and English).

3 Method: Named Entity Classification through Profiles

This NEC methodology is based on previous work (Lopes and Vieira, 2015), in which profiles where generated in an unsupervised manner for authorship detection. In addition to the purpose of the profiles usage, there are two main differences between their work and ours. On the one hand, Lopes and Vieira (2015) obtain profiles from a concept extractor system, while ours are derived from lemmas from nouns, verbs, adjectives and adverbs. On the other hand, categorization of Lopes and Vieira (2015) is performed by ranking possible entities using their own similarity measure, but ours calculates similarity between profiles and entities through a ML algorithm. This method consists of two main stages:

1. **Profile generation** process, whose main goal is to train a ML system to perform NEC, is an offline process that works as follows:

   (a) *Linguistic annotation*: a corpus previously annotated with NEs is tokenized, sentence-splitted, morphologically analysed and PoS-

tagged. In our case, Freeling (Padró and Stanilovsky, 2012) is used for Spanish whereas Treetagger (Schmid, 1994, 1995) is chosen for English.

(b) *Descriptors extraction*: For each entity instance, we extract descriptors[2] in a window and their frequency as the number of occurrences. The size of the window can be parametrized, but a window of a fixed length is used. In this work, length of the window was set to 10 lemmas (5 descriptors before and 5 after). Then, occurrences (occ) of each descriptor (d) are aggregated by entity type ($type$): $occ(d, type)$.

(c) *Split the training corpus*: For each NE type (e), the original training corpus is divided in two sets called target ($\mathcal{T}_e$) and contrasting ($\mathcal{G}_e$). The former set ($\mathcal{T}_e$) represents a fragment of the corpus capable to characterize that a given NE belongs to a certain class (i.e. positive examples of this NE type). While the latter ($\mathcal{G}_e$) is composed by a set of negative examples (i.e. examples of the remaining NE categories) aggregated by entity type.

(d) *Descriptors division*: For each NE type, the extracted descriptors are splitted in two list named unique (U_e) and common (C_e) descriptors list. The former (U_e) contains descriptors only present in the target set ($\mathcal{T}_e$), whereas the latter (C_e) includes common descriptors in both target ($\mathcal{T}_e$) and contrasting sets ($\mathcal{G}_e$) for a given entity type (e).

(e) *Relevance computation*: For both unique and common descriptors lists of each NE type, a relevance index is assigned to weight them and determine their importance for a given NE category. The *Term Frequency, disjoint corpora frequency* (TFDCF) index (Lopes and Vieira, 2015), defined in Equation 1, is applied to items from the unique descriptors list (U_e). Whereas the relevance common index, defined in Equation 2, is computed for each item in the common descriptors list (C_e) to penalize descriptors found in the contrasting set ($\mathcal{G}_e$) as well as in the target set ($\mathcal{T}_e$).

$$idx_{unique}(d, \mathcal{T}_e, \mathcal{G}_e) = \log(1 + \frac{occ(d, \mathcal{T}_e)}{\prod_{\forall g \in \mathcal{G}_e} 1 + \log(1 + occ(d, g))}) \quad (1)$$

$$idx_{common}(d, \mathcal{T}_e, \mathcal{G}_e) = \log(1 + occ(d, \mathcal{T}_e) - \frac{occ(d, \mathcal{T}_e)}{\prod_{\forall g \in \mathcal{G}_e} 1 + \log(1 + occ(d, g))}) \quad (2)$$

where d is a descriptor, $\mathcal{T}_e$ is the target set for an entity type e, g is a contrasting set, $\mathcal{G}_e$ contains all contrasting sets from the contrasting entities for an entity type e, and $occ(d, \mathcal{T}_e)$ is the occurrences of a term d in a set $\mathcal{T}_e$ for an entity type e.

This step produces a profile P_e for each entity type e. It would be constituted by its unique U_e and common C_e descriptors lists: $P_e = \{U_e, C_e\}$. In turn, each item from these lists is a pair $\{d,\ idx(d)\}$, where d represents a descriptor (i.e. term's lemma) and idx defines its relevance index. It is important to remember that relevance indexes are computed according to descriptors' occurrences extracted from both target $\mathcal{T}_e$ and contrasting $\mathcal{G}_e$. These lists only contain the most frequent descriptors. Specifically, this work employs up to 1000 most habitual descriptors in both lists.

(f) *Local features extraction*: Profiles are complemented with local information from NE itself, regardless of the category. Similarly, such data is obtained easily without requiring semantic or syntactic linguistic analysis or external knowledge. Concretely, three types of features are acquired from state-of-the-art NEC systems and extracted from training data: words of the entity[3](denoted as NE); entity length without stop-words (denoted as NElen); and affixes, distinguishing between suffixes and prefixes up to 4 characters from the first and last words (denoted as affix4).

(g) *Model training*: Our proposal creates a ML model for computing profile similarity between all NE classes, local NE features and its gold standard candidates. Thus, in this step, a multi-classification model is generated joining local features and profiles from all NE types. As a result, profiles are represented as follows: for each NE type, all descriptors' lemmas from the top list T_e are a feature, that has as value its relevance index, idx_{unique} (Equation 1). Similarly, all descriptors' lemmas from the common list

[2]Descriptors represent lemmas of content bearing terms that is nouns, verbs, adverbs and adjectives

[3]Please bear in mind that any special character is replaced with "_".

C_e of this NE type are a feature that has as value idx_{common} (Equation 2). Random Forest (RF) algorithm (Breiman, 2001) from Weka 3.6.7 (Hall et al., 2009) has been employed owing to the fact that it is able to deal with more than two classes. Moreover, its selection was motivated due to its fast training, its stability regarding data changes and its automatic variable selection. RF algorithm employs the default parameters, but he number of trees was set to 100.

2. **Profile application** process, whose aim is to classify a previously recognized NE in a set of predefined types, takes these steps:

(a) *Linguistic annotation*: text is tokenized, sentence-splitted, morphologically analyzed and PoS-tagged, as in the generation phase.

(b) *Candidate extraction*: these are gathered directly from the gold standard. It should be noted that candidates can be extracted by any NER module, but here the output of a "perfect NER" is used to avoid any bias.

(c) *Descriptors extraction*: For each candidate and each possible entity type, we extract descriptors that appear in a window using the same restrictions as in the generation phase.

(d) *Relevance computation*: The unique relevance index idx_{unique} (Equation 1) is computed for all candidates and all possible NE types.

(e) *Local features extraction*: For each candidate, local information is gathered (NE, NElen, affix4).

(f) *Similarity computation and classification*: Once an entity candidate has filled its profile and its local information, these data is compared against the ones generated from the training data, to compute their similarity. RF algorithm estimates similarity with a forest of trees that use as features local information as well as descriptors of all possible types of entities ($P = \{T, C\}$).

4 Datasets and Experimental Set-up

4.1 DrugSemantics Corpus and Set-up

DrugSemantics gold standard (Moreno et al., 2017a,b) is a collection of 5 Spanish Summaries of Product Characteristics (SPC) manually annotated, which contains 780 sentences and more than 2000 entities. This work uses the most frequent NEs from this gold standard: disease (724 entities), drug (657 entities) and unit of measurement (557 entities).

Evaluation uses 5-fold cross-validation at document-level (i.e. 4 SPCs to train and one to evaluate). It is a controlled environment that ensures unknown descriptors. In each fold (f, $F = 5$), the model is assessed for each entity (e, $E = 3$) in terms of traditional Precision (Pr), Recall (Re) and F-measure$_{\beta=1}$ (F1). Then, overall results for a fold are computed as the arithmetic-mean of all entities. Finally, the results of all iterations are averaged as the arithmetic-mean, thus obtaining Macro-averaged (M) figures for each entity e and globally. This decision is motivated to avoid any possible bias to the most frequent NE type.

4.2 English CoNLL2003 Corpus and Set-up

CoNLL2003 dataset (Sang and De Meulder, 2003) is a collection of English news stories from Reuters. It contains four entity types (person, organization, location and miscellaneous). However, miscellaneous is discarded because it has no practical application (Marrero et al., 2013). This corpus is divided in 3 sets: training (23499 entities), development (5942 entities) and testing (5648 entities). A ML model is inferred on the training set for all three NEs and this model is assessed on the test set. The development set is not used since no parameter tunning was done. The performance of the model is assessed for each entity e in terms of traditional Precision (Pr), Recall (Re) and F-measure$_{\beta=1}$ (F1). Last, overall *macro-averaged* results are calculated as the arithmetic-mean for all entities ($E = 3$) so as to avert a possible bias to the most frequent NE type.

5 Evaluation

The aim of our experiments is two fold. On one hand, to verify the appropriateness of our proposed method on two different domains and languages. On other hand, to find out the contribution of our local features (i.e. NE, NElen and affix4). For those reasons, first results on DrugSemantics corpus are shown (Section 5.1) and, then, performance on CoNLL English corpus are presented (Section 5.2). In both cases, local information is included gradually.

features		Pr	Re	F1
p	DR	50.60	47.40	48.39
	DI	62.34	76.14	67.41
	UM	56.42	44.86	49.29
	M	56.45	56.13	55.03
p + NElen	DR	57.17	52.04	53.50
	DI	67.61	75.47	70.19
	UM	56.52	56.29	55.07
	M	60.43	61.27	59.59
p + NE	DR	57.51	52.51	54.43
	DI	62.26	74.69	66.86
	UM	56.16	47.31	50.57
	M	58.65	58.17	57.29
p + affix4	DR	75.90	37.99	50.24
	DI	67.12	79.82	71.95
	UM	51.30	69.88	56.01
	M	64.77	62.56	59.40
p + NElen + NE	DR	59.40	52.07	54.94
	DI	65.89	80.81	71.70
	UM	63.02	55.51	58.30
	M	62.77	62.80	61.65
p + NE + affix4	DR	79.01	51.85	61.91
	DI	78.37	82.58	79.86
	UM	62.50	86.44	71.47
	M	73.29	73.63	71.08
p + NElen + affix4	DR	71.37	40.90	51.08
	DI	73.68	81.50	76.10
	UM	54.34	75.89	59.87
	M	66.47	66.09	62.35
p + NElen + NE + affix4	DR	81.95	57.90	66.97
	DI	82.28	84.00	82.62
	UM	66.52	89.22	75.17
	M	**76.92**	**77.04**	**74.92**

Note: (i) p: profile; (ii) NE: entity words; (iii) NElen: entity length without stopwords; (iv) affix4: entity suffixes and prefixs up to 4 characters; (v) DR: Drug; (vi) DI: Disease; (vii) UM: Unit of Measurement; and (viii) M: Overall Macro-average results.

Table 1: DrugSemantics Precision (Pr), Recall (Re) and $F_{\beta=1}$ (F1) results with different features

5.1 DrugSemantics Results

Table 1 collects overall results and results for each entity type. Disease is the entity type that always obtains the higher results for all measures. All types of entities, but especially Drug and UnitOfMeasurement, perform better when all local information is integrated in our pipeline. In fact, overall MF1 results improve 36.14% (19.89 points) if all local features are combined with context profiles.

5.2 CoNLL2003 Results

Table 2 collects overall results and results for each entity type. All classes, but especially Location, perform better when all local information is integrated in our pipeline. But, overall MF1 increases 50.49% (27.31 points) when profiles are combined with all local features.

6 Discussion

In view of the results, our NEC approach has demonstrated to be appropriated when combining profiles and local information. Using the same configuration, pharmacotherapuetic domain overall reaches almost 75% (MF1=74.92%), whereas general domain obtains an overall MF1 of 81.40%. In spite of this, our NEC system proves to be language as well as domain independent since a small difference is achieved in terms of global MF1 for the same configuration (RF 100 trees, profiles and local information), namely it is 6.48 (MF1: DrugSemantics 74.92%-CoNLL2003 81.40%).

A comparison between our system and NERCs presented in Section 2, it is not free of certain limitations. The corpora is different and, consequently, entities, domain and language also differs. Also, performance is assessed in various ways[4]. Besides, these systems do not provide results for NEC task alone. Still, such comparison is made in two steps. On the one hand, the appropriateness of our approach is compared to determine the extent of our contribution with systems trained on DrugDDI corpus, CoNLL datasets, and NERCs declared domain independent. On the other hand, a comparison in terms of absolute overall F1 difference between best and worst reported resultsis provided in order to prove that such difference is in line with them or lower. This data is summarized in Table 3 as follows: (i) for systems trained

[4]For example, CoNLL reported micro-averages; while this paper, macro-averages.

features		Pr	Re	F1
p	PER	55.91	43.84	49.15
	LOC	51.99	58.10	54.88
	ORG	55.82	60.92	58.26
	M	54.57	54.29	54.09
p + NElen	PER	62.61	51.55	56.55
	LOC	57.83	63.15	60.37
	ORG	57.62	62.19	59.82
	M	59.36	58.96	58.91
p + NE	PER	69.98	44.65	54.52
	LOC	61.06	80.55	69.46
	ORG	66.88	68.91	67.88
	M	65.97	64.70	63.95
p + affix4	PER	82.97	65.73	73.35
	LOC	80.65	75.57	78.03
	ORG	67.16	85.00	75.03
	M	76.93	75.43	75.47
p + NElen + NE	PER	76.07	54.35	63.40
	LOC	67.33	82.65	74.21
	ORG	66.19	69.39	67.75
	M	69.86	68.80	68.45
p + NE + affix4	PER	89.14	66.85	76.40
	LOC	81.04	80.17	80.60
	ORG	71.29	89.23	79.26
	M	80.49	78.75	78.75
p + NElen + affix4	PER	84.59	68.59	75.76
	LOC	85.15	76.77	80.74
	ORG	67.37	86.45	75.73
	M	79.04	77.27	77.41
p + NElen + NE + affix4	PER	91.99	68.53	78.55
	LOC	85.79	86.25	86.02
	ORG	71.86	89.29	79.63
	M	**83.21**	**81.36**	**81.40**

Note: (i) p: profile; (ii) NE: entity words; (iii) NElen: entity length without stopwords; (iv) affix4: entity suffixes and prefixs up to 4 characters; (v) PER: Person; (vi) LOC: Location; (vii) ORG: Organization; and (viii) M: Overall Macro-average results

Table 2: CoNLL2003 Precision (Pr), Recall (Re) and $F_{\beta=1}$ (F1) results

D	System	F1	Dif
P	Björne et al. (2013)	64.8	25.9
	Grego and Couto (2013)	65.6	20
	Rocktäschel et al. (2013)	71.5	29.7
	Profiles+local	**74.92**	**6.48**
G	**Profiles+local**	**81.40**	**6.48**
	Freeling	85	12.24
	Florian et al. (2003)	88.76	16.35
	Konkol et al. (2015)	89.19	15.10
	ixa-pipes	91.36	14.94
I	TkaSim	-	> 25
	DINERS	70.27	> 20

Note: (i) Dif: Absolute difference of overall F1 between best and worst results; (ii) D: Domain; (iii) G: general; (iv) P: Pharmacotherapeutic; (v) I: domain independent systems.

Table 3: Comparison with existing NERC in terms of F1 and difference between corpora. For each genre, systems are ranked by F1

on DrugDDI, the difference is computed between genres using STRICT measure, whose results are higher than the macro-average (MAVG) reported. And (ii) Whereas for systems trained on some CoNLL corpora, the difference is calculated from the best and the worst language from the results obtained, which are micro-averaged.

Regarding adequateness of our NEC, our system achieved the best results in the pharmacotherapeutic domain without external knowledge or complex linguistic analysis. On the contrary, it can be observed that although our results are slightly lower than the remaining systems trained on the CoNLL2003 corpus (F1 column in Table 3), it should be noted that these are obtained without using external knowledge or complex linguistic analysis.

Regarding domain and language independence (Dif column in Table 3), our profile-based NEC systems obtains the smaller difference (6.48 points) when compared to any of these systems. Furthermore, the two systems declared domain independent, TkaSim (Tkachenko and Simanovsky, 2012) and DINERS (Kitoogo and Baryamureeba, 2008), exhibit the greatest difference when domain or textual genre changes.

7 Conclusions and Future Work

This paper presented a Named Entity Classification system based on profiles and local informa-

tion. This proposal does not require external resources or complex linguistic analysis. It only needs an annotated corpus with the target NEs and a basic linguistic analyzer able to split sentences and tokens, lemmatize and POS tag this new language.

The evaluation has involved the adaptation of our approach to two different corpora: a Spanish pharmacotherapeutic (corpus MF1: 74.92%) and an English corpus from general domain (MF1: 81.40%). The results are in-line with the state of the art for the narrow domain. But, although our results are encouraging, classification needs to be improved to increase the general domain results. Nevertheless, the difference between the two corpora, employing the same configuration, shows a difference of only 6.48 points. This small change between languages and domains supports our hypothesis: profile-based entity classification is effective using minimal linguistic information (lemmatizer and PoStagger) and out of external knowledge resources in any domain and language.

As future work, we plan to enhance profiles with other features derived automatically from training data. We also consider the generation of profiles using only tokens (instead of lemmas) in order to remove the basic linguistic analyzer requirement. Additionally, aiming at reinforcing our hypothesis, our approach will be evaluated on other available corpora in other language-domain pairs, such as an English pharmacotherapeutic corpus.

Acknowledgements

This research work has been partially funded by the Spanish Government, Generalitat Valenciana, University of Alicante and Ayudas Fundación BBVA a equipos de investigación científica 2016 through the projects TIN2015-65100-R, TIN2015-65136-C2-2-R, PROMETEOII/2014/001, GRE16-01: "Plataforma inteligente para recuperación, análisis y representación de la información generada por usuarios en Internet" and "Análisis de Sentimientos Aplicado a la Prevención del Suicidio en las Redes Sociales" (ASAP).

References

Rodrigo Agerri and German Rigau. 2016. Robust multilingual Named Entity Recognition with shallow semi-supervised features. *Artificial Intelligence* 238:63–82. https://doi.org/10.1016/j.artint.2016.05.003.

Óscar Alcón and Elena Lloret. 2015. Estudio de la influencia de incorporar conocimiento léxico-semántico a la técnica de Análisis de Componentes Principales para la generación de resúmenes multilingües [Studying the influence of adding lexical-semantic knowledge to Principal Component Analysis technique for multilingual summarization]. *Linguamática* 7(1):43–53.

Jari Björne, Suwisa Kaewphan, and Tapio Salakoski. 2013. UTurku : Drug Named Entity Recognition and Drug-Drug Interaction Extraction Using SVM Classification and Domain Knowledge. In *Proceedings of the Seventh International Workshop on Semantic Evaluation*. volume 2, pages 651–659.

L. Breiman. 2001. Random Forests. *Machine Learning* 45(1):5–32. https://doi.org/10.1023/A:1010933404324.

X Carreras, L Marquez, and L Padró. 2002. Named entity extraction using adaboost. In *Proceeding of the 6th Conference on Natural Language Learning*. pages 152–155.

Xavier Carreras, Lluís Màrquez, and Lluís Padró. 2003. A simple named entity extractor using AdaBoost. In *Proceedings of the 7th Conference on Natural Language Learning*. pages 152–155.

R. Feldman and J. Sanger. 2007. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press, New York.

Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang. 2003. Named Entity Recognition through Classifier Combination. In *Proceedings of the 7th Conference on Natural Language Learning*. pages 168–171.

Tiago Grego and Francisco M Couto. 2013. LASIGE : using Conditional Random Fields and ChEBI ontology. In *Proceedings of the Seventh International Workshop on Semantic Evaluation*. volume 2, pages 660–666.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations* 11(1):10–18.

Nitin Indurkhya. 2014. Natural Language Processing. In Teofilo Gonzalez, Jorge Díaz-Herrera, and Allen Tucker, editors, *Computing Handbook, Third Edition: Computer Science and Software Engineering*, CRC Press, chapter 40, pages 40:1–17.

FE Kitoogo and Venansius Baryamureeba. 2008. Towards domain independent named entity recognition. In Janet Aisbett, Gibbon Greg, Anthony J. Rodriguez, Joseph Kizza Migga, Ravi Nath, and Gerald R Renardel, editors, *Strengthening the Role of of ICT in Development*, Fountain publishers, volume IV, chapter 5, pages 84 – 95.

Michal Konkol, T. Brychcín, Konopí, and Miloslav K. 2015. Latent semantics in Named Entity Recognition. *Expert Systems with Applications* 42(7):3470–3479. https://doi.org/10.1016/j.eswa.2014.12.015.

Lucelene Lopes and Renata Vieira. 2015. Building and Applying Profiles Through Term Extraction. In *X Brazilian Symposium in Information and Human Language Technology*. Natal, Brazil, pages 91–100.

Mónica Marrero, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato, and Juan Miguel Gómez-Berbís. 2013. Named Entity Recognition: Fallacies, challenges and opportunities. *Computer Standards and Interfaces* 35(5):482–489. https://doi.org/10.1016/j.csi.2012.09.004.

Isabel Moreno, Ester Boldrini, Paloma Moreda, and María Teresa Romá-Ferri. 2017a. DrugSemantics: A corpus for Named Entity Recognition in Spanish Summaries of Product Characteristics. *Journal of Biomedical Informatics* 72:8 – 22. https://doi.org/10.1016/j.jbi.2017.06.013.

Isabel Moreno, Ester Boldrini, Paloma Moreda, and María Teresa Romá-Ferri. 2017b. DrugSemantics Gold Standard. Mendeley Data, v1. https://doi.org/10.17632/fwc7jrc5jr.1.

Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference*. ELRA, Istanbul, Turkey.

Tim Rocktäschel, Torsten Huber, Unter Den Linden, and Tim Rockt. 2013. WBI-NER : The impact of domain-specific features on the performance of identifying and classifying mentions of drugs. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. pages 356–363.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the 7th Conference on Natural Language Learning*. pages 142–147.

H Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*. pages 44–49.

Helmut Schmid. 1995. Improvements In Part-of-Speech Tagging With an Application To German. In *Proceedings of the ACL SIGDAT-Workshop*. pages 47—-50.

I Segura-Bedmar, P Martínez, and M Herrero-Zazo. 2013. SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). In *Proceedings of the Seventh International Workshop on Semantic Evaluation*. pages 341–350.

Erik F Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task. In *Proceedings of the 6th Conference on Natural Language Learning*. pages 1–4.

M Tkachenko and A Simanovsky. 2012. Selecting Features for Domain-Independent Named Entity Recognition. In *Proceedings of KONVENS 2012*. pages 248–253.

Marta Vicente and Elena Lloret. 2016. Exploring Flexibility in Natural Language Generation throughout Discursive Analysis of New Textual Genres. In *Proceedings of the 2nd International Workshop Future and Emerging Trends in Language Technologies, Machine Learning and Big Data*. Sevilla, Spain.