

Robust Multilingual Named Entity Recognition with Shallow Semi-supervised Features (Extended Abstract)*

Rodrigo Agerri and German Rigau

IXA NLP Group, University of the Basque Country UPV/EHU, Donostia-San Sebastián
 {rodrigo.agerri,german.rigau}@ehu.eus

Abstract

We present a multilingual Named Entity Recognition approach based on a robust and general set of features across languages and datasets. Our system combines shallow local information with clustering semi-supervised features induced on large amounts of unlabeled text. Understanding via empirical experimentation how to effectively *combine* various types of *clustering features* allows us to seamlessly export our system to other datasets and languages. The result is a simple but highly competitive system which obtains state of the art results across five languages and twelve datasets. The results are reported on standard shared task evaluation data such as CoNLL for English, Spanish and Dutch. Furthermore, and despite the lack of linguistically motivated features, we also report best results for languages such as Basque and German. In addition, we demonstrate that our method also obtains very competitive results even when the amount of supervised data is cut by half, alleviating the dependency on manually annotated data. Finally, the results show that our emphasis on clustering features is crucial to develop robust out-of-domain models. The system and models are freely available to facilitate its use and guarantee the reproducibility of results.

1 Introduction

A named entity can be mentioned using a great variety of surface forms (Barack Obama, President Obama, Mr. Obama, B. Obama, etc.) and the same surface form can refer to a variety of named entities. For example, according to the English Wikipedia, the form ‘Europe’ can ambiguously be used to refer to 18 different entities, including the continent, the European Union, various Greek mythological entities, a rock band, some music albums, a magazine, a short story, etc.¹

*This paper is an extended abstract of the Artificial Intelligence Journal publication [Agerri and Rigau, 2016], <http://dx.doi.org/10.1016/j.artint.2016.05.003>

¹<http://en.wikipedia.org/wiki/Europe>

Most Named Entity Recognition (NER) taggers are supervised statistical systems that extract patterns and term features which are considered to be indications of Named Entity (NE) types using the manually annotated training data (extracting orthographic, linguistic and other types of evidence) and often external knowledge resources. As in other NLP tasks, supervised statistical NER systems are more robust and obtain better performance on available evaluation sets, although sometimes the statistical models can also be combined with specific rules for some NE types. For best performance, supervised statistical approaches require manually annotated training data, which is both expensive and time-consuming. This has seriously hindered the development of robust high performing NER systems for many languages but also for other domains and text genres in what we will henceforth call ‘out-of-domain’ evaluations.

Moreover, supervised NER systems often require fine-tuning for each language and, as some of the features require language-specific knowledge, this poses yet an extra complication for the development of robust multilingual NER systems. For example, it is well-known that in German every noun is capitalized and that compounds including named entities are pervasive. This also applies to agglutinative languages such as Basque, Korean, Finnish, Japanese, Hungarian or Turkish. For this type of languages, it had usually been assumed that linguistic features (typically Part of Speech (POS) and lemmas, but also semantic features based on WordNet, for example) and perhaps specific hand-crafted rules, were a necessary condition for good NER performance as they would allow to capture better the most recurrent declensions (cases) of named entities for Basque [Alegria *et al.*, 2006] or to address problems such as sparsity and capitalization of every noun for German [Faruqui *et al.*, 2010; Benikova *et al.*, 2014; 2015]. This language dependency was easy to see in the CoNLL 2002 and 2003 tasks, in which systems participating in the two available languages for each edition obtained in general different results for each language. This suggests that without fine-tuning for each corpus and language, the systems did not generalize well across languages [Nothman *et al.*, 2013].

This paper presents a multilingual and robust NER system based on simple, general and shallow features that heavily relies on word representation features for high performance. Even though we do not use linguistic motivated fea-

tures, our approach also works well for inflected languages such as Basque and German. We demonstrate the robustness of our approach by reporting best results for five languages (Basque, Dutch, German, English and Spanish) on 12 different datasets, including seven in-domain and eight out-of-domain evaluations.

2 Datasets

The CoNLL NER shared tasks focused on language independent machine learning approaches for 4 entity types: *person*, *location*, *organization* and *miscellaneous* entities [Tjong Kim Sang, 2002]. The 2002 edition provided manually annotated data in Dutch and Spanish whereas in 2003 the languages were German and English. In addition to the CoNLL data, for English we also use the formal run of MUC 7 and Wikigold for out-of-domain evaluation [Chinchor and Marsh, 1998; Nothman *et al.*, 2013]. The Wikigold corpus consists of 39K words of English Wikipedia manually annotated following the CoNLL 2003 guidelines. For Spanish and Dutch, we also use Ancora 2.0 [Taulé *et al.*, 2008] and SONAR-1 [Desmet and Hoste, 2014], respectively. In Basque the only gold standard corpus is Egunkaria [Alegria *et al.*, 2006]. In the datasets mentioned so far, named entities were assumed to be non-recursive and non-overlapping. The exceptions are the GermEval 2014 shared task data for German and MEANTIME, where nested entities are also annotated (both inner and outer spans). The GermEval 2014 NER shared task [Benikova *et al.*, 2014] annotated 12 NE types: *person*, *location*, *organization*, *other* plus their sub-types annotated at their inner and outer levels. Finally, the MEANTIME corpus [Minard *et al.*, 2016] is a multilingual (Dutch, English, Italian and Spanish) publicly available evaluation set annotated within the Newsreader project².

3 Related Work

Named entity recognition is a task with a long history in NLP. Therefore, we will summarize those approaches that are most relevant to our work. Since CoNLL shared tasks, the most competitive approaches have been supervised systems learning CRF, SVM, Maximum Entropy or Averaged Perceptron models, although the most recent approaches are based on deep-learning, [Lample *et al.*, 2016; Ma and Hovy, 2016]. In any case, while the machine learning method is important, it has also been demonstrated that good performance might largely be due to the feature set used [Clark and Curran, 2003]. Furthermore, and as argued by the CoNLL 2003 organizers, no feature set was deemed to be ideal for NER, although many approaches for English.

Linguistic features (POS, lemmas, chunks, but also specific rules or patterns) have been widely used, although these type of features was deemed to be most important for German, [Benikova *et al.*, 2014]. This is caused by the sparsity caused by the declension cases, the tendency to form compounds containing named entities and by the capitalization of every noun [Faruqui *et al.*, 2010]. For example, the best system among the 11 participants in GermEval 2014, ExB,

uses morphological features and specific suffix lists aimed at capturing frequent patterns in the endings of named entities [Hänig *et al.*, 2014]. In agglutinative languages such as Basque, which contains declension cases for named entities, linguistic features are considered to be a requirement. For example, the country name ‘Espainia’ (Spain in Basque) can occur in several forms, *Espainian*, *Espainiera*, *Espainiak*, *Espainiarentzat*, *Espainiako*, and many more.³ The only previous work for Basque developed Eihera, a rule-based NER system formalized as finite state transducers to take into account declension classes [Alegria *et al.*, 2006]. The features of Eihera include word, lemma, POS, declension case, capitalized lemma, etc. These features are complemented with semantic information from the Basque WordNet.

Global features have also been proposed: [Carreras *et al.*, 2002] used a function to inject features for the whole sentence whereas [Ratinov and Roth, 2009] developed three types of non-local features, analyzing global dependencies in a window of between 200 and 1000 tokens.

Recently, semi-supervised approaches leveraging unlabeled text had already been applied to improve results in various NLP tasks, including NER. [Ratinov and Roth, 2009] used Brown clusters, [Turian *et al.*, 2010] made a rather exhaustive comparison of word embeddings and clustering methods. They show that in some cases the combination of word representation features was positive but they did not manage to improve over the state of the art. Furthermore, they reported that Brown clustering features performed better than the word embeddings. [Passos *et al.*, 2014] extend the Skip-gram algorithm to learn 50-dimensional lexicon infused phrase embeddings from 22 different gazetteers and the Wikipedia and obtained 90.90 F1 score on the CoNLL 2003 English test set. [Faruqui *et al.*, 2010] trained the Stanford NER system with a variety of features, including lemma, POS tag, etc, including features based on Clark clusters [Clark, 2003] improving the results over the best CoNLL 2003 system by 4 points in F1. Finally, state of the art results have been obtained by deep learning approaches based on LSTMs and CNNs [Lample *et al.*, 2016; Ma and Hovy, 2016] and trained with word embeddings (91.21 F1 in CoNLL 2003 benchmark).

4 System Description

The design of *ixa-pipe-nerc* aims at establishing a simple and shallow feature set, avoiding any linguistic motivated features, with the objective of removing any reliance on costly extra gold annotations (POS tags, lemmas, syntax, semantics) and/or cascading errors if automatic language processors are used. The underlying motivation is to obtain robust models to facilitate the development of NER systems for other languages and datasets/domains while obtaining state of the art results. Our system consists of: (i) Local, shallow features based mostly on orthographic, word shape and n-gram features plus their context; (ii) three types of simple clustering features, based on unigram matching; (iii) publicly available gazetteers. *ixa-pipe-nerc* learns supervised models via

³English: in Spain, to Spain, Spain (in transitive clause), for Spain, in Spain.

²<http://www.newsreader-project.eu>

Features	P	R	F1
Local (L)	87.27	81.32	84.19
L + Brown reuters (BR)	90.28	86.79	88.50
L + Clark wiki 600 (CW600)	90.85	87.16	88.97
L + Word2vec giga 200 (W2VG200)	89.64	85.06	87.29
BR+CR600-CW600+W2VG200+dict	92.20	90.19	91.18
charngram 1:6 + en-91-18	92.16	90.56	91.36
Illinois NER	-	-	90.57
Turian et al. (2010)	90.10	90.61	90.36
Passos et al. (2014)	-	-	90.90
Lample et al. (2016)	-	-	90.94
Ma and Hovy (2016)	-	-	91.21

Table 1: CoNLL 2003 testb English results.

the Perceptron algorithm as described by [Collins, 2002]. To avoid duplication of efforts, *ixa-pipe-nerc* uses the Apache OpenNLP project implementation of the Perceptron algorithm⁴ customized with its own features. Specifically, *ixa-pipe-nerc* implements, on top of the local features, a combination of word representation features: (i) Brown [Brown *et al.*, 1992] clusters, taking the 4th, 8th, 12th and 20th node in the path; (ii) Clark [Clark, 2003] clusters and, (iii) Word2vec [Mikolov *et al.*, 2013] clusters, based on K-means applied over the extracted word vectors using the skip-gram algorithm. The implementation of the clustering features looks for the cluster class of the incoming token in one or more of the clustering lexicons induced following the three methods listed above. If found, then we add the class as a feature. The Brown clusters only apply to the token related features, which are duplicated.

So far the most successful approaches have only used one type of word representation. However, our simple baseline combined with one type of word representation features are not able to compete with previous, more complex, systems. Thus, instead of encoding more elaborate features, we have devised a simple method to *combine* and *stack* various types of clustering features induced over different data sources or corpora. In principle, our method can be used with any type of word representations. However, for comparison purposes, we decided to use word representations previously used in successful NER approaches: Brown, Word2vec, and Clark clusters. Our clustering features are placed in a 5 token window. The word representations obtained by different algorithms would capture different distributional properties of words in a given corpus or data source. Therefore, each type of clustering would allow us to capture different types of occurring named entity types. In other words, *combining* and *stacking* different types of clustering features induced over a variety of data sources should help to capture more similarities between different words in the training and test sets.

5 Experimental Results

The experiments are performed in 5 languages: Basque, Dutch, English, German and Spanish. The clustering features are obtained by processing the following clusters from publicly available corpora (described in detail in the origi-

⁴<http://opennlp.apache.org/>

Features	P	R	F1
Local	70.52	60.27	65.00
L + Brown eg (BE)	74.54	67.59	70.90
L + Clark eg 200 (CE200)	76.76	68.92	72.63
L + Clark wiki 200 (CW200)	75.57	65.60	70.23
L + Word2vec eg 300 (W2VE300)	74.04	62.71	67.91
L + Word2vec be 600 (W2WB600)	74.11	64.82	69.15
BE+C(EW)200+ W2V(E300+B600)	80.66	70.78	75.40
Alegria et al. (2006)	72.50	70.24	71.35

Table 2: Basque Egunkaria results.

nal paper): (i) 1000 Brown clusters; (ii) Clark and Word2vec clusters in the 100-600 range. To choose the best combination of clustering features we test the available permutations of Clark and Word2vec clusters with and without the Brown clusters on the development data. Table 1 shows the results obtained for English on CoNLL 2003 whereas Table 2 shows the results for Basque. The rest of the in-domain evaluations can be consulted in the original paper.

The *stacking* and *combining* effect of our clustering features manifests very clearly when we compare the single clustering feature models (BR, CW600, W2VG200 and W2VW400) with the best models which combine multiple clustering features from various data sources. The results for Basque show that our model clearly improves upon previous work by 4 points in F1 measure (75.40 vs 71.35). These results are particularly interesting as it had been so far assumed that complex linguistic features and language-specific rules were required to perform well for agglutinative languages such as Basque [Alegria *et al.*, 2006]. Finally, it is worth noting that our results increase the overall F1 score by 11.72 over the baseline, of which 10 points are gained in precision and 13 in terms of recall.

5.1 Out-of-domain Evaluations

NER systems are often used in out-of-domain settings, namely, to annotate data that greatly differs from the data from which the NER models were learned. These differences can be of text genre and/or domain, but also because the assumptions of what constitutes a named entity might differ. It is therefore interesting to develop robust NER systems across both domains and datasets. In this section we demonstrate that our approach, consisting of basic, general local features and the *combination* and *stacking* of clusters, produces robust NER systems in three out-of-domain evaluation settings: (i) Class disagreements: Named entities are assigned to different classes in training and test; (ii) Different text genre: The text genre of training and test data differs; (iii) Annotation guidelines: The gold annotation of the test data follows different guidelines from the training data. This is usually reflected in different named entity spans. Due to lack of space, in this paper we will only show the third type of out-of-domain evaluation, the other two can be consulted in the journal paper.

In this section the objective is studying not so much the differences in textual genre as the influence of substantially different annotation standards. We only use three classes (location, organization and person) to evaluate the best models presented for in-domain evaluations labeling ‘O’ every

Features	English				Spanish				Dutch			
	Outer		Inner		Outer		Inner		Outer		Inner	
	F1	T-F1										
Local	41.83	54.17	48.57	57.85	34.42	42.95	37.14	41.93	48.49	54.84	49.77	55.86
best-cluster	54.04	65.96	63.72	71.13	56.78	62.55	59.77	63.04	59.94	66.03	60.27	65.42
best-overall	55.48	67.36	64.95	71.98	58.94	65.63	62.14	65.54	63.40	70.68	63.93	70.24
Stanford NER	53.14	64.62	62.45	69.76	46.42	54.40	47.48	54.27	-	-	-	-
Illinois NER	53.24	65.68	62.72	71.04	-	-	-	-	-	-	-	-
Freeling 3.1	-	-	-	-	38.27	48.06	40.93	46.52	-	-	-	-
Sonar nerd	-	-	-	-	-	-	-	-	48.60	53.60	48.44	52.79

Table 3: MEANTIME out-of-domain evaluation. English systems trained on CoNLL data. Dutch systems trained with SONAR-1. Stanford NER Spanish model is trained with Ancora (20150130 version) whereas ixa-pipe-nerc is trained with CoNLL data. T-F1: token-based F1. *Local*: baseline system; *best-clusters*: nl-clusters, es-cluster and en-best-cluster; *best-overall*: best configuration previously presented for each language for the in-domain evaluations.

entity which is not LOC, ORG or PER. The text genre of MEANTIME is not that different from CoNLL data. However, differences in the gold standard annotation result in significant disagreements regarding the span of the named entities. For example, the following issues are markedly different with respect to the training data we use for each language: (i) Different criteria to decide when a named entity is annotated: in the expression “40 billion US air tanker contract” the MEANTIME gold standard does not mark ‘US’ as location, whereas in the training data this is systematically annotated; (ii) mentions including the definite article within the name entity span: ‘the United States’ versus ‘United States’, and (iii) common nouns modifying the proper name: ‘Spokeswoman Sandy Angers’ is annotated as a named entity of type PER whereas in the training data used the span of the named entity would usually be ‘Sandy Angers’. CoNLL NER phrase based evaluation punishes any bracketing error as both false positive and negative. Thus, these span-related disagreements make this setting extremely hard for models trained according to other annotation guidelines, as shown by Table 3. Our baseline models degrade around 40 F1 points and the cluster-based models around 35. Other systems’ results worsen much more, especially for Spanish and Dutch. The token-based scores are in general better but the proportion in performance between systems across languages is similar.

6 Concluding Remarks

We have shown how to develop robust NER systems across languages and datasets with minimal human intervention, even for languages with inflected named entities. This is based on adequately combining word representation features on top of shallow and general local features. Crucially, we have empirically demonstrate how to effectively combine various types of simple word representation features depending on the source data available. Despite the relative simplicity of our approach, we report state of the art results for Dutch, English, German, Spanish and Basque in seven in-domain evaluations. We also outperform previous work in eight out-of-domain evaluations, showing that our clustering features improve the robustness of NER systems across datasets. Finally, we have measured how much our system’s performance degrades when the amount of supervised data is drastically cut. This, together with the lack of linguistic features, facilitates

the easy and fast development of NER systems for new domains or languages.

Previous approaches to NER combining clusters or word embeddings have obtained mixed results [Turian *et al.*, 2010]. Up until now best results have been based on rather complex systems which also used one type clustering or embedding feature [Passos *et al.*, 2014; Ratinov and Roth, 2009; Faruqui *et al.*, 2010; Benikova *et al.*, 2014]. Our system displays two important differences with respect to previous approaches. First, the differences between our baseline system and the, for example, Clark features are much larger than in previous work (with the exception of [Faruqui *et al.*, 2010]), ranging from 2.2 and 5.5 points in F1 measure across the in-domain evaluations to 2-8 points for out-of-domain results. If we consider the combined clustering models, the differences over the baseline increase to 5-10 points of F1 measure for in-domain evaluations and between 4-22 in out-of-domain settings. Second, our combination of clustering features significantly increases the performance over the models using only one type of clustering feature. The improvements range over 2 to 6 points in F1 measure for in-domain and out-of-domain results. In our opinion, these results are quite interesting as previous experiments combining features of different word representations for NER [Turian *et al.*, 2010], while increasing the overall result, did not improve over the state of the art at the time [Ratinov and Roth, 2009]. The results also show that leaning heavily on the clustering features (instead of specific feature tuning) for performance proves very beneficial in out-of-domain settings.

Acknowledgments

This work has been supported by the European projects NewsReader, EC/FP7/316404 and QTLeap - EC/FP7/610516, and by the Spanish Ministry for Science and Innovation (MICINN) SKATER, Grant No. TIN2012-38584-C06-01 and TUNER, TIN2015-65308-C5-1-R.

References

[Alegria *et al.*, 2006] Iñaki Alegria, Olatz Arregi, Nerea Ezeiza, and Izaskun Fernández. Lessons from the development of a named entity recognizer for Basque. *Procesamiento del lenguaje natural*, 36:25–37, 2006.

- [Benikova *et al.*, 2014] Darina Benikova, Chris Biemann, Max Kisselew, and Sebastian Padó. Germeval 2014 named entity recognition shared task: Companion paper. In *Proceedings of the GermEval 2014 Named Entity Recognition Shared Task, KONVENS 2014*, pages 104–113, 2014.
- [Benikova *et al.*, 2015] Darina Benikova, Seid Muhie Yimam, Prabhakaran Santhanam, and Chris Biemann. Germaner: Free open german named entity recognition tool. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology (GSCL-2015)*, pages 31–38, 2015.
- [Brown *et al.*, 1992] Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479, 1992.
- [Carreras *et al.*, 2002] Xavier Carreras, Lluís Marquez, and Lluís Padro. Named entity extraction using AdaBoost. In *Proceedings of the 6th conference on Natural language learning-Volume 20*, pages 1–4, 2002.
- [Chinchor and Marsh, 1998] Nancy Chinchor and Elaine Marsh. Muc-7 information extraction task definition. In *Proceeding of the Seventh Message Understanding Conference (MUC-7), Appendices*, 1998.
- [Clark and Curran, 2003] Stephen Clark and James Curran. Language Independent NER using a Maximum Entropy Tagger. In *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-03)*, pages 164–167, Edmonton, Canada, 2003.
- [Clark, 2003] Alexander Clark. 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, 2003.
- [Collins, 2002] Michael Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 1–8, 2002.
- [Desmet and Hoste, 2014] Bart Desmet and Véronique Hoste. Fine-grained dutch named entity recognition. *Language resources and evaluation*, 48(2):307–343, 2014.
- [Faruqui *et al.*, 2010] Manaal Faruqui, Sebastian Padó, and Maschinelles Sprachverarbeitung. Training and evaluating a german named entity recognizer with semantic generalization. In *Proceedings of KONVENS*, pages 129–133, 2010.
- [Hänig *et al.*, 2014] Christian Hänig, Stefan Bordag, and Stefan Thomas. Modular classifier ensemble architecture for named entity recognition on low resource systems. *Proceedings of the KONVENS Germ-Eval Shared Task on Named Entity Recognition, Hildesheim, Germany*, 2014.
- [Lample *et al.*, 2016] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-2016)*, pages 260–270, 2016.
- [Ma and Hovy, 2016] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL-2016)*, pages 1064–1074, 2016.
- [Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.
- [Minard *et al.*, 2016] Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begoña Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. MEANTIME, the NewsReader Multilingual Event and Time Corpus. In *Proceedings of LREC 2016*, 2016.
- [Nothman *et al.*, 2013] Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175, 2013.
- [Passos *et al.*, 2014] Alexandre Passos, Vineet Kumar, and Andrew McCallum. Lexicon infused phrase embeddings for named entity resolution. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 78–86, Ann Arbor, Michigan, June 2014. Association for Computational Linguistics.
- [Ratinov and Roth, 2009] L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155, 2009.
- [Taulé *et al.*, 2008] Mariona Taulé, Maria Antònia Martí, and Marta Recasens. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of LREC 2008*, pages 96–101, 2008.
- [Tjong Kim Sang, 2002] Erik F. Tjong Kim Sang. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2002*, pages 155–158. Taipei, Taiwan, 2002.
- [Turian *et al.*, 2010] Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 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 2010. Association for Computational Linguistics.