# Conditional Random Fields for Spanish Named Entity Recognition using Unsupervised Features

Jenny Copara[1], Jose Ochoa[1], Camilo Thorne[2], Goran Glavaš[2]

[1]Universidad Católica San Pablo, Arequipa, Peru
{jenny.copara,jeochoa}@ucsp.edu.pe
[2]Data & Web Science Group, Universität Mannheim, Germany
{camilo,goran}@informatik.uni-mannheim.de

**Abstract.** Unsupervised features based on word representations such as word embeddings and word collocations have shown to significantly improve supervised NER for English. In this work we investigate whether such unsupervised features can also boost supervised NER in Spanish. To do so, we use word representations and collocations as additional features in a linear chain Conditional Random Field (CRF) classifier. Experimental results (82.44% F-score on the CoNLL-2002 corpus) show that our approach is comparable to some state-of-art Deep Learning approaches for Spanish, in particular when using cross-lingual word representations.

**Keywords:** NER for Spanish · Word Representations · Collocations · Conditional Random Fields

## 1 Introduction

Supervised Named Entity Recognition (NER) system are typically fed with supervised or manually engineered features [9] such as, e.g., word capitalization or domain-specific lexicons (lists of words related with named entity types) [5], [19], [17]. The performance of such techniques however depends on the availability, quality and size of annotated data, which can be scarce for NER for languages other than English. More recently, it has been shown that supervised NER can be boosted via unsupervised word features induced from corpora [23], such as **(i)** very large word clusters [3], [13], **(ii)** word collocations [13], and **(iii)** very large word embeddings [6], [7], [14], [15]. Such techniques show in particular that it is possible to take advantage of unlabeled data to enrich and boost supervised NER models learned over small gold standards.

For English NER, [11],[17] show that (large) word embeddings yield better results than clustering. However, when combined and fed as features to linear chain Conditional Random Field (CRF) sequence classifiers, they yield models comparable to state-of-the-art deep learning models.

In this paper we investigate whether these techniques can be successfully applied to NER in Spanish. In order to do so, we follow Guo's approach in [11], combining probabilistic graphical models in the form of CRFs learned from the CoNLL 2002 corpus with word representations learned from large unlabeled

Spanish corpora, while exploring the optimal setting and feature combinations that match state-of-the-art algorithms for NER in Spanish.

The paper is organized as follows. In Section 2 we provide a review of Spanish NER and unsupervised word features. Section 3 describes the structure of the word representations used. Section 4 shows our experimental setting and discusses our results. Section 5 presents our final remarks.

## 2 Related work

### 2.1 Spanish NER

The first results (CoNLL 2002 shared-task[1]) for supervised Spanish NER were obtained by Carreras[5]. A set of selected word features and lexicons (gazetteers) on an Adaboost learning model were used, obtaining an F-score of 81.39%. These results remained unbeaten until recently, and the spread of *Deep Learning* (currently achieving an F-score of 85.77%). The main algorithms that are currently used for NER in Spanish are: Convolutional Neural Networks with word and character embeddings [7], Recurrent Neural Networks (RNNs) with word and character embeddings [12], [25], and a character-based RNN with characters encoded as bytes [10].

### 2.2 Unsupervised Word features

Unsupervised features based on word representations and word collocations have been successfully used to boost many Natural Language Processing (NLP) tasks (e.g., language modeling[3], English NER [8], [11], [13], [17], [23], German NER[8], chunking[23], Chinese word segmentation[13]).

There are two main approaches used to induce word representations. One approach is to compute either clusters [3] or [13] Brown Clustering from unlabeled data and using them as features in NLP models (including NER). Another approach transforms each word into a continuous real-valued vector [6] of $n$ dimensions also known as a "word embedding" [14]. With Brown clustering, words that appear in the same or a similar sentence context are assigned to the same cluster. Whereas in word embeddings similar words occur close to each other in $\mathbb{R}^n$ (the induced $n$ dimensional vector space). Having more data is better for word representations. Cross-lingual datasets can be used to gather data, provided they overlap in vocabulary and domain. In this sense, cross-lingual word representations have been shown to improve several NLP tasks, such as model learning[1], [27]. This is because, among other things, they allow to extend the coverage of possibly limited (in the sense of small or sparsely annotated) resources with resources in other languages, such as: using English to enrich Chinese [27], or learning a model in English to solve a text classification task for German (also German-English, English-French and French-English) [1].

[1] http://www.cnts.ua.ac.be/conll2002/ner/

On the other hand, word collocations have also been used as additional word features to solve NLP tasks. In particular, Chinese word segmentation have been significantly improved by using them [13].

## 3 Unsupervised word features for Spanish NER

### 3.1 Brown clustering

Brown clustering is a hierarchical clustering of words that takes a sequence $w_1, \ldots, w_n$ of words as input and returns a binary tree as output (a dendogram). The binary tree's leaves are the input words. This clustering method is based on bigram language models [3], [13].

### 3.2 Clustering embeddings

A clustering method for embeddings based on *k-means* has been proposed by Yu [26]. In this method, different $k$ clusters values convey different clustering granularity levels. The toolkit Sofia-ml [20] [2] was used to compute such $k$ clusters.

### 3.3 Binarized embeddings

The idea behind this method is to "reduce" continuous word vectors $\boldsymbol{w}$ in standard word embeddings into discrete $bin(\boldsymbol{w})$ vectors that however preserve the ordering or ranking of the original vectors. To do this, we need to compute two thresholds per dimension (upper and lower) across the whole vocabulary. For each dimension (component) $i$ is computed the *mean* of positives values ($C_{i+}$, the upper threshold) and negative values ($C_{i-}$, the lower one). Thereafter, the following function is used over each component $C_{ij}$ of vector $\boldsymbol{w}_j$:

$$\phi(C_{ij}) = \begin{cases} U_+, & ifC_{ij} \geq mean(C_{i+}), \\ B_-, & ifC_{ij} \leq mean(C_{i-}), \\ 0, & otherwise. \end{cases} \tag{1}$$

### 3.4 Distributional Prototypes

This approach, proposed by Guo in [11] is based on the idea that each entity class has a set of words more likely to belong to this class than the other words (i.e., Maria, Jose are more likely to be classified as a *PERSON* entity). Thus, it is useful to identify a group of words that represent each class (*prototypes*) and select *some of them* in order to use them as word features. In order to compute prototypes two steps are necessary:

[2] https://code.google.com/archive/p/sofia-ml/

1. Generate a prototype for each class of an annotated training corpus. This step relies on Normalized Pointwise Mutual Information (NPMI) [2], as word-entity type relations can be modeled as a form of collocation. NPMI is a smoothed version of the Mutual Information measure typically used to detect word associations [24] and collocations[13]. Given an annotated training corpus, the NPMI is computed between labels $l$ and words $w$ using the following two formulas:

$$\lambda_n(l, w) = \frac{\lambda(l, w)}{-\ln p(l, w)}, \quad \lambda(l, w) \ = \ln \frac{p(l, w)}{p(l)p(w)}.$$

2. Map the prototypes to words in a (large) word embedding. In this step, given a group of prototypes for each class, we find out which prototypes in our set are the most *similar* to each word in the embeddings. *Cosine similarity* is used to do so and those prototypes above a threshold of usually 0.5 are chosen as the prototype features of the word.

### 3.5 Collocations

A collocation is defined as two or more lexical items that co-occur in a text or in a text corpus, whether or not they form a syntactic pattern[18]. Collocations are induced from unlabeled data by computing bigram counts and Pointwise Mutual Information [13].

## 4 Experiments and Discussion

Unlike previous approaches, our work focuses on using unsupervised word features in supervised NER for Spanish. We do it within a probabilistic graphical model framework: CRFs. We trained our (enriched) CRF model over the (Spanish) CoNLL 2002 corpus, and built our unsupervised word features over the Spanish Billion Corpus and English Wikipedia.

For Spanish this is a novel approach. The experimental results show it achieves competitive performance w.r.t. the current (Deep learning-driven) state-of-the-art for Spanish NER, in particular when using *cross-* or *multi-lingual* Word Representations.

### 4.1 NER Model

In order to perform our NER experiments, a linear chain CRF sequence classifier[3] has been used. Our classifier relies on a set of standard baseline features, that we have extended with additional features based on unsupervised word features and collocations. This setup is depicted in Figure 1. The classifier was implemented using *CRFSuite* [16], due to its simplicity and the ease with which one can add extra features. Additionally, we tried the Stanford CRF classifier for NER [9], for comparison purposes.

[3] http://github.com/linetcz/spanish-ner

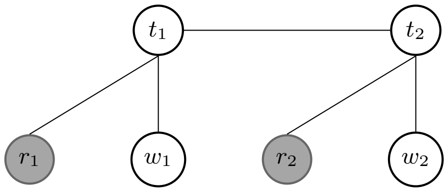

**Fig. 1.** Linear chain-CRF with word representations as features. The upper nodes are the label sequences, the bottom white nodes are the supervised word features in the model and the filled nodes are the unsupervised word features included in our model.

### 4.2 Baseline Features

The baseline features were defined over a window of ± *2 tokens*. The set of features for each word was:

- The word itself.
- Lower-case word.
- Part-of-speech tag.
- Capitalization pattern (e.g. from word "Twitter" we will get ULLLLLL) and type of character in the word(e.g. 'AllUpper', 'AllDigit', 'AllSymbol').
- Characters type information: capitalized, digits, symbols, initial upper case letter, all characters are letters or digits.
- Prefixes and suffixes: four first or latter letters respectively.
- Digit length: whether the current token has 2 or 4 length.
- Digit combination: which digit combination the current token has (alphanumeric, slash, comma, period).
- Whether the current token has just uppercase letter and period mark or contains an uppercase, lowercase, digit, alphanumeric, symbol character.
- Flags for initial letter capitalized, all letter capitalized, all lower case, all digits, all non-alphanumeric characters.

### 4.3 CoNLL 2002 Spanish Corpus

The CoNLL 2002 shared task [22] gave rise to a training and evaluation standard for supervised NER algorithms used ever since: the CoNLL-2002 Spanish corpus. The CoNLL is tagged using the so-called BIO format for NER gold standards. It covers four entities: *PERSON*, *ORGANIZATION*, *LOCATION*, *MISCELLANEOUS* and nine classes: B-PER, I-PER, B-ORG, I-ORG, B-LOC, I-LOC, B-MISC, I-MISC and O (B for "begin", I for "inside" and O for "outside" an entity mention of any of the four given types).

### 4.4 Unsupervised Word Features

*Spanish Dataset* In order to compute our word representations (viz., the Brown clusters, word embeddings) and word collocations, a large amount of unlabeled

**Table 1.** Brown cluster computed from SBW.

| **Brown Clusters** | **Word** |
|---|---|
| 011100010 | Française |
| 011100010 | Hamburg |
| 0111100011010 | latino |
| 0111100011010 | conservador |
| 0111111001111 | malogran |
| 0111111001111 | paralizaban |
| 011101001010 | Facebook |
| 011101001010 | Twitter |
| 011101001010 | Internet |

data is required. To this end we relied on the Spanish Billion Words (SBW) corpus and embeddings [4]. This dataset was gathered from several public domain resources[4] in Spanish: e.g., a Spanish portion of SenSem, the Ancora Corpus, the Europarl and OPUS Project Corpora, the Tibidabo Treebank and IULA Spanish LSP Treebank and dumps from Spanish Wikipedia, Wikisource and Wikibooks until September 2015 [4]. The corpora cover 3 817 833 *unique* tokens, and the embeddings 1 000 653 *unique* tokens with 300 dimensions per vector.

*Cross-lingual Dataset* Entity names tend to be very similar (often, identical) across languages and domains. This implies that word representation approaches should gain in performance when cross- or multi-lingual datasets are used. To test this hypothesis, we used an English Wikipedia dump from 2012 preprocessed by Guo[11], who removed paragraphs that contained non-roman characters and lowercased words. Additionally they removed very frequent words.

*Brown clustering* The number $k$ of word clusters for Brown clustering was fixed to 1000 following Turian[23]. Sample Brown clusters are shown in Table 1. The cluster is used as feature of each word in the annotated CoNLL 2002. As the reader can see Brown clustering tends to assign entities of same entity type to the same cluster.

*Binarized Embeddings* Table 2 shows a short view of word “equipo”. In the first column we can see each dimension of “equipo” and in the second its continuous value. The third column shows the binarized value. We used the binarized value as feature for each observed word (all dimensions with a *binarized value* different to *zero* will be considered).

*Clustering Embeddings* For cluster embeddings, 500, 1000, 1500, 2000 and 3000 clusters were computed, to model different levels of granularity [11]. As features for each word $w$, we return the cluster assignments at each granularity level. Table 3 shows the clusters of embeddings computed for word “Maria”. The first column denotes the level of granularity. The second column denotes the cluster assigned to “Maria” at each granularity level.

[4] http://crscardellino.me/SBWCE/

**Table 2.** Binarized embeddings from SBW for word "equipo".

| Dimension | Value | Binarized |
|---|---|---|
| 1 | -0.008255 | 0 |
| 2 | 0.145529 | U+ |
| 3 | 0.010853 | 0 |
| ⋮ | ⋮ | ⋮ |
| 298 | 0.050766 | U+ |
| 299 | -0.066613 | B- |
| 300 | 0.073499 | U+ |

**Table 3.** Clustering embeddings from SBW for word "Maria".

| Granularity $k$ | Cluster |
|---|---|
| 500 | 31 |
| 1000 | 978 |
| 1500 | 1317 |
| 2000 | 812 |
| 3000 | 812 |

*Distributional Prototypes* We extracted, for each CoNLL BIO label 40 prototypes (the topmost 40 w.r.t. NPMI).

Table 4 shows the top four prototypes per entity class computed from CoNLL-2002 Spanish corpus (training subset). These prototypes are instances of each entity class even non-entity tag(O) and therefore they are compound by entities or entity parts (i.e. *Buenos Aires* is a *LOCATION* so we see the word *Aires* as prototype of I-LOC).

*Collocations* Collocations were computed for each word in the the CoNLL 2002 corpus, and added as features. Table 5 depicts collocations for words: Estados" and General".

**Table 4.** CoNLL-2002 Spanish Prototypes.

| Class | Prototypes |
|---|---|
| B-ORG | EFE, Gobierno, PP, Ayuntamiento |
| I-ORG | Nacional, Europea, Unidos, Civil |
| I-MISC | Campeones, Ambiente, Ciudadana, Profesional |
| B-MISC | Liga, Copa, Juegos, Internet |
| B-LOC | Madrid, Barcelona, Badajoz, Santander |
| I-LOC | Janeiro, York, Denis, Aires |
| B-PER | Francisco, Juan, Fernando, Manuel |
| I-PER | Alvarez, Lozano, Bosque, Ibarra |
| O | que, el, en, y |

**Table 5.** Collocations computed for “Estados” and “General”

| Word | Collocations |
|---|---|
| Estados | los miembros Miembros Unidos |
| General | Asamblea Secretario |

### 4.5 Results

In order to evaluate our proposal the standard `conlleval`[5] script was used. Table 6 shows results achieved on CoNLL-2002 (Spanish), and compares them to Stanford and the state-of-the-art for Spanish NER. The baseline achieved 80.02% of F-score.

It is worth nothing that *Brown clustering* improves the baseline. The same holds for *clustering embeddings* and *collocations.* By contrast, *binarized embeddings* do worse than the *baseline.* This seems to be due to the fact that the process of binarization apparently discards information quite relevant for Spanish NER. The same holds for *prototypes* which, when taken alone, yield results also below the *baseline.*

Combining the features yields, on the other hand and in all cases, results above the baseline and above Brown clustering or clustering embeddings taken alone.

However, our best results were obtained by using a *cross-lingual combination* of Brown clusters computed from the English Wikipedia dump (2012) with clustered embeddings and prototypes computed from SBW. The same holds when combining Brown clusters, clustered embeddings and prototypes with collocations. The reason why cross-lingual combinations are good in this task is due to the high level of overlap among entities in Spanish and English. Put otherwise, many entities that share the same name and a similar context occur in texts from both languages, giving rise to features with higher predictive value.

### 4.6 Discussion

The first results for supervised Spanish NER using the CoNLL 2002 corpus considered a set of features with gazetteers and external knowledge [5] which turned out 81.39% F1-score (see Table 6). However, without gazetteers and external knowledge results go down to 79.28% (see Table 6).

It is worth noting that the knowledge injected to the previous learning model was *supervised.* We on the other hand have considered *unsupervised* external knowledge, while improving on those results. This is further substantiated by our exploring unsupervised features with the Stanford NER CRF model [9]. In this setting F-score of 81.44% was obtained, again above Carreras[5].

More importantly, our work shows that an English resource (Brown clusters computed from English Wikipedia) can be used to improve Spanish NER with

[5] http://www.cnts.ua.ac.be/conll2000/chunking/conlleval.txt

**Table 6.** CoNLL2002 Spanish Results. Top: results obtained by us. Middle: results obtained with previous approaches. Down: current Deep Learning-based state-of-the-art for Spanish NER.

| **Model** | **F1** |
|---|---|
| Baseline | 80.02% |
| +Binarization | 79.48% |
| +Brown | 80.99% |
| +Prototype | 79.82% |
| +Collocation | 80.23% |
| +Clustering | 80.24% |
| +Clustering+Prototype | 80.55% |
| +Brown+Collocation | 81.04% |
| +Brown+Clustering | 82.30% |
| +Brown+Clustering+Prototype | 81.19% |
| +Brown+Clustering+Prototype+Collocation | 80.96% |
| +Brown+Clustering+Prototype+Collocation$^\star$ | 82.23% |
| **+Brown+Clustering+Prototype*** | **82.44%** |
| Carreras[5]$^{\star\star}$ | 79.28% |
| Carreras[5] | 81.39% |
| Finkel[9] | 81.44% |
| dos Santos[7] | 82.21% |
| Gillick[10] | 82.95% |
| Lample[12] | 85.75% |
| Yang[25] | 85.77% |

$^\star$ Brown clusters from English resource

$^{\star\star}$ did not take into in account gazetteers

word representations as *(i)* entities in Spanish and English are often identical, and *(ii)* the resulting English Brown clusters for English entities correlate better with their entity types, giving rise to a better model.

Another point to note is that whilst binarization improves on English NER baselines Guo[11], the same does not work for Spanish. It seems that this approach adds instead noise to Spanish NER. Likewise, Collocations do not perform well for Spanish.

We also note that *word capitalization* has a distinct impact on our approach. With the following setting: English Brown clusters, Spanish cluster embeddings and *lower-cased* Spanish prototypes we got 0.78% less F-score than with upper-cased prototypes. This is because the lower-cased prototypes will ignore the real context in which the entity appears (since a prototype is an instance of an entity class) and will be therefore mapped to the wrong word vector in the embedding (when computing cosine similarity). This seems to suggest that while prototypes are globally speaking useful, using Spanish data alone is not.

Finally, when comparing our approach to the current state-of-the-art using Deep Learning methods [7], [10], [12], [25] (that extract features at the character, word and bytecode level to learn deep models), our work outperforms dos Santos[7] F-score and matches also Gillick[10].

## 5 Conclusions

This paper has explored unsupervised and minimally supervised features for Spanish NER, based on cross-lingual word representations within a CRF classification model. Our CRF model was trained over the Spanish CoNLL 2002 corpus, the Spanish Billion Word Corpus and English Wikipedia (2012 dump). This is a novel approach for Spanish. Our experiments show competitive results when compared to the current state-of-the-art in Spanish NER based on Deep Learning. In particular, we outmatched dos Santos[7].

Cross-lingual Word Representations have a positive impact on NER performance for Spanish. In the future, we would like to focus further on this aspect and consider more (large scale) cross-lingual datasets.

## Acknowledgments

We are grateful to the Data and Web Science Group at University of Mannheim. Special thanks to Heiner Stuckenschmidt and Simone Ponzetto for their contributions and comments. This work was supported by the Master Program in Computer Science at Universidad Católica San Pablo and the Peruvian National Fund of Scientific and Technological Development through grant number 011-2013-FONDECYT.

## References

1. Binod Bhattarai. 2013. Inducing cross-lingual word representations. Master's thesis, Multimodal Computing and Interaction, Machine Learning for Natural Language Processing. Universität des Saarlandes.
2. G. Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. In C. Chiarcos, E. de Castilho, and M. Stede, editors, *Von der Form zur Bedeutung: Texte automatisch verarbeiten / From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, pages 31–40, Tübingen. Gunter Narr Verlag.
3. Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Comput. Linguist.*, 18(4):467–479, December.
4. Cristian Cardellino. 2016. Spanish Billion Words Corpus and Embeddings, March. `http://crscardellino.me/SBWCE/`
5. Xavier Carreras, Lluís Màrques, and Lluís Padró. 2002. Named entity extraction using adaboost. In *Proceedings of CoNLL-2002*, pages 167–170. Taipei, Taiwan.
6. Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 160–167, New York, NY, USA. ACM.

7. Cicero dos Santos and Victor Guimarães. 2015. Boosting named entity recognition with neural character embeddings. In *Proceedings of the Fifth Named Entity Workshop*, pages 25–33, Beijing, China, July. Association for Computational Linguistics.
8. Manaal Faruqui and Sebastian Padó. 2010. Training and evaluating a german named entity recognizer with semantic generalization. In *Proceedings of KONVENS 2010*, Saarbrücken, Germany.
9. Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.
10. D. Gillick, C. Brunk, O. Vinyals, and A. Subramanya. 2015. Multilingual Language Processing From Bytes. *ArXiv e-prints*, November.
11. Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Revisiting embedding features for simple semi-supervised learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 110–120, Doha, Qatar, October. Association for Computational Linguistics.
12. Guillaume Lample, Miguel Ballesteros, Kazuya Kawakami, Sandeep Subramanian, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *In proceedings of NAACL-HLT (NAACL 2016).*, San Diego, US.
13. Percy Liang. 2005. Semi-supervised learning for natural language. Master's thesis, Department of Electrical Engineering and Computer Science. Massachusetts Institute of Technology.
14. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.
15. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In C.j.c. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119.
16. Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).
17. Alexandre Passos, Vineet Kumar, and Andrew McCallum. 2014. 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. Association for Computational Linguistics.
18. Sonja Poulsen. 2005. Collocations as a language resource. A functional and cognitive study in English phraseology. PhD dissertation, Institute of Language and Communication. University of Southern Denmark.
19. Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pages 147–155, Stroudsburg, PA, USA. Association for Computational Linguistics.
20. D. Sculley. 2010. Combined regression and ranking. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 979–988, New York, NY, USA. ACM.
21. Charles Sutton and Andrew McCallum. 2012. An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4(4):267–373.
22. Erik F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–4, Stroudsburg, PA, USA. Association for Computational Linguistics.

23. Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 384–394, Stroudsburg, PA, USA. Association for Computational Linguistics.
24. Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 412–420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
25. Zhilin Yang, Ruslan Salakhutdinov, and William Cohen. 2016. Multi-task cross-lingual sequence tagging from scratch. *CoRR*, abs/1603.06270.
26. Mo Yu, Tiejun Zhao, Daxiang Dong, Hao Tian, and Dianhai Yu. 2013. Compound embedding features for semi-supervised learning. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 563–568.
27. Mo Yu, Tiejun Zhao, Yalong Bai, Hao Tian, and Dianhai Yu. 2013a. Cross-lingual projections between languages from different families. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 312–317, Sofia, Bulgaria, August. Association for Computational Linguistics.