Exploring unsupervised features in Conditional Random Fields for Spanish Named Entity Recognition

Jenny Copara, Jose Ochoa
Research Center in Computer Science
Universidad Católica San Pablo
Arequipa, Peru
{jenny.copara,jeochoa}@ucsp.edu.pe

Camilo Thorne, Goran Glavaš
Data & Web Science Group, Germany
University of Mannheim
Mannheim, Germany
{camilo,goran}@informatik.uni-mannheim.de

***Abstract*—Unsupervised features such as word representations mostly given by word embeddings have been shown significantly improve semi supervised Named Entity Recognition (NER) for English language. In this work we investigate whether unsupervised features can boost (semi) 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 and 65.72% F-score on Ancora 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.**

***Index Terms*—NER for Spanish; Unsupervised features; Word Representations; Word embeddings; Conditional Random Fields.**

I. Introduction

Named Entity Recognition (NER) allows to identify and classify entities in a text [1], [2]. It has been used as a part of several Natural Language Processing (NLP) tasks (for instance Automatic summarization, information retrieval, machine translation, question answering, text mining [3]). NER is addressed as a sequential classification problem mostly through Conditional Random Fields [4].

CRF classifier is fed with features [4], [5] given by driven-knowledge (supervised features) and automatic learned knowledge (unsupervised features). A common practice has been to use domain-specific lexicon (list of words related with named entity types) [6], [7], [8]. More recently, it has been shown that supervised NER can be boosted via specific word features induced from very large unsupervised techniques such as word representations [5], and unsupervised knowledge as additional features. In particular, from **(i)** very large word clusters [9], [10], **(ii)** collocations [10], and **(iii)** very large word embeddings [11], [12], [13], [14].

Word features induced from supervised techniques require large amounts of (manually) labeled data to achieve good performance, data that is hard to acquire or generate. However, it is possible to take advantage of unlabeled data to enrich and boost supervised NER models learned over small gold standards.

For English NER, Passos [8] and Guo [15] show that word embeddings yield better results than clustering. However, when combined and fed as features to linear chain 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 [15] combining probabilistic graphical models learned from annotated corpora (CoNLL 2002 and Ancora), 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 II, we provide a review of Spanish NER, and its use of unsupervised word features. Section III describes the structure of the word representations used. Section IV shows our experimental setting and discusses results. Section V presents our final remarks.

II. Related work

A. Spanish NER

The first results (CoNLL 2002 shared-task[1]) for Spanish NER were obtained by Carreras [6] where a set of selected word features and lexicons[2] 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* (more detail in [2]). The state-of-the-art algorithms for this task (currently achieving an F-score of 85.77%) are mostly based on Deep Learning. Using Convolutional Neural Networks with word and character embeddings [14], Recurrent Neural Networks (RNNs) with word and character embeddings [2], [16], and a character-based RNN with characters encoded as bytes [17].

B. Unsupervised Word features

Among unsupervised word features, some techniques have shown improvement in several NLP tasks such as word repre-

[1]http://www.cnts.ua.ac.be/conll2002/ner/
[2]Also known as *gazetteers*

sentations [9], [10], [18], [5], [15], [8], and linguistic resources [10].

Word Representations have been shown to substantially improve several NLP tasks, among which NER for English and German [18]. There are two main approaches. One approach is to compute clusters [9], [10] (Brown Clustering). Another approach transforms each word into a continuous real-valued vector [11] of n dimensions also known as a "word embedding" [12]. With Brown clustering, words that appear in 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).

Word Representations work better the more data they are fed. One way to achieve this is to input them cross-lingual datasets, provided they overlap in vocabulary and domain. Cross-lingual Word Representations have been shown to improve several NLP tasks, such as model learning [19], [20]. This is because, among other things, they allow to extend the coverage of possibly limited (in the sense of small or sparsely annotated) resources through Word Representations in other languages. For instance, using English to enrich Chinese [20], or learning a model in English to solve a Text Classification task in German (also German-English, English-French and French-English) [19].

Linguistic resources can be effectively used as additional word features since they have shown improvement for Chinese Word Segmentation [10] through collocations.

III. Unsupervised word features for Spanish NER

A. 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. The binary tree's leaves are the input words. This clustering method is based on bigram language models [9], [10].

B. Clustering embeddings

A clustering method for embeddings based on *k-means* has been proposed in Yu [21]. Experiments have shown different numbers for *k's* which contains different granularity information. The toolkit Sofia-ml [22] [3] was used.

C. Binarized embeddings

The idea behind this method is to "reduce" continuous word vectors $\vec{w}$ into discrete $bin(\vec{w})$ 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 $\vec{w}_j$:

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

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

D. Distributional Prototypes

This approach 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 Guo [15] two steps are necessary:

1) Generate a prototype for each class of an annotated training corpus. This step relies on Normalized Pointwise Mutual Information (NPMI) [23]. 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. 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 word embeddings. 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.

E. Collocations

A collocation is given when two or more lexical items often co-occur in a text, or in a text corpus, whether or not they form a syntactic pattern [25]. Collocations are computed from unlabelled data and are induced by bigram counts using Pointwise Mutual Information [10].

IV. Experiments and Discussion

Unlike previous approaches, our work focuses on using unsupervised word features in supervised NER for Spanish. We do it within the probabilistic graphical model CRF. We have trained our model and built our unsupervised word features over the Spanish Billion Corpus (SBW) and English wikipedia. For Spanish this is a novel approach.

The experimental results have shown competitive performance with respect to the current state-of-the-art, in particular when using *cross-* or *multi-lingual* Word Representations.

A. NER Model

We used a linear chain CRF sequence classifier which is a discriminative probabilistic graphical model that estimates the conditional probability of label sequence $\boldsymbol{t}$ given word sequence (sentence) $\boldsymbol{w}$:

$$p(\boldsymbol{t}|\boldsymbol{w}) = \frac{1}{Z} \exp\left(\sum_{i=1}^{|\boldsymbol{t}|} \sum_{j=1}^{\#(F)} \theta_j f_j(t_{i_1}, t_i, \boldsymbol{w}_i) \right)$$

where Z is a normalization factor that sums the body (argument) of the exponential over all sequences of labels $\boldsymbol{t}$. f_js are feature functions and $\boldsymbol{w}_i$ is the word window observed at input position i. θ_j parameters are estimated via gradient minimization methods. The computational cost is $O(hn+af)$, where h is the average number of features that are relevant to each token, n is the number of tokens, f is the number of features and a is the learning rate.

Our classifier relies on a set of baseline features which were extended with additional features based on unsupervised word features. This use of unlabeled data is depicted in Figure 1. The classifier was implemented using *CRFSuite* [26], due to its simplicity and the ease with which one can add extra features. Additionally, we experimented with the Stanford CRF classifier for NER [4], for comparison purposes.

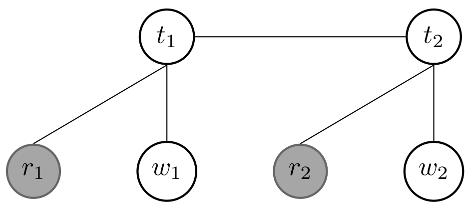

Fig. 1. Linear chain-CRF with word representations as features. The upper nodes are the label sequences, the bottom white nodes are the supervised word features in the model and the filled nodes are the unsupervised word features included in our model.

B. Baseline Features

The baseline features minimally supervised 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 and type of character in the word.
- Characters type information: capitalized, digits, simbols, initial upper case letter, all characters are letters or digits.
- Prefixes and suffixes of token: Since one to four first and 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,

C. Spanish Corpora

On one hand, the CoNLL 2002 shared task [1] gave rise to a training and evaluation standard for supervised NER algorithms used ever since: the CoNLL-2002 Spanish corpus. The CoNLL is tagged with 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. On the other hand, AnCora corpus (for Catalan and Spanish languages) is compound by multilevel annotations [27]. Named entities are annotated manually. It has six entities: *DATE*, *LOCATION*, *NUMBER*, *ORGANIZATION*, *OTHER* and *PERSON*. The IOB-style has also been used for entity annotations. Therefore, there are thirteen classes (corresponding with entity classes). To Coreference Resolution the AnCora by SemEval Shared-Task [28] has been used due to training, development and test sets are provided.

TABLE I
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

D. Unsupervised Word Features

a) Spanish Dataset: In order to compute our word representations (Brown clusters, word embeddings) and collocations a large amount of unlabeled data is required. To this end we relied on the SBW corpus and embeddings [29]. This dataset was gathered from several public domain resources[4] in Spanish. The corpora covers 3 817 833 *unique* tokens, and the embeddings 1 000 653 *unique* tokens with 300 dimensions per vector.

b) Cross-lingual Dataset: Entity names tend to be very similar (often, identical) across languages and domains. This should imply 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 [15], who removed paragraphs that contained non-roman characters and lowercased words. Additionally they removed frequent words.

c) Brown clustering: The number *k* of word clusters for Brown clustering was fixed to 1000 according Turian [5]. Sample Brown clusters are shown in Table I. The cluster is used as feature of each word in the annotated corpora. As can be observed, Brown clustering tends to assign same type entities to the same cluster.

d) Binarized Embeddings: Table II shows a short view of word "equipo" vector. In the first column we can see each dimension of "equipo", in the second its continuous value and the next shows the binarized value. It is worth noting that we just took *binarized values* (third column) with values between $\{U+, B-\}$.

e) Clustering Embeddings: For cluster embeddings, 500, 1000, 1500, 2000 and 3000 clusters were computed, to model different levels of granularity [15]. As features for each word w, we return the cluster assignments at each granularity level. Table III 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 II
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 III
CLUSTERING EMBEDDINGS FROM SBW FOR WORD "MARIA".

Granularity	k
500	31
1000	978
1500	1317
2000	812
3000	812

f) Distributional Prototypes: Regarding prototypes, we extracted the topmost 40 prototypes with respect to NPMI, for each class in CoNLL-2002 corpus whereas 80 prototypes in AnCora corpus.

Table IV 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). It is worth noting that a token could belong to more than one entity in computation of NPMI, however all the words selected as prototypes are taken into account, including repeated. This fact does not have effect to compute of prototypes since they are working as a set (without tag entities).

g) Collocations: Computed from SBW associated with the corresponding words in each corpora and taken as features. Table V shows instances of words "Estados" and "General".

TABLE IV
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 V
COLLOCATIONS COMPUTED OF THE WORDS: "ESTADOS AND "GENERAL

Word	Collocations	
Estados	los	miembros
	Miembros	Unidos
General	Asamblea	Secretario

TABLE VI
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*	82.23%
+Brown+Clustering+Prototype*	**82.44%**
Carreras [6]†	79.28%
Carreras [6]	81.39%
Finkel [4]	81.44%
Finkel [4]‡	81.02%
dos Santos [14]	82.21%
Gillick [17]	82.95%
Lample [2]	85.75%
Yang [16]	85.77%

*Brown clusters from English resource
†did not take into in account gazetteers
‡using an unsupervised feature

E. Results

In order to evaluate our models we used the standard `conlleval`[5] script. Table VI shows the 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. In Table VII shows results on AnCora Spanish corpus, and compares them with Stanford CRF NER.

It is worth nothing that in CoNLL results *Brown clustering* improves the baseline as well as *Collocations*. The same holds for *Clustered embeddings*. By contrast, *Binarization embeddings* does worse than the *Baseline*. This seems to be due to the fact that binarized embeddings by grouping vector components into a finite set of discrete values throw away information relevant for Spanish NER. The same goes for *Prototypes*, which when taken alone yield results also below the *Baseline*.

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

However, our best results in this corpus were obtained by using a *cross-lingual combination* between Brown clusters computed from the English Wikipedia dump (2012) with clustered embeddings and prototypes computed from SBW. The same holds combining Brown clusters, clustered embeddings and prototypes with Collocations. The reason Brown clusters 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

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

TABLE VII
ANCORA SPANISH RESULTS. TOP: RESULTS OBTAINED BY US. DOWN: RESULTS OBTAINED WITH PREVIOUS APPROACHES.

Model	**FB1**
Baseline	62.76%
+Brown	63.49%
+Prototypes	63.22%
+Collocation	62.79%
+Clustering	65.23%
+Clustering+Prototype	64.86%
+Brown+Clustering	64.57%
+Clustering+Collocation	64.20%
+Brown+Collocation	63.57%
+Prototype+Collocation	62.68%
+Brown+Clustering+Prototype*	64.19%
+Brown+Clustering+Prototype	64.19%
+Brown+Clustering+Collocation	64.30%
+Brown+Clustering+Prototype+Collocation	64.39%
+Brown+Clustering+Prototype+Collocation*	**65.72%**
Finkel [4]	61.84%
Finkel† [4]	62.36%

*Brown clusters from English resource
†using an unsupervised feature

texts from both languages, giving rise to features with higher predictive value.

Whereas results on AnCora corpus, all the approaches outperform Baseline, however combination of *prototype* and *collocations* perform worse. It is worth noting that *clustering embeddings* approach shows high performance with respect to *Baseline* and results given by Stanford CRF NER [4]. But different to CoNLL 2002, in AnCora the use of Collocations in combination with *Brown clustering* (computed from in English resource), *clustering embeddings*, *prototype embeddings* give rise to our best results in this corpus.

F. Discussion

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

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 significantly improving on those results. This is further substantiated by our exploring unsupervised features with the Stanford NER CRF model [4]. In this setting F-score of 81.44% was obtained, again above Carreras [6].

More importantly, our work shows that an English resource (Brown clusters computed from English Wikipedia) can be used to improve Spanish NER with Word Representations as *(i)* entities in Spanish and English are often similar, 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 while binarization improves on English NER baselines Guo [15], the same does not work for Spanish. It seems that this approach adds instead noise to Spanish NER. Likewise, combinations with *collocations* do not improve results.

We also note that *word capitalization* has a distict impact on our approach. With the following setting: English Brown clusters, Spanish cluster embeddings and *lowercased* Spanish prototypes we got 0.78% less F-score than with uppercased prototypes. This is because the lowercased 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). Despite using collocations as features, they provide complementary information for NER however we can see this approach directly applied adds noise.

Finally, when comparing our approach to the current state-of-the-art using Deep Learning methods [14], [17], [2], [16] (that extract features at the character, word and bytecode level to learn deep models), our work outperforms dos Santos [14] F-score and matches also Gillick [17].

Additional experiments on AnCora corpus confirm that using *cross-lingual word representations* bring us complementary information to recognize entities(even when there are nested entities). As the reader can see in Table VII the best combination reached 65.72% of F-score, this is because in nested entities in this corpus can be compound by collocations.

V. CONCLUSIONS

This paper has explored unsupervised and minimally supervised features, based on cross-lingual Word Representations mostly, within a CRF classification model for Spanish NER, trained over the Spanish CoNLL 2002 corpus, AnCora 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 [14].

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

ACKNOWLEDGMENT

We would like to thank to Data and Web Science Group in particular Heiner Stuckenschmidt and Simone Ponzetto for useful help. This work was supported by the Master Program in Computer Science of the Universidad Católica San Pablo and the Peruvian National Fund of Scientific and Technological Development through grant number 011-2013-FONDECYT.

REFERENCES

[1] E. F. Tjong Kim Sang, "Introduction to the conll-2002 shared task: Language-independent named entity recognition," in *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, ser. COLING-02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 1–4. [Online]. Available: http://dx.doi.org/10.3115/1118853.1118877

[2] G. Lample, M. Ballesteros, K. Kawakami, S. Subramanian, and C. Dyer, "Neural architectures for named entity recognition," in *In proceedings of NAACL-HLT (NAACL 2016).*, San Diego, US, 2016.

[3] G. Szarvas, "Feature engineering for domain independent named entity recognition and biomedical text mining applications," Ph.D. dissertation, Research Group on Artificial Intelligence of the Hungarian Academy of Sciences and the University of Szeged, June 2008.

[4] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ser. ACL '05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 363–370. [Online]. Available: http://dx.doi.org/10.3115/1219840.1219885

[5] J. Turian, L. Ratinov, and Y. 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*, ser. ACL '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 384–394. [Online]. Available: http://dl.acm.org/citation.cfm?id=1858681.1858721

[6] X. Carreras, L. Màrques, and L. Padró, "Named entity extraction using adaboost," in *Proceedings of CoNLL-2002*. Taipei, Taiwan, 2002, pp. 167–170.

[7] L. Ratinov and D. Roth, "Design challenges and misconceptions in named entity recognition," in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, ser. CoNLL '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 147–155. [Online]. Available: http://dl.acm.org/citation.cfm?id=1596374.1596399

[8] A. Passos, V. Kumar, and A. McCallum, "Lexicon infused phrase embeddings for named entity resolution," in *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. Ann Arbor, Michigan: Association for Computational Linguistics, June 2014, pp. 78–86. [Online]. Available: http://www.aclweb.org/anthology/W/W14/W14-1609

[9] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based n-gram models of natural language," *Comput. Linguist.*, vol. 18, no. 4, pp. 467–479, Dec. 1992. [Online]. Available: http://dl.acm.org/citation.cfm?id=176313.176316

[10] P. Liang, "Semi-supervised learning for natural language," Master's thesis, Department of Electrical Engineering and Computer Science. Massachusetts Institute of Technology, 2005.

[11] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th International Conference on Machine Learning*, ser. ICML '08. New York, NY, USA: ACM, 2008, pp. 160–167. [Online]. Available: http://doi.acm.org/10.1145/1390156.1390177

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013. [Online]. Available: http://arxiv.org/abs/1301.3781

[13] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., 2013, pp. 3111–3119. [Online]. Available: http://media.nips.cc/nipsbooks/nipspapers/paper_files/nips26/1421.pdf

[14] C. dos Santos and V. Guimarães, "Boosting named entity recognition with neural character embeddings," in *Proceedings of the Fifth Named Entity Workshop*. Beijing, China: Association for Computational Linguistics, July 2015, pp. 25–33. [Online]. Available: http://www.aclweb.org/anthology/W15-3904

[15] J. Guo, W. Che, H. Wang, and T. Liu, "Revisiting embedding features for simple semi-supervised learning," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, October 2014, pp. 110–120. [Online]. Available: http://www.aclweb.org/anthology/D14-1012

[16] Z. Yang, R. Salakhutdinov, and W. Cohen, "Multi-task cross-lingual sequence tagging from scratch," *CoRR*, vol. abs/1603.06270, 2016. [Online]. Available: http://arxiv.org/abs/1603.06270

[17] D. Gillick, C. Brunk, O. Vinyals, and A. Subramanya, "Multilingual Language Processing From Bytes," *ArXiv e-prints*, Nov. 2015.

[18] M. Faruqui and S. Padó, "Training and evaluating a german named entity recognizer with semantic generalization," in *Proceedings of KONVENS 2010*, Saarbrücken, Germany, 2010.

[19] B. Bhattarai, "Inducing cross-lingual word representations," Master's thesis, Multimodal Computing and Interaction, Machine Learning for Natural Language Processing. Universität des Saarlandes, 2013.

[20] M. Yu, T. Zhao, Y. Bai, H. Tian, and D. Yu, "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)*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 312–317. [Online]. Available: http://www.aclweb.org/anthology/P13-2056

[21] M. Yu, T. Zhao, D. Dong, H. Tian, and D. Yu, "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*, 2013, pp. 563–568. [Online]. Available: http://aclweb.org/anthology/N/N13/N13-1063.pdf

[22] D. Sculley, "Combined regression and ranking," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '10. New York, NY, USA: ACM, 2010, pp. 979–988. [Online]. Available: http://doi.acm.org/10.1145/1835804.1835928

[23] G. Bouma, "Normalized (pointwise) mutual information in collocation extraction," in *Von der Form zur Bedeutung: Texte automatisch verarbeiten / From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, C. Chiarcos, E. de Castilho, and M. Stede, Eds. Tübingen: Gunter Narr Verlag, 2009, pp. 31–40.

[24] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proceedings of the Fourteenth International Conference on Machine Learning*, ser. ICML '97. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997, pp. 412–420. [Online]. Available: http://dl.acm.org/citation.cfm?id=645526.657137

[25] S. Poulsen, "Collocations as a language resource. a functional and cognitive study in english phraseology," Ph.D. dissertation, Institute of Language and Communication. University of Southern Denmark., 2005.

[26] N. Okazaki, "Crfsuite: a fast implementation of conditional random fields (crfs)," 2007. [Online]. Available: http://www.chokkan.org/software/crfsuite/

[27] M. T. M. A. M. M. Recasens, "Ancora: Multilevel annotated corpora for catalan and spanish," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-08)*. Marrakech, Morocco: European Language Resources Association (ELRA), May 2008, aCL Anthology Identifier: L08-1222. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2008/pdf/35_paper.pdf

[28] M. Recasens, L. Màrquez, E. Sapena, M. A. Martí, M. Taulé, V. Hoste, M. Poesio, and Y. Versley, "Semeval-2010 task 1: Coreference resolution in multiple languages," in *Proceedings of the 5th International Workshop on Semantic Evaluation*, ser. SemEval '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 1–8. [Online]. Available: http://dl.acm.org/citation.cfm?id=1859664.1859665

[29] C. Cardellino, "Spanish Billion Words Corpus and Embeddings," March 2016. [Online]. Available: http://crscardellino.me/SBWCE/