Spanish NER with Word Representations and CRFs

J. Copara[1] J. Ochoa[1] C. Thorne[2] G. Glavaš[2]

[1]UCSP (Arequipa, Peru)
jenny.copara@ucsp.edu.pe

[2]UMa (Mannheim, Germany)
camilo@informatik.uni-mannheim.de

NEWS 2016, Berlin, August 12, 2016

Motivation

(cf. El Comercio, 27.7.2016)

Pedro Pablo Kuczynski (PPK) recibió esta mañana en el Teatro Municipal de Lima las credenciales que lo ratifican como presidente del Perú para el período 2016–2021. En la ceremonia, presidida por el titular del Jurado Nacional de Elecciones (JNE), Francisco Távara, también recibieron sus credenciales de vicepresidentes Martín Vizcarra y Mercedes Aráoz.

Motivation

(cf. El Comercio, 27.7.2016)

Pedro Pablo Kuczynski (PPK) recibió esta mañana en el Teatro Municipal de Lima las credenciales que lo ratifican como presidente del Perú para el período 2016–2021. En la ceremonia, presidida por el titular del Jurado Nacional de Elecciones (JNE), Francisco Távara, también recibieron sus credenciales de vicepresidentes Martín Vizcarra y Mercedes Aráoz.

Motivation

(cf. El Comercio, 27.7.2016)

Pedro Pablo Kuczynski (PPK) recibió esta mañana en el Teatro Municipal de Lima las credenciales que lo ratifican como presidente del Perú para el período 2016–2021. En la ceremonia, presidida por el titular del Jurado Nacional de Elecciones (JNE), Francisco Távara, también recibieron sus credenciales de vicepresidentes Martín Vizcarra y Mercedes Aráoz.

detect
- persons
- locations
- organizations

CoNLL-2002 Shared Task

- In the CoNLL-2002 shared task [SF02] we seek to learn and predict 4 kinds of entities:
  1. persons (PER)
  2. organizations (ORG)
  3. locations (LOC)
  4. miscellaneous (MISC)
- Gave rise to the CoNLL-2002 Spanish corpus [CMP02], a collection of news wire articles made available by the Spanish EFE News Agency from May 2000
- IOB2 tag format: B-PER, I-PER, B-ORG, I-ORG, B-LOC, I-LOC, B-MISC, I-MISC and O (9 labels)

LOC	MISC	ORG	PER	Tokens	Sentences
6 983	2 958	10 490	6 278	74 683	11 755

Spanish NER - State of the Art

- The first results on Spanish NER were obtained by Carreras et al. in [CMP02]
  1. AdaBoost: 79.20% F-score
  2. AdaBoost with gazetteers: 81.30% F-score
- Current approaches are based on deep learning
  1. Recurrent neural networks (RNNs): 85.75% F-score by Lample et al. in [LBK+16]
  2. RNNs at bytecode level: 85.77% F-score by Yang et. al in [YSC16]

Spanish NER - State of the Art

- The first results on Spanish NER were obtained by Carreras et al. in [CMP02]
  1. AdaBoost: 79.20% F-score
  2. AdaBoost with gazetteers: 81.30% F-score
- Current approaches are based on deep learning
  1. Recurrent neural networks (RNNs): 85.75% F-score by Lample et al. in [LBK$^+$16]
  2. RNNs at bytecode level: 85.77% F-score by Yang et. al in [YSC16]
- Problem: models have good performance over CoNLL, but: **(1)** are monolingual and **(2)** come with huge computational costs


Spanish NER - State of the Art

- The first results on Spanish NER were obtained by Carreras et al. in [CMP02]
  1. AdaBoost: 79.20% F-score
  2. AdaBoost with gazetteers: 81.30% F-score
- Current approaches are based on deep learning
  1. Recurrent neural networks (RNNs): 85.75% F-score by Lample et al. in [LBK+16]
  2. RNNs at bytecode level: 85.77% F-score by Yang et. al in [YSC16]
- Problem: models have good performance over CoNLL, but: **(1)** are monolingual and **(2)** come with huge computational costs

Question

Build "lighter" models with cross-lingual word representations?


Word Representations

- By word representation we understand a word space induced from a corpus and any structure thereof computed
- Three main kinds:
  1. word clusters (e.g., Brown Clustering)
  2. word embeddings (e.g., Word2Vec)
  3. mixtures thereof
- Can be used to define unsupervised features to boost supervised NLP tasks
- Work better the more data they are fed
- Can combine cross-lingual data


Linear Chain Conditional Random Field (CRF)

We used for our experiments a linear chain CRF sequence classifier, that works by estimating the conditional probability of label sequence $\overline{c}$ given word sequence (sentence) $\overline{w}$:

$$p(\overline{c}|\overline{w}) = \frac{1}{Z} \exp\left(\sum_{i=1}^{|\overline{c}|} \sum_{j=1}^{\#(F)} \theta_j f_j(c_{i-1}, c_i, \overline{w}_i)\right)$$

- Z is a normalization factor (sums the exponential over all $\overline{c}$s)
- the f_js are feature functions
- $\overline{w}_i$ is the word window observed at the i-eth position of $\overline{w}$
- the θ_js are the parameters

NER Baseline Features (Window of ± 2 Tokens)

For each token w in the window we observe:

- token w "as is" and in lower case
- the POS tag of w
- prefixes and suffixes: the first and last four characters of w
- the digit combinations in w and the number of digits in w
- character information: the digits, symbols and initial upper case letter of w
- character type information: do all characters in w are letters or digits, are they capitalized, does it contain alphanumeric characters, symbols, punctuation marks?

Inducing Word Representations

- In order to compute our word representations large amounts of unlabeled data are required:
  1. English Wikipedia dump (2012)
  2. Spanish Billion Words (SBW) corpus and embeddings
- The SBW corpora cover $3,817,833$ unique tokens and the embeddings $1,000,653$ unique tokens with 300 dimensions per vector (cf. http://crscardellino.me/SBWCE/) collected/built from
  - 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
- English Wikipedia dump of 2012 ($5,800,280$ unique tokens)

Clustering Words (Brown Clustering [BdM+92]) I

Brown clustering, proposed by Brown et al. in [BdM+92], assigns to each word w a sequence $c_1, \ldots, c_n$ of hierarchical clusters

1. Use a large corpus (raw text)
2. Compute a bigram language model
3. Compute a hierarchical clustering/dendogram T using
   - single link clustering (minimize similarity)
   - similarity among word pairs w_i and w_j leveraging on bigram probabilities
4. For each word w, return the sequence of all (hierarchical) clusters to which they belong (path in dendogram T)

Clustering Words (Brown Clustering [BdM+92]) II

Brown Cluster ID	Word	
011100010	Française	location
011100010	Hamburg	
0111100011010	latino	adjective
0111100011010	conservador	
0111111001111	malogran	verb
0111111001111	paralizaban	
011101001010	Facebook	organization
011101001010	Twitter	
011101001010	Internet	

(similar words = similar clusters)

Clustering Word Embeddings [YZD+13]

A possible distributional feature is to consider the word clusters of word w in a word embedding following Yue et al. in [YZD+13]

1. Use k-means to compute k clusters
2. Consider different levels of granularity (number of clusters) by setting different values for k

Clusters ($= k$)	**Cluster ID**
500	31
1000	978
1500	1317
2000	812
3000	812

(clusters for "Maria")

Distributional Prototypes [GCWL14] I

Word-entity type relations can be modeled as a form of collocation following Guo et al. in [GCWL14]

1. Compute a prototype for each NER type in a training corpus can be computed via normalized Pointwise Mutual Information (PMI):

$$PMI_n(l, w) = \frac{PMI(l, w)}{-\ln p(l, w)}, \quad PMI(l, w) = \ln \frac{p(l, w)}{p(l)p(w)}$$

2. Map the prototypes to words in a (large) word embedding by checking which prototype is the the most similar to each word w

We used as input corpora CoNLL 2002 and the Spanish Billion corpus

Distributional Prototypes [GCWL14] II

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

(top 4 per class)

Binarized Embeddings [GCWL14] I

The idea behind this method proposed also by Guo et al. is to "reduce" continous word vectors $\bar{w}$ in to discrete $\phi(\bar{w})$ preserving ranking

1. Compute two thresholds $i+$ (upper) and $i-$ (lower) per dimension i for each w
2. For each dimension i compute the positive mean $\mu(i+)$ and the negative mean $\mu(i-)$
3. For each component C_{ij} of vector $\bar{w}_j$ apply

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

4. Return as features for w the values $\neq 0$ (U_+, B_-) of $\phi(\bar{w})$

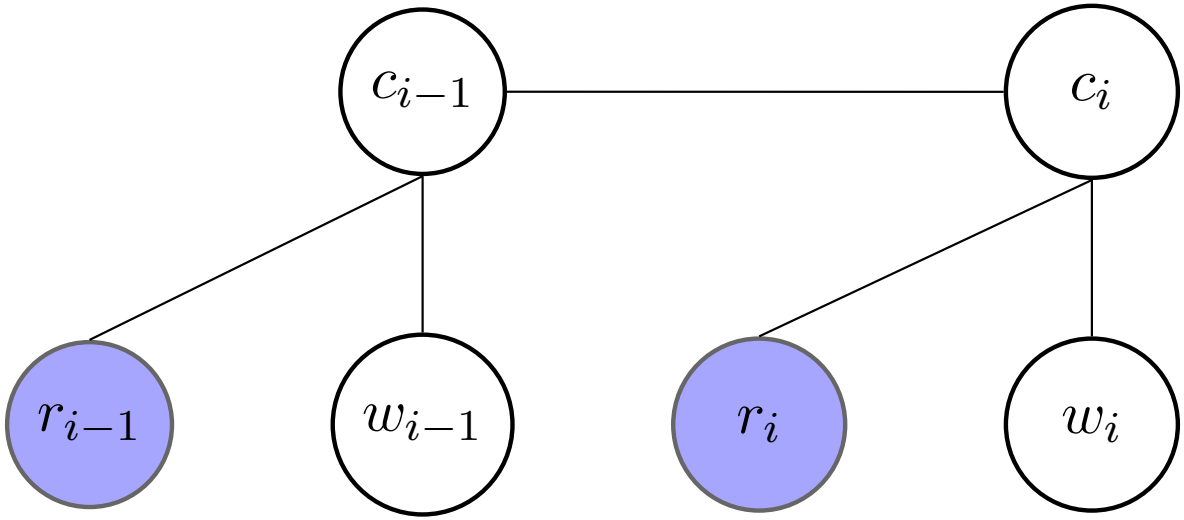
Binarized Embeddings [GCWL14] II

Dimension	Value	Binarized
1	-0.008255	0
2	-0.013051	0
3	0.145529	U+
4	0.010853	0
⋮	⋮	⋮
295	0.050766	U+
296	-0.066613	B-
297	0.073499	U+
298	-0.034749	0
299	-0.023611	0
300	-0.025693	0

(binarization for “equipo”)

NER Global Model (CRF with Distributional Features)

Linear chain-CRF with word representations as features: the upper nodes are the label sequences, the bottom white nodes are the word features in the model and the gray nodes are the word representation features



Results on the CoNLL2002 Dataset

Model	F1
Baseline	80.02%
+ Binarization	79.48%
+ Brown	80.99%
+ Prototype	79.82%
+ Clustering	80.24%
+ Clustering + Prototype	80.55%
+ Brown + Clustering	82.30%
+ Brown + Clustering + Prototype	81.19%
+ Brown* + Clustering + Prototype	**82.44%**
Carreras et al. [CMP02]†	79.28%
Carreras et al. [CMP02]	81.39%
Finkel et al. [FGM]	81.44%
dos Santos et al. [dSGa15]	82.21%
Gillick et al. [GBVS15]	82.95%
Lample et al. [LBK+16]	85.75%
Yang et al. [YSC16]	85.77%

*Brown clusters from English resource

†did not take into in account gazetteers

Discussion

- Our work outperforms dos Santos et al.'s F-score and matches also Gillick et al.'s
- It shows that cross-lingual resources (Brown clusters computed from English Wikipedia) can be used to further improve Spanish NER as
  1. entities in Spanish and English are often identical
  2. the resulting English Brown clusters for English entities correlate better with their entity types, giving rise to a better model
- While binarization improves on English NER baselines [GCWL14], this doesn't hold for Spanish (it seems that it adds noise)
- Finally, capitalization has a distinct impact on our approach: English Brown clusters, Spanish cluster embeddings and lower-cased Spanish prototypes yield 0.78% less F-score


Conclusions

- This paper has explored unsupervised and minimally supervised features for Spanish NER
  1. features based on cross-lingual word representations
  2. fed to a CRF classification model for Spanish
- Our experiments show competitive results w.r.t. state-of-the-art in Spanish NER (based on deep learning)
- 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


Thank You!

References I

[BdM+92] 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. *Comput. Linguist.*, 18(4):467–479, 1992.

[CMP02] Xavier Carreras, Lluís Màrques, and Lluís Padró. Named entity extraction using adaboost. In *Proceedings of CoNLL-2002*, 2002.

[dSGa15] Cicero dos Santos and Victor Guimarães. Boosting named entity recognition with neural character embeddings. In *Proceedings of NEWS 2015*, 2015.

[FGM] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of ACL 2005*.

[GBVS15] D. Gillick, C. Brunk, O. Vinyals, and A. Subramanya. Multilingual Language Processing From Bytes. *ArXiv e-prints*, 2015.

[GCWL14] Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. Revisiting embedding features for simple semi-supervised learning. In *Proceedings of EMNLP 2014*, 2014.

References II

[LBK+16] Guillaume Lample, Miguel Ballesteros, Kazuya Kawakami, Sandeep Subramanian, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of NAACL 2016*, 2016.

[SF02] Tjong Kim Sang and Erik F. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of COLING 2002*, 2002.

[YSC16] Zhilin Yang, Ruslan Salakhutdinov, and William Cohen. Multi-task cross-lingual sequence tagging from scratch. *CoRR*, abs/1603.06270, 2016.

[YZD+13] Mo Yu, Tiejun Zhao, Daxiang Dong, Hao Tian, and Dianhai Yu. Compound embedding features for semi-supervised learning. In *Proceedings of NAACL 2013*, 2013.