Two Models for Named Entity Recognition in Spanish. Submission to the CAPITEL Shared Task at IberLEF 2020

Elena Álvarez-Mellado[a]

[a]*Information Sciences Institute, University of Southern California*

Abstract

This paper documents two sequence-labeling models for NER in Spanish: a conditional random field model with handcrafted features and a BiLSTM-CRF model with word and character embeddings. Both models were trained and tested using CAPITEL (an annotated corpus of newswire written in European Spanish) and were submitted to the shared task on Spanish NER at IberLEF 2020. The best result was obtained by the CRF model, which produced an F1 score of 84.39 on the test set and was ranked #6 on the shared task.

Keywords

Spanish NER, CRF, BiLSTM-CRF

1. Introduction and Previous Work

Named entity recognition (NER) is the task of extracting relevant spans of text (such as person names, location names, organization names) from a given document. NER has usually been framed as a sequence labeling problem [1]. Consequently, different sequence labeling approaches have been applied to NER, such as hidden Markov models [2], maxent Markov models [3] or conditional random field [4, 5]. More recently, different neural architectures have also been applied to NER [6, 7].

In this paper, we explore two sequence-labeling models to perform NER on CAPITEL, an annotated corpus of journalistic texts written in European Spanish. The explored models are a CRF model with handcrafted features and a BiLSTM-CRF model with word and character embeddings. Both models were submitted to the CAPITEL shared task on Spanish NER at IberLEF 2020 [8].

2. Models

In this section we describe the two models that were explored for Spanish NER on the CAPITEL corpus: a CRF model with handcrafted features and a BiLSTM-CRF model with word and

Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)

EMAIL: elena@isi.edu (E. Álvarez-Mellado)
URL: https://lirondos.github.io/ (E. Álvarez-Mellado)
ORCID: 0000-0001-5952-6902 (E. Álvarez-Mellado)

character embeddings.

2.1. CRF model

A CRF model was built using `pycrfsuite`[1] [9], a Python wrapper for `crfsuite`[2] [10] that implements CRF for labeling sequential data. The model also used the `Token` and `Span` classes from `spaCy`[3] library [11].

2.1.1. Feature engineering

The following handcrafted binary features were used for the model:

- **Bias feature:** a feature that is active on every single token to support setting per-class bias weights.
- **Token feature**: the string of the token (`features[token=“agua”]=1.0`).
- **Uppercase feature** (y/n): active if all characters in the token are uppercase (*ONG*, *IBM*) (`features[is_uppercase]=1.0`).
- **Titlecase feature** (y/n): active if only the first character of the token is capitalized (*Gobierno*, *Alcalá*) (`features[is_titlecase]=1.0`).
- **Character trigram feature**: an active feature for every trigram contained in the token (`features[trigram=“agu”]=1.0`).
- **Punctuation feature** (y/n): active if the token contains any type of punctuation mark (,-/) (`features[is_punctuation]=1.0`).
- **Word suffix feature**: last three characters of the token (`features[suffix=“dad”]=1.0`).
- **POS tag feature**: part-of-speech tag of the token provided by `spaCy` library (`es_core_news_md` model) (`features[postag=“VERB”]=1.0`).
- **Digit feature** (y/n): active if the token contains a number (*2014*, *salud2*) (`features[is_digit]=1.0`).
- **Word shape**: shape representation of the token provided by `spaCy` library (`es_core_news_md` model) (`features[shape=“Xxxx”]=1.0`).
- **Lemma feature**: lemma of the token provided by the `es_core_news_md` model from `spaCy` (`features[lemma=“reír”]=1.0`).
- **Word embedding** (see Table 1).

[1] https://github.com/scrapinghub/python-crfsuite

[2] https://github.com/chokkan/crfsuite

[3] https://spacy.io/

Table 1
Embeddings used in experiments.

Type	# Vectors	Dimensions	Reference
FastText	985,667	300	[12]
FastText	1,313,423	300	[13]
FastText	2,000,001	300	[14]
FastText	855,380	300	[15]
GloVe	855,380	300	[16]
word2vec	1,000,653	300	[17]
word2vec	534,000	50	[11]

2.1.2. Model tuning

The model was trained on the training set of the CAPITEL corpus and tuned on the validation set using grid search. The following hyperparameters were explored during tuning: c1 (L1 regularization coefficient: 0.01, 0.05, 0.1, 0.5, 1.0), c2 (L2 regularization coefficient: 0.01, 0.05, 0.1, 0.5, 1.0), embedding scaling (scaling factor applied to each dimension of the embedding: 0.5, 1.0, 2.0, 4.0), and embedding type (from a set of different Spanish word embeddings; see Table 1) [12, 13, 17, 14, 11, 15, 16]. A window of two tokens in each direction was set for the feature extractor. Optimization was performed using L-BFGS. The threshold for the stopping criterion delta was set to delta = $1e-3$.

The best results were obtained with c1 = 0.01, c2 = 0.5, scaling = 0.5 and FastText Spanish embeddings [13] that were trained on the Spanish Unannotated Corpora[4] [18]. These hyperparameters produced an F1 score of 83.60 on the validation set (precision = 84.23, recall = 83.03). The lowest F1 result obtained on the validation set during grid search experiments was 78.90, which illustrates the substantial impact that hyperparameter tuning can have on the model's performance.

2.1.3. Feature ablation study

A feature ablation study was then performed on the tuned model by removing one feature at a time and testing on the validation set. The results obtained during the feature ablation study were consistently worse than those obtained with the full set of features, which demonstrates that all features contribute positively to the model's performance. The embedding feature seemed to be the one that contributed the most to the result (see Table 2).

2.2. BiLSTM-CRF model

A BiLSTM-CRF model was also trained for the same task using the CAPITEL corpus. This neural model was created using the library `NCRF++`[5] [19]. `NCRF++` is a PyTorch-based framework that implements a neural sequence-labeling model in three layers (character sequence layer, word sequence layer and inference layer). The architecture chosen for our neural NER model

[4] https://github.com/josecannete/spanish-corpora
[5] https://github.com/jiesutd/NCRFpp

Table 2
Ablation study results on the validation test.

Features	Precision	Recall	F1 score	F1 change
All features	**84.23**	**83.03**	**83.60**	
− Bias	83.78	82.89	83.29	−0.31
− Token	83.74	82.74	83.20	−0.40
− Uppercase	83.88	82.95	83.37	−0.23
− Titlecase	83.90	82.96	83.38	−0.22
− Char trigram	83.87	82.75	83.26	−0.34
− Punctuation	83.86	82.87	83.31	−0.29
− Suffix	83.90	82.90	83.36	−0.24
− POS tag	83.97	82.45	83.15	−0.45
− Digit	83.90	82.93	83.37	−0.23
− Shape	83.94	82.66	83.24	−0.36
− Lemma	83.42	82.51	82.92	−0.68
− Embedding	81.39	80.51	80.83	**−2.77**

was a character CNN + word LSTM + CRF model. This neural architecture has previously demonstrated to be succesful on other NER tasks [7].

For this model, we used FastText Spanish embeddings with 300 dimensions from [14]. The dimensions of the character embeddings were set to 50. The learning rate was 0.015, the batch size was 20. L2 regularization coefficient was 1e-08, the number of hidden dimensions was 200, dropout was 0.5, the optimization algorithm used was SGD. Iterations were set to 100, the best results on the validation set were obtained at iteration 81, which produced an F1 score of 85.04 (precision = 85.66, recall = 84.42), almost two points higher than the best result obtained by the non-neural model on the validation set.

3. Results and Discussion

3.1. Results

The best performing version of both the CRF and the BiLSTM-CRF model were then run on the test set of the CAPITEL corpus. The CRF model produced an F1 score of 84.39 on the test set (precision = 84.75, recall = 84.12, ranked #6 on the shared task), which is almost 1 point higher than the best result obtained on the validation set. The BiLSTM-CRF model, however, obtained an F1 score of 83.01 (precision = 84.33, recall = 81.82, ranked #8 on the shared task), two points lower than the best result obtained on the validation set and 1 point less than the CRF model. This gap between the neural and non-neural model (the non-neural model outperforming the neural model) can perhaps be explained by the differences between the test set and the training set, but also by the lack of tuning on the BiLSTM-CRF model (as the neural model could have benefited from a more exhaustive tuning). Either way, both results are modest compared to the best performing model of the shared task (which produced an F1 score of 90.30) [20].

3.2. Error analysis

In this subsection we will document some of the errors that were produced by the CRF and the BiLSTM-CRF model:

The tag OTHER was behind many of the disagreements in tag selection: for instance, the name of newspapers and TV channels (such as *La Vanguardia* or *TVE*) were labeled on the gold standard as ORG when they referred to the institution, but labeled as OTH when referring to the actual publication or channel (see guideline 5.4.1.7 from the task annotation guidelines [21]). These subtleties in meaning were not captured by the presented systems, as both models tended to tag these entities as ORG, regardless of the context.

Similarly, the name of countries, regions and other geopolitical units (that have sometimes been considered under GPE in other annotation schemes [22, 23]) were a frequent source of error. According to the annotation guidelines, countries were labeled as LOC or ORG depending on whether the name was referring to the political institution or the geographical location (see guideline 5.2.1. on the annotation guidelines [21]). These nuances were not well-captured by our models and also produced frequent tag selection errors.

Nested entities also produced some issues, particularly to the CRF model: *Semana de la Moda de China* was labeled as one single OTH entity in the gold standard, but was labeled as two separate entities by the CRF model (*Semana de la Moda* as OTH, *China* as LOC).

The scope of the entities was also a common source of error, as when tagging *Satélite GOES-16* (instead of just *GOES-16*).

Additionally, film and book titles (such as *La llegada* o *La peste*) should have been labeled as OTH but were consistently ignored by our models (this entities could perhaps have been captured by the CRF model had a quotation feature been included).

Finally, person names with unusual shapes were sometimes mislabeled, as in *El Bigotes*. Likewise, the actress and singer *Imperio Argentina* (that appeared once in the background set) was labeled as OTH by the CRF and as ORG by the neural model, and is a good example of the difficulty that Spanish artistic nicknames can pose to NER systems.

4. Conclusions

In this paper we have presented two different sequence-labeling models for Spanish NER: a CRF model with handcrafted features and a BiLSTM-CRF model with word and character embeddings. These models were applied to the CAPITEL corpus, an annotated corpus of journalistic texts written in European Spanish. Both models were submitted to the CAPITEL shared task on NER at IberLEF 2020. The CRF model produced an F1 score of 84.39 on the test set and was ranked #6 on the shared task, while the BiLSTM-CRF model obtained an F1 score of 83.01 and was ranked #8.

References

[1] E. F. Tjong Kim Sang, F. De Meulder, Introduction to the CoNLL-2003 shared task: Language-independent Named Entity Recognition, in: Proceedings of the Seventh Con-

ference on Natural Language Learning at HLT-NAACL 2003, 2003, pp. 142–147. URL: https://www.aclweb.org/anthology/W03-0419.

[2] D. M. Bikel, S. Miller, R. Schwartz, R. Weischedel, Nymble: a high-performance learning name-finder, in: Proceedings of the fifth conference on Applied Natural Language Processing, Association for Computational Linguistics, 1997, pp. 194–201.

[3] A. McCallum, D. Freitag, F. C. Pereira, Maximum Entropy Markov models for Information Extraction and segmentation, in: ICML, volume 17, 2000, pp. 591–598.

[4] A. McCallum, W. Li, Early results for Named Entity Recognition with Conditional Random Fields, feature induction and web-enhanced lexicons, in: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, 2003, pp. 188–191. URL: https://www.aclweb.org/anthology/W03-0430.

[5] C. Sutton, A. McCallum, et al., An introduction to Conditional Random Fields, Foundations and Trends in Machine Learning 4 (2012) 267–373.

[6] J. P. Chiu, E. Nichols, Named entity recognition with bidirectional LSTM-CNNs, Transactions of the Association for Computational Linguistics 4 (2016) 357–370.

[7] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. 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, Association for Computational Linguistics, San Diego, California, 2016, pp. 260–270. URL: https://www.aclweb.org/anthology/N16-1030.

[8] J. Porta-Zamorano, L. Espinosa-Anke, Overview of CAPITEL Shared Tasks at IberLEF 2020: NERC and Universal Dependencies Parsing, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), 2020.

[9] M. Korobov, T. Peng, Python-crfsuite, 2014. https://github.com/scrapinghub/python-crfsuite.

[10] N. Okazaki, CRFsuite: a fast implementation of Conditional Random Fields (CRFs), 2007. http://www.chokkan.org/software/crfsuite/.

[11] M. Honnibal, I. Montani, spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing, 2017. https://spacy.io/.

[12] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146.

[13] J. Cañete, Spanish Word Embeddings, 2019. https://doi.org/10.5281/zenodo.3255001.

[14] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, Learning word vectors for 157 languages, in: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[15] J. Pérez, Fasttext embeddings from SBWC, 2017. Available at https://github.com/dccuchile/spanish-word-embeddings#fasttext-embeddings-from-sbwc.

[16] J. Pérez, Glove embeddings from SBWC, 2017. Available at https://github.com/dccuchile/spanish-word-embeddings#glove-embeddings-from-sbwc.

[17] C. Cardellino, Spanish Billion Words Corpus and Embeddings, 2019. https://crscardellino.github.io/SBWCE/.

[18] J. Cañete, Compilation of Large Spanish Unannotated Corpora, 2019. URL: https://doi.org/10.5281/zenodo.3247731.

[19] J. Yang, S. Liang, Y. Zhang, Design challenges and misconceptions in neural sequence labeling, in: Proceedings of the 27th International Conference on Computational Linguistics (COLING), 2018. URL: http://aclweb.org/anthology/C18-1327.

[20] R. Agerri, G. Rigau, Projecting Heterogeneous Annotations for Named Entity Recognition, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), 2020.

[21] J. Porta-Zamorano, J. Romeu Fernández, Esquema de anotación de entidades nombradas de CAPITEL, 2019.

[22] Linguistic Data Consortium, ACE (Automatic Content Extraction) English annotation guidelines for entities, Version 5 (2005) 2005–08.

[23] R. Weischedel, S. Pradhan, L. Ramshaw, M. Palmer, N. Xue, M. Marcus, A. Taylor, C. Greenberg, E. Hovy, R. Belvin, et al., Ontonotes release 4.0, LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium (2011).