spaCy Performance on Named Entity Recognition with Code-Mixed Data

BN Project Abstract for Seminar Computational Modelling (WiSe 2023/24)

Hanxin Xia | 3417418
hanxin.xia@uni-duesseldorf.de

1 Introduction

The term "named entity" was first defined in the Sixth Message Understanding Conference (MUC-6) (Grishman and Sundheim, 1996). In the publicly available description, at least three types of language expressions were categorized as such entities: "unique identifiers" of entities (organizations, persons, locations), times (dates, times), and quantities (monetary values, percentages). The extraction of such entities in a sentence is however not a trivial job. Outside the most obvious dictionary matching (Higashinaka et al., 2012; Shang et al., 2018), several other hybrid approaches have been proposed over the years, especially for bio-medical other domain specific terms (Rocktäschel et al., 2012; Lou et al., 2020). Nowadays, with pre-trained language models spaCy, StanfordNLP as such, named entity recognition tasks can be easily carried out as a part of standard pipeline during text annotation.

This research aims to investigate the multi-purposed language model spaCy's performance on named entity recognition tasks when it comes to multilingual data, above all where two languages are used interchangeably in one sentence, the so-called insertional code-switching, as follows:

(1) It's just pretending *tengo una ventana aquí.* (Arias and Lakshmanan, 2005, 104)

The sentence starts with an English matrix sentence. The language in the embedded clause however, is switched to Spanish (italicized). We want to know how difficult it is for a monolingual spaCy model to extract named entities from such sentences.

2 Research Questions

3 Extracts from BN Proposal

Proposal 1 — spaCy performance on named entity recognition task with code-mixed data

The goal of the research is to evaluate how well spaCy performs on NER tasks when it comes to multilingual data, above all where two languages are used interchangeably in one sentence. The English-Spanish code-mixed data used here comes directly from CALCS 2018 (Computational Approaches to Linguistic Code-Switching) shared task (Aguilar et al., 2018). Since spaCy out of the box is language specific, the transformer variants of the both involved languages will be used to tag the data. The results (two versions) will be compared to the gold labels given in the CoNLL-U file to retrieve accuracy and possibly other evaluation metrics.

Possible expansion for AP One interesting type of code-switching is the so-called "insertional code-switching", where a single token or a short phrase of the embedded language L2 appears inside a sentence structure of the matrix language L1. One could imagine that at least in some cases a language model trained only on L1 would falsely identify inserted L2 elements as special names. My question is, whether there is a correlation between the length of the inserted words and the possibility of them being regarded as named entities. For this the results from above will be re-analysed. The cases of L2 insertions which are not named entities themselves will be extracted with the corresponding word lengths and the annotations returned by spaCy (namely NE == True/False). The percentage of falsely annotated insertions on each word length can then be calculated. Using a correlation test, whether there is a correlation between inserted L2 word length and error rate can then be answered easily.

Proposal 2 — NLP model performance on sentiment analysis tasks with code-mixed data

The dataset collected for this proposal is Dataset for Sentiment Analysis on Code-Mix Telugu-English Text (Kusampudi et al., 2021). The goals and procedures of this BN base research will be largely similar to the previous one except for two critical points: 1) Instead of NER results, the sentiment pipeline provided by `SpacyTextBlob` will be called up. The results are encoded as `polarity`; 2) spaCy does not support Telugu. So only the English model can be used. Fundamentally this is a text classification task but with tricky data. Two aspects will be evaluated: 1) How well spaCy English model performs on sentences with English as L1, Telugu as L2; 2) How well the same model performs on sentences with Telugu as L1, English as L2. The results from the latter condition are expected to be much worse than the first one. Because the model itself is not trained on the matrix language, it should not be surprising if the sentiment information of the "foreign language" can not be captured effectively.

Possible expansion for AP The unpredictability of the classification results using spaCy lies in that spaCy models are trained solely on monolingual data. A possible way to run a "real" monolingual analysis would be integrating multilingual word vectors in the classifier training process. The retrieving of this kind of vectors is doable using language-agnostic models like those from Smith et al. (2017), Devlin et al. (2018) and more recently Conneau et al. (2019). My design for this project is to compare the performances of each embedding model on the list on sentiment analysis to determine how well they can capture the semantics of multilingual sentences. The procedure goes as follows: First the data will be split into train and test set. The embeddings of all sentences in both sets will be retrieved using different models. Train different classification models with train embeddings using scikit-learn. Get prediction accuracies on test set embeddings. The results will be organized in tabular form to provide an overview. It will be of great interest to see whether there is a consistent increase of accuracy with embeddings generated by newer models. spaCy's performance could also be brought into comparison as the baseline. One aspect worth further considering is whether a subset of spaCy predictions should be used as the baseline. Since its monolingual nature, including predictions on data with Telugu as matrix language will certainly bring unfair disadvantages to its results. One approach would be by reducing the data to English as L1 only and let spaCy English model compete against other models trained on multilingual embeddings of the same subset of sentences.

References

Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Mona Diab, Julia Hirschberg, and Thamar Solorio. 2018. Named Entity Recognition on Code-Switched Data: Overview of the CALCS 2018 Shared Task. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 138–147, Melbourne, Australia. Association for Computational Linguistics.

Raquel Arias and Usha Lakshmanan. 2005. Code switching in a spanish-english bilingual child: A communication resource. In *ISB4: Proceedings of the 4th International Symposium on Bilingualism*, pages 94–109. Cascadilla Press Somerville, MA.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Ralph Grishman and Beth M Sundheim. 1996. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Ryuichiro Higashinaka, Kugatsu Sadamitsu, Kuniko Saito, Toshiro Makino, and Yoshihiro Matsuo. 2012. Creating an extended named entity dictionary from wikipedia. In *Proceedings of COLING 2012*, pages 1163–1178.

Siva Subrahamanyam Varma Kusampudi, Preetham Sathineni, and Radhika Mamidi. 2021. Sentiment analysis in code-mixed Telugu-English text with unsupervised data normalization. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 753–760, Held Online. INCOMA Ltd.

Yinxia Lou, Tao Qian, Fei Li, and Donghong Ji. 2020. A graph attention model for dictionary-guided named entity recognition. *IEEE Access*, 8:71584–71592.

Tim Rocktäschel, Michael Weidlich, and Ulf Leser. 2012. ChemSpot: a hybrid system for chemical named entity recognition. *Bioinformatics*, 28(12):1633–1640.

Jingbo Shang, Liyuan Liu, Xiang Ren, Xiaotao Gu, Teng Ren, and Jiawei Han. 2018. Learning named entity tagger using domain-specific dictionary.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax.