Alan Perez
AXP200075
CS4395.001

ACL Paper Summary: Detecting Unassimilated Borrowings in Spanish: An Annotated Corpus and Approaches to Modeling

The title of the paper is “Detecting Unassimilated Borrowings in Spanish:
An Annotated Corpus and Approaches to Modeling”. The authors of this study are Elena Álvarez-Mellado who is a computational linguist and is affiliated with the European University of Madrid and Constantine Lignos who is affiliated with the Michtom School of Computer Science In this paper the problem that addresses, is detecting unassimilated lexical borrowings in Spanish text. Lexical borrowings are words from one language that are introduced into another without properly adapting, leading to out-of-vocabulary words. It’s crucial for various natural language processing tasks such as parsing, text-to-speech, and machine translation. This paper presents a new annotated corpus of rich Spanish in unassimilated lexical borrowings and analyzes the performance and errors of several sequence labeling models on this task. The models evaluated include a Conditional Random Field (CRF), Bi-directional Long Short-Term Memory with CRF (BiLSTM-CRF), and Transformer-based models. The corpus, containing 370,000 tokens, is larger, more borrowing-dense, OOV-rich, and topic-varied than previous corpora available for this task. By introducing new and larger corpus of Spanish text it aims to address the problem by going past its limitations.

The author of this paper Elena Álvarez-Mellado has spent a decade working on different language technology projects with different organizations such as Information Science Institute at University of Southern California, Molino de Ideas, McLean Hospital and UNED Digital Humanities Lab. Some of the previous work includes introducing and improving chunk-based models for borrowing detection in Spanish media. Constantine Lignos expertise lies in computational linguistics and some of the papers he’s worked on include ParaNames: A

Alan Perez
AXP200075
CS4395.001

Massively Multilingual Entity Name Corpus which demonstrates the training of a multilingual model for canonical name translation to and from English. These two have a rich background in computational linguistics and research.

The unique contribution of this paper from the authors introduced a new annotated corpus of Spanish text in assimilated lexical borrowings. The text is both rich and dense and they were able to provide a comprehensive analysis of the performance of several models on the new corpus. The paper demonstrates improved performance when the BiLSTM-CRF model was fed subword embeddings. Also, the fact that the data set is rich and designed to be difficult it makes the results more reliable and informative. The way the authors evaluated their work was by comparing the performance of several sequence labeling models on the new Spanish corpus. They assessed each model using precision, recall, and the F1 score along with the error analysis to further understand the limitations of the model.

The lead author, Elena Álvarez-Mellado has 93 citations, and the last author, Constantine Lignos, has 679 citations. The reason why this work is important is that it shows a new highly annotated corpus that detects unassimilated borrowings in Spanish. The contribution is valuable in the field of natural language processing, and it shows a thorough evaluation of the models along with the performance.