Part of Speech Tagging for Low-Resource Languages

by Deniz Beser
Advised by Jordan Kodner, Mitch Marcus

Introduction

This project aims to to provide **unsupervised algorithms to extract information** from low-resource languages. Current natural language processors are predominantly supervised machine learning systems that exploit vast amount of text data available to train models. These methods are remarkably accurate on languages that have sufficient training data, such as English.

There is **insufficient human-annotated data** to train models with high accuracy for many languages, and construction of properly annotated data through human translation is costly and takes years. Lack of "big data", therefore, has become barrier for processing most languages.

Using universal linguistic properties in an unsupervised fashion, it is possible **to derive information and linguistic structures** from such low-resource languages,. The linguistic information we focused on to extract is **part of speech (POS) tags**, such as nouns and verbs, classification of words based on syntactic functions.

Test Languages

We have tested the models primarily on English and Korean, mainly due to their differences in linguistic structures. For instance, while English is heavily dependent on prepositions, Korean depends on suffixes to enhance meaning.

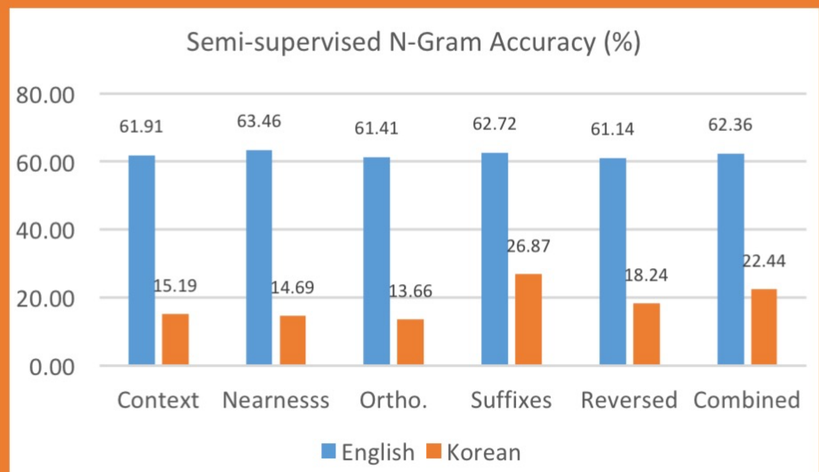
English	Korean
movie theater (topic)	영화관은 (yeonghwagwan-eun)
movie theater (object)	영화관을 (yeonghwagwan-eul)
in a movie theater	영화관에서 (yeonghwagwan-eseo)

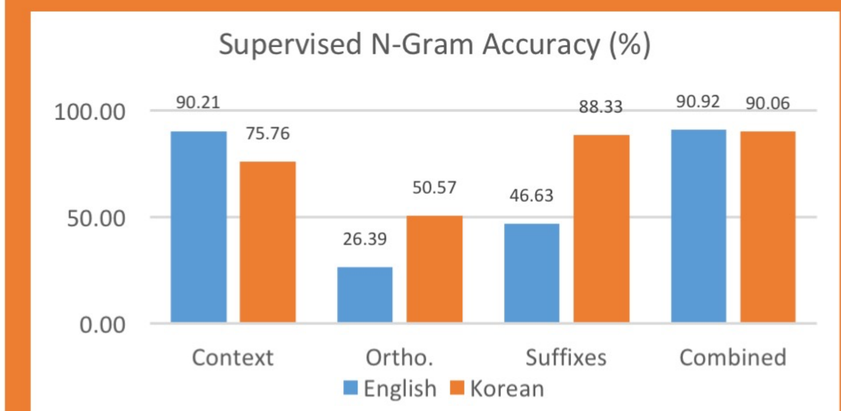
N-Gram POS Tagger

The n-gram tagging is a semi-supervised Markov model approach that focuses on the **context** of a word**;** maximizing **probability a tag appearing given the previous tags**[1]:

$$P(T|W)P(W) = p(t_0)\prod_i p(t_i|t_{i-1}, t_{i-2})p(w_i|t_i)$$

Using a small number of "seed" words, we generated a set of assignments to train the model. For unknown words, we used orthographic features, such as periods, hyphens, and suffixes.





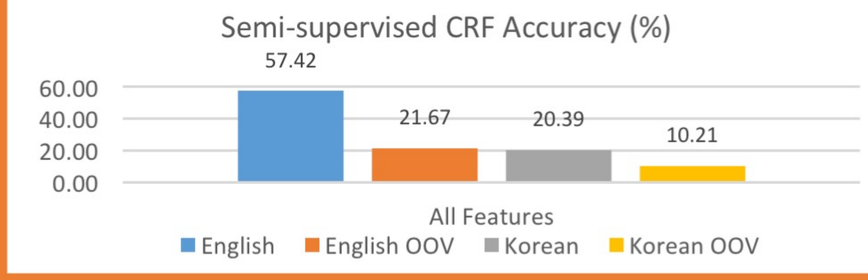
References

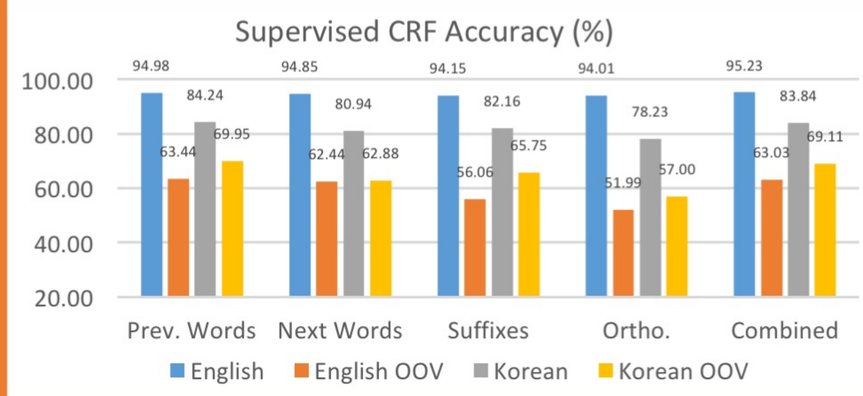
1. Meteer, M., Schwartz, R., & Weischedel, R. (1991, February). Studies in part of speech labelling. In *Proceedings of the workshop on Speech and Natural Language* (pp. 331-336). Association for Computational Linguistics.
2. Wallach, H. M. (2004). Conditional random fields: An introduction. *Technical Reports (CIS)*, 22.

Conditional Random Fields

The CRF learns weights of various **feature functions** f_j across sentences of training data to maximize the probability of **labels given the data**[2]:

$$P(t|w, \lambda) = \frac{\exp(\sum_j \sum_i \lambda_j f_j(i, w, t_i, t_{i-1}, t_{i-2}))}{Z(w)}$$





Conclusion

Based on low rates on Korean, limited **seed selection** seems to be inadequate for languages that do not have many **standalone function words**, such as determiners "the" and "a" which are frequent in English. Similarly, affix dependence causes a **high type/token ratio** in the data, weakening usefulness and accuracy of distributional information.

The **contextual linguistic properties** are not universal enough; systems that emphasize **generality over locality** by exploiting entire sentences instead of immediate context may capture more information.

Comparing supervised methods, **weighting** seems to be effective. Iterative models that can **selectively eliminate or emphasize features after determining some properties** may further improve accuracy.