Ben Christensen and Jacob Dunefsky 11 Nov 2021
LING 380 Final Project Proposal

For our final project, we plan on investigating whether generative language models contain knowledge of linguistic case across languages. We will investigate the following three questions:

1. When a language model is trained on data from a language which has morphological case (e.g. Russian, Icelandic, etc), then is the relationship between word embeddings in one case and word embeddings in another case similar for all words?
2. When two language models are trained on two separate languages with morphological case, are the relationships between cases in one language model similar to the relationships between cases in another language?
3. If a noun is modified due to its morphological case, to what extent does a model learn different features or produce results with greater accuracy when the noun is left whole or tokenized (e.g. in transliterated Russian, *menya koshka* versus *men ##ya kosh ##ka*)?

For the first question, we plan on using a pre-trained language model (likely a Transformer). We will then choose a pair of grammatical cases (c1, c2). While there are many ways to compute degrees of similarity (such as calculating variances), we will compute the mean vector *m* as $\{[[w_1]] - [[w_2]] \mid w_1 \text{ has case } c_1 \text{ and } w_2 \text{ has case } c_2\}$. Then, we will compute the mean cosine similarity between *m* and [[w_1]] - [[w_2]] for all pairs (w_1, w_2). If this mean cosine similarity is high, then we can infer that the model has developed a consistent notion of how case c1 relates to case c2. This process can be repeated for different pairs of cases.

The second question will be investigated if the first question produces a positive answer. We can then consider two different pre-trained language models, each trained on a different language with morphological case. Let $m_{i,j}$ denote the mean of all word embeddings of words belonging to case *j* in language *i*. Then, we want to see if there exists a linear transformation that (approximately) maps $m_{1,j}$ to $m_{2,j}$ for all cases *j* where the same case exists in both languages (e.g. the "accusative" exists in both Russian and Icelandic, but the "instrumental" only exists in Russian). Such an approximate linear mapping can be found with least squares. If this approximate mapping is good, then it suggests that the two language models, operating in different languages, were both able to agree on some cross-linguistic notion of these cases. It will also be interesting to observe differences in how cases are expressed. For instance, German and Turkish share the genitive case, but Turkish does not have definite articles to be modified.

The third question will be partially answered by the results of the first two, as the language datasets used will or will not tokenize words in some particular fashion. However, a separate dataset will be needed for the opposite method of tokenization. It is possible that this will be difficult to find, and that the dataset will contain paired sentence translations that differ in quality and length from the first, thus yielding unbalanced conclusions. However, results from this question will illuminate the advantages, or lack thereof, to teaching the model a little bit about general, underlying linguistic morphology before training, instead of representing each cased noun as a distinct embedding.