Leveraging English Translations for an automatic Semantic Classification of Plains Cree Verbs and Nouns

Following up on Harrigan and Arppe (2019), I present an updated semi-automated semantic classification of Plains Cree verbs and nouns. The previously presented method of automatic classification produced some valid classes, but the resulting clusters required an immense amount of manual adjustment, likely due to the sparsity of the data owing to a corpus of only ~75,000 Cree tokens.

To address this issue, I leveraged a Cree-to-English dictionary (Wolvengrey 2011) and a pre-trained English vector model of 100 billion words (Google News Vectors). Each English word in the Google News Vectors collection is represented by a 300-dimension vector, resulting from a Word2Vec model (Mikolov et al. 2013), while each Plains Cree word is defined by a set of English terms. Focusing on Plains Cree verb and noun definitions in Wolvengrey (2011), English definitions were replaced with their respective vectors from the Google News Vectors collection. All vectors for each Plains Cree word were averaged, resulting in each Plains Cree word being 'defined' by a single 300-dimesnion 'sentence' vector. This process can be seen in (1) below, where a) represents the Wolvengrey (2011) definitions, b) represents the intermediate replacement of words with their vectors, and c) represents the final averaged vector for a Plains Cree word:

(1)

```
a) awâsisihkânis: small
b) awâsisihkânis: [0.159, 0.096, -0.125, ...] [0.108, 0.031, -0.034, ...]
c) awâsisihkânis: [0.134, 0.064, -0.080, ...]
```

These vectors were then clustered using hierarchical agglomerative clustering (Ward method). Resulting clusters were manually adjusted as needed. Clustering was split by verb subclass (VII, VAI, VTI, VTA) and noun subclass (NA, NI, Dependent NA, and Dependent NI). The results of this study are compared to Harrigan and Arppe (2019). Over all, clusters produced through this method are far more cohesive with less post processing needed, especially for nouns. Implications for other Algonquian languages are discussed.

References

Google News Vectors (2013). A collection of 100 billion words found in Google's Google News dataset. Words are defined by 300-dimension vectors trained on Word2Vec. Available from: GoogleNews-vectors-negative300.bin.gz

Harrigan, A., & Arppe, A. (2019) Automatic Semantic Classification of Plains Cree Verbs. Paper presented at the 51st Algonquian Conference in Montreal, Canada, 24–27 October.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Wolvengrey, A. (2011). Cree Words. University of Regina Press.