kêtiski-kotahâskwâtam: The Effectiveness of Various Hypernymic Levels of WordNet Synsets as Vector Semantic Classification Categories

Daniel Dacanay, Antti Arppe, Jolene Poulin

University of Alberta

Despite being among the more well-documented extant Algonquian languages, current dictionaries of Plains Cree/*nêhiyawêwin* (ISO: crk) largely lack any ontologically-based semantic classifications. Semantic classifications of this nature have a wide variety of potential uses, ranging from improving digital dictionary searches, creating intelligent spell-checkers or automatic text-completion programs (Giménez et al., 2005), use as pedagogical tools (Lemnitzer and Kunze 2003), and even potential future use in machine translation (Klyueva 2007). However, traditional, manual methods of semantic classification are slow, laborious, and largely repetitive, requiring months of work to classify most large to mid-sized dictionaries (Dacanay et al. 2021a).

Given the repetitious nature of much of this task, semantic classification is well-suited to computerisation with sporadic human oversight, in particular, making use of vector semantics to assign classifications to dictionary entries within some existing ontological framework. So far, this basic process has been applied on Arok Wolvengrey's *nêhiyawêwin: itwêwina* (*Cree: Words*) (2011) dictionary using two separate semantic ontologies (Princeton WordNet and SIL's Rapid Word Collection Methodology) as classification schemes (Dacanay et al. 2021a; Dacanay et al. 2021b), with promising results compared to manual semantic classification in both. Although *nêhiyawêwin* nouns tend to have more humanlike classifications than verbs, the majority of vector classifications for both are, if not exactly correspondent to manual classifications, sufficiently semantically related to create well-defined semantic domains. However, the vector method has consistently provided overly-specific classifications (in both ontologies) for semantically general *nêhiyawêwin* words.

In general, we have seen a trend towards smaller, simpler ontologies providing (to a certain extent) superior vector semantic classifications over larger, complex ones. As such, we intend to modify WordNet so as to reduce its size, taking advantage of its linear hypo-hypernymic structure to use only broad, less-specific categories as classification labels, operating on the hypothesis that a smaller, simpler WordNet will yield more 'human-like' semantic classifications than its fully-elaborated variant.

References

Dacanay, Daniel, Atticus Harrigan & Antti Arppe. 2021a. Computational Analysis versus Human Intuition: A Critical Comparison of Vector Semantics with Manual Semantic Classification in the Context of Plains Cree. *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages*, doi:https://computelworkshop.org/wpcontent/uploads/2021/02/2021.computel-1.5.pdf

Dacanay, Daniel, Atticus Harrigan, Arok Wolvengrey & Antti Arppe. 2021b. The More Detail, the Better? – Investigating the Effects of Semantic Ontology Specificity on Vector Semantic Classification with a Plains Cree / nêhiyawêwin Dictionary. *Proceedings of the 1st Workshop on NLP for American Indigenous Languages*, doi: 10.18653/v1/2021.americasnlp-1.15

Giménez, Jesus, Lluis Marquez, & German Rigau. 2005. "Automatic Translation of WordNet Glosses". *TALP Research Center*, January 2005,

Klyueva, Natalia. Semantics in Machine Translation. 2007. *WDS'07 Proceedings of Contributed Papers*, Part I, pp. 141-44, doi:https://www.mff.cuni.cz/veda/konference/wds/proc/pdf07/WDS07_123_i3_Klyueva.pdf.

Lemnitzer, Lothar & Claudia Kunze. 2003. Using WordNets in Teaching Virtual Courses of Computational Linguistics. *Seminar für Sprachwissenschaft*, Universität Tübingen.

Wolvengrey, Arok. 2011. *Cree: Words*. 11th ed., University of Regina Press.