The More Detail, the Better? – Investigating the Effects of Semantic Ontology Specificity on Vector Semantic Classification with a Plains Cree / *nêhiyawêwin* Dictionary

Daniel Benedict Dacanay (dacanay@ualberta.ca)
Atticus Harrigan (galvin@ualberta.ca)
Arok Wolvengrey (awolvengrey@firstnationsuniversity.ca)
Antti Arppe (arppe@ualberta.ca)

University of Alberta
4-32 Assiniboia Hall,
Edmonton, Alberta, Canada T6G 2E7

Abstract

One problem in the task of automatic semantic classification is the problem of determining the level on which to group lexical items. This is often accomplished using already existing, hierarchical semantic ontologies. The following investigation explores the computational assignment of semantic classifications on the contents of a dictionary of *nêhiyawêwin* / Plains Cree (ISO: crk, Algonquian, Western Canada and United States), using a semantic vector space model, and following two semantic ontologies, WordNet and SIL's Rapid Words, and compares how these computational results compare to manual classifications with the same two ontologies.

1 Introduction

Despite the benefits and usages of semantically organised lexical resources such as dictionaries, ranging from uses as pedagogical tools (Lemnitzer and Kunze 2003) to aids for machine translation (Klyueva 2007), fully elaborated semantic dictionaries remain less common than those assembled with more routine alphabetical ordering systems. Aside from the reason of convention, one prominent dissuasive factor towards creating semantic dictionaries is the sheer amount of effort necessary to create them if their lexical content is not already organised along some ontologically principled semantic lines; the manual semantic classification of even relatively small dictionaries of this nature frequently takes months. This may be a prohibitively costly procedure in situations where resources for linguistic analysis, be they temporal or economic, are limited. Thus, a dilemma faced by the prospective compiler of a semantic dictionary is that of selecting an ontology, that is, a principled system of semantic categories, typically (but not universally) arranged hierarchically, into which lexical items may be grouped. The following investigation aims to address potential remedies to both of these limitations, with vector semantics as a first-pass alternative to manual semantic classification, and with Princeton WordNet and SIL's Rapid Words as two practical contenders for pre-existing semantic ontologies. In practice, these methods are to be demonstrated on an existing bilingual dictionary of Plains Cree (*nêhiyawêwin*), with results compared against human-made semantic classifications in both ontologies.

2 Vector Semantics

The first, and perhaps most daunting, obstacle in the process of creating a semantic dictionary (or indeed any semantically organised lexical resource) is the issue of time; even with a well-defined ontology and ample resources, manual semantic classification is a lengthy and expensive process, with teams of linguists and native speakers often requiring years to produce fully annotated semantic dictionaries (Bosch and Griesl 2017). Even with a more reduced ontology, semantically classifying an already existing full dictionary by hand takes months, and requires a thorough understanding of the chosen ontology (Dacanay et al. 2021). Although the process of manually assigning semantic categories or correspondences to

Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, pages 143–152

dictionary entries is generally not an exceptionally difficult task for a human annotator (Basile et al. 2012), the length of dictionaries, and the existence of highly polysemous lexical items, both complicate and lengthen the process of manual classification. As such, the mechanisation of the process of semantic classification assignment (or semantic annotation) appears to be one of the most direct routes to increasing overall efficiency with respect to time and resources, and to that end, the method of vector semantic classification is an alluring and well-attested alternative (Turney and Pantel 2010).

In short, vector semantic classification is a method of computationally determining the semantic similarity between any two given lexical units based on commonalities in the usage contexts of those units in large corpora. This is accomplished by representing the meaning of a lexical unit (primarily a word) as a vector in multidimensional space, which is based on the co-occurrences of this lexical unit with other lexical units in its context, followed by a reduction of dimensionality using some heuristic to result in a compact, dense vector space (typically with several hundred dimensions). Since this vector space is based on common contextual features, one may compare the multidimensional vector of one word with that of another, calculating their cosine distance to determine similarity; the closer this value is to 1, the more similar the average contexts of those two words are, and thus the more similar those words are semantically. In this way, the model functions largely on the assumptions of the Distributional Hypothesis as put forth by Firth and Harris in the 1950s (Jurafsky and Martin 2019; Firth 1957; Harris 1954), that semantic similarity begets distributional similarity, and vice versa. Vector generation is not monolithic, and various tools using various methods exist in common use, including frequency-weighted techniques such as tf-idf and Latent Semantic Analysis. In the context of this investigation, *word2vec*, a tool which makes use of prediction-based models rather than concurrence matrices to generate clusterable vector sets, has been used to generate all vectors; this decision was motivated chiefly by word2vec being readily available, easily applicable without lengthy training, and being able to leverage extensive, pre-existing pretraining on large English corpora, all advantages which largely offset the primary disadvantage of word2vec, being that it is a purely word-level vector generation tool, lacking the ability to model polysemy and contextual variances, a shortcoming which may possibly be addressed by using a sentence-level model such as BERT (see Section 5 and 7).

The vector method is not a novelty, and its utility as a practical method of semantic classification assignment has been demonstrated on numerous occasions (Brixey et al. 2020; Vecchi et al. 2017). However, useful as the method may be, in order to use vector semantics to classify entries in a dictionary, one requires a principled structure of semantic relationships into which to classify them. To this end, pre-existing semantic ontologies are a widespread and convenient solution.

3 Semantic Ontologies

Although it is possible to computationally generate sets of semantic hierarchies, the results of such attempts generally indicate that human-made, preset ontologies are preferable (Koper et. al 2015). Many such premade ontologies exist, serving a wide variety of different classificational purposes; however, we will compare here only two, being a slightly modified version of the Princeton WordNet and SIL's Rapid Word Collection Method, both popular, general-purpose ontologies intended to cover the breadth of most semantic reference in a largely language-neutral fashion. A visual representation of the structures of both is detailed in Figure 1 (see next page).

3.1 Princeton WordNet

The Princeton WordNet is one of the oldest and most widely-used semantic classification systems, originating in the 1990s at Princeton University as a hierarchically organised structure wherein contextually synonymous word-senses (or individual word-senses) are grouped into 'synsets', each of which has a hypernymic

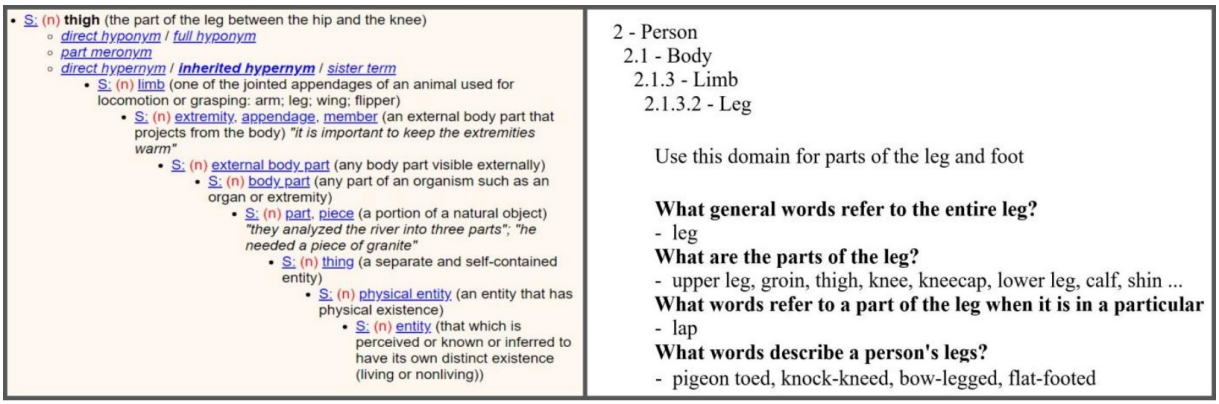

Figure 1, a visual demonstration of the differences in structure and specificity between WordNet (left) and Rapid Words (right).

synset above it in the hierarchy and possibly one or several hyponymic synsets below it (for example, the words *(n) cod#2* and *(n) codfish#1* form a synset with the definition "lean white flesh of important North Atlantic food fish; usually baked or poached"; this synset is a hyponym of the synset *(n) saltwater fish#1*, and is hypernymic to the synset *(n) salt cod#1*.). In this way, WordNet is essentially a hierarchy of hypernyms and hyponyms, with each level of hypernym and hyponym being populated by various contextually synonymous words. Although other semantic relations such as antonymy are also modelled in a 'full' WordNet, the three relations of hypernymy, hyponymy, and synonymy form the "central organizing principle" of WordNet as a whole (Miller 1993), and a structurally complete, albeit semantically basic, WordNet can be constructed using only these three relationships; in Dacanay et al. (2021) we referred to this core-level WordNet as a 'skeletal WordNet'.

3.2 Rapid Words

An alternative semantic classification scheme is the Rapid Word Collection Method of SIL, created as a framework for collecting native speaker vocabulary elicitations for dictionary creation, rather than the organisation of finished dictionaries (Moe 2003). Despite this, the structure of Rapid Words is broadly similar to that of WordNet, consisting of various numbered, hierarchically organised, roughly hyper/hyponymic semantic domains, each of which is populated by highly semantically related (although in Rapid Words, not necessarily contextually synonymous) sets of words, which may be spread across various parts of speech. Broadly speaking, these domains are less specific than WordNet synsets. There are five 'tiers' of domains in RW, with the highest being the most general (e.g. *5 Daily Life*, *7 Physical Actions*, etc) and the lowest being the most specific (e.g. *5.2.3.3.3 Spice*, *7.2.1.1.1 Run*); for our purposes, only domains on the fourth tier (or level) were used for the vector classifications (see Section 5). These semantic domains are sub-organised into specific elicitation questions, each of which has a set of potential vocabulary items in English; for example, the domain *2.1.1.5 Tooth* contains the elicitation question 'What are the parts of a tooth?', which would have with it the list of potential English answers as prompts 'enamel, root, crown, ivory'. Although not explicitly designed for it, Rapid Words has been used successfully for after-the-fact dictionary classification in the past (Reule 2018).

4 Plains Cree / *nêhiyawêwin*

Plains Cree (*nêhiyawêwin*) is an Indigenous language of the Algonquian family, spoken by ~30 000 throughout Saskatchewan, Alberta, and Northern Montana. Although slightly less critically endangered in comparison with other Canadian Indigenous languages, the majority of speakers are elderly, and intergenerational transmission remains low. Various revitalisation efforts have been undertaken in Cree communities, including bilingual education and the creation of online lexical resources (Arppe et al. 2018); however, digital resources for Cree remain limited overall. Like most Algonquian languages, Plains Cree is highly polysynthetic,

with extensive morphology, particularly on verbs, which make up the bulk of the lexicon (e.g., Wolfart 1973).

The lexical resource used for this investigation was a fully digitised copy of the database underlying *nêhiyawêwin: itwêwina*/Cree: Words (CW), a continually-updated bilingual Cree-English dictionary compiled by Arok Wolvengrey across the late 20th and early 21st centuries (Wolvengrey 2001). Consisting currently of 21,347 words with morphological notes and PoS-tagging, CW is the most extensive modern dictionary of Plains Cree, and its contents may be accessed through the University of Alberta's online Cree dictionary, *itwêwina*.

5 Method

Word vectors were obtained for every Cree entry in CW using *word2vec*, a popular off-the-shelf vector generation tool (Mikolov et al. 2013). We used the pretrained Google News Corpus, which contains 3 million word embeddings trained on 3 billion tokens. Cree word (or rather, dictionary entry) vectors were obtained as a simple, dimension-wise average of the individual English word vectors as extracted from the English definition phrases/sentences (glosses) of their respective entries, rather than the Cree words themselves, as existing Cree corpora (Arppe 2020) are too small for meaningful dimensional vectors to be obtained (Harrigan and Arppe 2019). For example, the vector for the Cree noun *mahkahk* (glossed in CW as 'tub, barrel; box') would be generated by averaging the vectors for the English words 'tub', 'barrel', and 'box', treated as a bag of words. Similarly, for the Cree verb *nâtwânam* (glossed as 's/he breaks s.t. apart; s/he breaks s.t. off by hand'), the vector would be derived from the average of the vectors for 's/he', 'breaks', 's.t.', 'apart', 'off', and 'hand'. CW noun glosses tend to be either single words or extremely curt noun phrases, and verb glosses are usually brief, utilitarian verb phrases, with no grammatical or derivational information included in the gloss itself; this fact is a further justification for using a word-level vector generation tool such as word2vec rather than a sentence-level tool like BERT, as the pieces of linguistic information on which the CW vectors are based are typically either non-sentential or highly simplistic and formulaic, seemingly making the context-sensitivity of tools such as BERT much less useful.

The Google News Corpus and word2vec were similarly used to generate the vectors for the WordNet synsets, using the head words and synset description (definitions and example sentences) as context to create the vectors, and the head word(s) of the synset as labels (Dacanay et al. 2021). For example, the vector for the synset *(n) barrel#2* (glossed as "barrel, cask (a cylindrical container that holds liquids)") would be the average of the vectors for the words 'barrel', 'cask', 'cylindrical', 'container', 'holds', and 'liquids'. The vectors for Rapid Words were created using the semantic domain levels as labels, with all example words and elicitation questions contained therein as context. For example, for the word 'barrel' in Rapid Words, which is contained in the semantic domain *6.7.7 Container*, the vector would be the average of the vectors for all of the English words in each elicitation question (i.e. "What words refer to a container", "What words refer to what is in a container", etc.), as well as all of the words listed as possible examples (such as 'container', 'vessel', 'bowl', 'pot', 'contents' etc.).

These sets of vectors were then compared against the CW vectors using cosine distance, and for every CW entry, two lists were created. For each entry on the first list (the WordNet list), all WordNet synsets were listed, ordered by cosine similarity to that entry. On the second list (the four-level Rapid Words list), for each CW entry, all Rapid Words semantic domains at the fourth tier of the hierarchy were ordered by similarity. To provide an example for the second list, even if the manually-selected RW domain for the Cree word *acihkos* ('caribou calf, fawn') was *1.6.1.1.3 Hoofed Animal*, because, on this list, the vector method would only have access to the fourth hierarchy level, the ideal, most 'human-like' vector classification would instead be *1.6.1.1 Mammal*, as this domain is at the fourth level of the hierarchy and is identical to the manual classification up to the fourth level (*1.6.1.1*). The reasoning behind limiting the RW

domains to the fourth level of the hierarchy in the vector method was threefold; firstly, tests in which the vector method was allowed to select domain classifications from any of the five levels returned notably poorer results than those which limited the choice to only one tier. (see Table 1 Any-Level (AL) columns), secondly, the fourth level of the hierarchy had the largest number of domains (at 983 compared to the fifth level with 311 and the second level with 68), and thirdly, RW did not always provide fifth level domains throughout the hierarchy. Additionally, the fourth level of the hierarchy provided a useful middleground in terms of specificity compared with the other RW levels; fourth level domains are moderately, rather than highly, specific, and thus allow for a more informative comparison with WordNet's highly specific and complex synset structure. Still, investigating whether using the most specific Rapid Words domains as labels would provide more or less accurate results than the moderately specific four-level domains would be a worthwhile avenue of future study, as would be using the individual elicitation questions as labels instead of domains.

In total, applying the vector semantic method to this end requires access to a fully digitised copy of the target dictionary (with entries and their glosses clearly delineated), access to WordNet, Rapid Words, and word2vec (all of which are freely available online), and a computer capable of both generating vectors for the dictionary entries and comparing those vectors with the pre-existing ontology vectors. To this end, a 2-core laptop with 8gb RAM is able to complete the cosine comparisons for the ~16k CW entries with the ~117k WordNet synsets in 4-5 days, and the same entries with the Rapid Words domains in no more than one and half days. On a highly parallelised computing cluster, such as ComputeCanada's Cedar (using 64 cores, each having 4-8gb RAM), performing all of the cosine comparisons takes less than 90 minutes. The computational cost of the actual vector cosine comparisons is fairly negligible, and the lengthy runtime of this operation on more basic machines is likely due to the inefficiency of retrieving each vector from large matrices.

To assess their quality, these vector classifications were compared against a gold standard of manual classifications for each entry in CW. These manual classifications were done following both WordNet and Rapid Words, with one or several synsets or RW elicitation questions assigned to each CW entry based on the meaning of the Cree word. For the WordNet classifications, the part of speech of the English WN synset was ignored; for example, the manual classification of the Cree verb *mihkwâw* ("it is red") was given in WordNet as the adjectival synset *(adj) red#1*. For Rapid Words classifications, given that RW elicitation questions do not have hard-coded parts of speech, whichever domain-internal elicitation question(s) were most semantically related to the target Cree word were used, regardless of their domain level in the hierarchy. For example, for *mihkwâw*, the question *8.3.3.3.4.3 What are the shades of red?* in the domain *8.3.3.3.4 Colors of the Spectrum* was used. More information on the manual classification method used is detailed in Dacanay et al. (2021).

	Verbs, 4L-RW top	Verbs, 4L-RW median	Nouns 4L-RW top	Nouns 4L-RW median	Verbs, AL-RW, top	Nouns, AL-RW, top	Verbs, WN, top	Verbs, WN, median	Nouns, WN, top	Nouns, WN, median
0%	1	1	1	1	1	1	1	1	1	1
10%	1	1	1	1	6	3	5	11	1	2
20%	1	2	1	1	19	7	18	51.7	2	4

30%	3	5	1	1	59	14	51.6	166.3	4	8
40%	6	15	1	2	118	23	136.8	448.8	7	16.1
50%	**15**	**36**	**2**	**3**	**222**	**36.5**	**333**	**1045**	**15**	**30.5**
60%	38	73	4	6.5	354	66	762.2	2057.3	28	60
70%	80	130.5	10	14	519	126	1633.9	4096.4	59	139
80%	161	225	24	33	717	256	3553.8	8036.9	164	375.4
90%	327	369	69	102.1	993	501	9553.8	17488.6	864.2	1670.4
100%	983	983	976	976	1739	1760	137352	137352	121883	121883

Table 1, the vector assigned ranks of manual WN and RW classifications in percentiles, for both the top-ranked manual classification and the median if there were several. '4L' indicates four-level domains, and 'AL' indicates any level of domain. Medians are written in bold.

6 Comparison of WordNet and Rapid Words Results

Statistics: Overall, although the results of both ontologies are comparable, semantic classifications using Rapid Words appear noticeably more human-like than those with WordNet, with 'human-like' here referring to how high the rank of the manual classification(s) is among the total vector classifications for each entry on average. For the vector classifications of Cree nouns, the median position of the top manual classification was 2 for four-level RW domains (with 983 possible classes) and 36.5 when the vector method was allowed to choose from any level of domain (with 1789 possible classes). For Cree verbs, the median position of the top manual classification was 15 for the four-level domains and 222 for any-level domains. In cases where there was more than one manual RW classification, the median position of the median of the multiple classes for CW nouns was 3 for four-levels, and for CW verbs, the median of the medians of multiple classes was 36 for four-levels. For the WordNet vector classifications, the median computationally selected position for the top manual classification was 15th for Cree nouns and 333rd for verbs, and the median position of the manual classifications when there were several was 30.5 for the nouns and 1045 for the verbs.

From this, it is clear that vector classifications with Rapid Words domains are, on average, much more human-like than their WordNet counterparts, being up to 22 times more accurate in the case of Cree verbs, and that limiting the vector methods' potential selections to a single, moderately specific RW hierarchy level provides much more human-like results than allowing it to select from all domains at all levels. However, it is prudent to keep in mind that even with all of its domains, Rapid Words still has far fewer potential correspondences than WordNet (1789 total RW domains (with 983 four-levels) compared to 117,659 WN synsets), and in relative terms, relevant manual classifications occur on average in a higher position proportionate to the total number of possible choices in WN vector classifications than in those with RW; with four-level RW vector classifications, the median position of the top manual classification is in the top 0.203% for the nouns (2nd out of 983) and in the top 1.53% for the verbs (15th out of 983), compared with the top 0.0127% (15th out of 117659) and 0.283% (333rd out of 117659) respectively for WN.

In general, the reduced specificity of Rapid Words, by virtue of both its inherently less detailed structure and its restriction here to a single hierarchical level of specificity, seemed to lend itself well to resolving a particular ill in the vector method, being its propensity to preferentially assign overly specific classifications to the high ranks of 'umbrella-terms', rather than the more appropriate general vocabulary. In this sense, Rapid Words semantic domains often represent concepts several steps higher in the hypernymic hierarchy than their WordNet equivalents. For example, with the WordNet classifications, the top classification for *môhkomân* (glossed as 'knife') was *(n) knife blade#1*, and the top 15 classifications consisted almost entirely of either specific types of knives or parts of knives, with the more appropriate generic term *(n) knife#1* not appearing until 18th place. By contrast, in Rapid Words, in which such specific classifications are by nature impossible at the domain-level, the top ranking classifications are more appropriately general, with the any-level list, for example, having the appropriate *6.7.1 Cutting Tool* as the top classification, and the similarly relevant *4.8.3.7 Weapon, Shoot* in second place.

The 'regift' problem: The in-built simplicity of Rapid Words also seems to have partially remedied, if not entirely solved, the so-called 'regift problem' which was prevalent in WordNet classifications; we discuss this problem in more detail in Dacanay et al. (2021), but simply put, a small number of extremely low frequency WordNet synsets occurred disproportionately frequently in the high-ranking classifications of target Cree words. The problem was so named due to such one low-frequency synset, *(v) regift#1*, being present in the top 1000 computational classifications of 65% of all Cree verbs, despite almost always being entirely unrelated semantically to the target Cree word. *(v) regift#1* is not the only WordNet entry to exhibit this behaviour, and other words, such as *(n) Rumpelstiltskin#1* occurred in as many as 72% of the top 1000 vector classifications of Cree verbs; other common regift words include *(n) Dido#1*, *(n) gumption#1*, and *(n) dingbat#1*. As a rule, these 'regift' words were both low frequency in corpora and highly specific, often being proper nouns, however, there did not appear to be any pattern in the formatting or content of these entries' glosses. The Rapid Words vector classifications also exhibited this problem to an extent; for example, subdomains of the domain *4.1.9 Kinship* occurred in the top 1000 vector classifications of CW nouns and verbs an average of ~12 times, and appeared in the top 10 classifications 33.9% and 35.7% of the time for CW nouns and verbs respectively. However, as a whole, the regift problem was markedly less notable with RW classifications of both types than with WN classifications, with both fewer different regift words (or domains) and fewer occurrences of these words/domains overall. This broadly supported our initial theory that the 'regift' problem was at least partially caused by the excessive degree of specificity in WordNet synsets overwhelming the vector method and providing it with a large number of potential classification choices with poorly defined vectors (due the low frequency of 'regift' words in the Google News Corpus) which muddy the optimal, human-like choices.

By contrast, since Rapid Words generally lacks highly specific vocabulary and is instead structured by more generic categories or 'domains', fewer of these low-frequency words are factored into the Rapid Words vectors, and these vectors are thus, in general, based on higher frequency, more contextually attested vocabulary, and are therefore (in theory) more accurate. In general, the lack of highly specific vocabulary in Rapid Words seems to contribute both to diminishing the number of semantically-related, but overly specific correspondences in the computational classifications, as well as to reducing the prominence of semantically-unrelated, overly specific 'regift' words (or in the case of Rapid Words, domains). One potential method to imitate this degree of simplicity in WordNet could involve using the hypernymic synsets of the current WordNet correspondences as labels, in essence, shifting all classifications one or more levels up in the WordNet hierarchy. This would appear to at least partially resolve the over-specificity issue (although it would do nothing to reduce the number of outright irrelevant classifications), despite incurring an obvious cost in terms of semantic richness.

Vector Content: Broadly speaking, the improved results with Rapid Words seem to be due not only to its simpler hierarchical structure and reduced level of specificity, but also due to its domain internal structure, in which domains generally include fewer irrelevant content words than WordNet synsets do. WordNet synsets frequently include example sentences in their glosses; although useful for human clarification, these inclusions inevitably lead to large amounts of semantically unrelated vocabulary influencing the respective synset vectors. As an example, the gloss for the synset *(v) drive#2* (defined as "travel or be transported in a vehicle") includes the example sentences "We drove to the university every morning" and "They motored to London for the theater". As such, the semantically irrelevant words "university", "morning", "London", and "theater" are all factored equally into the vector for *(v) drive#2* as the semantically relevant terms "drive", "motor", "vehicle". While the inclusion of these less relevant words may more accurately simulate natural linguistic contexts, the otherwise terse nature of WordNet synset glosses means that they introduce a potentially significant amount of distracting information, possibly skewing synset vectors towards the contexts of their irrelevant example sentence vocabulary rather than their relevant gloss vocabulary. By contrast, with the exception of infrequent descriptions of lexicalisation patterns, Rapid Words domains and questions contain only semantically related vocabulary, lessening potential 'distractions' for their vectors.

6.1 Utility of Results

Given the state of current results, it remains unfeasible to fully replace manual semantic annotators using the vector method; even with the best possible RW results, the vector method still only selects the most human-like classification as the top classification less than 50% of the time for Cree nouns, and less than 30% of the time for Cree verbs. Rather, the vector method in its present state seems most immediately usable as an accessory to manual classification, with the method being applied on dictionary resources as a preparatory step for manual annotators, who would then select the best classification for each entry based on the pre-generated vector classification lists. Using only the top 15 vector selected four-level RW classifications, the most human-like classification would be present on this list 50% of the time for Cree verbs, and over 70% of the time for nouns, preventing the annotator from needing to search through the entire ontology every time they wished to classify a word. In this way, present vector results are best suited as a time-saving addition to manual semantic annotation, rather than as a replacement for it.

7 Conclusion

The vector semantic method is a significantly faster and cheaper alternative to manual semantic annotation for tasks of semantic classification. However, the method is not yet capable of producing reliably human-like results across target-language parts of speech, and struggles to match natural levels of semantic specificity. To this end, using a consistent hierarchical level of a simpler, more generalistic semantic ontology, such as Rapid Words, seems to make vector semantic classifications appear more human-like, as restricting the breadth of choices available to the method as labels for correspondences seems to both reduce the number of potentially unrelated classifications and make the remaining classifications general enough that a less precise vector is necessary to generate a human-like correspondence.

Future avenues of research into dictionary vector semantics include the use of sentence-based vector generation tools such as BERT which can more accurately model polysemy, although it should be kept in mind that even a model like BERT cannot be expected to generate human-like results for dictionary glosses if those glosses are non-sentential or otherwise overly brief. It may also prove productive to experiment with the further modification of existing semantic ontologies such as WordNet and Rapid Words (such as reducing the specificity of WN by using only synsets one or several levels higher in the hypernym hierarchy as correspondences), with one of the ultimate goals of this being the integration of the results of automatic vector classifications into online dictionaries in a form which is easily navigable and understandable to an untrained user.

Acknowledgements

We would like to thank the Roger S. Smith Undergraduate Researcher Award for the generous funding they provided for the manual classification process and the early stages of this study.

References

Arppe, Antti, Atticus Harrigan, Katherine Schmirler & Arok Wolvengrey. 2018. A morphologically intelligent online dictionary for Plains Cree, Presentation conducted at the meeting of Stabilizing Indigenous Languages Symposium (SILS), University of Lethbridge, Lethbridge, Alberta.

Arppe, Antti, Katherine Schmirler, Atticus G. Harrigan & Arok Wolvengrey. 2020. A Morphosyntactically Tagged Corpus for Plains Cree**. In M. Macaulay & M. Noodin (eds), Papers of the 49th Algonquian Conference (PAC49), 49: 1-16. East Lansing, Michigan: MSU Press.

Basile, Valerio, Johan Bos, Kilian Evang & Noortje J. Venhuizen. 2012. "Developing a large semantically annotated corpus." *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, May 2012, 3196-200, doi:http://www.lrec-conf.org/proceedings/lrec2012/pdf/534_Paper.pdf.

Boerger, Brenda H. 2017. "Rapid Word Collection, dictionary production, and community well-being." *5th International Conference on Language Documentation & Conservation*, Mar. 2017, doi:https://scholarspace.manoa.hawaii.edu/bitstream/10125/41988/41988-b.pdf.

Bosch, Sonja E & Marissa Griesel. 2017. "Strategies for building wordnets for underresourced languages: The case of African languages." Literator - Journal of Literary Criticism, Comparative Linguistics and Literary Studies, vol. 38, no. 1, 31, 8, doi:https://literator.org.za/index.php/literator/article/view/1351/2294. Accessed 12 Sept. 2020.

Brixey, Jacqueline, David Sides, Timothy Vizthum, David Traum & Khalil Iskarous. 2020. "Exploring a Choctaw Language Corpus with Word Vectors and Minimum Distance Length." *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, May 2020, 2746-53.

Dacanay, Daniel, Atticus Harrigan & Antti Arppe. 2021. "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://computel-workshop.org/wpcontent/uploads/2021/02/2021.computel-1.5.pdf

Fellbaum, Christiane. 1998, ed. WordNet: *An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Firth, J. R. A Synopsis of Linguistic Theory, 1930–1955. 1957. In: Firth, J. R. 1968. Selected Papers of J. R. Firth 1952-1959. London: Logmans, 168-205.

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

Harris, Zellig S. 1954. "Distributional Structure." *Word*, vol. 10, no. 2-3, 146-62, doi:https://www.tandfonline.com/doi/pdf/10.1080/00437956.1954.11659520.

Jurafsky, Dan & James H. Martin. 2019. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd ed., 94-119

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

Koper, Maximilian, Christian Scheible & Sabine Schulte im Walde. 2015. "Multilingual Reliability and "Semantic" Structure of Continuous Word Spaces." *Proceedings of the 11th International Conference on Computational Semantics*, 15 Apr. 2015, doi:https://www.aclweb.org/anthology/W15-0105.pdf.

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

Li, Wei, Yunfang Wu & Xueqiang Lv. 2018. *Improving Word Vector with Prior Knowledge in Semantic Dictionary*. Beijing, Key Laboratory of Computational Linguistics, Peking University.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. *Efficient Estimation of Word Representations in Vector Space*. arxiv.org/pdf/1301.3781.pdf.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. "Distributed representations of words and phrases and their compositionality." *Advances in Neural Information Processing Systems*, https://arxiv.org/pdf/1310.4546.pdf.

Miller, George A. 1995. “WordNet: A Lexical Database for English”.Communications of the ACM, vol. 38, no. 11: 39-41.

Miller, George, Richard Beckwith, Christiane Fellbaum, Derek Gross & Katherine Miller. 1993. *Introduction to WordNet: An On-line Lexical Database*. Princeton University, 1-9

Moe, Ronald. 2003. Compiling dictionaries using semantic domains. *Lexikos* 13, 215-223, doi:http://lexikos.journals.ac.za/pub/article/view/731

Reule, Tanzi. 2018. *Elicitation and Speech Acts in the Maskwacîs Spoken Cree Dictionary Project*. Department of Linguistics, University of Alberta.

Tous, Ruben & Jaime Delgado. 2006. "A vector space model for semantic similarity calculation and OWL ontology alignment." *Proceedings of the 17th International Conference on Database and Expert Systems Applications*, Sept. 2006, 307-16, doi:https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.9727&rep=rep1&type=pdf.

Turney, Peter D. & Patrick Pantel. 2010. "From Frequency to Meaning:Vector Space Models of Semantics." *Journal of Artificial Intelligence Research*, vol. 37, 141-1888, doi:https://www.jair.org/index.php/jair/article/view/10640/25440.

Vecchi EM, Marelli M, Zamparelli R & Baroni M. 2017. “Spicy Adjectives and Nominal Donkeys: Capturing Semantic Deviance Using Compositionality in Distributional Spaces. “ *Cognitive Science* 41, 102–136

Wolfart, H. Christoph. 1973. "Plains Cree: A Grammatical Study." *Transactions of the American Philosophical Society, New Series*, vol.63, no. 5, Nov. 1973, 1-90.

Wolvengrey, Arok. 2001. *nêhiyawêwin: itwêwina - Cree: Words*. 11th ed., University of Regina Press.