On the Universality of Lexical Concepts

Bradley Hauer **Grzegorz Kondrak**
Alberta Machine Intelligence Institute
Department of Computing Science
University of Alberta, Edmonton, Canada
{bmhauer,gkondrak}@ualberta.ca

Translational Equivalence is Multilingual Synonymy

We posit that lexicalized concepts are universal, and thus can be annotated cross-linguistically in parallel corpora. This is one of the implications of a novel theory that formalizes the relationship between words and senses in both monolingual and multilingual settings (Hauer and Kondrak, 2020). The theory is based on a unifying treatment of the notions of synonymy and translational equivalence as different aspects of the relation of sameness of meaning within and across languages.

In prior work, the notions of senses, concepts, synonymy, and translation, are often left undefined, and theoretical assumptions about them are left unstated. Our theory provides a clear and consistent theoretical framework for reasoning about these phenomena. Building on a set of clearly formulated axioms, we are able to state and prove theorems that characterize the relationship between synonymy and translational equivalence at the level of both words and senses. Our results allow us to reassess previous approaches in terms of their consequences and implications, and make progress towards resolving open issues. Many of the formal propositions reflect unstated intuitions discernible in prior work, but their explicit statement and derivation from first principles is novel.

Our theory has important implications for lexical semantics. First, word senses are determined by word synonymy, and therefore, sense granularity cannot be substantially reduced without violating the fundamental properties of wordnets. Second, the expand model of multilingual wordnet construction has the potential to preserve those properties in a multilingual setting, at the cost of increased sense granularity. Most surprisingly, the existence of an exact matching between synsets across wordnets implies the universality of lexicalized concepts in natural languages.

Synset Properties

Wordnets, such as Princeton WordNet[1] and their multilingual generalizations, *multi-wordnets*, such as BabelNet[2], are central to our work. The basic units of their ontologies, synsets and multi-synsets, are defined using the notions of synonymy and translation, respectively. Synonymy is the relation of sameness of meaning, which can be conditional (i.e. near-synonymy) or absolute, and synsets are sets of near-synonymous words (equivalently, sets of synonymous senses), with each synset corresponding to a lexicalized concept. Synsets induce a sense inventory in which there is a one-to-one correspondence between the senses of a word and the synsets which contain the word.

We formulate the following synset properties, which we use to formally prove a variety of propositions:

1. A word is monosemous iff it is in a single synset. A word is polysemous iff it is in multiple synsets.

2. Words are near-synonyms iff they share at least one synset. Words are absolute synonyms iff they share all their synsets.

3. Word senses are synonymous iff they are in the same synset.

4. Every word sense belongs to exactly one synset.

5. Every sense of a polysemous word belongs to a different synset.

Multi-wordnets are multilingual wordnets, which consist of multi-synsets (multilingual synsets). They are constructed either by adding words from other languages to the monolingual synsets of a preexisting wordnet, or linking synsets

[1] https://wordnet.princeton.edu
[2] https://babelnet.org

from multiple wordnets in different languages. A multi-synset can be viewed as a set of words, each associated with a language, that express a single lexicalized concept. Thus, there exists a one-to-one correspondence between concepts and multi-synsets, even though not all concept distinctions are necessarily lexicalized in any given language.

As sources of lexical knowledge, wordnets and multi-wordnets are extensively used in many state-of-the-art NLP systems. In particular, they serve as the standard sense inventories for semantic tasks such as word sense disambiguation (WSD). A wordnet facilitates the enumeration of the senses of a word, by identifying the concepts associated with the synsets containing the word. A multi-wordnet further enables the enumeration of the translations of a specific sense of a word, which are the words in the corresponding multi-synset. We refer to this important property as the *multi-wordnet assumption*.

Implications

Our theory has several interesting implications. It demonstrates that word senses in wordnets are objectively determined by the relation of near-synonymy between words. Synsets are equivalence classes of synonymous senses, which represent lexicalized concepts. These concepts are discrete and disjoint. Unlike dictionary senses defined by lexicographers independently for each word, multi-synsets are induced by monolingual synonymy and translational equivalence.

Since monolingual wordnet synsets are induced by near-synonymy relations between words, the number of senses in a wordnet cannot be substantially reduced without violating the synset properties formulated above. In particular, synset property #2 implies that each non-absolute synonym word pair must involve multiple distinct word senses. As a consequence, the coarse-grained sense inventories created by clustering wordnet senses cannot be assumed to preserve the synset properties.

According to our theory, all senses that are synonymous or translationally equivalent share the same multi-synset. This implies a one-to-one mapping between synsets across languages, with lexical gaps represented by empty synsets. If we view a pair of wordnets as a bipartite graph in which nodes are non-empty synsets and edges represent the relation of translational equivalence, then every node has a degree of at most one. Since every synset represents a different lexicalized concept, a concept in one language cannot correspond to more than one concept in another language. We refer to this implication of our theory as the *concept universality principle*.

In practical terms, the concept universality principle dictates that any differences in coverage between concepts across languages must be resolved by increasing the granularity of the corresponding multi-wordnets. For example, consider the following three concepts: A) "father's brother," B) "mother's brother," and C) "aunt's husband." If one language makes a lexical distinction between A and B/C, and another language has different words for A/B and C, then A, B, and C are three distinct, universal concepts, which need to be represented by distinct multi-synsets in a multi-wordnet. While splitting monolingual synsets into multi-synsets may increase polysemy, it is indispensable to preserve the multi-wordnet assumption, which ensures that multi-synsets encode correct lexical translation pairs.

The concept universality principle, which follows logically from the fundamental assumptions of wordnets, provides a theoretical justification for avoiding bias towards English lexicalization patterns, which has its roots in the practice of founding new multi-wordnets on the synset structure of the original Princeton WordNet. Because of the lack of an accepted procedure for adding new synsets, existing multi-wordnets such as BabelNet are bounded by the set of concepts that was manually created for English. We hope that awareness of the concept universality principle will lead to the incorporation of conceptual distinctions from other languages, thus guiding the evolution of multi-wordnets away from the hegemony of English, and toward greater linguistic diversity.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Alberta Machine Intelligence Institute (Amii).

References

Bradley Hauer and Grzegorz Kondrak. 2020. Synonymy = translational equivalence. *arXiv preprint arXiv:2004.13886*.