University *of Ljubljana*
Faculty *of Computer and Information Science*

Extending SloWNet using Machine Translation and Digital Dictionary Database

Miha Bombek, Luka Dragar, David Ocepek

Abstract

For our project we propose to extend the current version of SloWNet (43,460 synsets) by translating missing synsets from the Princeton Wordnet (117,000 synsets) using the pre-trained Nvidia NeMo sequence-to-sequence transformer. We then intend to further improve the extended SloWNet by linking each individual synset with all the synonimous literals from the slovenian Digital Dictionary Database (DDD) thereby adding additional information such as use cases. Finally, we intend to manually evaluate a random subsample of the added synsets which we will then use to statistically estimate the performance of our approach on all of the added synsets.

Keywords
WordNet, machine translation, entity linking

Advisors: Assist. Prof. Dr. Slavko Žitnik, Assist. Aleš Žagar

Introduction

Wordnets are lexical knowledge graphs (KG) with many practical use cases, such as information retrieval, automatic text classification, machine translation (MT), etc. Despite their wide range of use cases Wordnet construction is a difficult task often requiring manual annotation of tens of thousands words by expert. Because of the aforementioned difficulties Wordnets are frequently built using MT models [2, 7]. For our project we will focus on extending the Slovene Wordnet (SloWNet) by translating and adding the synsets from the english Princeton Wordnet (PWN). We will then add additional information, obtained from the DDD [1], such as use cases to each synset. Under our proposed approach one can potentially almost triple the number of synsets in SloWNet while simulteneously using DDD for anomaly detection and quality control therefore utilizing already existing manually annotated Slovene resources to achieve a construction accuracy that is likely to be above that of heuristic methods that are used to construct Wordnets for low resource languages.

Corpus analysis

In general, the construction or extension of wordnets relies on leveraging the information from existing wordnets in order to guide the process and provide the basis for newly constructed synsets and their connections. As was the case during the construction of the SloWNet [2], which relied heavily on the work done on the Serbian wordnet and Princeton WordNet, which will be the primary basis for our extension. The structure of most wordnets is represented as an undirected graph connecting synsets, consisting of a group of words representing the same concept, i.e. synonyms. In the current version, the SloWNet is represented as a subgraph of the Princeton Wordnet, containing the translations of some of the english synsets and connecting them with olny a subset of all the relations present in the PWN, which consists of more than 117.000 synsets, out of which only 43.460 are present in the SloWNet. Together it contains almost 72.000, over 33.000 of them manually annotated. Besideds literals, english synsets also include a definition of the concept and in some cases one or more examples of these literals used in a sentence. While PWM contains 32.880 examples and definitions for all the synsets, SloWNet contains only 3.178 examples and 517 definitions. Besides translating existing english synsets, adding and linking new definitions and examples therefore presents one of the biggest oppurtunities for the extension of the SloWNet.

The Digital Dictionary Database (DDD) [1] is an ongoing project trying to unify the morphological structure and semantic meaning of Slovene words. Each single word in DDD is equipped with one or more senses that represent different meanings the word can have. They are interlinked, forming different structures, such as phrases, collocations and relations such as hyponym/hypernym, synonym and antonym. Senses are also equipped with definitions explaining it's meaning. Every structure also has a number of examples of its usage

in a sentence, which were taken from the Gigafida 2.0 corpus [3], a collection of written Slovene, containing newspapers, magazines, selected online texts, textbooks and other material. As such, DDD offers a great starting place for the extension of the SloWNet, because of it's similar graph like structure.

Related work

The original PWN was manually constructed in the 1995 [6]. Since then Wordnets for many other languages have been constructed semi-automatically; most Wordnets were constructed using either the expand or the merge approach [9]. Under the expand approach synsets are directly translated into the target language and if words from two different synsets map to the same word they are considered to be different senses of the word. In comparison under the merge approach, the Wordnet is built first for the target language and then the synsets are linked with the synsets from PWN. The expand approach is more common because it does not require additional resources such as parallel corpora which are frequently limited to a specific domain. In recent years, new approached have been proposed for the Wordnet merge approach, mainly focusing on the topic of more robust resource linking and word sense disambiguation [8].

Proposed solution

Machine translation

For translating synsets from the English WordNet to Slovene, we will utilize the NeMo toolkit [4], developed by NVIDIA. It includes state-of-the-art MT models, that are based on the Transformer sequence-to-sequence architecture. For our purposes, we will use the pre-trained model for Slovene-English translation [5], which was trained on approximately 33 million translation pairs.

When translating single literals, we can encounter a problem with homonymous words, since a translation of one word sense isn't necessarily valid for other senses of that word and as such while we will use literal translations as our baseline we intend to investigate the possibility of adding additional context for our MT model.

SloWNet extension

Our primary goal is to translate some of the missing synsets from WordNet and adding them to SloWnet. Afterwards we will link the synonyms from the Digital Dictionary Database (DDD) with the corresponding synsets in SloWNet. Each synset is expected to form cliques with its synonyms [8]; while the clique problem is NP-Complete many aproximate algorithms exist to solve this problem. We propose to use one of these clique detecting algorithm such as Monte Carlo Clustering. After linkage with DDD we will add any missing information relations or nodes not present SloWNet that are present in DDD as well as flag nodes or relations not present in DDD. Furthermore, we will add sense definitions from DDD and examples from the Gigafida 2.0 corpus to the linked synsets. By doing this, we will also enrich already existing synsets from SloWNet 3.1 with potential new definitions and examples.

Evaluation

Due to the prohibitively large amount of time it would take to evaluate the quality of all our added concepts we propose to instead evaluate a sufficiently large (≥ 30) random sample (S) of concepts which we will then use to estimate the performance of our approach on all our concepts. Given a concept we will calculate the precision as the ratio of words that we consider to accurately represent that concept (synset) relative to the number of all words assigned to that concept. Using the calculated precisions for the set of sampled concepts we will then calculate the bootstraped mean and standard deviation of our expected precision (1)

$$\forall s \in S : \text{precision}(s) = \frac{\textit{words belonging to synset } s}{\textit{words assigned to synset } s},$$
$$\forall \hat{S} \in \text{bootstrap(S)} : \hat{\mu}_{\text{precision}} = \frac{1}{|\hat{S}|} \sum_{s \in \hat{S}} \text{precision}(s),$$
$$\mu_{\text{precision}} = E[\hat{\mu}_{\text{precision}}], \quad \sigma_{\text{precision}} = D[\hat{\mu}_{\text{precision}}]. \quad (1)$$

References

[1] CJVT. *Digital Dictionary Database*. Accessed on 16. 3. 2023. 2023. URL: https://wiki.cjvt.si/books/digital-dictionary-database.

[2] Darja Fišer and Tomaž Erjavec. *sloWNet: construction and corpus annotation*.

[3] Simon Krek et al. *Corpus of Written Standard Slovene Gigafida 2.0*. Slovenian language resource repository CLARIN.SI. 2019. URL: http://hdl.handle.net/11356/1320.

[4] Oleksii Kuchaiev et al. "NeMo: a toolkit for building AI applications using Neural Modules". In: *CoRR* abs/1909.09577 (2019). arXiv: 1909.09577. URL: http://arxiv.org/abs/1909.09577.

[5] Iztok Lebar Bajec et al. *Neural Machine Translation model for Slovene-English language pair RSDO-DS4-NMT 1.2.6*. Slovenian language resource repository CLARIN.SI. 2022. URL: http://hdl.handle.net/11356/1736.

[6] George A Miller. "WordNet: a lexical database for English". In: *Communications of the ACM* 38.11 (1995), pp. 39–41.

[7] Roberto Navigli and Simone Paolo Ponzetto. "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network". In: *Artificial intelligence* 193 (2012), pp. 217–250.

[8] Dmitry Ustalov, Alexander Panchenko, and Chris Biemann. "Watset: Automatic induction of synsets from a graph of synonyms". In: *arXiv preprint arXiv:1704.07157* (2017).

[9] Piek Vossen, Francis Bond, and John Philip McCrae. "Toward a truly multilingual GlobalWordnet Grid". In: *Proceedings of the 8th Global WordNet Conference (GWC)*. 2016, pp. 424–431.