Automatic Word Sense Mapping from Princeton WordNet to Latvian WordNet

Laine Strankale and Madara Stāde
Institute of Mathematics and Computer Science, University of Latvia, Riga, Latvia

Keywords: WordNet, Latvian, Automatic Extension.

Abstract: Latvian WordNet is a resource where word senses are connected based on their semantic relationships. The manual construction of a high-quality core Latvian WordNet is currently underway. However, text processing tasks require broad coverage, therefore, this work aims to extend the wordnet by automatically linking additional word senses in the Latvian online dictionary Tēzaurs.lv and aligning them to the English-language Princeton WordNet (PWN). Our method only needs translation data, sense definitions and usage examples to compare it to PWN using pretrained word embeddings and sBERT. As a result, 57 927 interlanguage links were found that can potentially be added to Latvian WordNet, with an accuracy of 80% for nouns, 56% for verbs, 67% for adjectives and 66% for adverbs.

1 INTRODUCTION

WordNets are an important tool for modern linguistic research enabling in-depth semantic analysis of synonymic, hyponymic and meronymic relations between word senses in Latvian, as well as corresponding interlingual semantic relations. Additionally, it is an essential resource in other NLP tasks such as word sense disambiguation (WSD).

Until now, the focus of Latvian WordNet construction has been on manually developing a small but qualitative core wordnet. This paper aims to expand the coverage of the wordnet by automatic means. More specifically, we attempt to automatically find equivalence links between word senses in an existing Latvian language dictionary Tēzaurs with synsets in the Princeton WordNet (PWN) which allows us to transfer semantic links to Latvian and to combine Latvian word senses into new synsets thus significantly expanding the coverage of Latvian WordNet.

2 RELATED WORK

The first wordnet for English named Princeton WordNet (PWN) (Fellbaum, 1998) heralded the era of wordnet constructions. It was created manually, however since then multiple projects (Vossen, 1998)(Tufis et al., 2004) have tried to exploit semi-automatic or automatic methods and existing resources to accelerate the process.

A common approach for both initial construction and extension is to essentially copy the structure of PWN and then translate the synsets to the target language, for instance, in FinnWorNet (Lindén and Carlson, 2010) it was done by employing professional translators who translated around 200 000 senses completely manually. Open Dutch WordNet (Postma et al., 2016), Persian WordNet (Montazery and Faili, 2010), WN-Ja (Bond et al., 2008) and many other projects used existing bilingual dictionaries. The French WOLF (Sagot and Fišer, 2012) also added translation data from Wikipedia and the Slovene sloWNet (Fišer and Sagot, 2015) extracted word pairs from parallel texts.

The translation step is usually followed by a filtering step. For sloWNet and WOLF a classifier was developed that used hand-crafted features such as semantic distance and translation pair origin. Whereas an unsupervised method (Khodak et al., 2017) (further called *the embedding method*) was tested on Russian and French which used similarity metrics calculated from word embeddings to rank candidate links. The filtering step seems essential for a large coverage otherwise the translation step is limited only to a small subset of highly reliable translations.

In contrast to the copying method, core DanNet used a monolingual construction approach wherein they extracted semantic link information from an existing language resource: a dictionary (it mainly had homonym links) (Pedersen et al., 2009). Similarly RuWordNet (Loukachevitch and Gerasimova, 2019) used existing sense level translations in RuThes to

Strankale, L. and Stāde, M.
Automatic Word Sense Mapping from Princeton WordNet to Latvian WordNet.
DOI: 10.5220/0011006000003116
In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022) - Volume 1*, pages 478-485
ISBN: 978-989-758-547-0; ISSN: 2184-433X

link with PWN.

Although the monolingual approach produces linguistically higher quality results, the major disadvantage of it is that the resource cannot be used in any multilingual settings whereas after the copying approach the wordnet automatically gets linked to other wordnets in the Open Multilingual WordNet (OMW) (Bond and Paik, 2012).

Therefore, the monolingual wordnets often still require subsequent linking to PWN. In the merging of DanNet and PWN it was noted that the two resources differ significantly in both structure and vocabulary and, thus, a perfect merge is improbable (Pedersen et al., 2019). Additionally, it should be noted that the average inter-annotator agreement rate for PWN is only 71% (Palmer et al., 2004).

From this, we can conclude that any alignment technique be it done before or after initial wordnet construction cannot produce very high precision results. However, an alignment process is unavoidable if we want a highly applicable multilingual resource, therefore, we have to be careful about how the alignment is generated and used to append data to the existing wordnet.

3 CORE LATVIAN WORDNET

Given the previously outlined problems with both the copying and monolingual methods, in Latvian WordNet construction, we have aimed to combine them both.

In the first phase we are manually constructing a core wordnet of 5000 word senses. We largely base our wordnet on the sense data from a pre-existing resource Tēzaurs which is a digital compilation of legacy dictionaries maintained by the Institute of Mathematics and Computer Science of the University of Latvia (IMCS UL) and contains more than 381 000 entries (as of September 2021). In this phase we take the most popular words (as determined by parsing The Balanced Corpus of Modern Latvian), check and edit the sense inventories and add usage examples (for future WSD tasks). A particular challenge was developing a methodology for separating verb senses in a systematic but language appropriate manner(Lokmane and Rituma, 2021). These new synsets have both inner and outer links, that is, they are connected to each other and they have manually found links to PWN synsets. Inner links have types:

- hyponymy
- meronymy
- approximate synonymy (weaker than the criteria for inclusion in the synset)
- antonymy
- related words (only when the semantic relation is unclear)

PWN links have three types:

- $l_{=}$ - exact match
- $l_{<}$ - narrower than Latvian WordNet sense
- $l_{>}$ - broader than Latvian WordNet sense

Currently around 1700 Latvian synsets have PWN links, of those 74% are with the type $l_{=}$

In the second phase - the topic of this paper - we are using an automatic expansion method to copy synsets from PWN. We believe this approach allows us to maximize both quality and resources since we know that the manually-created core wordnet already includes the most common and highly-polysemous words which would have been the most problematic for automatic methods. This allows us to speed up the process without a significant decrease in the quality of the wordnet.

4 METHOD

As previously noted, the core Latvian WordNet is built on top of an existing dictionary Tēzaurs. Since we want the core wordnet and the new data to be compatible we are using the Tēzaurs sense inventory also in our extension phase.

4.1 Selection Criteria

To find the best approach for extending the Latvian WordNet we looked at three factors: (1) quality of results (precision and coverage); (2) easy of implementation; (3) resource availability for Latvian. The criteria were chosen so as to minimize the manual resources needed and account for the specifics of Latvian resource availability.

The chosen method is an adaptation of the unsupervised method which used word embeddings to construct vector representations of synsets and rank them by calculating similarity metrics. Our method is adapted in the following ways: (1) the automatic sense disambiguation step is skipped because we have access to sense inventories in Tēzaurs; (2) the vector representation of a synset uses BERT sentence embeddings in addition to word embeddings.

The method was chosen because, firstly, it doesn't necessitate or heavily rely on language resources, such as Wikipedia or parallel texts, which are poor for

Latvian, secondly, the data preparation can be largely automatized (no manual translations, checks, or handcrafted features).

Finally, there are two important points that should be noted concerning the chosen method:

(1) This work is concerned with the information that can be extracted by finding commonalities between the Princeton WordNet and Tēzaurs sense dictionary. Thus only concepts that exist in both languages, to be exact, in both resources, can potentially be found and added to the Latvian WordNet. This is a limitation but it still allows us to get a large number of word senses and, importantly, it produces synsets that are linked to a PWN equivalent, thus, making them a useful resource in future multilingual applications.

(2) PWN and Tēzaurs are sense inventories with different development principles and levels of granularity. Therefore, there is an upper limit to the precision that can be achieved with this method. As already noted, to compensate for the differences the manually-set interlingual links had three different labels.

4.2 Overview

Fundamentally, we are trying to align Tēzaurs and PWN by automatically getting all possible links between a Tēzaurs sense and a PWN synset (further referred to just as links), scoring the links using a similarity metric calculated from embeddings, and picking the best links.

This is done in three steps (see figure 1):

1. Link Generation
2. Link Scoring
3. Link Curation

4.3 First Step: Link Generation

In the first step we prepare data and generate the set of all possible links for each word sense.

4.3.1 Description

Firstly, how do we generate all possible links for a set of Tēzaurs senses? The most obvious way is to produce a list of all possible English translations for entry lemmas and match them with lemmas in PWN. We have chosen to use a combination of three translation sources: bilingual dictionary, machine translation (MT), and links between Wikipedia article titles in different languages. The hand-crafted bilingual dictionary yields the highest quality translations while MT and Wikipedia allow to extended the vocabulary significantly.

Secondly, which subset of data from Tēzaurs is worth analyzing? An alignment with PWN allows us only to extract wordnet data for words and concepts which are common across languages. Thus old and regional words should be excluded from the dataset and left for future research. Additionally, MWEs might behave differently and would necessitate additional processing, therefore, at the current stage we also exclude those.

4.3.2 Implementation

The Tēzaurs dictionary is split into entries which have lexemes (usually only one) and a sense inventory. Word senses are structured into a two-level hierarchy with main senses and their subsenses (here a subsense indicates a slight shift in meaning.) The entries often (but not always) include the POS tag. All senses have a gloss and a few have examples of use. Glosses do not always follow the same structure and some include unprocessed textual information about its origin or synonymy.

For this task we only look at single-word entries. As the word associated to the sense we choose the main lexeme with the exception of some two-lexeme adjective+adverb entries, for instance, ”energetic” and ”energetically”, where from a single sense inventory we generate two sets of sense lists. The Tēzaurs word list is further filtered down to exclude obsolete word, regional words, proper names that are specific to Latvia, slang, etc.

PWN synsets are divided by their POS: noun, verb, adjective or adverb. Only the Tēzaurs lexemes belonging to one of those categories are included. If a lexeme does not have POS information then it is found using a morphological analyzer for Latvian[1].

We translate the glosses to English using Google Translate. The lemmas are translated if possible with Tilde’s bilingual dictionary otherwise with Google Translate and translations extracted from Wikipedia entry interlanguage link data[2]. Note that a single lemma can have multiple translations.

Now comes the main step: generation of all possible *(PWN synset)-(Tēzaurs sense)* links. For each English translation t of lemma l we find all PWN synsets s with same POS that include t in their synset lemma list. To each l sense we add s as a potential link. For instance, lemma ”ceļš” has 15 translations including ”path”, ”road”, ”way” and 17 PWN synset include one of them in their lemma list like the synset *path,*

[1]https://github.com/LUMII-AILab/Webservices

[2]https://dumps.wikimedia.org/

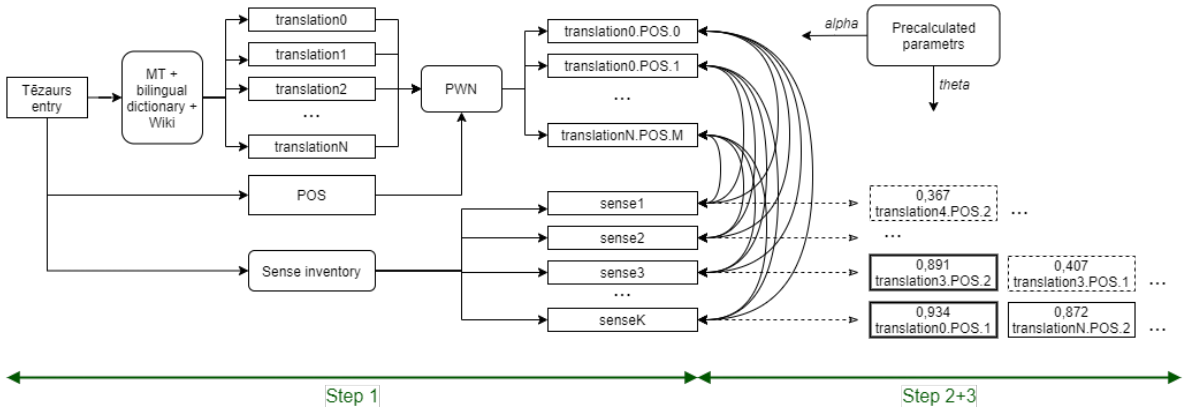

Figure 1: Schema for the extension method for a single Tēzaurs entry (word). For each word all possible English translations are found and looked up in PWN. Then all Tēzaurs senses for the given word are combined with all the found PWN synsets to produce candidate links. Finally, for each sense the final link is found by scoring all its candidate links, discarding those below a threshold θ and taking the highest (if has any).

route, itinerary - an established line of travel or access. For each 13 senses and subsenses of "ceļš" we add all the 17 synset as possible equivalence links. As can be seen correctly determining the equivalence is not trivial.

4.4 Second Step: Link Scoring

In the second step we create a vector representation for each entity, that is, each Tēzaurs word sense and each PWN synset, which allows us to score each potential link.

4.4.1 Description

To score links we have chosen to employ a vector similarity metric. The results of the original embedding method indicate that a combination of word embeddings can encompass meaningful information about a synset and thus can in our task. Additionally, due to more recent developments we have chosen to augment their technique with BERT data, which we expect could improve the rather simplistic construction of sentence embeddings in the original paper (they used a sum of word vectors).

Representation for a PWN synset rep_{PWN} is constructed as follows:

1. Calculate $v_L = \sum_{l \in L} v_l$, where L is a list of lemmas in the PWN synset
2. Calculate v_D where D is PWN synset definition (gloss)
3. Calculate $v_E = \frac{1}{|E|} \sum_{e \in E} v_e$ where E is a list of usage examples for the PWN synset
4. $rep_{PWN} = \alpha v_L + (1-\alpha) avg(v_D, v_E)$ where avg is the element-wise average and α is a pre-computed coefficient.

Representation for a Tēzaurs sense rep_T is similar:

1. L is a list with one element: entry lemma for the sense
2. D is the word sense definition
3. E is a list of usage examples for the sense (most do not have this information)
4. same calculation for rep_T

Then we use the representations rep_{PWN} and rep_T for each link to calculate its similarity score. To further interpret the scores we have chosen to use a simple ranking algorithm wherein we gather a list of all links for a single Tēzaurs sense and sort the list based on the score. The sense gets assigned the link with the highest score. This link is equivalent to the manually added interlanguage link of type $l_=$ (note: in the manual case these are synset-to-synset links but here we have sense-to-synset link; we assume that all senses that link to the same PWN synset should form a new Latvian synset).

4.4.2 Implementation

To create a PWN synset representation we calculate v_L using pre-computed word embedding resource based on the corpus data from Google News articles[3] (Mikolov et al., 2013). v_D and v_E is calculated using sentence-BERT (sBERT) (Reimers and Gurevych, 2019)[4] and the pre-trained BERT model all-MiniLM-L12-v2[5]. The Tēzaurs sense representations are cal-

[3]https://code.google.com/archive/p/word2vec/

[4]https://github.com/UKPLab/sentence-transformers

[5]https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2

culated similarly except we use the English translation we obtained in the first step and not the original Latvian.

For each link we calculate a lemma similarity score and a definition similarity score via a simple vector dot product.

4.5 Third Step: Link Curation

In the third step we determine the most equivalent PWN synset for each Tēzaurs word sense and further filter the results.

4.5.1 Description

In the previous step we assumed that all the highest scoring links are valid $l_{=}$ links. This is not the case because it is possible that, firstly, no such link exists or, secondly, our step 1 did not generate the valid link as one of the possibilities since word translations can be lacking especially for less common word senses. Therefore, we calculate and use a score threshold θ below which all links are considered invalid.

The calculation for θ as well as α (from Step 2) requires a correctly labeled dataset of Tēzaurs-PWN links. When parameters are calculated we can use them to directly generate results for the complete wordlist.

4.5.2 Implementation

To obtain the coefficients α and β we extract a dataset of interlanguage links from the manual Latvian WordNet. Note that the core word list differs from the word list used in this extension phase since here we are working with rarer words. However, this is the only available labeled dataset and manual linking is time consuming.

We process these senses as detailed in the first and second steps to obtain the lemma and definition similarity scores for each. Then we calculate the final scores with gradually incremented values of α, choose the highest-scoring synset and check whether it matches with the one indicated in the core Latvian WordNet. The parameter that yields the highest precision are further used for the rest of the data set.

The score threshold - the cutoff point under which we considered that the link does not represent a valid equivalence - is calculated by maximizing the F_1 metric and looking at thresholds in the range $[0,1]$ incremented by 0.01. We used the same test data set from the core Latvian WordNet to get the final value.

When all parameters are determined we run the ranking algorithm for all word senses in the data set, get the highest scoring and discard the link if its score is below the threshold.

Finally, the new links are used to create new Latvian synsets in Tēzaurs.lv by combining senses that link to the same PWN synset. In addition for each link we also save the generated score, which allows future users of the Latvian WordNet to filter data by the precision level which is desired.

5 EVALUATION OF RESULTS

WordNet evaluation methods vary widely which makes it difficult to have cross-wordnet comparisons. Some results only make sense in the context of the language, the specific method chosen and their initial staring point, and there is no standardization of evaluation methods. Therefore, we have chosen to mostly focus on evaluating our results in the context of Latvian and our specific needs.

We evaluate our results by looking at the coverage (total link count) and the precision (how many of the generated links are valid $l_{=}$ links).

In evaluation we are using two different data sets. Firstly, we are comparing it to the core Latvian WordNet, which has high quality, independently chosen PWN links. Secondly, we are taking a random sample of 400 produced links (100 of each POS) and manually checking whether they are valid. The two data sets were chosen to show the method's performance on data sets of differing complexities which gives a better sense of the real precision (see figure 1).

Table 1: Average word polysemy (including words with one sense) in each Princeton WordNet, core Latvian WordNet, and the Tēzaurs wordlist used for automatic extension.

POS	PWN	Core Latvian WordNet	Our wordlist
Noun	1.24	3.18	1.26
Verb	2.17	5.99	2.23
Adj	1.40	4.89	1.72
Adv	1.25	2.92	1.93

Latvian and English differ linguistically in how words are formed and used depending on the POS. Therefore, we evaluate each POS separately. Additionally, this lets us avoid the issue wherein the more common POS (noun) or more polysemous POS (verb) skew or occlude the results when viewed in aggregate.

5.1 Evaluation against Core Latvian WordNet

The details of the extracted test set can be seen in table 2. Significant portion of those links are of types $l_>$ and $l_<$. We have chosen to exclude those from our evaluation since our method aims to find $l_=$ links.

Table 2: Interlanguage link counts in the test set extracted from the core Latvian WordNet.

POS	$l_=$	$l_>$ or $l_<$
Noun	1144	402
Verb	495	310
Adj	134	95
Adv	101	23

At first we compare the results with the data set from core Latvian WordNet. Here we look at the link rankings produced in step 2 and measure whether the correct link appeared in the top 1, 3 or 5 highest scoring links. As seen in table 3 the precision is the highest for nouns and lowest for verbs, as we would expect. Given that, for instance, nouns have a median of 22 candidates per sense and verbs 44 candidates, the top 3 and top 5 metrics are significant and indicate that although the method is not powerful enough to distinguish between all those cases, the data could be useful if further processed.

In addition we experimented with a setup where the vector representations were calculated entirely using word embeddings (more similar to the setup in the original embedding method) and as can also been seen in Table 4, the results are better across all POS.

Table 3: Precision of the generated data after the application of the threshold θ when compared to the links in core Latvian WordNet.

POS	Top 1	Top 3	Top 5
Noun ($\alpha = 0.34$)	49.0%	65.7%	68.9%
Verb ($\alpha = 0.50$)	37.3%	46.5%	47.2%
Adj ($\alpha = 0.59$)	39.5%	50.8%	52.4%
Adv ($\alpha = 0.86$)	47.4%	64.9%	68.0%

Table 4: Precision comparison for two methods: the original method that uses only word embedding and our method that supplements them with sBERT.

POS	Only word embeddings	With BERT (our method)
Noun	49.0%	43.2%
Verb	37.3%	29.1%
Adj	39.5%	27.8%
Adv	47.4%	38.5%

5.2 Evaluation of Errors

To helps highlight any discrepancies in the automatically generated links and make the necessary adjustments we manually evaluated a subset of 100 samples, which were distributed evenly throughout the four main lexical categories: nouns, verbs, adjectives, and adverbs (25 samples in each).

The automatic links of adjectives and verbs have been the most difficult to form, probably due to more specific and distinct meanings that are less interchangeable with each other and more situationally used than other parts of speech.

5.2.1 Nouns

In 11 samples the manually selected link matched the first choice of automatically generated links. In one case the manually selected link matched third-best choice. In some cases, the system failed to differentiate between more general and specific notions, for example, by selecting “morality” (a set of perceived values) instead of “moral” (a lesson). A similar tendency could be seen in the Latvian term “frontier”, for which the system instead had selected the semantically broader term “boundary”. In other cases, however, the algorithm succeeded at selecting more appropriate and nuanced links, surpassing the manually selected data. This was seen in the following pairs: “mother” – “ma”; “poetry” – “verse”, where the first option was manually selected, whereas the latter was more semantically appropriate and selected by the system.

5.2.2 Verbs

In 11 samples the manually selected link matched the first choice of generated links, in two cases it matched the third-best choice. Most discrepancies were connected to verbs describing verbal exchange of information, e.g. “say”, “tell”, “assure”, “verify”. This could indicate that a broader set of samples is necessary to identify, separate and correctly link the more nuanced notions of verbal communication, which are slightly different in Latvian and English. Forming links for verbs describing the production of sound also proved to be problematic, especially when dealing with figurative meanings, as in “sing” (when talking about instruments, not people).

5.2.3 Adjectives

In 10 samples the manually selected links matched the first choice of automatically generated links. In four samples it matched second-best choice, and two

matched third or lower options. The best results were yielded by less ambiguous adjectives the meaning of which is not particularly nuanced, e.g. "Olympic" or "central", whereas the adjective "dear" proved to be unexpectedly challenging, as the system could not differentiate between the financial and sentimental meanings of this term. On the other hand, the system could successfully differentiate between the closely similar terms "accomplishable" and "achievable" and the link it generated matched the manually created one.

5.2.4 Adverbs

In 14 samples the manually selected links matched the first choice. In four samples the manually selected link matched second-best choice. In the case of adverbs, the system generally seems to favour uncommon terms with narrower, more specific meanings, for example, by selecting "afresh" instead of "again" or "synchronously" instead of "simultaneously". As expected, the system also faced some difficulty selecting the right option for ambiguous Latvian terms that are highly situational, for example "reiz" (once) and "reiz" (finally), but it should be noted that such meanings are the primary reason for using specialists in combination with automatic linking.

5.3 Evaluation of All Generated Links

We have shown an evaluation against the core wordnet data. However, in reality we are interested in the performance on the larger Tēzaurs wordlist, which contains more Latvian-specific and less common words. The full extension data was evaluated with the 400 link test set and as can be seen in table 5 has significantly higher precision, probably, mostly due to reduced polysemy levels.

The final extension step was to remove highly dubious and invalid links. We used a simple threshold metric which as shown in table 5 was effective at reducing the proportion of invalid links in our resulting data set. The link counts before and after this step can be seen in table 6 and, as expected, there is a significant decrease in the wordnet link counts, however, given our experience from the core wordnet development we know that we should not expect all Latvian senses to have a perfect alignment in PWN. A careful manual evaluation of the links in our sample supports this.

It revealed that we are mainly dealing with three types of invalid links: (1) the Latvian sense does not have an equivalent in PWN (wordlist selection failure); (2) the full candidate list for a sense does not contain the valid PWN link (translation failure); (3) the full candidate list contains the valid candidate but it is not the highest scoring link (scoring failure). The most common type, especially for verbs, was the first and second.

It could be possible to further clean-up our wordnet by developing additional heuristics about which words are unlikely to have a PWN link or a good translation. However, we leave this as a research direction for future work.

Table 5: Precision of the generated data before and after the application of the threshold θ as determined by manual evaluated of 400 links.

POS	Precision (before θ)	Precision (after θ)
Noun	52%	80%
Verb	35%	56%
Adj	49%	67%
Adv	47%	66%

Table 6: Count of the generated links before and after the application of the threshold θ.

POS	Count (before θ)	Count (after θ)
Noun ($\theta = 0.49$)	51 487	28 644
Verb ($\theta = 0.54$)	35 181	20 667
Adj ($\theta = 0.55$)	10 828	7 609
Adv ($\theta = 0.56$)	3 667	1 007
Total	101 163	57 927

Our method achieves results that are similar to those with more complex methods and for well-resourced languages. The original embedding method used on Russian (we chose to compare to Russian as opposed to French since its language characteristics should be more similar to Latvian) achieved 73.4% precision with approx. 51 000 synsets. However, their precision metric excludes all cases where the correct synset was not in the generated list of synsets (in our case 20%). WOLF found 55 159 pairs which in manual evaluation and before thresholding and clean up where 52% correct, after clean up 81% (almost 80% of the new WOLF is nouns). sloWNet used a similar method and achieved 25% initial precision and 82% after clean up. Finally, a method applied to ajz, asm, arb, dis and vie had an output of on average 53 000 and the average precision evaluated on a 5-point scale was 3.78 (Lam et al., 2014)

6 CONCLUSIONS

In this paper we have described the method used for the automatic extension of Latvian WordNet. First, we outlined the current state of core Latvian WordNet, a manually constructed high-quality wordnet of 5000 word senses for the most common words which

contains both interlanguage links and links to PWN. Then we described the automatic extension technique used to increase the coverage of Latvian WordNet. In it we attempt to align the Latvian dictionary Tēzaurs with PWN by, first, translating senses to English, second, constructing a vector representation for each Latvian sense and PWN synset using word embeddings and sBERT, and, third, score links and filter them using a threshold.

The results were evaluated in terms of precision and coverage by, first, comparing them to core Latvian WordNet and, second, manually checking a sample of 400 senses. Ultimately we found 57 927 new sense-synset pairs with precision of 80% for nouns, 56% for verbs, 67% for adjectives and 66% for adverbs.

The focus of this paper was on how to extract information from other wordnets, namely, PWN. However, as mentioned in Section 2, it is also possible to extract interlanguage link information from a resource in the same language, if such a resource exists. Tēzaurs sense glosses contain some textual information about word formation and synonyms. However, this data has not yet been processed and the existing errors fixed, therefore, we have chosen to exclude it for now but it is a fruitful future research direction which could be explored.

We have shown that automatic extension of a wordnet requiring only a target-language dictionary and translation resources is possible. Our results will be added to the current core Latvian WordNet and merged into the online resource Tēzaurs, where they will be available also to the public. Additionally, the data will be used to develop word sense disambiguation (WSD) capabilities for Latvian.

ACKNOWLEDGEMENTS

This research work was supported by the Latvian Council of Science, project "Latvian WordNet and word sense disambiguation", project No. LZP-2019/1-0464.

REFERENCES

Bond, F., Isahara, H., Kanzaki, K., and Uchimoto, K. (2008). Boot-strapping a wordnet using multiple existing wordnets. In *LREC*.

Bond, F. and Paik, K. (2012). A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 64–71.

Fellbaum, C. (1998). Wordnet: An electronic lexical databasemit.

Fišer, D. and Sagot, B. (2015). Constructing a poor man's wordnet in a resource-rich world. *Language Resources and Evaluation*, 49(3):601–635.

Khodak, M., Risteski, A., Fellbaum, C., and Arora, S. (2017). Automated wordnet construction using word embeddings. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 12–23.

Lam, K. N., Al Tarouti, F., and Kalita, J. (2014). Automatically constructing wordnet synsets. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 106–111.

Lindén, K. and Carlson, L. (2010). Finnwordnet–finnish wordnet by translation. *LexicoNordica–Nordic Journal of Lexicography*, 17:119–140.

Lokmane, I. and Rituma, L. (2021). Verba nozīmju nošķiršana: teorija un prakse verb sense distinction: theory and practice. *Valoda: Nozīme un forma 12. Rīga: LU Akadēmiskais apgāds*.

Loukachevitch, N. and Gerasimova, A. (2019). Linking russian wordnet ruwordnet to wordnet. In *Proceedings of the 10th Global Wordnet Conference*, pages 64–71.

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

Montazery, M. and Faili, H. (2010). Automatic persian wordnet construction. In *Coling 2010: Posters*, pages 846–850.

Palmer, M., Babko-Malaya, O., and Dang, H. T. (2004). Different sense granularities for different applications. In *Proceedings of the 2nd International Workshop on Scalable Natural Language Understanding (ScaNaLU 2004) at HLT-NAACL 2004*, pages 49–56.

Pedersen, B. S., Nimb, S., Asmussen, J., Sørensen, N. H., Trap-Jensen, L., and Lorentzen, H. (2009). Dannet: the challenge of compiling a wordnet for danish by reusing a monolingual dictionary. *Language resources and evaluation*, 43(3):269–299.

Pedersen, B. S., Nimb, S., Olsen, I. R., and Olsen, S. (2019). Merging DanNet with Princeton Wordnet. In *Proceedings of the 10th Global Wordnet Conference*, pages 125–134, Wroclaw, Poland. Global Wordnet Association.

Postma, M., van Miltenburg, E., Segers, R., Schoen, A., and Vossen, P. (2016). Open dutch wordnet. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 302–310.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Sagot, B. and Fišer, D. (2012). Automatic extension of wolf. In *GWC2012-6th International Global Wordnet Conference*.

Tufis, D., Cristea, D., and Stamou, S. (2004). Balkanet: Aims, methods, results and perspectives. a general overview. *Romanian Journal of Information science and technology*, 7(1-2):9–43.

Vossen, P. (1998). Introduction to eurowordnet. In *EuroWordNet: A multilingual database with lexical semantic networks*, pages 1–17. Springer.