Classifier-based Polarity Propagation in a Wordnet

Jan Kocoń, Arkadiusz Janz, Maciej Piasecki
Faculty of Computer Science and Management
Wrocław University of Science and Technology, Wrocław, Poland
{jan.kocon, arkadiusz.janz, maciej.piasecki}@pwr.edu.pl

Abstract

In this paper we present a novel approach to the construction of an extensive, sense-level sentiment lexicon built on the basis of a wordnet. The main aim of this work is to create a high-quality sentiment lexicon in a partially automated way. We propose a method called Classifier-based Polarity Propagation, which utilises a very rich set of wordnet-based features, to recognize and assign specific sentiment polarity values to wordnet senses. We have demonstrated that in comparison to the existing rule-base solutions using specific, narrow set of semantic relations, our method allows for the construction of a more reliable sentiment lexicon, starting with the same seed of annotated synsets.

Keywords: sentiment analysis, polarity propagation, wordnet

1. Introduction

Sentiment analysis of natural language utterances is continuously increasing its importance as one of the most expected techniques. The best results have been obtained with supervised approaches trained on the basis of annotated texts from a selected domain, e.g. movie or product reviews. However, cross-domain applications show a significant drop in the performance of classifiers trained on a corpus built from texts from the other domains. This can be attributed to a high correlation of a classifier with words and phrases that are specific for the positive and negative utterances of the given domain. However, language expresses some lexical means of conveying sentiment polarity in a way that is shared across different domains. A compromise between performance and domain adaptability can be achieved using hybrid methods. A lexicon of sentiment polarity could be a very useful basis for constructing such a domain independent, hybrid system, if such a lexicon is large, comprehensive and reliable enough.

plWordNet 3.1 emo[1] (Maziarz et al., 2016; Zaśko-Zielińska and Piasecki, 2018) is a very large lexical semantic network for Polish, in which more than 190,000 different lemmas and 285,000 Polish lexical meanings are described by the lexico-semantic relations. plWordNet has become one of the largest Polish dictionaries ever built, and the largest wordnet in the world. A substantial part of plWordNet was manually described by emotive annotation (Zaśko-Zielińska and Piasecki, 2018). In this work we would like to expand this emotive annotation to a very large scale and make plWordNet a basis for a large hybrid emotive lexicon of Polish, as well as for the development of a hybrid system for sentiment and emotion analysis in Polish texts. Existing solutions for sentiment propagation over a wordnet are mostly based on a simple label propagation algorithm, starting with a relatively small initial seed. However, generally they do not take into account more complex wordnet structures, thus we may lose some information which can be a good indicator of sentiment polarity. Exploiting the full wordnet structure gives us an opportunity to capture the polarity of senses in a more accurate way. We may want to consider not only a wider network context of a word sense, but also a richer set of lexical relations to propagate the sentiment polarity. Yet another problem with the existing approaches is that many solutions depend on handcrafted propagation rules that cannot be easily transferred to a wordnet built for another language. Here we propose a method which allows for automated discovery of propagation rules by using the wordnet structure in a more extensive way to recognise the sentiment polarity of senses.

2. Related Works

SentiWordNet, one of the most commonly used sentiment resources for English, was introduced in (Esuli and Sebastiani, 2006). The main goal of the authors was to construct a large lexical resource with sentiment polarity assigned to meanings, rather than words[2]. There were many attempts to construct sense-level sentiment lexicons, but most of them were evaluated only for English. The easiest way to create a sense-level sentiment lexicon for another languages is simply to map SentiWordNet annotations to a non-English wordnet via existing mappings between the two wordnet or even translating first Princeton WordNet (Fellbaum, 1998) into another language. However, wordnets for different languages may differ significantly, e.g., in the number of relation instances and a different semantic structure. One of the most commonly used techniques for this task is relational label propagation using a random walk algorithm.

(Vossen et al., 2008) compared a lexicon constructed by applying a simple polarity transfer from SentiWordNet and a lexicon built by label propagation on Dutch WordNet (Vossen et al., 2013). The simple transfer of annotations resulted in a general decrease of performance in comparison to SentiWordNet. The second approach was based on random walk with propagation rules exploiting a narrow set of lexical relations from Dutch WordNet. The authors noted, that the factors such as the seed size, seed composition and the number of iterations had a great impact on propagation performance. Thus, they evaluated their approach on

[1] http://plwordnet.pwr.edu.pl

[2] However, the annotation in SentiWordNet is still done on the level of synsets, not individual word senses.

three datasets of different quality: high-quality, low-quality and mixed. The best results were achieved using the largest dataset of mixed quality, derived from the General Inquirer (Stone, 1966) – a sentiment lexicon. The conclusion was, that the size is the most important factor. The authors also proposed a third approach combining this transfer method with label propagation, with almost the same result. The results may also suggest, that simple transfer methods are not perfect, but combining multiple approaches with transfer methods may bring us promising results.

The authors of (Maks et al., 2014) expanded research on the sentiment propagation to non-English wordnets. They applied the same propagation method to five wordnets of different languages. Words and their polarity were acquired from the well-known sentiment lexicon – the General Inquirer Lexicon, and then translated with a machine translation service to five languages. The words were manually mapped to their corresponding synsets in particular wordnets, and used as a seed for propagation. The resulting lexicons varied significantly with respect to their size and precision. The authors concluded, that the way a given wordnet has been built seems to affect the propagation performance. Thus, we should not apply the same propagation scheme for every wordnet.

(Mahyoub et al., 2014) was the first attempt to build an Arabic sentiment lexicon on the basis of Arabic WordNet (Black et al., 2006). They introduced two steps in their procedure: the *expansion step* – a sentiment lexicon is expanded by iteratively reaching concepts of the wordnet; the *scoring step* – the sentiment score of the reached concepts is computed according to their distance from the seeds. A task-based evaluation was applied to evaluate this solution. The acquired polarity scores were incorporated into features for the sentiment classification task, next evaluated on the Arabic corpora.

There were several attempts to construct a large sentiment lexicon for Polish in an automated way e.g. (Haniewicz et al., 2013), (Haniewicz et al., 2014). (Haniewicz et al., 2013) attempted to build a polarity lexicon from web documents. They utilized an older version of plWordNet, so still without sentiment annotation, as a general lexical resource in order to develop domain-aware polarity lexicons. plWordNet was utilized to identify semantic relations between the acquired terms. To determine their polarity, a supervised learning with Naive Bayes and SVM was applied. This approach was extended in (Haniewicz et al., 2014), where the semantic lexicon was expanded up to 140,000 terms, using simple rule-based propagation method based on an adaptation of the random walk algorithm.

SentiWordNet construction in its recent stages was generally based on glosses from Princeton WordNet. (Misiaszek et al., 2013) proposed a lexicon construction method for wordnets, in which a simple transfer method could not be easily applied, or external sources of knowledge, such as tagged and disambiguated glosses, are not available. They used relational propagation scheme with local, collective classification method to determine polarity of a synset. The training features for the classifier were obtained using only a close neighbourhood of annotated synsets, consisting of nodes with known polarity. They manually annotated specific synsets in a wordnet and used them as seeds for the propagation process. However, the details of the extraction of features were not specified, and there was no evaluation for their approach.

In (Kulisiewicz et al., 2015) the propagation was performed by using an adaptation of Loopy Belief Propagation (*LPA*) on Princeton WordNet 3.0. Three different variants of the *LPA* have been tested and evaluated. First, the authors compared their results with polarity scores from SentiWordNet (*Mean Square Error*), but without the *Objective* class. Second evaluation was a comparison with polarity of words existing in the General Inquirer Lexicon. The resultant performance was ambiguous and the main conclusion was, that semantic relations within wordnet may not be well correlated with the sentiment relations.

3. Emotive Annotation in plWordNet

3.1. plWordNet model

plWordNet in brief, follows generally the main ideas of Princeton WordNet(Fellbaum, 1998), consists of *lexical units* linked by *lexico-semantic relations* and grouped into *synsets*. A lexical unit (LU) represents a lexical meaning and is a triple: lemma, Part of Speech and sense identifier. Contrary to WordNet, LUs, not synsets, are the basic building blocks of plWordNet. Use examples are in a natural way assigned to LUs, as well as glosses. Lexico-semantic relations are defined by detailed guidelines including substitution tests and referring to the use examples that can be observed in text corpora. Moreover, plWordNet is developed by a corpus-based wordnet development method in which corpus exploration and the work with examples of the use of different lemmas and their potential senses (not synsets or concepts shared in lexical meanings) are crucial for wordnet editing. Finally, it is also worth to notice that the construction of plWordNet follows in this aspect the long term tradition of the lexicography.

3.2. Emotive annotation scheme

Thus, following this fundamental construction decisions, *emotive annotation* in plWordNet has been also defined on the level of LUs. LUs are the natural targets of the emotive annotation which is strongly associated with the use of LUs. Initially, as a result of a pilot project (Zaśko-Zielińska et al., 2015), emotive annotations have been manually added for a selected subset of more than 31,000 LUs in plWordNet 2.3 emo. LUs were described, see also (Zaśko-Zielińska and Piasecki, 2018), by:

- *markedness*,
- *intensity* of sentiment polarity,
- *basic emotions*,
- *fundamental human values*,
- *usage examples*.

The annotation goes beyond a typical sentiment polarity annotation and that is why it is called *emotive*.

First of all, LUs are dived into *neutral* vs *marked* with respect to sentiment *polarity*.

In the case of the *intensity*, we assumed a rather modest scale for sentiment polarity of five grades, namely: *strong* or *weak* vs *negative* and *positive*, plus *neutral* LUs in the middle of the scale. We keep the number of grades limited, as the annotation is performed by two annotators per one LU. Each LU is annotated by a linguist and a psychologist. The work of annotators is controlled and verified by a super-annotator and is based on a strict lexicographic procedure and detailed guidelines, see (Zaśko-Zielińska et al., 2015; Zaśko-Zielińska and Piasecki, 2018). In general, the procedures combines work on the corpus data, several linguistic tests and analysis of glosses, relation structure, as well as definitions in traditional dictionaries. We were afraid that with a larger number of intensity grades the inter-annotator agreement could be low. This assumption was not experimentally verified, but the achieved IAA for the applied scale, see Sec. 3.3. is high.

Sentiment analysis often uses sets of basic emotions proposed by Ekman (Ekman, 1992) or Plutchik (Plutchik, 1980). In order to make plWordNet emo compatible with a number of other resources, we used the set of eight basic emotions recognised by Plutchik. It contains Ekman's six basic emotions (Ekman, 1992): *joy* , *fear*, *surprise*, *sadness*, *disgust*, *anger*, complemented by Plutchik's *trust* and *anticipation*. Annotators are allowed to assigned more than one emotion per LU. In this way complex emotions can be also expressed.

From the very beginning of the pilot project, we use the set of *fundamental human values* postulated by Puzynina (Puzynina, 1992) later followed in many works on lexicography and derivation, as a tool for the analysis of the evaluative attitude of a hearer or speaker, see (Zaśko-Zielińska and Piasecki, 2018). The set of the fundamental human values includes: *użyteczność* 'utility', *dobro drugiego człowieka* 'another's good', *prawda* 'truth', *wiedza* 'knowledge', *piękno* 'beauty', *szczęście* 'happiness' (all of them positive), *nieużyteczność* 'futility', *krzywda* 'harm', *niewiedza* 'ignorance', *błąd* 'error', *brzydota* 'ugliness', *nieszczęście* 'misfortune' (all negative) (Puzynina, 1992).

We do not expect perfect agreement on, both, basic emotions and fundamental human values assigned by the annotators. Moreover, assignment of basic emotions and fundamental human values is a tool supporting annotators in making the final decision about the grade of the sentiment polarity, which is done after the emotions and values are assigned. Nevertheless, the overlap of both types of sets is very high in the case of almost annotated LUs.

Use examples are sentences provided for the analysed LUs by annotators in order to justify their decisions and to illustrate the assigned annotation. The annotators select use examples from a corpus, if the source texts are available on an open licence, otherwise an use example is created in a way similar to the sentences observed in the corpus.

Below we present examples of annotation: ***dziad*** 1 gloss:"stary mężczyzna" 'an old man'

⟨ Annot.:A1, BE: {*złość* 'anger', *wstręt* 'disgust'}, FHV:{*nieużyteczność* 'futility', *niewiedza* 'ignorance'}, SP:$-s$

Exam: "Stary dziad nie powinien podrywać młodych dziewczyn."

'An old geezer should not pick up young girls.' ⟩

⟨ Annot.:A2, BE: {*wstręt* 'disgust'}, FHV:{*nieużyteczność* 'futility', *brzydota* 'ugliness'}, SP:$-w$

Exam: "Jakiś dziad się dosiadł do naszego przedziału i wyciągnął śmierdzące kanapki z jajkiem." 'An old geezer joined our compartment and took out stinky egg sandwiches.' ⟩

⟨ Annot.:A3, BE: {*wstręt* 'disgust'}, FHV:{*nieużyteczność* 'futility', *brzydota* 'ugliness'}, SP:$-s$

Exam:"Kilkanaście lat minęło i zrobił się z niego stary dziad."

'Several years have passed and he has become an old geezer' ⟩

gość 3 '≈a fellow, ≈a man' gloss:"z podziwem o kimś godnym szacunku, kto się czymś wykazał" 'with admiration about someone who is worth respect, who showed something exceptional'

⟨ Annot.:A1, BE: {*zaufanie* 'trust', *radość* 'joy'}, FHV:{*wiedza* 'knowledge', *dobro* 'another's good', *użyteczność* 'utility'}, SP:$+w$

Exam: "Mój pracodawca jest świetnym gościem."

'≈My employer is a very good man.' ⟩

⟨ Annot.:A2, BE: {*zaufanie* 'trust', *radość* 'joy'}, FHV:{*szczęście* 'happiness', *wiedza* 'knowledge', *dobro* 'another's good'}, SP:$+s$

Exam: "Paweł to jest dopiero gość!" 'Paweł, he is a really good man!' ⟩

⟨ Annot.:A3, BE: {*zaufanie* 'trust', *radość* 'joy'}, FHV:{*szczęście* 'happiness', *wiedza* 'knowledge'}, SP:$+s$

Exam:"Boże, ale z niego gość, potrafił taką sprawę załatwić w pięć minut."

'My God! What a man is he, he has been able to solve such a problem in five minutes' ⟩

szalbierski 2 'deceitful'

⟨ Annot.:A1, BE: {*smutek* 'sadness', *złość* 'anger'}, FHV: {*krzywda* 'harm', *błąd* 'error' }, SP:$-s$,

Exam: "Nie chciałam brać udziału w tym szalbierskim planie, którego pomyślność zależała od stopnia naiwności nieświadomych klientów."

'I did not want to take part in this deceitful plan, whose success depended on the level of naiveness of the unaware clients.'⟩

⟨A2, BE: {*smutek* 'sadness', *złość* 'anger'}, FHV: {*krzywda* 'harm', *błąd* 'error'}, SP:$-s$,

Exam: "Mam szalbierski pomysł, który pomoże nam naciągnąć paru idiotów."

'I have a deceitful idea which might help us to con a couple of idiots. ' ⟩

Following the approach of the pilot project, and keeping the annotation scheme unchanged, in June 2017 we have started work on large scale expansion of this pilot project. Annotation procedure, guidelines and tools have been improved and expanded and the target size is adding emotive annotations to ≈100,000 more LUs, so the expected target number of annotated LUs by June 2018 is ≈130,000 LUs.

3.3. Statistics

At the time when the experiments were carried out, there was more than $83k$ annotations, covering more than $54k$

LUs and $41k$ synsets (Janz et al., 2017). This data has been successfully applied to PolEval 2017 Sentiment Analysis Task (Ptaszyński et al., 2017). In this previous plWordNet version, about $22k$ of the polarity annotations are different than neutral and these annotations cover $13k$ LUs and $9k$ synsets (22% of all synsets that include annotated LUs). We found, that $1.5k$ of these synsets were annotated with differences among the polarity values assigned to their synset members. If we exclude neutral LUs, only 345 of them have diversified polarity intensity (e.g. synset that contains two LUs annotated as *strong positive* and one annotated as *weak positive*). If we exclude both neutral and ambiguous annotations, there are only 41 synsets expressing potential conflicting, opposite polarity values of their LUs, i.e. synsets that include both positive and negative LUs. However they comprise only 3.8% of all marked synsets, i.e. synsets that do not contain any neutral LUs, namely 9164 in total. The contemporary intermediate state of the process of the emotive annotation of plWordNet is illustrated in Table 1. The overall numbers of annotations, as well as distribution of the polarity intensity is shown.

As our annotators work in a completely independent way, we were able to measure the inter-annotator agreement (IAA) with respect only to the sentiment polarity by using the Cohen's Kappa measure (Cohen, 1960), see Tab. 2. Due to the large number of annotators, we simplified the problem of IAA a little bit, and we have calculated the agreement between the first and the second decision registered in the system for a LUs. All LUs with at least one annotation from the pilot project were excluded from this analysis. The observed IAA values, both, 0.78 for all decisions and around 0.75 for different sentiment polarity values, are very high. The value for the neutral polarity is in fact a value for the decision: marked vs unmarked (or polarised vs non-polarised) LUs. It can show that the annotators are quite confident about the neutrality of LUs. However, also it can be biased by the fact that describing a LU as a neutral can be easier than by other values. The evaluation of the neutrality of a given LU is made in the first step of the annotation procedure and LUs decided as neutral are not further analyses. This issue needs further investigation.

As the neutral annotations dominate (more than 70% of all decisions, in the case of nouns even more than 80%), we have calculated an estimated IAA value for the marked LUs only, all LUs with neutral tags were excluded. The obtained values are much higher than for all decisions, so we can conclude that neutral values do not increase artificially the general IAA.

Negative polarity values dominate in annotation: 17.1% vs 8.51% in Tab. 2. This correlates with the observed dominance of the negative basic emotions, i.e. 76.48% emotions of noun LUs and 70.13% of adjective LUs are negative. A similar dominance of words marked negatively could be also observed in the dictionary of the colloquial Polish language (Anusiewicz and Skawiński, 1996). For instance, if we compare two thematic fields of this dictionary, namely: acting towards somebody's harm – enforcing some particular behaviours (id:2.3.2) and acting towards somebody's profit (id.: 2.3.3), we can notice that the former includes 324 entries while the latter only 20 (Zaśko-Zielińska and Piasecki, 2018).

4. Sentiment Polarity Propagation

We propose a method called *Classifier-based Polarity Propagation* (henceforth CPP), which utilises a rich set of features. This richness arises from their construction, as they take into account even broad neighbourhood context of synsets (up to 2 levels around the synset, see Figure 1), and refer to an extended set of semantic relations.

In Section 5. we compare the results obtained by CPP with the rule-based and relation-based method called *Seed Propagation*, using the best configuration presented by Maks and Vossen (2011).

4.1. Polarity Transfer from Units to Synsets

We have analysed the contemporary annotation of plWordNet from the perspective of the diversification of polarity intensity of LUs belonging to one synset, see Sec. 3.3.. As we could notice, unless, cases of significant differences of the polarity values in one synset are rare, still we can find a number of cases in which the values express smaller differences, e.g. between strong vs weak or weak vs neutral. In contrast to SentiWordNet the manual annotation in plWordNet is done only on the level of LUs (Zaśko-Zielińska et al., 2015) and synsets are not manually assigned sentiment polarity values.

The acquired statistics show, that synsets are relatively homogeneous in terms of the polarity values of their member LUs. Thus, we decided that moving the polarity intensity annotations from the LU-level to the synset-level can be meaningful and profitable[3]. In order to simplify the polarity transfer problem, we decided to project these values onto only three coarse-grained values: *positive*, *negative*, and *neutral*. In order to do this, in the first step, each original polarity value is assigned an heuristic weight: 2 for *strong* variants and 1 for weak variants, neutral and ambiguous. Next, the weights are summed up for each synset. For example, for a synset with a set of LUs annotated with the following values: {*strong negative, negative, strong positive, neutral*}, the total weight for the *positive* coarse-grained value equals 2, for the *negative* one equals 3 (2 + 1) and for the *neutral* is 1. Finally, a generalised sentiment polarity value for a synset is determined on the basis of a simple heuristic: coarse-grained synset polarity value is set to the one which have obtained the highest total weight inside the given synset. In the case of the above example the synset is assigned the *negative* polarity value. If two polarity values have the same total weight, we apply the following rules to solve this discrepancy:

- $\{positive, neutral\} \rightarrow positive$
- $\{negative, neutral\} \rightarrow negative$
- $\{positive, negative\} \rightarrow neutral$

[3] This can simplify some applications or facilitate comparison of even mapping of the sentiment polarity annotation between plWordNet and other wordnets with the help of the manually built interlingual mapping of good coverage.

PoS	# Comp	# Sing	-s	-w	n	+w	+s	amb
N	43,883	1,251	6.45	6.09	80.34	2.82	1.63	2.67
Adj	23,035	49	8.42	14.41	58.26	9.33	4.35	5.24
Verb	5,084	2191	5.11	21.32	48.08	14.97	1.09	9.43
Adv	28	731	7.64	16.73	52.31	12.65	5.14	5.53
All	72,030	4,222	6.93	10.17	70.30	6.05	2.44	4.12

Table 1: Contemporary (Feb. 2018) sentiment polarity annotation of plWordNet 4.0 in progress (Comp – completed, Sing – one annotator only so far); -s, -w, n, +w, +s, amb (negative strong/weak, neutral, positive weak/strong, ambiguous) are shown in percentage points.

PoS	All	-s	-w	n	+w	+s	amb
All	0.78	0.77	0.78	0.82	0.74	0.73	0.65
Mrk.	0.84	0.80	0.84	–	0.89	0.80	0.86

Table 2: Inter-annotator agreement (IAA), measured in Cohen's' κ, for different sentiment polarities: -s, -w, n, +w, +s, amb (negative/positive vs strong/weak, neutral, ambiguous). *All* describes agreement for all decisions, *Mrk* – estimated IAA for marked LUs only.

4.2. Features

From the structure of plWordNet we selected only the most frequent relations as a basis for features. As a result, the selected relations cover more than 95% of all relation instances (occurrences) in plWordNet: *hyponymy*, *hypernymy*, *fuzzynymy*, *similar_to*, *feature_value*, *meronymy*, *holonymy*, *collection_meronym*, *collection_holonym*, *type*, *member*, *taxonomic_meronym*, *taxonomic_holonym*.
Each synset is described by a set of feature values that all together form a kind of *bag-of-words* representation. This representation refers both to synsets and their polarities. Elements of the bag are constructed on the basis of the following four components each.

- *Relation* – one of the 13 selected wordnet relations.
- *Direction* – the direction of the relation, expressing whether the synset described by the feature is the *source* or *target* of the relation instance (i.e. outgoing vs ingoing relation instance).
- *Target* – it can be one of the two: a *synset_ID* (identifier) or *synset_polarity*, coarse-grained of the target synset encoded as: $-1, 0, 1$ for, respectively: *negative, neutral, positive*).
- *Level* – represents the distance in which the given relation instance is on the path connected to the synset being described, it can a *directly* linked to this synset (level=0), but also in some further distance, always only one link is expressed without the information about the rest of the path, so the level informs how broad is the context, see the example presented in Figure 1.

In total, we use 13 wordnet relations (types), 2 directions, 2 types of targets (exclusively) and up to $m = 2$ levels, so elements of the bag-of-word representation can be constructed in $13 \cdot 2^3 = 104$ ways. For example a feature of the type *hyponym_source_id_level_2* introduces into the

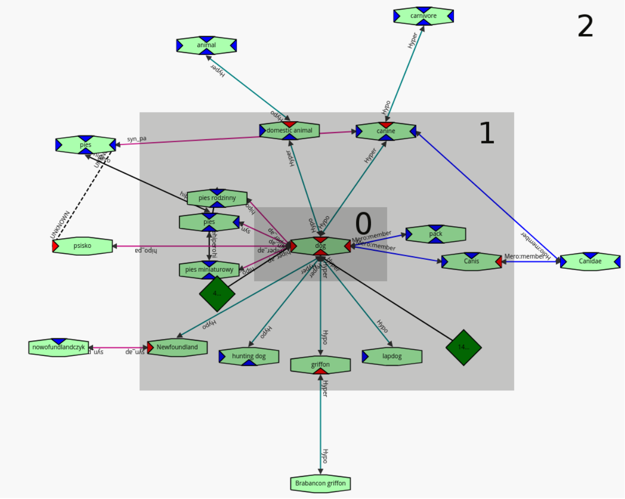

Figure 1: Example of synsets at the specific level (1 and 2), with respect to the synset at level 0.

representation all IDs of synsets which are sources of all hyponym relation instances, for which the target is a synset indirectly accessible (by the third link), see Figure 1.
In order to produce a single bag for the classifier, we concatenate the values of this bag into a single document and convert it with `TfidfVectorizer` into a vector representation. Acquired vectors for a bag of synset IDs and a bag of polarities are different in terms of their vector spaces, due to the different type of collected elements. However, if we consider only a single type of a bag, e.g. a bag of synset IDs collected for the hyponymy relation on level k, the constructed vectors for this type of a bag should be represented always in the same vector space.

4.3. Classifier and Propagation Method

To train a model for a classifier we need a set of manually annotated synsets with sentiment polarity annotations (i.e. *seeds*). Next, each synset is automatically described with 104 different bags of words as a complex representation, which takes into account its broader relational context of linked synsets and their annotations. The constructed bags-of-words are input to the `TfidfVectorizer` module from `scikit-learn`[4] Python machine learning package for building vectors. This feature extraction method allows to convert a collection of elements into a matrix of tf.idf features. Each synset belongs to one of the three fol-

[4] `http://scikit-learn.org`

lowing coarse-grained classes: *positive, negative, neutral.* Finally all vectors representing all bag-of-words are concatenated. The final model contains $38,108$ vectors – one vector represents one synset. The size of the final vector space is $38,108 \times 49,0170$. Transformed vectors are used to train a predictive model. We used Logistic Regression from `scikit-learn` package as a classifier.

A classifier with a trained model is applied in propagating annotations to the unlabelled part of plWordNet. At the beginning we treat our seeds as a set of synsets at level-0 (see Figure 1). Each next iteration is a classification of unlabelled synsets at the 1st level, using information from annotated synsets from the surrounding context. We tested two approaches to performing propagations, namely in each iteration:

- *naive* – we preserve the graph depth order of the remaining synsets to be classified,
- *sorted* – before each iteration we sort synsets at the 1st level by the number of relations with synsets that already have been assigned a polarity value (descending order). Then we start propagation with synsets at the top of this list.

5. Evaluation

The proposed method assumes that the propagation is performed only for synsets. However, the existing polarity annotations in the plWordNet refer only to LUs, so preprocessing was required. First, we used a simple generalization function, to assign the polarity to the synsets, depending on the polarity of their LUs (see Section 4.1.). As a result the original a 5-degree scale of sentiment polarity intensity was projected onto the coarse-grained 3-degree scale. Next, we prepared a large graph of plWordNet, consisting of synsets with sentiment polarity annotation transferred from their members, as a basis for wordnet-based evaluation of the method.

5.1. Wordnet-based Evaluation

During evaluation we included complete plWordNet annotation with $43k$ synsets annotated with sentiment polarity (positive, negative, neutral) in this particular version (from Oct. 2017). For each method and configuration we performed 10-fold cross-validation. Annotated synsets were divided into 10 parts, where 9 parts (about 40,400 synsets in total) were treated as a seed for the baseline (or a training set for CPP) and the 10th part (about 3,600 synsets) as a test set.

We implemented a simple rule-based seed-driven propagation method described in (Maks and Vossen, 2011) to obtain a *baseline* (henceforth BASE). Then we compared results of its application with the CPP method in two variants, described in Section 4.: naive (CPP-N) and sorted (CPP-S).

5.2. Task-based Evaluation

In (Qian et al., 2017) the authors proposed a simple, yet effective solution to recognize sentiment polarity of sentences using Bidirectional LSTM network. The proposed solution is based on additional regularization terms incorporating linguistic knowledge into the network. The regularization terms $L_{t,k}$ have been combined with the original cross entropy loss:

$$C(\theta) = -\sum_i \hat{y_i} \log y_i + \alpha \sum_i \sum_t L_{t,i} + \beta \|\theta\|^2 \quad (1)$$

The authors called their solution Linguistically Regularized LSTM (henceforth LR-LSTM) which is a model of LSTM network but expanded with a set of regularizers to better reflect the linguistic role of sentiment, negation and intensity of words (Qian et al., 2017). Here we have applied all four proposed regularization terms namely non-sentiment regularizer (NSR), sentiment regularizer (SR), negation regularizer (NR) and intensity regularizer (IR). In case of SR, the idea is that the model should restrict the polarity distribution of adjacent words in text to drift in the same way, especially in case of sentiment words included in provided lexicon:

$$p_{t-1}^{(SR)} = p_{t-1} + s_{c(x_t)} \quad (2)$$

$$L_t^{(SR)} = \max\left(0, D_{KL}(p_t \| p_{t-1}^{(SR)}) - M\right) \quad (3)$$

LR-LSTM was prepared and evaluated on English, thus we needed to adapt their work to the Polish language. To show the impact of the lexicon on the accuracy in sentiment analysis tasks three different variants of a lexicon (of the same size) were prepared:

1. a lexicon with randomly assigned sentiment scores,
2. a sentiment lexicon after rule-based propagation,
3. a sentiment lexicon constructed with CPP.

Dataset

We have collected a corpus of $4,039$ user reviews from Trip Advisor[5]. Table 3 presents the distribution of Users' Ratings assigned to the collected reviews. During the evaluation of our lexicon we needed to convert values of Users' Ratings into sentiment polarity values. Thus we automatically replaced rating values of reviews with the coarse grained polarity values according to the following schema: $\{1, 2\} \rightarrow negative$, $\{3\} \rightarrow neutral$, $\{4, 5\} \rightarrow positive.$ The reviews were manually revised once again, just to correct assigned sentiment if necessary. The final classification was limited only to the recognition of $positive$ and $negative$ reviews.

Experimental Setting

In this section, we present our experimental setting for task-based evaluation, especially the characteristics of evaluated lexicons and the training procedure for LR-LSTM. On a basis of a collected corpus, we have prepared three different sentiment lexicons with the same distribution of words. The first one contains words with randomly assigned polarities (RAND). The second (BASE) and third (CPP-N) were

[5] `https://www.tripadvisor.com`

Rating	# Reviews	# Words	# Sentences	Words per Rev.	Words per Sent.
rate-1	298	25809	2249	86.61	11.47
rate-2	274	25044	2079	91.40	12.04
rate-3	659	50030	4712	75.92	10.62
rate-4	1419	94203	9123	66.39	10.32
rate-5	1389	80178	8419	57.72	9.52

Table 3: The distribution of Users' Ratings and polarity classes assigned to the reviews in the corpus.

derived directly from plWordNet using a simple averaging procedure, i.e. for every word appearing in the corpus we collect its synsets and average their polarities to derive the final polarity assignment.

For the training procedure, we plugged constructed lexicons to the LR-LSTM network. The parameters proposed in (Qian et al., 2017) were modified in order to adapt the network to our task. The number of training mini-batches was increased to $4,000$, each with 15 samples. To train the model we used *adaGrad* with the learning rate $lr = 0.05$, and the coefficients for all regularizers were the same as in (Qian et al., 2017), $\alpha = 0.5$ and $\beta = 0.0001$ respectively. To ensure that the network will be able to achieve the highest performance, we performed this training procedure multiple times for each lexicon. Vector representation for LR-LSTM was computed with *FastText* (Joulin et al., 2016) using *SkipGram* model with the size of a vector $dim = 300$ and the minimal frequency $minCount = 50$.

5.3. Results and Discussion

Table 4 presents the results obtained during experiments in the wordnet-based evaluation. We calculated precision (P), recall (R) and F-measure (F) for different coarse-grained polarity value separately: negative (NEG), positive (POS) and neutral (NEU). We compared differences between the two pairs: {BASE, CPP-N} and {CPP-N, CPP-S}. In Tab. 4 results for which differences were statistically significant are highlighted. We analysed the statistical significance of differences using paired-differences Student's t-test with a significance level $\alpha = 0.05$ (Dietterich, 1998).

Measure	BASE	CPP-N	CPP-S
P-NEG	84.01	84.58	84.73
P-NEU	92.18	**93.75**	93.66
P-POS	69.20	**83.11**	82.95
R-NEG	68.63	**75.82**	75.90
R-NEU	95.80	**97.02**	96.97
R-POS	64.64	**68.41**	67.80
F-NEG	75.52	**79.91**	79.81
F-NEU	93.95	**95.34**	95.35
F-POS	66.77	**74.99**	74.61

Table 4: Precision (P), recall (R) and F-score (F) for separate coarse-grained polarity values. BASE results are compared to CPP-N and CPP-S. Statistically significant differences are emphasised.

The naive solution (CPP-N) is significantly better than BASE in all test cases except the precision for the *negative* value. The order of neighbours classified in each iteration is not important in this case, because there was no significant difference between CPP-N and CPP-S variants. These three approaches were evaluated once again, but this time we decided to incorporate also the instances of the *ambiguous* class, which seems to be more realistic scenario. Table 5 presents precision, recall and F-score obtained in this experiment. CPP approaches outperformed baseline solution even for the most difficult *ambiguous* class, but in this scenario the resulting performance was slightly higher for CPP-S solution, which suggests that in some cases sorting has a positive effect on the final propagation.

Measure	BASE	CPP-N	CPP-S
P-NEG	**79.21**	73.83	74.60
P-NEU	90.45	94.37	**94.53**
P-POS	**60.52**	59.32	59.01
P-AMB	40.89	**54.41**	53.73
R-NEG	60.83	**75.35**	74.83
R-NEU	**95.24**	94.51	94.45
R-POS	57.19	**64.98**	66.78
R-AMB	33.76	41.88	**42.18**
F-NEG	68.81	74.58	**74.71**
F-NEU	92.78	94.44	**94.49**
F-POS	58.80	62.02	**62.65**
F-AMB	36.98	**47.33**	47.25

Table 5: Precision (P), recall (R) and F-score (F) for separate classes of polarity extended with propagation for *ambiguous* units.

To investigate the impact of a lexicon in sentiment recognition task we compared the precision (P), recall (R), and F-score (F) of LR-LSTM for different polarity classes – the results for this experiment are presented in table 6. Unfortunately, the convergence of adapted network was quite unstable. We decided to select the model with the highest performance on the validation dataset, in the same way as it was conducted in original work (Qian et al., 2017). The final scores for the best model were determined by averaging the values obtained from multiple executions of this network on the validation dataset.

The resulting accuracy for the models was in many cases similar (due to the class imbalance in our corpus), that is why we also decided to use more specific measures for evaluation. An observed precision and recall for *positive* and *negative* reviews is slightly different, especially when we compare a model using randomly generated lexicon (RAND) with the models using lexicons constructed in a controlled way (BASE, CPP-N). However, the difference between rule-based propagation and CPP is small which may suggest that hybrid methods combining neural approaches with language resources are still imperfect for this

Measure	RAND	BASE	CPP-N
P-NEG	0.761	0.821	0.880
R-NEG	0.910	0.842	0.837
F-NEG	0.828	0.831	0.858
P-POS	0.957	0.921	0.951
R-POS	0.875	0.931	0.930
F-POS	0.914	0.926	0.940

Table 6: Precision (P), recall (R) and F-score (F) for specific polarity classes, in the task of sentence-level sentiment recognition with LR-LSTM.

task and are not able to fully utilize the potential of such resources.

6. Further Works

By June 2018 we plan to complete and publish the emotive annotation of plWordNet 4.0 emo on an open licence (the intermediate results discussed here). The target size is more than 130k manually annotated LUs from all Parts of Speech. We presented an intermediate version of more than 76k manually annotated LUs in a way expressing high Inter-annotator Agreement (i.e. good consistency between annotators was achieved). This version have been already published as a part of plWordNet 3.1 emo. Next, the annotation will be automatically spread to the rest of plWordNet LUs. In parallel, the experiment-based emotive lexicon in Sentimenti[6] will be built. The method of automated selection of LUs proposed by us will be used to prepare the subsequent batches of LUs for the experiments. plWordNet descriptions of all selected LUs will be supplemented with possibly missing glosses and use examples, but not with emotive annotations, because we expect still to achieve some complementarity. We need also to solve the problem of appropriate prompting of LUs to the experiment participants, i.e. to find a way in which a certain meaning of a lemma is clearly targeted.

7. Acknowledgements

Work co-financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education and in part by the National Centre for Research and Development, Poland, under grant no POIR.01.01.01-00-0472/16, and as a result of works carried out within the project funded by the National Science Centre, Poland under the grant agreement No UMO-2015/18/M/HS2/00100.

8. Bibliographical References

Anusiewicz, J. and Skawiński, J. (1996). *Słownik polszczyzny potocznej*. Wrocław.

Black, W., Elkateb, S., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A., and Fellbaum, C. (2006). Introducing the Arabic wordnet project. In *Proceedings of the Third International WordNet Conference*, pages 295–300. Global WordNet Association.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:1895–1923.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, 6(3):169–200.

Esuli, A. and Sebastiani, F. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of 5th Conference on Language Resources and Evaluation LREC 2006*, pages 417–422.

Christiane Fellbaum, editor. (1998). *WordNet – An Electronic Lexical Database*. The MIT Press.

Haniewicz, K., Rutkowski, W., Adamczyk, M., and Kaczmarek, M., (2013). *Towards the Lexicon-Based Sentiment Analysis of Polish Texts: Polarity Lexicon*, pages 286–295. Springer, Berlin, Heidelberg.

Haniewicz, K., Kaczmarek, M., Adamczyk, M., and Rutkowski, W., (2014). *Polarity Lexicon for the Polish Language: Design and Extension with Random Walk Algorithm*, pages 173–182. Springer.

Janz, A., Kocoń, J., Piasecki, M., and Monika, Z.-Z. (2017). plWordNet as a Basis for Large Emotive Lexicons of Polish. In *LTC'17 8th Language and Technology Conference*, Poznań, Poland, November. Fundacja Uniwersytetu im. Adama Mickiewicza w Poznaniu.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Kulisiewicz, M., Kajdanowicz, T., Kazienko, P., and Piasecki, M., (2015). *On Sentiment Polarity Assignment in the Wordnet Using Loopy Belief Propagation*, pages 451–462. Springer International Publishing, Cham.

Mahyoub, F. H., Siddiqui, M. A., and Dahab, M. Y. (2014). Building an arabic sentiment lexicon using semi-supervised learning. *Journal of King Saud University - Computer and Information Sciences*, 26(4):417 – 424. Special Issue on Arabic NLP.

Maks, I. and Vossen, P. (2011). Different approaches to automatic polarity annotation at synset level. In *Proceedings of the First International Workshop on Lexical Resources*, pages 62–69.

Maks, I., Izquierdo, R., Frontini, F., Agerri, R., Vossen, P., and Azpeitia, A. (2014). Generating polarity lexicons with wordnet propagation in 5 languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Maziarz, M., Piasecki, M., Rudnicka, E., Szpakowicz, S., and Kędzia, P. (2016). plwordnet 3.0 – a comprehensive lexical-semantic resource. In N. Calzolari, et al., editors, *Proc. of COLING 2016, 26th Inter. Conf. on Computational Linguistics*, pages 2259–2268. ACL.

Misiaszek, A., Kajdanowicz, T., Kazienko, P., and Piasecki, M., (2013). *Relational Propagation of Word Sentiment in WordNet*, pages 137–140. Springer Berlin Heidelberg, Berlin, Heidelberg.

[6] Project supported by the National Centre for Research and Development, Poland, under grant no POIR.01.01.01-00-0472/16.

Plutchik, R. (1980). *EMOTION: A Psychoevolutionary Synthesis*. Harper & Row.

Ptaszyński, M., Masui, F., Janz, A., Kocoń, J., Piasecki, M., Monika, Z.-Z., and Dybała, P. (2017). Three attempts in PolEval 2017 Sentiment Analysis Task. In *LTC'17 8th Language and Technology Conference*, Poznań, Poland, November. Fundacja Uniwersytetu im. Adama Mickiewicza w Poznaniu.

Puzynina, J. (1992). *Język wartości [The language of values]*. Scientific Publishers PWN.

Qian, Q., Huang, M., Lei, J., and Zhu, X. (2017). Linguistically regularized lstm for sentiment classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1679–1689. Association for Computational Linguistics.

Stone, P. J. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press.

Vossen, P., Maks, I., Segers, R., and van der Vliet, H. (2008). Integrating lexical units, synsets and ontology in the cornetto database. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Vossen, P., Maks, I., Segers, R., van der Vliet, H., Moens, M., Hofmann, K., Tjong Kim Sang, E., and de Rijke, M. (2013). Cornetto: a lexical semantic database for Dutch. In P. Spyns et al., editors, *Speech and Language Technology for Dutch, Results by the STEVIN-programme*, Theory and Applications of Natural Language Processing. Springer Berlin Heidelberg.

Zaśko-Zielińska, M. and Piasecki, M. (2018). Towards emotive annotation in plWordNet 4.0. In *Proceedings of the The 9th Global WordNet Conference GWC'18*.

Zaśko-Zielińska, M., Piasecki, M., and Szpakowicz, S. (2015). A large wordnet-based sentiment lexicon for Polish. In R. Mitkov, et al., editors, *Proc. of the Conf. Recent Advances in Natural Language Processing – RANLP'2015*, pages 721—730.