Building a Sentiment Lexicon for Swedish

Bianka Nusko
Dept of Philosophy, Linguistics
and Theory of Science,
University of Gothenburg
gusnusbi@student.gu.se

Nina Tahmasebi
Språkbanken,
University of Gothenburg
nina.tahmasebi@gu.se

Olof Mogren
Dept of Computer Science
and Engineering,
Chalmers University of
Technology
mogren@chalmers.se

Abstract

In this paper we will present our ongoing project to build and evaluate a sentiment lexicon for Swedish. Our main resource is SALDO, a lexical resource of modern Swedish developed at Språkbanken, University of Gothenburg. Using a semi-supervised approach, we expand a manually chosen set of six core words using parent-child relations based on the semantic network structure of SALDO. At its current stage the lexicon consists of 175 seeds, 633 children, and 1319 grandchildren.

1 Introduction

Sentiment lexicons have proven a valuable resource for opinion mining tasks. With large amounts of data readily available on the Internet, gathering user sentiments and opinions is a relatively effortless and inexpensive undertaking. As an example, it is possible to easily check whether or not a product is positively received. This information is useful to both potential new customers and the manufacturers or the suppliers of said product. While customers may wish to inform themselves whether the product lives up to the desired quality or value, the company has the opportunity to quickly gather the general consensus on the product and is therefore able to react accordingly. Other examples include politicians or political parties that are able to quickly gather their voters' opinions by using opinion mining.

So far, this field of research is mainly restricted to anglophone data and hence the majority of sentiment lexicons in existence today are in English. Some of the most well-known lexicons include SentiWordNet (Esuli and Sebastiani, 2006), the Bing Liu Opinion Lexicon (Hu and Liu, 2004) and the General Inquirer (Stone et al., 1966). In recent years there has been increasing interest in building opinion lexicons for other languages as well. For German, Remus et al. (2010) built the SentimentWortschatz Lexicon, short SentiWS, using semi-automatic translations of English sentiment resources combined with information about word co-occurrences and word collocations. Banea et al. (2008) use raw data and a bootstrapping method to construct a subjectivity lexicon for languages with scarce resources such as Romanian and Wan (2009) exploits the large amount of annotated English data available to classify Chinese reviews. The approach is supervised and allows classification without any annotated Chinese data.

To the best of our knowledge there are no publicly available lexical resources for sentiment analysis for the Swedish language. Our goal is therefore to lay the groundwork for a Swedish sentiment lexicon.

2 Related Work

Strategies for creating a sentiment lexicon range from purely manual, over semi-supervised, to more or less automatic machine learning approaches. Well known manually constructed sentiment lexicons are e.g. MPQA (Wiebe et al., 2005) and Bing Liu's Opinon Lexicon (Hu and Liu, 2004). However, manual approaches are expensive and difficult to adapt to new domains, therefore, we refrain from using a manual approach and instead rely on a manually created resource.

Severyn and Moschitti (2015) use hashtags and emoticons contained in a Twitter corpus as sentiment information to automatically construct an opinion lexicon. An SVM classifier was trained using words as well as multi-word expressions taken from a distantly-supervised corpus of tweets as features, a corpus that was obtained by using hashtags or emoticons as sentiment indicators. The authors state that the corpus size compensates for the noisiness of the data. While previous approaches like Mohammad et al. (2013) used statistical measures such as PMI to calculate word sentiment association for building the lexicon, Severyn and Moschitti (2015) rely on machine learning techniques and as a consequence achieve higher results. The drawback is that the created lexicons are not necessarily human-interpretable, they include good features for classification and are not meant as stand-alone sentiment lexicons.

A semi-supervised approach that links conjunctions to infer opinion was introduced by Hatzivassiloglou et al. (1997). They used a labelled seed set of adjectives taken from the 1987 Wall Street Journal (WSJ) corpus provided by the ACL Data Collection Initiative (`https://catalog.ldc.upenn.edu/LDC93T1`) and expanded them to pairs of adjectives that are linked by the conjunctions *and* or *but*. Adjectives linked by *and* were assigned the same polarity and adjectives linked by *but* were assigned opposing polarity. A supervised log-linear-regression model was used for this task.

Esuli and Sebastiani (2005), who were involved in building SentiWordNet, used textual term descriptions called *glosses* to determine the polarity of a sense by assuming that, generally, terms with similar glosses also have similar polarity. The method is used to expand a set of seed words by means of the lexical relations specified in a thesaurus. For each term in the expanded set the gloss is extracted from an online dictionary. After conversion into a vectorial format, a binary classifier is trained on this data and finally applied to a new data set. An advantage of this method is that it is not restricted to classifying adjectives, like Hatzivassiloglou et al. (1997)'s approach, but allows classification of all terms. A disadvantage is that the method can produce noisy results.

3 The Resource: SALDO

SALDO (Borin et al., 2013) is a publicly available electronic resource for the written modern Swedish language. It was developed and is continuously expanded at Språkbanken, the Swedish Language Bank, at the University of Gothenburg. SALDO includes semantic as well as morphological information about words. It is organised as a lexical-semantic network linking word senses by their association to other word senses and does not purely rely on synonymy relations like the synset structure of Princeton WordNet (`http://wordnet.princeton.edu`) for example. Furthermore, SALDO contains words from all word classes, including closed-class words, although the full scope of the word classes are not exploited for the sentiment lexicon described in this paper.

All entries, i.e. word senses, in SALDO are arranged hierarchically around the dummy core *PRIM*. Each word sense, except the dummy, has a primary semantic descriptor and optionally one or more secondary semantic descriptors – all of which are word senses in SALDO as well. While some word senses listed in SALDO have more than one secondary descriptor, most of them being names, this does not apply to the words included in our lexicon. Therefore, there is no strategy for dealing with words that have two or more secondary descriptors. In this study we use the semantic descriptors assigned to a word sense to determine the sentiment of that sense. Additionally, we also extract the lemgram and part-of-speech information for each word sense from SALDO.

primary descriptor / secondary descriptor The primary descriptor "(1) [...] is a semantic neighbor of the entry to be described and (2) it is more central than it" (Borin et al., 2013, p. 1195). Secondary descriptors give more information about the particular sense of the word and can also include words that negate or indicate the strength of the primary descriptor. Therefore they can also modify the polarity of a word[1]. The descriptors link all the words in SALDO to form a network.

[1] Due to the modifying qualities of certain words, we specifically removed *lagom* "just the right amount" and all words that have *sjukdom* "disease" in them from our data sets to reduce noise.

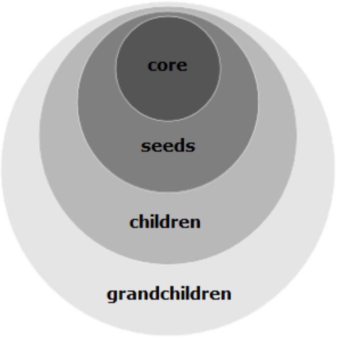

Figure 1: structure of the data sets.

lemgram "A *Lemgram* in SALDO is a combination of a base form and an inflectional pattern" (Borin et al., 2013, emphasis added). That is, a lemgram can basically be seen as denoting an inflection table. It should not be confused with *lemma*. Although lemma can refer to the headword of a set of inflectional forms, which is very similar to the notion of lemgram, in most cases it simply connotes the base form of a word.

4 Methodology

Similarly to Esuli and Sebastiani (2005), we start with a small set of core words that is expanded semi-automatically using the lexical-semantic relations of the network structure in SALDO. Figure 1 shows how our data sets are structured. We first introduce the format of our data in section 4.1 and explain the expansion process in the subsequent section 4.2.

4.1 Format

Each entry in the data set is represented in the format *(word, polarity, strength, word sense, written form(s), part-of-speech, confidence, lemgram frequency, example)*.

polarity We manually assign a polarity to the six core words. The polarity of the seed words then depends on the polarity of the core word that is its primary descriptor. The same is true for the children and grandchildren which are assigned the polarity of their primary descriptor as well. In cases where the secondary descriptor is *inte* "not" or *motsats* "the opposite" the polarity of a word is reversed.

strength The polarity strength is a discrete number in the range [1,4] and [-4,-1]. Positive numbers indicate positive word polarity and negative numbers express negative polarity. All words are initially assigned a value of 2 or -2 respectively. A higher number indicates a stronger polarity. Certain secondary descriptors can increase or decrease these values so that words with one of the secondaries shown in Table 1 get the respective strength as specified below.

lite "a little"	1, -1
none	2, -2
mycket "very"	3, -3
enastående "outstanding", värdelös "worthless"	4, -4

Table 1: secondary descriptors and corresponding strength values.

word form(s) and word sense A word sense is the semantic sense or meaning of a word recorded in SALDO. A sense is written in the form *wordsense..sensenumber*, i.e. *bra..1*. The word form is the written form of the word sense, i.e *bra*.

part-of-speech (pos) The part-of-speech tag of a word extracted from SALDO. We only use words that are either adjectives, adverbs or verbs, i.e having one of the following tags: *av, avh, avm, ava, vb, vbn, vba, ab*. All other words, especially nouns, were removed since they did not add polarised words to our data.

confidence The confidence score, ranging [0,1], indicates how certain we are that the assigned polarity is correct. The higher the score, the more confident we are in the polarity. Seed words get a value of 1, children a value of 0.75 and grandchildren a value of 0.5. So the value decreases the further away we get from the seed words.

lemgram frequency Using a large data set of modern Swedish, the Gigaword corpus (Eide et al., 2016), we measure the frequency of the lemgram corresponding to the word sense.

example For each word we intend to provide an example sentence taken from the same data set used for computing the above frequencies. This will allow us to examine the context in which the word occurs. Deciding on how to extract the most useful example sentence for each entry in our lexicon is future work and will be our next focus.

4.2 Expanding the Seed Set into a Sentiment Lexicon

The core of our data set consists of the six words *bra* "good", *glad* "happy", *frisk* "healthy", *dålig* "bad", *ledsen* "sad" and *sjuk* "ill". We expand this core set into the different seed sets shown in Figure 1 as described below:

1. Extract all words from SALDO that have one of the core words as their primary descriptor.
2. Extract and add children, i.e. words having one of the seed words as primary descriptor.
3. Extract and add grandchildren, i.e. words having one of the children as their primary descriptor.

A minimal expansion example is the following (The two underlined items are the primary and secondary descriptor of the word in first position of the entry.):

core word: *bra*
seed: *(god, +, 2, god..1, ['god'], bra..1, None, av, 1, LEMGRAMFREQ, EXAMPLE)*
child: *(elak, -, -2, elak..1, ['elak'], god..1, inte..1, av, 0.75, LEMGRAMFREQ, EXAMPLE)*
grandchild: *(sarkastisk, -, -2, sarkastisk..1, ['sarkastisk'], elak..1, None, av, 0.5, LEMGRAMFREQ, EXAMPLE)*

The polarity, its strength, and also the certainty value of a word sense depend on the values of its primary and secondary descriptors. We add *god* "kind" to our set of seed words because it has the core word *bra* "good" as a primary descriptor. A child of *god* is *elak* "mean" with *god* itself as its primary and the negation *inte* "not" as its secondary descriptor. Because of this secondary descriptor we give the word a negated polarity. *Elak* in turn has *sarkastisk* "sarcastic" as a child. The polarity of this child is the same as the polarity of its parent because it lacks a secondary descriptor that negates its polarity. Being the child of *elak*, *sarkastisk* is at the same time the grandchild of *god*.

5 The Sentiment Lexicon and Its Quality

The seeds extracted for the core words amounted to 233 words in the seed set. The two expansion steps yielded 1344 children and finally 3848 grandchildren. After removing all words that have part-of-speech tags other than those listed under the *part of speech* section, the sentiment lexicon consists of 175 seeds, 633 children, and 1319 grandchildren.

To test the quality of the sentiment lexicon, we performed an evaluation on 100 words, 50 positive and 50 negative, from each of our three seed sets. All 300 words were shuffled and manually labelled by at least two native speakers of Swedish. Since determining the overall sentiment of a word is easier than deciding on its strength, we only evaluated the polarity. The annotators were asked to assign each word to one of the following four categories: positive, negative, neutral or unknown. A Fleiss' Kappa (Fleiss, 1971) score of 0.70[2] indicates *substantial agreement* (Artstein and Poesio, 2008) between the annotators.

Out of the 150 words that were evaluated by all three evaluators, 113 were fully agreed upon. For 107 of them, the manually assigned polarity corresponds to the respective sentiment listed in our lexicon giving us a precision of 71%. Partial agreement, i.e. at least two evaluators assigned the same label to a word, was reached in 145 cases. 128 out of these are in accordance with our automatically assigned labels, resulting in a precision of 85%. While one of the annotators classified 13 words as having unknown polarity, each of the remaining two evaluators only marked 5 words words as unknown. Although the evaluation showed that parts of the data were difficult to classify, the agreement score $\kappa = 0.70$ and the low number of words labelled as unknown show that the annotators had a good intuition about word sentiment in the majority of cases. The polarity of the 22 incorrect words (as agreed by at least two annotators) has been reversed in the sentiment lexicon.

6 Outlook

Currently, the entries of our lexicon are not supported by example sentences. An important task is therefore to find a representative example for each entry which shows the usage as well as the context of the respective word sense and can be used for e.g. classification. These sentences should also serve as a source for extracting possible new words to further expand our data set beyond SALDO. Since the final lexicon should be genre-independent, we must ensure that the extracted examples are representative for the general language or include examples for different genres such as newspapers and social media.

A future step is to refine the strategies for assigning confidence scores and polarity strength. Confidence is assigned using a heuristics based on the word's placement in the parent-child-grandchild hierarchy but at present does not extend beyond words included in SALDO. Determining basic polarity is, for the time being, a binary decision between the discrete values *positive* (plus sign) and *negative* (minus sign). A separate strength value is given to supplement the polarity and further differentiate between the words within the same category. The main reason for this is that strength is much harder to determine than the basic polarity and current strength assignment can be seen as an approximation at best. Also, since we have more confidence in the specified polarities than the respective strength, keeping these values separated simplifies working with the data. In the current version of the lexicon, only words listed in SALDO will get a strength value and it is future work to find a method for also assigning strength to words that cannot be found in the resource.

The data set is available under the following link:
`https://spraakbanken.gu.se/swe/resurs/sentimentlex/xml`

Acknowledgements

We would like to thank Richard Johansson and Markus Forsberg for sharing ideas and expertise. The authors would also like to acknowledge the project "Towards a knowledge-based culturomics" supported by a framework grant from the Swedish Research Council (2012– 2016; dnr 2012-5738).

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Carmen Banea, Janyce M Wiebe, and Rada Mihalcea. 2008. A bootstrapping method for building subjectivity lexicons for languages with scarce resources.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. Saldo: a touch of yin to wordnets yang. *Language resources and evaluation*, 47(4):1191–1211.

[2] Since one of the annotators only labelled the first 150 items in the evaluation data, the agreement calculation is based on this set of words.

Stian Rødven Eide, Nina Tahmasebi, and Lars Borin. 2016. The Swedish Culturomics Gigaword Corpus: A One BillionWord Swedish Reference Dataset for NLP. In *From Digitization to Knowledge 2016, D2K16 in conjunction with Digital Humanities 2016.*

Andrea Esuli and Fabrizio Sebastiani. 2005. Determining the semantic orientation of terms through gloss classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 617–624. ACM.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422. Citeseer.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Vasileios Hatzivassiloglou and Kathleen R McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics*, pages 174–181. Association for Computational Linguistics.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242.*

Robert Remus, Uwe Quasthoff, and Gerhard Heyer. 2010. Sentiws-a publicly available german-language resource for sentiment analysis. In *LREC.*

Aliaksei Severyn and Alessandro Moschitti. 2015. On the automatic learning of sentiment lexicons. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2015).*

Philip J Stone, Dexter C Dunphy, and Marshall S Smith. 1966. The general inquirer: A computer approach to content analysis.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 235–243. Association for Computational Linguistics.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.