Defining a Gold Standard for a Swedish Sentiment Lexicon: Towards Higher-Yield Text Mining in the Digital Humanities

Jacobo Rouces, Lars Borin, Nina Tahmasebi, Stian Rødven Eide

Språkbanken, University of Gothenburg, Sweden
{jacobo.rouces, lars.borin, nina.tahmasebi, stian.rodven.eide}@gu.se

Abstract. There is an increasing demand for multilingual sentiment analysis, and most work on sentiment lexicons is still carried out based on English lexicons like WordNet. In addition, many of the non-English sentiment lexicons that do exist have been compiled by (machine) translation from English resources, thereby arguably obscuring possible language-specific characteristics of sentiment-loaded vocabulary. In this paper we describe the creation from scratch of a gold standard for the sentiment annotation of Swedish terms as a first step towards the creation of a full-fledged sentiment lexicon for Swedish.

1 Introduction

As the amounts of digital textual data available to scholars grow beyond all bounds, forever eluding all hope of being able to deal with them in time-honored "close-reading" fashion, *text mining* (TM; also "text data mining" or "text analytics") is seeing increasing use as a research tool in the humanities and social sciences. TM relies heavily on linguistic processing of the texts in order to produce reliable results. In other words: text mining for a particular language will be limited by the accuracy of the natural language processing (NLP) tools available for that language.[1]

2 Sentiment Analysis and its Uses in Digital Humanities

The NLP subfield known as *sentiment analysis* or *opinion mining* is an important component technology of TM, which has seen an explosive expansion over the last decade or so. Since the publication of the comprehensive overview of the field by Pang and Lee (2008), we have seen hundreds of papers as well as dedicated workshops on this topic in NLP conferences.

Even though sentiment analysis has become a standard item in the NLP toolbox, there still remain many theoretical and methodological questions to be answered and resource gaps to be filled. For the latter, we note that most work on automated sentiment analysis has been done on English and a few other languages; for most of even the written languages of the

[1] The language dependence of NLP tools makes up a complex and sorely underresearched area; see, e.g., the insightful discussion in (Bender, 2011).

world,[2] this tool is not available. All sentiment analysis methods in the literature rely on lexical knowledge in one way or another, often in the form of a sentiment lexicon, i.e., a list of words (lemmas or text words) and multi-word expressions annotated with sentiment information. This of course must be a language-specific resource. The present paper describes the first steps towards the development of an extensive sentiment lexicon for written (standard) Swedish.

There is an increasing demand for multilingual sentiment analysis, as well as – in particular in the digital humanities – for sentiment analysis tools for historical texts, while most published work deals with contemporary English, more often than not texts from product and service review websites. In fact, many of the non-English sentiment lexicons that do exist have been compiled by (machine) translation from English resources,[3] thereby arguably obscuring possible language-specific characteristics of sentiment-loaded vocabulary.

The theoretical and methodological issues arising in connection with sentiment analysis of texts are at least partly due to the position of this field at the intersection of the linguistic subfields of pragmatics and lexical semantics. In practice this means that we find many different proposals in the literature both for how sentiment information should be represented in the lexicon, to which kinds of lexical entities it should be attached (lemmas, lexemes or word senses), and how contextual information should be encoded and used in calculating the sentiment of a text passage from its constituent parts.

The methodological position taken here is that *prior sentiment* (or *polarity*) forms part of a word's sense, and that a word sense only has one prior polarity.[4] Connotations are considered to form part of the word sense (as opposed to, e.g., the practice in Princeton WordNet; Fellbaum, 1998). From this follows that if a word appears in text with two different sentiment values, it must either represent two senses of this lexeme or, alternatively, reflect a contextual effect, to be accounted for by invoking the venerable linguistic device of compositionality.

3 Towards a Swedish Sentiment Lexicon

In this paper we describe the creation of a gold standard (GS) for the sentiment annotation of Swedish terms as a first step towards the creation of a full-fledged sentiment lexicon for Swedish – i.e., a lexicon containing information about prior sentiment values of lexical items. For this purpose, we use human annotations of items sampled from a general-purpose computational lexical resource. More specifically, we employ a multi-stage approach combining

[2] According to a standard reference, *Ethnologue* (Simons and Fennig, 2017), there are about 7,000 languages in the world. A fair estimate would be that at the most 1,000 of these have a tradition of writing (Borin, 2009). Sentiment analysis tools are available for far fewer languages than this.

[3] E.g., the NRC Emotion Lexicon: `http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm`

[4] Notably, our use of *word sense* is to be construed as 'lexical word sense', which also is intended to cover lexicalized multi-word expressions.

corpus-based frequency sampling, direct annotation and Best–Worst Scaling (BWS) (Kiritchenko and Mohammad, 2016).

The remainder of this paper is structured as follows:

In Section 4 we describe SALDO, the Swedish lexical resource forming the basis for both the GS and the sentiment lexicon under construction. In Section 5 we describe our approach to compiling the GS. Section 6 is devoted to an analysis of the GS in order to arrive at a suitable sentiment model to be encoded in a Swedish sentiment lexicon. In the literature we find different proposed ways of modeling sentiment for a word sense or unit of text. The simplest model is the bipolar model, which assigns to each lexical unit a scalar, often normalized in the interval $[-1, +1]$ (with -1 representing the most negative possible sentiment, and $+1$ the most positive). SentiWordNet (Baccianella et al, 2010) and its gold standard Micro-WNOp (Cerini et al, 2007) use a model with two degrees of freedom. Each semantic unit in WordNet is assigned a three-dimensional vector (pos, neg, neu) with positive, negative and neutral components, normalized so that pos + neg + neu $= 1$ (this effectively gives 2 degrees of freedom). This model can be trivially converted to the previous one using sen $=$ pos $-$ neg.

In Section 7 we wrap up and point to future research directions.

Most technical details of our work have been left out of the present exposition. The companion papers Rouces et al (forthcoming-a) and Rouces et al (forthcoming-b) provide detailed technical information pertaining to the compilation of the GS and the construction of the sentiment lexicon, respectively.

4 SALDO

Both our GS and the sentiment lexicon under construction are based on SALDO, which is an existing large Swedish lexical-semantic computational resource (Borin et al, 2013). For the work described here, we use the current stable version SALDO v. 2.3, which contains 131,020 word senses.[5]

SALDO is organized as a lexical-semantic network of word senses, whose topology reflects semantic distance among the word senses. It is superficially similar to WordNet, but quite different from it in the principles by which it is structured. The basic organizational principle of SALDO is hierarchical. Every entry in SALDO – representing a word sense[6] – is supplied with one or more semantic descriptors, which are themselves also entries in the dictionary. All entries in SALDO are actually occurring words or conventionalized or lexicalized multi-word expressions (MWEs) of the language. The primary – obligatory – descriptor is the entry which better than any other entry fulfills two requirements: (1) it is a semantic neighbor of the entry to be described; and (2) it is more central than it.

[5] SALDO is freely available (under a CC-BY license) at `https://spraakbanken.gu.se/eng/resource/saldo`.

[6] Each word sense in SALDO is additionally connected to one or more form units (lemmas plus part of speech and full inflectional and compounding information).

That two entries are semantic neighbors means that there is a direct semantic relationship between them, for instance synonymy, hyponymy, argument–predicate relationship, etc. Centrality is determined by means of several criteria, the most important being frequency: a frequent entry is more central than an infrequent entry.[7] The basic linguistic idea underlying SALDO is in effect that, semantically speaking, the whole vocabulary of a language can be described as having a center – or core – and (consequently) a periphery. In SALDO, the higher levels in the hierarchy contain simpler and more basic entries. Contrast this with WordNet, where the higher nodes in the hierarchy contain very abstract vocabulary (e.g. 'entity').

5 Compiling the Gold Standard from SALDO

We aim to have a GS that assigns a sentiment to each SALDO entry. The bipolar sentiment model should be supported, but we also want to investigate the feasibility of using the SentiWordNet model. We have used a three-stage procedure for compiling the GS.

5.1 Corpus-Based Sampling

First, an initial sampling from SALDO was done following the distribution given by the estimated frequency of each word sense in the Gigaword corpus (Eide et al, 2016), which is a one-billion-word mixed-genre corpus of written Swedish.[8] Due to the Zipfian distribution of many kinds of linguistic items (Baayen, 2001), the GS would otherwise include mostly words that occur very rarely in written text, including rather obscure and outdated terms, as the lexicon has been designed to cover a time period from the mid-20th century until today.

We used the subset of the corpus covering the period from 1990 to the present (∼940 MW). Because the tokens in the corpus are not sense-disambiguated, we followed a simple heuristic. The different word senses for a given lemma are not annotated for their corpus frequency in SALDO, but the first sense is by design the most common one. Because the most common sense for a lemma in SALDO tends to occur around 70% of the time in corpus data (Nieto Piña and Johansson, 2016), we assume a distribution where the first of a lemma's n senses is given a probability of $\hat{p} = 0.7$, and each of the $n - 1$ remaining ones are given $\hat{p} = 0.3/(n-1)$. Then, for every polysemous lemma in the corpus, an associated word sense is sampled according to $\hat{p}$, and a count c for that word sense is increased. By using a sampling

[7] The actual work on SALDO relies mainly on the lexicographical experience and linguistic intuition of the compilers, who use clues such as stylistic value, word-formation complexity, the type of semantic relation holding between an entry and its primary descriptor, acquisition order in first-language acquisition, etc. Frequency correlates highly with these, however: It turns out that about 90% of the SALDO entries have primary descriptors which are at least as frequent as the entries themselves in a corpus of more than one billion words of Swedish. A more detailed description and discussion of the semantic organization of SALDO can be found in Borin et al (2013, 1196–1200).

[8] The corpus is freely available (under a CC-BY license) at `https://spraakbanken.gu.se/eng/resource/gigaword`.

based on a large corpus of the last two decades, the GS becomes more representative of modern written language. Namely, it is equivalent to sampling the tokens (sense-disambiguated lemmas) directly from modern text. By filtering out obscure and dated terms, we also reduce the proportion of terms that the annotators may not understand.

5.2 Best-Worst Scaling Filtered by Direct Annotation

Having annotators directly assign continuous sentiment scores to lexicon entries has several issues. It is difficult for annotators to remain consistent throughout their own annotation and across themselves. *Best-Worst Scaling* (BWS) annotation (Kiritchenko and Mohammad, 2016) has been proposed as an alternative. With BWS, annotators are presented tuples (usually 4-tuples) of items to annotate, and they select the highest and lowest according to the score at hand (in this case, the most positive and the most negative). If certain statistical properties are ensured about the appearance of elements in the tuples, then the number of times an element is chosen as most positive minus the number of times it is chosen as most negative can be used as a sentiment score.

However, we experienced that if the items are chosen by direct sampling from the lexicon or from a general corpus, most 4-tuples would not contain any items with a clear non-neutral polarity, let alone one most positive and one most negative item. Increasing the size of the tuples could solve this, but would imply a higher cognitive load for the annotator. Our solution to this problem is pre-filtering the initial set of terms by means of a preceding direct, but coarse-grained annotation that allows us to feed into the BWS annotation a subset of word senses with a more even distribution of sentiment values.

Using the corpus-derived distribution described above, we independently sampled 1998 word senses from SALDO, creating the set of words that would be annotated directly, W_{DA}. The sampling was filtered in order to avoid having too many difficult-to-judge non-content items (SALDO contains all parts of speech) in the annotation set. We also left out all multiword expressions and single-letter lemmas (typically corresponding to the names of letters of the alphabet, musical notes, or units of measurement). Thus only single-word adjectives, interjections, nouns, and verbs, having a lemma two letters or longer were sampled.[9]

We also sampled 200 additional word senses that were used for a joint annotation exercise across all annotators of W_{DA}, with the purpose of standarizing the annotation criteria.

Each of the three annotators then independently assigned a label to each word sense in W_{DA}. The possible labels are "positive","negative" or "neutral". All three annotators – coauthors of the present paper – are NLP researchers with formal backgrounds in linguistics and computer science, and native-level knowledge of Swedish.

For the BWS annotation, we selected those elements from W_{DA} that had been labeled as non-neutral by at least two annotators (278 items in total), which ensured that most 4-

[9] Lexical adverbs were not included, since this set holds too many function words. There are very few deadjectival adverbs in Swedish of the type *quickly*. These are instead normally rendered by the neuter singular indefinite form of the adjective.

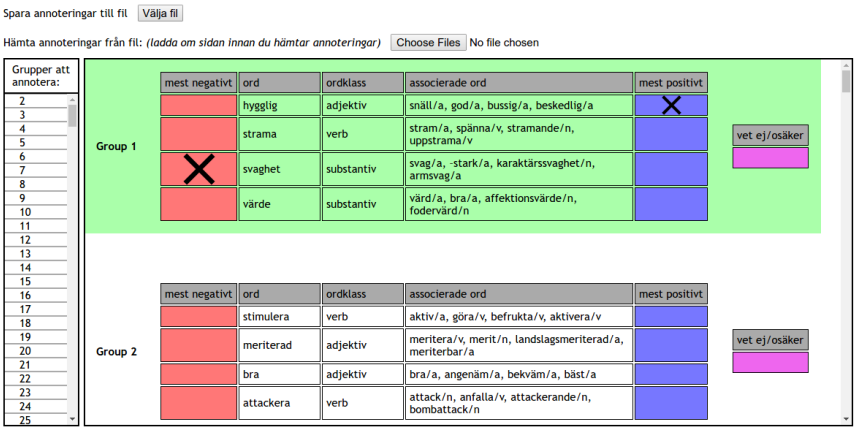

Fig. 1: Screenshot for the Best-Worst Scaling annotation interface. The labels for each group are 'most negative', 'word', 'part of speech', 'associated words', 'most positive', 'don't know/uncertain' from left to right.

tuples had clear candidates for most positive and most negative. From this set, we generated 572 4-tuples, in order to get a sufficient number of annotations per item (Kiritchenko and Mohammad, 2016).

We developed a web application (see Figure 1) that allows annotators to assign sentiments to SALDO word senses, using Best-Worst Scaling. The user can select the most positive and most negative item in each tuple, and also has an 'I don't know' option. It includes an interactive menu of pending groups, and the ability to save and load partial annotations to and from local files, allowing the annotators to organize their work over several sessions. We employed 4 annotators, who were different from the previous ones but also had formal background in (computational) linguistics and/or computer science, as well as native-level knowledge of Swedish.

6 Annotation Outcomes and Choice of Sentiment Model

We calculated interannotator agreement and other statistics for the annotations. In brief, the interannotator agreement was higher for BWS than for DA. See Rouces et al (forthcoming-a) for a detailed discussion.

The following table shows some representative scores obtained by BWS annotation.

w	gloss	$\mathbf{pos}_{\mathbf{BWS}}(w)$	$\mathbf{neg}_{\mathbf{BWS}}(w)$	$\mathbf{neu}_{\mathbf{BWS}}(w)$	$\mathbf{sen}_{\mathbf{BWS}}(w)$
svår..1	'difficult'	0.0500	0.3250	0.6250	-0.2750
slippa..1	'be spared'	0.2500	0.1944	0.5556	0.0556
depression..2	'depression	0.0000	0.4688	0.5312	-0.4688
stimulera..1	'stimulate'	0.1250	0.0000	0.8750	0.1250
absurd..1	'absurd'	0.0625	0.4375	0.5000	-0.3750

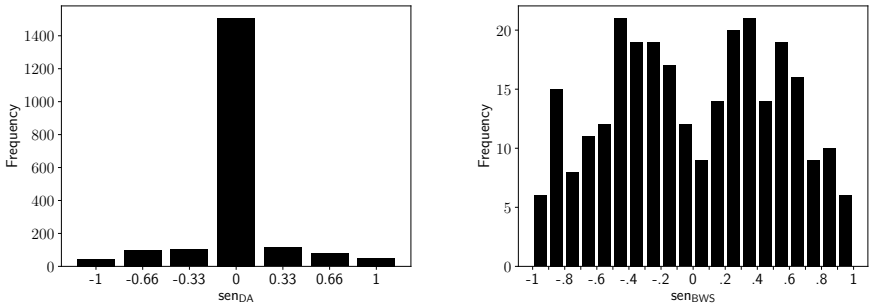

Fig. 2: Histograms of the sen values resulting from direct (left) and BWS (right) annotation

The histograms in Figure 2 shows the distributions of the (bipolar) sentiment values obtained with the two kinds of annotation, illustrating the effectiveness of the preliminary filtering steps in ensuring that the BWS annotators were presented mainly non-neutral items.

The output of the BWS annotation could be used both for the SentiWordNet and the bipolar model. From the results of the BWS annotation, 86 of 278 items have $\text{pos}_{\text{BWS}}(w) > 0$ and $\text{neg}_{\text{BWS}}(w) > 0$, but in many cases one of these components is small and a strong bias is common. The average over w of the value $\min(\text{pos}_{\text{BWS}}(w), \text{neg}_{\text{BWS}}(w))$, which reflects the overlap between the positive and negative components, is 0.022. In contrast, for Micro-WNOp, the GS used for SentiWordNet, which uses the same model but was obtained from direct annotation of the two variables 'pos' and 'neg', it is 0.015. Our higher value is probably due to the fact that we made W_{BWS} with a high proportion of non-neutral word senses, and therefore, a non-negligible proportion of the BWS 4-tuples contained elements that either were all negative or all positive, making the choice for most positive or most negative a sort of "lesser evil" or "lesser good", respectively. As an example, *absurd* from the table in Section 6, appeared in the annotation interface in a tuple containing [*dålig* 'bad', *utplåna* 'obliterate', *irriterad* 'irritated', *absurd*].

7 Towards a Sentiment Lexicon for Swedish

At the moment we are putting the resulting GS to the use for which it was intended: to train and compare different lexicon-based algorithms for creating a complete sentiment lexicon for Swedish. We have made initial experiments using both pure lexicon-based methods and methods combining lexical data and corpus information. This work is described in Rouces et al (forthcoming-b). The resulting resource – SenSALDO – will contribute significantly to the development of higher-yield TM tools in support of digital humanities research targeting Swedish data.

Acknowledgements

This work has been supported by a framework grant (*Towards a knowledge-based culturomics*;contract 2012-5738) as well as funding to Swedish CLARIN (*Swe-Clarin*;contract 2013-2003), both awarded by the Swedish Research Council, and by infrastructure funding granted to Språkbanken by the University of Gothenburg.

References

Baayen RH (2001) Word frequency distributions. Kluwer Academic Publishers, Dordrecht.

Baccianella S, Esuli A, Sebastiani F (2010) SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of LREC 2010, ELRA, Valletta, pp 2200–2204.

Bender EM (2011) On achieving and evaluating language-independence in NLP. Linguistic Issues in Language Technology 6(3).

Borin L (2009) Linguistic diversity in the information society. In: Proceedings of the SALTMIL workshop 2009, University of the Basque Country, Donostia, pp 1–7.

Borin L, Forsberg M, Lönngren L (2013) SALDO: A touch of yin to WordNet's yang. Language Resources and Evaluation 47(4):1191–1211.

Cerini S, Compagnoni V, Demontis A, Formentelli M, Gandini C (2007) Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. In: Sansò A (ed) Language Resources and Linguistic Theory, Franco Angeli, Milano, pp 200–210.

Eide SR, Tahmasebi N, Borin L (2016) The Swedish culturomics gigaword corpus: A one billion word Swedish reference dataset for NLP. In: Proceedings of the *From Digitization to Knowledge* workshop at DH 2016, Kraków, LiUEP, Linköping, pp 8–12.

Fellbaum C (ed) (1998) WordNet: An Electronic Lexical Database. The MIT Press, Cambridge, Mass.

Kiritchenko S, Mohammad SM (2016) Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In: Proceedings of NAACL 2016, ACL, San Diego, pp 811–817.

Nieto Piña L, Johansson R (2016) Embedding senses for efficient graph-based word sense disambiguation. In: Proceedings of TextGraphs-10, ACL, San Diego, pp 1–5.

Pang B, Lee L (2008) Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1–2):1–135.

Rouces J, Tahmasebi N, Borin L, Eide SR (forthcoming-a) Generating a gold standard for a Swedish sentiment lexicon. In: Proceedings of LREC 2018, ELRA, Miyazaki.

Rouces J, Tahmasebi N, Borin L, Eide SR (forthcoming-b) SenSALDO: Creating a sentiment lexicon for Swedish. In: Proceedings of LREC 2018, ELRA, Miyazaki.

Simons GF, Fennig CD (eds) (2017) Ethnologue: Languages of the world, twentieth edn. SIL International, Dallas, online version: `http://www.ethnologue.com`.