Generating a Gold Standard for a Swedish Sentiment Lexicon

Jacobo Rouces, Nina Tahmasebi, Lars Borin, Stian Rødven Eide
Sprakbanken, University of Gothenburg
{jacobo.rouces, nina.tahmasebi, lars.borin, 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 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 – i.e., a lexicon containing information about *prior* sentiment (also called polarity) values of lexical items (words or disambiguated word senses), along a scale negative–positive. We create a gold standard for sentiment annotation of Swedish terms, using the freely available SALDO lexicon and the Gigaword corpus. For this purpose, we employ a multi-stage approach combining corpus-based frequency sampling and two stages of human annotation: direct score annotation followed by Best-Worst Scaling. In addition to obtaining a gold standard, we analyze the data from our process and we draw conclusions about the optimal sentiment model.

Keywords: sentiment analysis, Swedish, gold standard, lexical resource

1. Introduction

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 (Fellbaum, 1998). In addition, many of the non-English sentiment lexicons that do exist have been compiled by (machine) translation from English resources, e.g., by Mohammad and Turney (2010)[1] and Chen and Skiena (2014), thereby arguably obscuring possible language-specific characteristics of sentiment-loaded vocabulary.

In this paper we describe the creation 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 – i.e., a lexicon containing information about *prior* sentiment (also called polarity) values of lexical items (words or disambiguated word senses), along a scale negative–positive. 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 corpus-based frequency sampling, direct score annotation and Best- Worst Scaling (BWS) (Kiritchenko and Mohammad, 2016).

2. State of the art

We base our gold standard on *SALDO* (Språkbanken, 2015a), which is an existing Swedish lexical-semantic computational resource (Borin et al., 2013). It is organized as a lexical-semantic network of word senses, whose topology reflects semantic distance among the word senses. 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). These are formally organized as an independent lexical resource – *SALDO's Morphology* (Språkbanken, 2015b) – which consequently can be used in NLP applications independently of SALDO, e.g., for lemmatization and morphological analysis of Swedish text. For the work described here, we use SALDO v. 2.3, which contains 131,020 word senses. SALDO is freely available (under a CC-BY license).

Different ways of modeling sentiment for a word sense or unit of text are possible.

The simplest (but not necessarily the less appropriate) model is the bipolar model, which assigns to each lexical unit a scalar, which is 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 (Fellbaum, 1998) is assigned a three-dimensional vector (pos, neg, neu) with positive, negative and neutral components, normalized so that $\text{pos}+\text{neg}+\text{neu} = 1$ (this effectively gives 2 degrees of freedom). This model can be trivially converted to the previous one using $\text{sen} = \text{pos} - \text{neg}$.

3. Annotation

We aim to have a gold standard that assigns a sentiment to each SALDO entry. The bipolar sentiment model should be supported, but we also want to investigate the feasibility and convenience of using the SentiWordNet model.

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 (Språkbanken, 2016).[2] Due to the Zipfian distribution of many kinds of linguistic items (Baayen, 2001), the gold standard would include, like the underlying lexicon SALDO does, 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. Thus, by using a sampling based on corpus of the last two decades, the gold standard

[1] http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

[2] The corpus is freely available under a CC-BY license.

becomes more representative of modern written language. By filtering out obscure and dated terms, we also reduce the proportion of terms that the annotators may not understand.

Having annotators directly assign continuous sentiment scores (be it (sen) or (pos,neg,neu)) to lexicon entries has several issues. It is difficult for annotators to remain consistent throughout their own annotation and across themselves; this is a rationale for instead relying on *Best-Worst Scaling* BWS annotation (Kiritchenko and Mohammad, 2016). With BWS, annotators are presented tuples (usually 4-tuples) of elements 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 BWS annotation is performed by directly sampling from the lexicon (or from a general corpus, for that matter), 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.

3.1. Direct Annotation

The initial sampling from SALDO was performed following the distribution given by the estimated frequency of each word senses in the Gigaword corpus (Eide et al., 2016), which is a one-billion-word corpus of Swedish text comprising newspaper and scientific articles, government reports, fiction and social media.[3] We used the subset of the corpus containing text written from 1990 to the present. 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 be the referred sense around 70% of the time (Nieto Piña and Johansson, 2016) (this figure is also a good approximation for other sense-disambiguation tasks in general (Gale et al., 1992; Kilgarriff, 2004)), we assume a distribution where the first sense is given a probability of $\hat{p} = 0.7$, and each of the n remaining ones are given $\hat{p} = 0.3/n$. This provides a reasonable approximation for the zipfian distribution, whose biggest differential is between the first and the second most common elements.

Using this distribution, we independently sampled 1998 word senses from SALDO, creating the set of words that would be annotated directly, W_{DA}. Each sample was performed independently, without replacement, using as probability distribution the normalized count, $c/\sum c$. 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 multi-word 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.

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 to a reasonable degree.

After the joint annotation, each of the three annotators independently assigned a label to each word sense in W_{DA}. The possible labels are "positive","negative" or "neutral". All three annotators are language technology/ natural language processing researchers with formal backgrounds in linguistics and computer science, and native-level knowledge of Swedish.

The annotated value of each word $w \in W_{\text{DA}}$ was calculated as for Equation 1, where A_{DA} is the set of DA annotators.

$$\text{sen}_{\text{DA}}(w) = \frac{\sum_{a \in A_{\text{DA}}} l_{\text{DA}}(a, w)}{|A_{\text{DA}}|}$$

$$l_{\text{DA}}(a, w) = \begin{cases} 1 & \text{if } a \text{ annotated } w \text{ as positive} \\ 0 & \text{if } a \text{ annotated } w \text{ as neutral} \\ -1 & \text{if } a \text{ annotated } w \text{ as negative} \end{cases} \quad (1)$$

3.2. Best-Worst Scaling Annotation

For the BWS annotation, we selected the set W_{BWS} defined as those elements from W_{DA} that had been labeled as non-neutral (positive or negative) by at least two annotators (Equation 2), which ensured that most 4-tuples had clear candidates for most positive and most negative.

$$W_{\text{BWS}} = \{w : w \in W_{\text{DA}} \wedge |\text{sen}_{\text{DA}}(w)| \geq 2/3\} \quad (2)$$

Since $|W_{\text{BWS}}| = 278$, we generated 572 4-tuples, which is greater than $|W_{\text{BWS}}| \cdot 2$ and thus largely sufficient for BWS (Kiritchenko and Mohammad, 2016).

We developed a web application that allows annotators to assign sentiments to SALDO word senses, using Best-Worst Scaling. The user can select the most positive and most negative SALDO entry for each tuple, including 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 spread their work over several sessions.

Word senses in SALDO do not have definitions or glosses, so in order for the user to be able to distinguish different senses of a lemma, we include a list of other lemmas that correspond to word senses which are associated to the item at hand. These are obtained using the semantic-network structure of SALDO.

Figure 1 includes a screenshot of the application. It is publicly available.[4]

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

[4]http://demo.spraakdata.gu.se/jacobo/bws-annotation/main.html

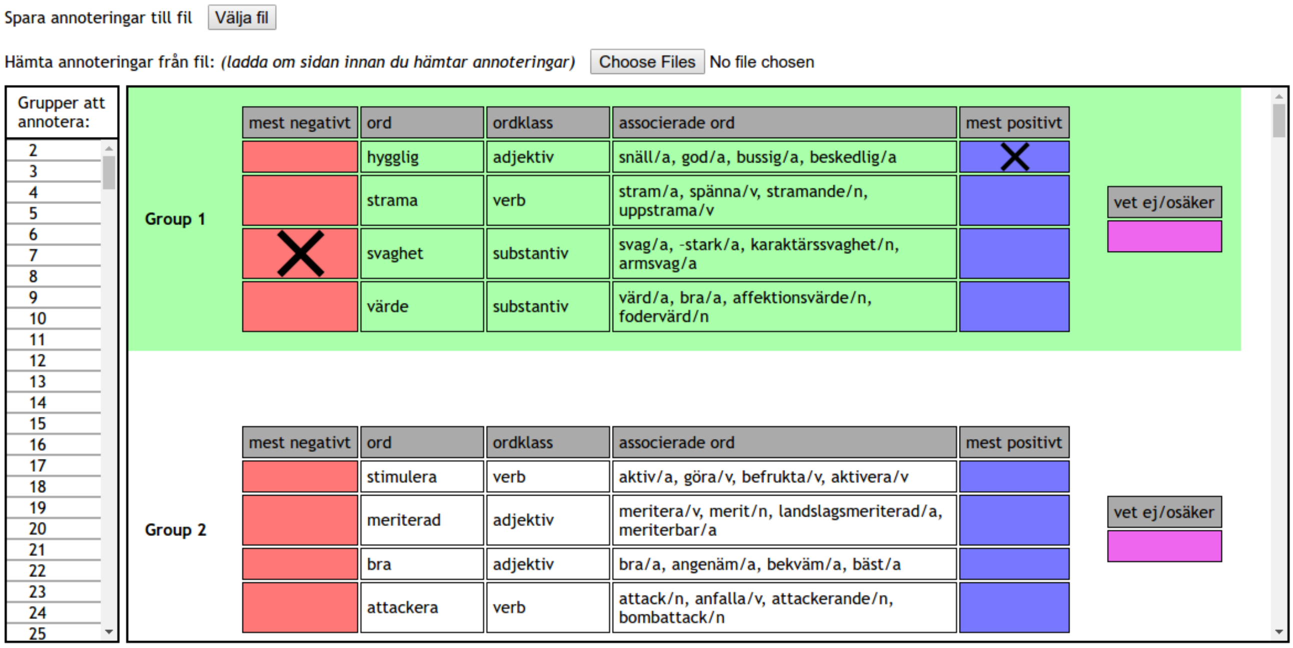


Figure 1: Screenshot for the Best-Worst Scaling annotation interface. The labels for each group are 'most negative', 'word', 'part of speech', 'associated words', 'most positive', 'don't know/uncertain' from left to right.

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.

The annotated values of each word $w \in W_{\text{BWS}}$ were calculated as for Equation 3, where A_{BWS} is the set of BWS annotators and $T(w)$ is the set of 4-tuples that contain the SALDO entry w. sen_{BWS} corresponds to the typical output used when applying BWS (either for sentiment analysis or for any other type of scaling). However, in Section 4. we will also analyze pos_{BWS}, neg_{BWS}, and neu_{BWS}, in order to determine whether they could be used to obtain a representation under the SentiWordNet model.

$$
\begin{aligned}
\text{pos}_{\text{BWS}}(w) &= \frac{\sum_{a \in A_{\text{BWS}}} \sum_{t \in T(w)} l^{\text{pos}}_{\text{DA}}(a,t,w)}{|A_{\text{BWS}}| \cdot |T(w)|} \\
l^{\text{pos}}_{\text{BWS}}(a,t,w) &= \begin{cases} 1 & \text{if } a \text{ annotated } w \text{ as most positive in } t \\ 0 & \text{otherwise} \end{cases} \\
\text{neg}_{\text{BWS}}(w) &= \frac{\sum_{a \in A_{\text{BWS}}} \sum_{t \in T(w)} l^{\text{neg}}_{\text{DA}}(a,t,w)}{|A_{\text{BWS}}| \cdot |T(w)|} \\
l^{\text{neg}}_{\text{BWS}}(a,t,w) &= \begin{cases} 1 & \text{if } a \text{ annotated } w \text{ as most negative in } t \\ 0 & \text{otherwise} \end{cases} \\
\text{neu}_{\text{BWS}}(w) &= 1 - \text{pos}_{\text{BWS}}(w) - \text{neg}_{\text{BWS}}(w) \\
\text{sen}_{\text{BWS}}(w) &= \text{pos}_{\text{BWS}}(w) - \text{neg}_{\text{BWS}}(w)
\end{aligned}
\tag{3}
$$

4. Results

4.1. Interannotator agreement

Table 2 shows the interannotator agreement for the two annotation stages: direct and BWS.

We chose Krippendorff's alpha instead of Cohen's or Fleiss's kappa, because in addition to the nominal metric, it allows for an interval metric as well. The interval metric takes into account that some annotation labels are more similar than others, i.e. the set of labels can be bijected to some metric space.

Under the interval metric, the inter-annotator agreement for BWS outperforms the one obtained from direct annotation, specially for the sen_{BWS}, which is the most comparable variable to direct annotation.

	nominal	**interval**
$\text{sen}_{\text{DA}}(w)$	0.480	0.529
$\text{pos}_{\text{BWS}}(w)$	0.551	0.889
$\text{neg}_{\text{BWS}}(w)$	0.621	0.893
$\text{neu}_{\text{BWS}}(w)$	0.446	0.744
$\text{sen}_{\text{BWS}}(w)$	0.462	0.927

Table 2: Interannotator agreements (Krippendorff's alpha, nominal and interval) for scores obtained from best-worst scaling (BWS) and direct annotation (DA). Since we used three annotators for sen_{DA}, in order to make the Krippendorff's alpha values comparable, we take the first 3 of the 4 annotators we used for BWS.

4.2. Annotation results

Figure 3 shows the histogram of the values obtained from the direct annotation stage. Figures 4a and 4b show the histogram of the values obtained from the BWS annotation, and Table 3 shows basic statistics. The data is publicly available under an open-source CC-BY 4.0 license.[5]

[5]http://demo.spraakdata.gu.se/jacobo/bws-annotation/data

w	gloss	$\mathbf{pos_{BWS}}(w)$	$\mathbf{neg_{BWS}}(w)$	$\mathbf{neu_{BWS}}(w)$	$\mathbf{sen_{BWS}}(w)$
medalj..1	‘medal’	0.4722	0.0000	0.5278	0.4722
lugna..1	‘calm (v)’	0.3333	0.0000	0.6667	0.3333
svår..1	‘difficult’	0.0500	0.3250	0.6250	-0.2750
gnista..3	‘spark, spunk’	0.5938	0.0000	0.4062	0.5938
slippa..1	‘be spared’	0.2500	0.1944	0.5556	0.0556
depression..2	‘depression	0.0000	0.4688	0.5312	-0.4688
möda..1	‘difficulty’	0.0312	0.2500	0.7188	-0.2188
stimulera..1	‘stimulate’	0.1250	0.0000	0.8750	0.1250
världsmästare..1	‘world champion’	0.5312	0.0000	0.4688	0.5312
absurd..1	‘absurd’	0.0625	0.4375	0.5000	-0.3750
rätt..1	‘correct (a)’	0.3125	0.0000	0.6875	0.3125
tryck..4	‘pressure’	0.1389	0.0000	0.8611	0.1389
protest..1	‘protest (n)’	0.1875	0.3125	0.5000	-0.1250
kraftfull..1	‘powerful’	0.3750	0.0000	0.6250	0.3750
överdrift..1	‘exaggeration’	0.1786	0.1429	0.6786	0.0357
mista..1	‘lose’	0.0833	0.5833	0.3333	-0.5000
trög..2	‘sluggish’	0.0500	0.2250	0.7250	-0.1750
misstanke..1	‘suspicion’	0.0750	0.3250	0.6000	-0.2500
hyllning..1	‘tribute’	0.8750	0.0000	0.1250	0.8750
förbud..1	‘prohibition’	0.0000	0.4286	0.5714	-0.4286

Table 1: Examples of sentiment scores obtained from BWS.

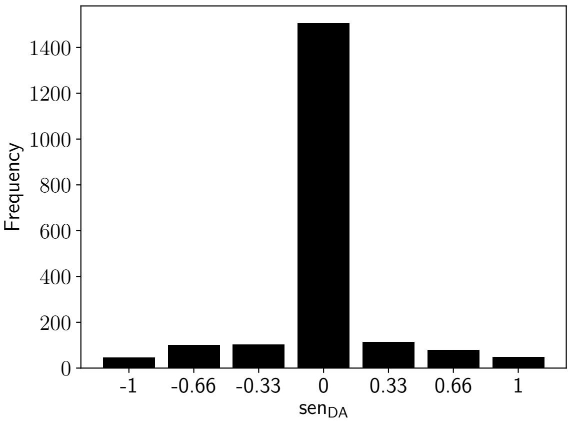
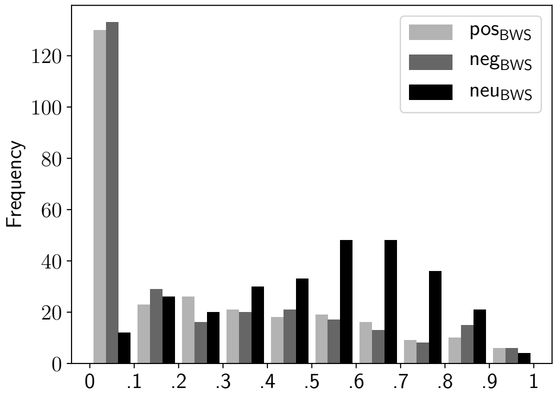


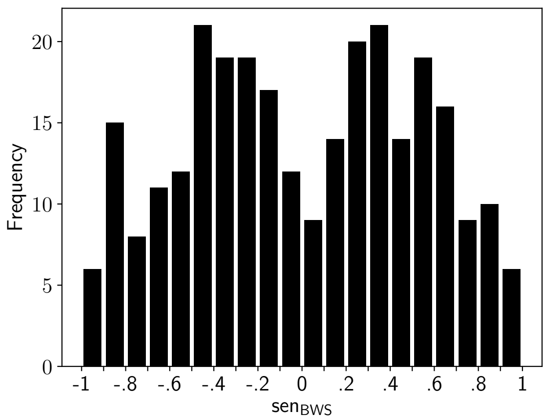
Figure 3: Histogram of the values obtained from the direct annotation stage.

	pos_{BWS}	neg_{BWS}	neu_{BWS}	sen_{BWS}
mean	0.248	0.246	0.505	0.002
std.	0.278	0.288	0.228	0.518

Table 3: Mean and standard deviation for values obtained from the BWS annotation.

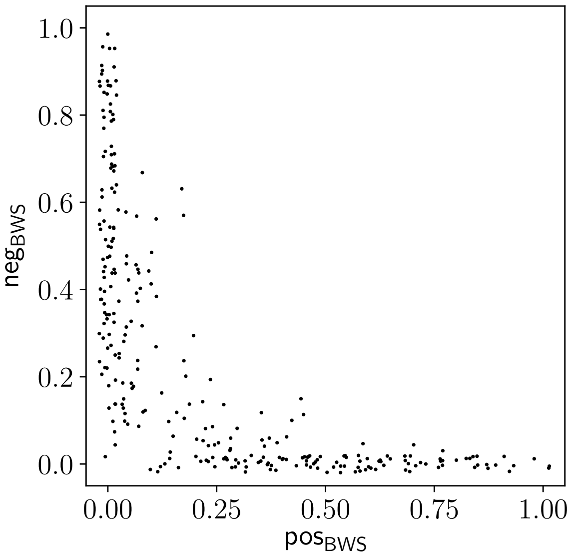


(a) Histogram of the (pos,neg,neu) values

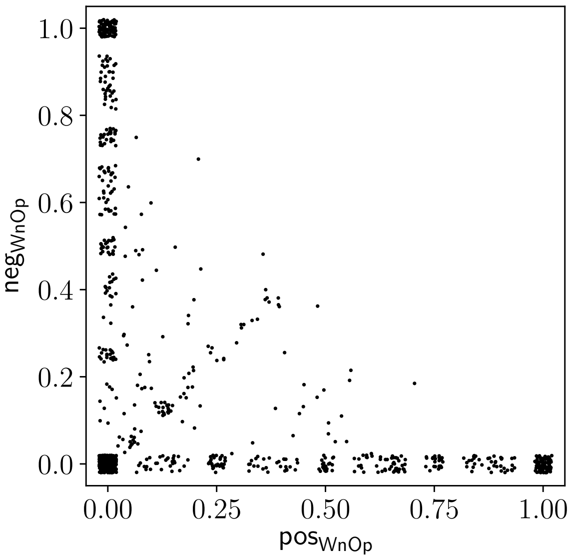


(b) Histogram of the sen value

Figure 4: Histogram of the values obtained from the BWS annotation



(a) Scatterplot of the BWS annotation



(b) Scatterplot of Micro-WNOp

Figure 2: Scatterplots of (pos,neg) values

4.3. Choice of Sentiment Model

The output of the BWS annotation could be used both for the SentiWordNet and the bipolar model ($\text{sen}(w) = \text{sen}_{\text{BWS}}(w), \text{pos}(w) = \text{pos}_{\text{BWS}}(w), \ldots$). In this section we analyze our data and the data from the gold standard used for building SentiWordnet, in order to determine whether the SentiWordNet model offers some advantage in return for its added complexity.

From the results of the BWS annotation, 86 of 278 SALDO ids 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 gold standard 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 Figure 1, appeared in the annotation interface in a tuple containing [*dålig* 'bad' , *utplåna* 'obliterate' , *irriterad* 'irritated' , *absurd*].

Figure 2a shows a scatterplot of the BWS annotation results adapted to the SentiWordNet model, using uniform ($[-0.02, 0.02]$) and independently distributed dithering. Figure 2b shows the equivalent plot for Micro-WNOp. The consistently low overlap between negative and positive components seems to indicate that the multi-dimensional SentiWordNet model is not necessary, or at least does not offer sufficient advantages to outweigh the simplicity and efficiency of the bipolar model.

5. Extensions and Future Work

The obtained gold standard is now being used to train and compare different lexicon-based algorithms for creating a complete sentiment lexicon for Swedish. In particular, we have experimented with growing sentiment lexicons based on a set of initial items and the lexical-semantic network structure of resources such as SALDO, plus contextual information from large corpora. We describe this work in more detail in Rouces et al. (2018). The resulting lexical resource, *SenSALDO* (Språkbanken, 2018), is freely available under a CC-BY license.

In parallel, we are also considering translating sentiment lexicons (from English). This is still future work, which will provide us with an opportunity to compare the results of the two approaches of building a Swedish sentiment lexicon from scratch based on monolingual resources, or of basing it on translation of an existing sentiment lexicon for another language.[6]

The more linguistic aspects of our work on the gold standard are treated in a companion publication to the present paper (Rouces et al., forthcoming), where we also discuss potential applications of (Swedish) sentiment lecicons in text mining for digital humanities research,

Acknowledgements

This work has been supported by a framework grant (*Towards a knowledge-based culturomics*[7] – contract 2012-5738) as well as funding to Swedish CLARIN (*Swe-Clarin*[8] – contract 2013-2003), both awarded by the Swedish Research Council, and by infrastructure funding granted to Språkbanken by the University of Gothenburg.

6. Bibliographical References

Baayen, R. H. (2001). *Word frequency distributions*. Kluwer Academic Publishers, Dordrecht.

[6] In the literature, these two approaches are often referred to as the *merge* (monolingual) and *expand* (translation) approach, respectively (Vossen, 1998).

[7] `https://spraakbanken.gu.se/eng/culturomics`

[8] `https://sweclarin.se/eng`

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

Borin, L., Forsberg, M., and 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., and Gandini, C. (2007). Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. In Andrea Sansò (ed), *Language Resources and Linguistic Theory*, pages 200–210. Franco Angeli, Milano.

Chen, Y. and Skiena, S. (2014). Building sentiment lexicons for all major languages. In *Proceedings of ACL 2014*, pages 383–389, Baltimore. ACL.

Eide, S. R., Tahmasebi, N., and 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*, pages 8–12, Linköping. LiUEP.

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

Gale, W., Church, K. W., and Yarowsky, D. (1992). Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings of ACL 1992*, ACL '92, pages 249–256, Newark. ACL.

Kilgarriff, A. (2004). How dominant is the commonest sense of a word? In Petr Sojka, et al. (eds), *Text, Speech and Dialogue*, pages 103–111, Berlin. Springer.

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

Mohammad, S. and Turney, P. (2010). Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, Los Angeles. ACL.

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

Rouces, J., Tahmasebi, N., Borin, L., and Eide, S. R. (2018). SenSALDO: Creating a sentiment lexicon for Swedish. In *Proceedings of LREC 2018*, Miyazaki. ELRA.

Rouces, J., Borin, L., Tahmasebi, N., and Eide, S. R. (forthcoming). Defining a gold standard for a Swedish sentiment lexicon: Towards higher-yield text mining in the digital humanities. In *Proceedings of DHN 2018, Helsinki*, Aachen. CEUR WS.

Vossen, P. (1998). Introduction to EuroWordNet. *Computers and the Humanities*, 32(2–3):73–89.

7. Language Resource References

Språkbanken. (2015a). *SALDO*. Språkbanken, University of Gothenburg, Sweden, Språkbanken lexical resources: SALDO, v. 2.3. A lexical-semantic resource for modern Swedish (∼130k entries), URL `https://spraakbanken.gu.se/eng/resource/saldo`.

Språkbanken. (2015b). *SALDO's Morphology*. Språkbanken, University of Gothenburg, Sweden, Språkbanken lexical resources: SALDO's Morphology, v. 2.3. A morphological lexicon (full-form and inflectional paradigms) for modern Swedish (∼128k entries), URL `https://spraakbanken.gu.se/eng/resource/saldom`.

Språkbanken. (2016). *The Swedish Culturomics Gigaword Corpus*. Språkbanken, University of Gothenburg, Sweden, Språkbanken corpus resources: Swedish Culturomics Gigaword Corpus, v. 1.0. A mixed-genre corpus of modern written Swedish comprising one billion words, URL `https://spraakbanken.gu.se/eng/resource/gigaword`.

Språkbanken. (2018). *SenSALDO*. Språkbanken, University of Gothenburg, Sweden, Språkbanken lexical resources: SenSALDO, v. 0.1. A word-sense prior-polarity sentiment lexicon for modern Swedish (∼7.6k entries), URL `https://spraakbanken.gu.se/eng/resource/sensaldo`, `http://hdl.handle.net/10794/sensaldo`.