SenSALDO: Creating a Sentiment Lexicon for Swedish

Jacobo Rouces, Nina Tahmasebi, Lars Borin, Stian Rødven Eide
Språkbanken, University of Gothenburg
{jacobo.rouces, nina.tahmasebi, lars.borin, stian.rodven.eide}@gu.se

Abstract

The natural language processing subfield known as *sentiment analysis* or *opinion mining* has seen an explosive expansion over the last decade or so, and sentiment analysis has become a standard item in the NLP toolbox. Still, many theoretical and methodological questions remain unanswered and resource gaps unfilled. Most work on automated sentiment analysis has been done on English and a few other languages; for most written languages of the world, this tool is not available. This paper describes the development of an extensive sentiment lexicon for written (standard) Swedish. We investigate different methods for developing a sentiment lexicon for Swedish. We use an existing gold standard dataset for training and testing. For each word sense from the SALDO Swedish lexicon, we assign a real value sentiment score in the range [-1,1] and produce a sentiment label. We implement and evaluate three methods: a graph-based method that iterates over the SALDO structure, a method based on random paths over the SALDO structure and a corpus-driven method based on word embeddings. The resulting sense-disambiguated sentiment lexicon (SenSALDO) is an open source resource and freely available from Språkbanken, The Swedish Language Bank at the University of Gothenburg.

Keywords: sentiment analysis, Swedish, lexicon, lexical resource

1. Introduction

The natural language processing (NLP) subfield known as *sentiment analysis* or *opinion mining* 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 implement in the NLP toolbox, many theoretical and methodological questions remain unanswered and resource gaps need to be filled. Most work on automated sentiment analysis has been done on English and a few other languages; for most of the written languages of the world,[1] this tool is not available. This paper describes the development of an extensive sentiment lexicon for written (standard) Swedish, an essential component in sentiment analysis.

The theoretical and methodological issues that arise in connection with sentiment analysis of texts lie partly in the intersection of the linguistic subfields of pragmatics and lexical semantics. Depending on your view of the scope of these subdisciplines, you may end up with very different thoughts about the prior-polarity – i.e., lexical-semantic – and contextual – i.e., pragmatic – elements of sentiment information, and how these are combined in concrete text analysis. An added complication is that sentiments and emotions are central objects of study also in other fields, notably psychology.

In practice this means that we find many different proposals in the literature, for how prior sentiment polarity should be represented in the lexicon, which kinds of lexical entities should be attached (lemmas, lexemes or word senses), and how contextual information is to be encoded and used when calculating the sentiment of a text passage from its constituent parts.

The methodological position taken in this paper is, in brief, that prior sentiment polarity forms part of a word's sense, and that a word sense only has one prior polarity. In our case the polarity is expressed as a real number in the range $[-1, 1]$, with higher positive values associated to more positive sentiments. Connotations are considered to form part of the word sense (as opposed to, e.g., the practice in WordNet). From this it 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.

The focus on word senses as bearers of prior polarity is in line with our general view on lexical-semantic resources for NLP, where the word sense takes center stage.[2] Thus, our point of departure in this paper is the Swedish SALDO lexical resource (Språkbanken, 2015b). SALDO is an onomasiological lexicon, i.e., organized by content (lexical entries are word senses), rather than by form (lemmas or lexemes). For a detailed description of the organization of SALDO and a discussion of the underlying theoretical and methodological principles, we refer the reader to Borin et al. (2013).

However, one aspect of SALDO's organization will be important in the context of what follows below, namely the basic lexical-semantic relations defining the network structure of SALDO, which provide important information for creating SenSALDO. 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[3] – is supplied with one or more semantic descriptors,

[1] According to a standard reference, *Ethnologue* (Simons and Fennig, 2017), there are about 7,000 spoken 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.

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

[3] Each word sense in SALDO is additionally connected to one or more form units (lemmas plus part of speech and full inflec-

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 close semantic neighbor of the entry to be described; and (2) it is more central than it.

In defining SALDO, criterion (1), semantic neighborhood, is understood as involving a direct semantic relationship between lexical items,[4] for instance synonymy, hyponymy, argument–predicate relationship, etc. Since there can be only one primary descriptor for any given entry in SALDO, one of these relationships must be chosen in every case, but this will not necessarily be the same. The predecessor of SALDO was characterized as an "associative thesaurus" (Lönngren, 1998), and its basic structure can still be said to be 'micro-thesaural', i.e., more similar to what we find in Roget's *Thesaurus* (Roget, 1852) or its Swedish counterpart by Bring (Bring, 1930; Borin et al., 2014; Borin et al., 2015) (Språkbanken, 2015a; Språkbanken, 2017a) rather than the straightforward taxonomical structure exemplified by WordNet.

As for criterion (2), centrality is determined by means of several criteria, e.g., 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. In practice, however, frequency is among the best predictors of centrality even when the latter has been determined by these other means. 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.

Since the primary descriptor must be another actual lexical entry, in reality SALDO forms 40-some different hierarchies, where no more suitable primary descriptor can be found.[5] Here, an artificial lexical item (called *PRIM*) is used in order to make a single rooted tree for the primary-descriptor relation.

In addition to the obligatory primary descriptor, any number of secondary decsriptors can be added, whose main purpose is to disambiguate or further group entries with the same primary descriptor. Their usage is much more pragmatic and less consistent than in the case of the primary descriptors.

The lexical-semantic organization of SALDO is predicated on the idea of the vocabulary of a language having a core part and a more peripheral part. Consequently, the higher levels in the lexical-semantic hierarchy of SALDO contain simpler and more basic entries. Contrast this with WordNet, where the higher nodes in the hierarchy contain very abstract vocabulary (e.g. 'entity').

2. State of the Art

Many methods have been developed and tested for building sentiment lexicons, English being the most popular language for these. The methods may rely on corpus analysis (making use of word co-occurrence, syntactic patterns, or distant-supervision signals) or on existing lexicons (usually using some sort of label propagation exploiting the structure of the lexicon), although both approaches can be combined (Devitt and Ahmad, 2013; Hamilton et al., 2016). The different methods can also involve a varying degree of manual annotation.

Among the English lexicons built with mostly-automatic lexicon-driven methods, SentiWordNet (Baccianella et al., 2010) has become a popular resource. It is created by combining a semi-supervised learning step that uses existing relations between WordNet 3.0 entries (Fellbaum, 1998), (such as *synonymy*, *antonymy*, and *related with*), and a random-walk step over a graph built using the implicit *definiens-definiendum* relation between words in the entries and words in the glosses of the entries (Esuli and Sebastiani, 2007). However, these relations require WordNet, or an equivalent lexicon, which in turn requires a big amount of manual work by trained lexicographers. Therefore, this kind of approach has severe limitations for languages with fewer resources than English.

Among the English lexicons using corpus-driven approaches, SENTPROP (Hamilton et al., 2016) is a recent state-of-the-art approach that builds a directed weighted graph of terms using the nearest neighbors in the space of word embeddings obtained from applying singular value decomposition to the positive pointwise mutual information matrix obtained from the corpus. Then, it uses random walks in a similar fashion to SentiWordNet.

Given a set of labeled training words annotated as positive and negative, Rothe et al. (2016) find an orthogonal transformation of the embedding space that maximizes the distance among those with different labels and minimizes the distance among those with the same label.

Amir et al. (2015) train different linear regression models (least squares and regularized variants) over different word embeddings (GloVe, CBOW, skip-gram, struct skip-gram) **add something like: comparable to our word2vec method**. Bar-Haim et al. (2017) expand an already existing sentiment lexicon by training a linear SVM. They obtain an accuracy of 90.5%. Both of these methods are applied exclusively on English and produce sentiment labels for non-disambiguated lemmas.

For Swedish, two openly available sentiment lexicons exist (Nusko et al., 2016; Rosell and Kann, 2010). In addition, there are some Swedish sentiment lexicons or word lists produced by automatic translation of corresponding English resources, e.g., by Mohammad and Turney (2010)[6] and Chen and Skiena (2014).

Rosell and Kann (2010) developed a Swedish sentiment lexicon using random walks over a graph of synonyms and

tional and compounding information). These are formally organized as an independent lexical resource – *SALDO's Morphology* (Språkbanken, 2015c) – which consequently can be used in NLP applications independently of SALDO, e.g., for lemmatization and morphological analysis of Swedish text.

[4] SALDO contains all parts of speech, not only the open lexical classes. Thus the noun *yta* 'surface' has as its primary descriptor the preposition *på* 'on'.

[5] For instance, the preposition *på* 'on' has no primary descriptor.

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

4 positive and 4 negative seed words. The graph was built using the *Synlex/People's Dictionary of Synonyms* (Kann and Rosell, 2005), which used Swedish-English lemma pairs concatenated with their inverse relation to generate candidate synonym pairs. The pairs were filtered by grading and then averaging the grades. The result of Synlex was 16006 words with 18920 weighted pairs, which were used as edges of the graph in the random walks. The random walk algorithm is described in Algorithm 1 (with some notational changes from Kann and Rosell (2005) to describe our extensions later in Section 3.).

Algorithm 1 Random walk Algorithm

Require: Similarity weighted directed graph G with set W words as vertices and weighted edges defined by similarity function $0 \leq \text{sim}(w_1, w_2) \leq 1$.

function $\text{prob}(w_1, w_2) = \frac{\text{sim}(w_1,w_2)}{\sum_{w_3 \in W} \text{sim}(w_1,w_3)}$
for all word w **do**
  for all $i \in [0, 1, ..., I]$ **do** ▷ $I = 100$
    $v_+ \leftarrow 0$ and $v_- \leftarrow 0$ and $w_{\text{now}} \leftarrow w$ and $l \leftarrow 0$
    for all $j \in [0, 1, ..., M]$ **do** ▷ $M = 250$
      sample w_{next} with $\text{prob}(w_{\text{now}}, w_{\text{next}})$
      $w_{\text{now}} \leftarrow w_{\text{next}}$
      $l \leftarrow 1/\text{sim}(w_{\text{now}}, w_{\text{next}})$
      if $w_{\text{now}} \in S_+$ **then**
        $v_+ \leftarrow v_+ + m/l$
      if $w_{\text{now}} \in S_-$ **then**
        $v_- \leftarrow v_- + m/l$
  sentiment$(w) \leftarrow \frac{v_+ - v_-}{I}$

Nusko et al. (2016) propose a tree traversal method on the tree defined by the primary descriptor relation between SALDO entries. This method starts with 6 seeds with a manually assigned polarity and recursively calculates the sentiment of children based on the sentiment of the parent.[7] The algorithm calculates a confidence score for each sentiment, which decreases at a constant rate from the distance to the original seed (steps of -0.25 from a confidence of 1 for the descendants of the core words), and sets a threshold of 0.5 as the lowest acceptable confidence. It also uses secondary descriptors, but only when the secondary descriptor is *inte* (Swedish negation 'not'), which indicates that the child and parent have opposite semantic values and therefore the sign on the sentiment value should also be inverted, or a strength modifier like *lite* 'a little', or *enastående* 'outstanding'. It obtains a sentiment for 2133 entries. Three annotators labeled 150 entries as positive, negative or neutral, and for 117 of the entries the annotators were in full agreement. From these a 71% precision was obtained. The original 150 entries were sampled using equally sized stratification over the three confidence levels.

3. Methods

We model the sentiment associated to a word sense using a real value in the interval $[-1, 1]$. After first considering using a three-dimensional model like that of SentiWordNet (Baccianella et al., 2010), we collected some experimental evidence indicating that this was largely unnecessary since the additional degree of freedom was all but unused in practice (Rouces et al., 2018).

We implement different methods, which we describe below, extending the methods in Rosell and Kann (2010) and Nusko et al. (2016) and also try a corpus-oriented approach similar to the one in Hamilton et al. (2016). For all methods, we produce continuous scores and discrete labels (positive, neutral, negative). What is relevant about the scores is not their magnitudes but the relative order that they produce. The values and their distributions depend on idiosyncrasies of the methods employed and do not necessarily resemble what would be produced by direct human annotations, but instead can be fit to any desired distribution. The discrete labels are less fine-grained, but may be more appropriate for certain applications.

3.1. Inheritance over Graph

Our first method is a modified and extended version of the tree traversal method presented by Nusko et al. (2016), where sentiment of a word sense is inherited from the primary descriptor (which defines a tree structure). We extended it such that the traversal occurs over the directed acyclic graph defined by using both primary and secondary descriptors. The secondary descriptors of an entry are used not only for polarity inversion or intensification, but their sentiment value is also used, although with a lower weight.

The algorithm cannot use a simple breadth-first exploration, because for a given node, in general, some incoming neighbors will be at a different distance from the seed set than others, and the node will be reached before all the incoming neighbors have been calculated. This prevents all elements in the frontier to be expanded in a single iteration.

In addition, even when attempting different partially-successful passes over the frontier, the algorithm would stagnate easily because some secondary descriptors are not reachable from the given seed words. For this reason, the algorithm incorporates a best-effort mechanism for stagnation cases, whereby a sentiment is calculated for a node with the lowest possible number of unreached secondary-descriptor incoming nodes, ignoring these, and a new pass is performed over the frontier. Primary descriptors are never ignored. A priority queue is used for the nodes with unreached secondary-descriptor incoming nodes. If it is still not possible to calculate new nodes from all their parents, the process is repeated until it is possible, or the queue and the frontier are empty.

We used the same seeds as Nusko et al. (2016).

The method outputs sentiment scores for each sense, so in order to obtain discrete labels we apply thresholds. The thresholds are obtained from the percentiles of each class in a training set obtained from sampling two thirds of the gold standard described in section 4.. For example, the positive class in the gold standard corresponds to 6.34%, so we classify the 6.34% most positive senses as *positive*. The other third is used for testing. This is the only learning needed by this algorithm.

[7] In Nusko et al. (2016) the seeds and their children are referred as "core words" and "seeds" respectively.

3.2. Random Paths over Graphs

Our second method is an adaptation of the method developed by Rosell and Kann (2010) and presented in the algorithm in Algorithm 1. Our adaptation allows it to be used on SALDO sense-disambiguated entries instead of Swedish lemmas, and extends the edges of graph G (by creating new non-zero sim() relations) in order to make it more dense and prevent the existence of isolated nodes, which would be too common using only Synlex (which has $131,020$ nodes connected by only $25,516$ relations).
For this, we use the union of several graphs.

- A manual conversion of Synlex to SALDO. Rather than the original Synlex, which is a crowd-sourced resource with many questionable entries, and which is not word-sense disambiguated, we use a manually curated version which forms part of an emerging Swedish enriched wordnet, *Swesaurus* (Språkbanken, 2017b),[8] where an experienced lexicographer has (i) removed incorrect entries and (ii) replaced each remaining entry with its corresponding SALDO word sense. In some cases the degree of synonymy (the weight) has also been modified. We use the original weights in the $[0, 1]$ interval.
- The edges defined by primary descriptors in SALDO.
- The edges defined by secondary descriptors in SALDO.
- The edges that connect SALDO entries that have the same primary descriptor.

Since the new graphs are unweighted, their edges were assigned a weight of 0.5. This is a simple heuristic that reflects that they represent a certain level of similarity, but not the level of synonymy represented by a value of 1 in Synlex.

3.3. Classification over word2vec

As opposed to the previous methods, which are purely lexicon-driven, this approach is partly corpus-based. We used already existing vector representations of SALDO entries (which are sense-disambiguated) that were derived from *word2vec* lemma embeddings (Johansson and Nieto Piña, 2015). This was done by solving a constrained optimization problem where each lemma embedding is a linear combination of the embeddings of the senses associated to that lemma, and the distance between neighboring senses (i.e. neighbors in SALDO's descriptor graph) is minimized. The corpus size was 1 billion words, and the vector space dimensionality was 512. We trained a logistic regression (logit) classifier and a support vector classifier with a radial basis function (RBF) kernel. All the classifiers used a one-vs-rest approach of the three-class classification. For the classifiers we used 5-fold cross-validation stratified by the (pos,neu,neg) classes. For each fold, the SVM/RBF meta-parameters (C, γ) were estimated using 5-fold cross-validation over the training set. Although not equivalent, the linear nature of the logit classifier makes it comparable to the method in Rothe et al. (2016).
These methods output labels, but scores are obtained computing $p((pos) - p((neg))$, where p is the probability for a given entry to belong to the positive or negative classes. For the logit classifier, p is straightforward. For the support vector classifier, an extension of Platt scaling for multiple classes is used (Wu et al., 2004).

[8] Swesaurus contains information from several sources, but the curated Synlex data can be extracted from the LMF XML file by finding `SenseRelation` elements (inside `Sense` elements) containing `<feat att="source" val="fsl" />`.

4. Results

For training and testing the different methods, we used the direct annotation gold standard developed in Rouces et al. (2018), composed of 1998 entries from SALDO entries labeled as negative (value -1), neutral (value 0), or positive (value $+1$). The values were averaged over three annotators (so if an entry is labeled as positive by two annotators and as neutral by one, the final value would be $2/3$).
Table 1 shows the results for each method. We employed two different sets of measures for measuring the quality of the gold standard: one is based on ranks and other is based on discrete labels.

- The rank-based measures are the Spearman rank-order correlation coefficient (ρ) (Kokoska and Zwillinger, 2000), in the interval $[-1, 1]$, the p-normalized Kendall tau distance (τ_p) (Fagin et al., 2004) in the interval $[0, 1]$ (the one used in Baccianella et al. (2010)), and Kendall's tau-b (τ_b) (Kendall, 1945) (the one used in Rothe et al. (2016)). Both τ_p and τ_b are suited to handle ties —which in our case means word senses with equal sentiment values— but they do so in different ways. In addition to the direct annotation values in the test set, we also used more fine-grained sentiment values of 278 entries that are available as part of the same gold standard (Rouces et al., 2018), which were obtained using Best-Worst Scaling (BWS) and also comprised in the $[-1, 1]$ range. The reason for this is that these values are more fine-grained than the Direct Annotation (DA) values (which due to the use of 3 annotators range over only 7 possible values), and therefore ties are less common in the gold standard, making some ranking comparison algorithms more suitable. Since the BWS values were created only for the entries annotated as non-neutral by the DA scoring ($|\text{value}| \geq 0.5$), they cannot all be used for testing (or else the training set would be too biased towards neutral elements). Therefore, the intersection of the DA test set and the entries with BWS value is used for applying the rank-based measures.
- The measures based on discrete labels are the precision and recall values for each label, derived from the confusion matrix, which is also included.

'Graph inheritance' corresponds to the method in Nusko et al. (2016), although the results are not completely equivalent because of the stratification used in the evaluation in Nusko et al. (2016). 'Graph inheritance ext' corresponds to the extended version described in Section 3.1.. 'Graph

	DA									BWS
	ρ	τ_p	τ_b	precision	recall	acc.	confusion matrix			τ_b
							GS	SL		
								pos	neu	neg	
graph inheritance	0.39	0.39	0.38	pos: 0.28 neu: 0.91 neg: 0.33	pos: 0.26 neu: 0.90 neg: 0.42	0.82	pos neu neg	10 23 3	28 391 12	1 21 11	0.49
graph inheritance ext	0.33	0.42	0.32	pos: 0.22 neu: 0.90 neg: 0.27	pos: 0.21 neu: 0.89 neg: 0.35	0.81	pos neu neg	8 26 2	30 386 15	1 23 9	0.46
graph random paths	0.30	0.31	0.24	pos: 0.25 neu: 0.90 neg: 0.39	pos: 0.23 neu: 0.90 neg: 0.50	0.82	pos neu neg	9 26 1	29 390 12	1 19 13	0.46
word2vec +logit	0.47	0.21	0.38	pos: 0.37 neu: 0.93 neg: 0.46	pos: 0.54 neu: 0.88 neg: 0.52	0.84	pos neu neg	15 25 1	13 301 11	0 15 13	0.61
word2vec +svc /rbf	**0.55**	**0.15**	**0.45**	pos: 0.65 neu: 0.92 neg: 0.65	pos: 0.46 neu: 0.96 neg: 0.44	**0.89**	pos neu neg	13 7 0	15 328 14	0 6 11	**0.62**

Table 1: Results for evaluating the different methods for constructing the sentiment lexicon in Swedish. Note that the Kendall tau τ_p is a distance, and therefore it is inversely related to the Spearman correlation ρ. GS and SL stand for gold standard and sentiment lexicon respectively.

random paths' corresponds to the method described in Section 3.2.. The last three rows correspond to the results of the different classifiers used in Section 3.3..

word sense ID	gloss	value	label
ond..4	'bad'	-0.9959	neg
farlig..1	'dangerous'	-0.9919	neg
kriminalitet..1	'criminality'	-0.9838	neg
skrämma..1	'frighten'	-0.9797	neg
problem..1	'problem'	-0.9716	neg
angrepp..1	'attack'	-0.9594	neg
förhållande..1	'relationship'	-0.0345	neu
radio..1	'radio'	-0.0264	neu
sälja..1	'sell'	-0.0223	neu
surdeg..1	'sourdough'	0.0426	neu
god..2	'tasty'	0.9675	pos
riktig..2	'genuine'	0.9716	pos
hjälpa..1	'help (v)'	0.9797	pos

Table 2: Examples of sentiment values and labels obtained with the word2vec-svc-rbf method. The values have been fitted to the uniform distribution in $[-1, +1]$

The extension of the original graph inheritance method by using all the secondary descriptors in SALDO slightly reduces the quality of the results, which seems to indicate that the semantic connection behind the secondary descriptors in general is too weak and not useful for this task.

The method word2vec-svc-rbf performs consistently better than the rest. Table 2 shows some examples obtained from this method. SentiWordNet is reported to have τ_p values of 0.281 and 0.231 for positive and negative dimensions (their sentiment model has 2 degrees of freedom). All our embeddings-based methods outperform both measures (τ_p is a distance, and therefore lower values are desired). Rothe et al. (2016) reports $\tau_b = 0.654$. We obtain $\tau_b = 0.45$ when testing against the DA values, which is significantly lower. However, this probably owes to τ_b penalizing the big amount of ties in the DA values (61.95% of the possible pairs), as the method obtains $\tau_b = 0.63$ (a very close value) when testing against the BWS values, where ties are much less common (0.63%).

The resulting sentiment lexicon – SenSALDO (Språkbanken, 2018) is an open-source resource and freely available from Språkbanken, The Swedish Language Bank at the University of Gothenburg. In a companion publication (Rouces et al., forthcoming), we discuss applications of this resource in text mining for digital humanities research.

Acknowledgements

This work has been supported by a framework grant (*Towards a knowledge-based culturomics*[9] – contract 2012-5738) as well as funding to Swedish CLARIN (*SweClarin*[10] – contract 2013-2003), both awarded by the Swedish Research Council, and by infrastructure funding granted to Språkbanken by the University of Gothenburg.

5. Bibliographical References

Amir, S., Astudillo, R., Ling, W., Martins, B., Silva, M. J., and Trancoso, I. (2015). Inesc-id: A regression model for large scale twitter sentiment lexicon induction. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 613–618.

[9] https://spraakbanken.gu.se/eng/culturomics

[10] 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.

Bar-Haim, R., Edelstein, L., Jochim, C., and Slonim, N. (2017). Improving claim stance classification with lexical knowledge expansion and context utilization. In *Proceedings of the 4th Workshop on Argument Mining*, pages 32–38.

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.

Borin, L., Allwood, J., and de Melo, G. (2014). Bring vs. MTRoget: Evaluating automatic thesaurus translation. In *Proceedings of LREC 2014*, pages 2115–2121, Reykjavík. ELRA.

Borin, L., Nieto Piña, L., and de Johansson, R. (2015). Here be dragons? The perils and promises of inter-resource lexical-semantic mapping. In *Semantic resources and semantic annotation for Natural Language Processing and the Digital Humanities. Workshop at NODALIDA 2015*, pages 1–11, Linköping. Linköping Electronic Conference Proceedings.

Borin, L. (2009). Linguistic diversity in the information society. In *Proceedings of the SALTMIL 2009 workshop on Information Retrieval and Information Extraction for Less Resourced Languages*, pages 1–7, Donostia. University of the Basque Country.

Bring, S. C. (1930). *Svenskt ordförråd ordnat i begreppsklasser*. Hugo Gebers förlag, Stockholm.

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

Devitt, A. and Ahmad, K. (2013). Is there a language of sentiment? An analysis of lexical resources for sentiment analysis. *Language Resources and Evaluation*, 47(4):475–511.

Esuli, A. and Sebastiani, F. (2007). Random-walk models of term semantics: An application to opinion-related properties. *Proceedings of LTC 2007*, pages 221–225.

Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D., and Vee, E. (2004). Comparing and aggregating rankings with ties. In *Proceedings of the Twenty-third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '04, pages 47–58, New York. ACM.

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

Hamilton, W. L., Clark, K., Leskovec, J., and Jurafsky, D. (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora. *arXiv preprint arXiv:1606.02820*.

Johansson, R. and Nieto Piña, L. (2015). Embedding a semantic network in a word space. In *Proceedings of NAACL-HLT 2015*, pages 1428–1433, Denver. ACL.

Kann, V. and Rosell, M. (2005). Free construction of a free Swedish dictionary of synonyms. In *Proceedings of NODALIDA 2010*, Joensuu. University of Eastern Finland.

Kendall, M. G. (1945). The treatment of ties in ranking problems. *Biometrika*, pages 239–251.

Kokoska, S. and Zwillinger, D. (2000). *Standard Probability and Statistics Tables and Formulae*. Chapman & Hall / CRC.

Lönngren, L. (1998). A Swedish associative thesaurus. In *Euralex '98 proceedings, Vol. 2*, pages 467–474, Liège. EURALEX.

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.

Nusko, B., Tahmasebi, N., and Mogren, O. (2016). Building a sentiment lexicon for Swedish. In *Proceedings of the* From Digitization to Knowledge *workshop at DH 2016, Kraków*, pages 32–37, Linköping. LiUEP.

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

Roget, M. P. (1852). *Thesaurus of English Words and Phrases*. Longman, London.

Rosell, M. and Kann, V. (2010). Constructing a Swedish general purpose polarity lexicon: Random walks in the People's dictionary of synonyms. In *Proceedings of SLTC 2010*, pages 19–20, Stockholm. KTH.

Rothe, S., Ebert, S., and Schütze, H. (2016). Ultradense word embeddings by orthogonal transformation. *arXiv preprint arXiv:1602.07572*.

Rouces, J., Tahmasebi, N., Borin, L., and Eide, S. R. (2018). Generating a gold standard for a Swedish sentiment lexicon. 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.

Simons, G. F. and Fennig, C. D. (eds). (2017). *Ethnologue: Languages of the world*. SIL International, Dallas, 20th edition. Online version: `http://www.ethnologue.com`.

Wu, T.-F., Lin, C.-J., and Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5(Aug):975–1005.

6. Language Resource References

Språkbanken. (2015a). *Bring*. Språkbanken, University of Gothenburg, Sweden, Språkbanken lexical resources: Bring, v. 1.1. A Swedish version of Roget's Thesaurus from 1930 (∼149k entries), URL `https://spraakbanken.gu.se/eng/resource/bring`.

Språkbanken. (2015b). *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. (2015c). *SALDO's Morphology*. Språkbanken, University of Gothenburg, Sweden, Språk-

banken 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. (2017a). *Bring*. Språkbanken, University of Gothenburg, Sweden, Språkbanken lexical resources: Blingbring, v. 0.3. A Swedish version of Roget's Thesaurus from 1930, where outdated entries have been removed and the (lemma) entries have been partially linked to SALDO word senses (∼149k entries), URL https://spraakbanken.gu.se/eng/resource/blingbring.

Språkbanken. (2017b). *Swesaurus*. Språkbanken, University of Gothenburg, Sweden, Språkbanken lexical resources: Swesaurus, v. 0.2. A Swedish wordnet combining curated information from a number of free lexical resources (∼72k entries), URL https://spraakbanken.gu.se/eng/resource/swesaurus.

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.