SenSALDO: a Swedish Sentiment Lexicon for the SWE-CLARIN Toolbox

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

Språkbanken, University of Gothenburg

`{jacobo.rouces, nina.tahmasebi, lars.borin, stian.rodven.eide}@gu.se`

Introduction

Lexicons and resources for sentiment analysis in languages other than English are still scarce.

We implement, test and evaluate different methods to automatically create SenSALDO, a sentiment lexicon in Swedish. We manually curate the result and make it freely available.

SenSALDO is based on SALDO, an open-source computational lexical-semantic computational resource for Swedish. SALDO is composed, among other components, of 131,020 word senses connected pairwise by lexical-semantic descriptor relations.

Construction

We assign each SALDO word sense a label: negative (-1), neutral (0) or positive ($+1$), as well as a continuous score between -1 and $+1$. We implement and compare different methods:

- Inheritance of sentiment value using descriptor relations.
- Simulating random paths over a graph whose edges are descriptor relations as well as synonyms obtained using a curated collection of synonyms. The sentiment score of an element is the difference between the average number of times that a random path connects it with a positive seed, minus the same average with a negative seed.
- Training classifiers (logit, SVM+rbf) using dimensions from word embeddings as features. We use an adaptation of word embeddings to SALDO word senses.

Evaluation

For training and testing we use a gold standard we developed previously with 1998 terms. To evaluate the discrete labels, we use precision and recall; for the continuous scores we use ranking scores (Spearman rank-order correlation $\rho \in [-1, 1]$, p-normalized Kendall tau distance $\tau_p \in [0, 1]$, Kendall's tau-b τ_b).

The method using word embeddings with a support vector machine using a RBF kernel is the one performing consistently better. For these results, we manually curate the labels of all non-neutral items, plus the top 2,500 neutral items as determined by corpus frequency in the Gigaword Corpus (7,618 word senses in total).

	ρ	τ_p	τ_b	precision	recall	acc.
graph inheritance prim	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
graph inheritance prim+sec	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
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
embeddings +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
embeddings +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**

SenSALDO contains 7,618 word senses as well as a full-form version containing 65,953 items (text word forms), and is available as a research tool in the SWE-CLARIN toolbox under an open-source (CC-BY) license at:

`https://spraakbanken.gu.se/eng/resource/sensaldo`

References

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

Stian Rødven Eide, Nina Tahmasebi, and Lars Borin. 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, 2016. LiUEP.

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

Jacobo Rouces, Nina Tahmasebi, Lars Borin, and Stian Rødven Eide. Generating a gold standard for a Swedish sentiment lexicon. In *Proceedings of LREC 2018*, pages 2689–2694, Miyazaki, 2018. ELRA.

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.