Constructing a Swedish General Purpose Polarity Lexicon Random Walks in the People's Dictionary of Synonyms

Magnus Rosell, Viggo Kann

KTH CSC
100 44 Stockholm, Sweden
rosell@csc.kth.se, viggo@csc.kth.se

1. Introduction

In *opinion mining* or *sentiment analysis* one task is to assign polarity to a text or a segment (Pang and Lee, 2008). Methods for this can be helped by lexical resources with polarity assigned to words and/or phrase. We aim to construct a large free Swedish general purpose polarity lexicon.

There are many available polarity resources for English and several descriptions of how to create them, see Pang and Lee (2008). Many such methods use some other lexical resource, such as thesauruses and lexicons, viewed as a graph of word relatedness. Hassan and Radev (2010) use random walks on such a graph and achieve better and comparable results to previous work. Random walks on the graph consider all paths between two words, as opposed to only the shortest.

Velikovich et. al. (2010) derive a large polarity lexicon from web-documents, which is not limited to specific word classes and contains slang and multi-word expressions. They find that it gives better performance in sentence polarity classification than lexicons constructed from ordinary lexical resources such as WordNet.

2. The People's Dictionary of Synonyms

The People's Dictionary of Synonyms (Kann and Rosell, 2005) contains words from different stylistic classes, both slang and formal words appear. It also does not distinguish between different word classes. Synonymity is defined by the users.

The dictionary was constructed in two steps. In the first a list of possible synonyms was created by translating all Swedish words in a Swedish-English dictionary to English and then back again using an English-Swedish dictionary. The generated pairs contained lots of non-synonyms. The worst pairs were automatically removed using Random Indexing.

In the second step every user of the popular dictionary Lexin on-line was given a randomly chosen pair from the list, and asked to judge it. An example (translated from Swedish): "Are 'spread' and 'lengthen' synonyms? Answer using a scale from 0 to 5 where 0 means I don't agree and 5 means I do fully agree, or answer I do not know." Users could also propose pairs of synonyms, which subsequently were presented to other users for judgment.

All responses were analyzed and screened for spam. The good pairs were compiled into the dictionary. Millions of contributions have resulted in a constantly growing dictionary of more than 80 000 Swedish pairs of synonyms. Since it is constructed in a giant cooperative project, the dictionary is a free downloadable language resource.

Each synonym pair in the dictionary has a grade. It is the mean grading by the users who have judged the pair. The available list contains 16 006 words with 18 920 pairs that have a grading of 3.0 to 5.0 in increments of 0.1. The dictionary can be considered an undirected weighted graph. It has 2 268 connected components, the second largest of which consists of 35 words and 46 pairs. In the following work we only use use the largest component, which we call Synlex. It consists of 9 850 words and 14 801 pairs.

3. Method

We use a method very similar to Hassan and Radev (2010). However, in Synlex we have weights on the edges, a measure of relatedness, which we exploit.

Synlex is a graph $G = (V, E)$, where $V = \{i\}_{i\in[1,\dots n]}$ is the set of n words, and $E = \{(i, j)\}_{i,j\in V}$ is the set of edges or links between the words, corresponding to the synonym pairs of Synlex. With each edge in E we associate three values. First, the synonymity level of Synlex: $\mathrm{syn}(i, j) \in [3.0, 5.0]$. We define the length of an edge as $\mathrm{len}(i, j) = 5.0/\mathrm{syn}(i, j)$, i.e. we consider words with high synonymity to be close to each other. Finally, we define the transition probability associated with each edge:

$$\mathrm{prob}(s, d) = \frac{\mathrm{syn}(s, d)}{\sum_{(s,j)\in E} \mathrm{syn}(s, j)}. \quad (1)$$

Thus the random walk we use takes the synonymity level of Synlex into account in deciding to which node to go next, and the length of each edge.

See Figure 1 for the random walk method. We have used $I = 100$ and $m = 250$ and the following seedwords:

- positive: $S_+ =$ *{ positiv, bra, glad, rolig }*
- negative: $S_- =$ *{ negativ, dålig, ledsen, tråkig }*

The random walk may result in different values everytime. To study this we repeat the method 20 times for each word and calculate mean values and standard deviations.

4. Results and Discussion

In Table 1 we give som examples of words with their polarity values after applying the method to Synlex. We present the words that were deemed most positve and negative, as well as some of those deemed neutral, and some further positve and negative examples.

Most Positive		Neutral		Most Negative		More Examples	
Word	Value	Word	Value	Word	Value	Word	Value
på bra humör	249.9 (0.2)	...		...		hoppingivande	98.1(34.8)
inte dåligt	229.9 (0.2)	kropp	0.0 (0.2)	tråkig	−74.5 (30.9)	säll	25.5(16.0)
positivt	204.9 (0.3)	skråla	0.0 (0.2)	tradig	−74.9 (42.9)	godtagbart	25.1(20.8)
fryntlig	199.9 (0.3)	medicin	0.0 (0.3)	tristess	−78.8 (30.5)	duktig	24.5(23.3)
på gott humör	189.9 (0.2)	pinne	0.0 (0.2)	grå	−87.7 (29.5)	euforisk	23.4(12.6)
på topp	179.8 (0.3)	tips	0.0 (0.2)	sårad	−104.8 (42.9)	läckert	23.5(15.7)
optimist	179.9 (0.2)	notering	0.0 (0.2)	suger	−107.9 (37.3)	kalas	22.4(20.8)
optimism	174.9 (0.2)	kommunicera	0.0 (0.2)	illa	−110.9 (36.6)	superbra	20.8(17.5)
suveränt	164.9 (0.3)	lönelyft	0.0 (0.2)	inte bra	−112.9 (48.4)	artilleripjäs	20.8(14.1)
gladsint	154.9 (0.4)	klassindelning	0.0 (0.2)	mossig	−119.0 (63.2)	sprallig	20.1(8.9)
uppåt	154.8 (0.2)	flacka	0.0 (0.2)	negativ	−137.3 (45.3)	nedgången	−8.7(10.5)
förträffligt	149.9 (0.2)	tjoa	0.0 (0.1)	utråkande	−149.8 (0.3)	tungsinne	−8.8(5.0)
jovialisk	149.9 (0.2)	handledning	0.0 (0.3)	ointressant	−149.8 (0.4)	åldrig	−9.0(7.9)
lajban	137.1 (37.1)	falsifiera	0.0 (0.1)	trälig	−157.9 (53.9)	inkompetent	−9.3(6.6)
festlig	135.6 (34.4)	erektion	0.0 (0.2)	sorgset	−199.8 (0.2)	ålderstigen	−9.3(5.5)
lattjo	131.0 (42.0)	gilla	0.0 (0.1)	sorgen	−199.9 (0.2)	flum	−9.4(5.6)
roande	120.9 (42.6)	lyft	0.0 (0.1)	neråt	−199.9 (0.2)	fatal	−9.5(3.7)
positiv	117.1 (27.0)	utmåla	0.0 (0.3)	boring	−214.8 (0.3)	skruttig	−10.1(8.1)
uppsluppen	111.3 (23.9)	inrymma	0.0 (0.1)	ofördelaktig	−219.8 (0.3)	matt	−11.0(6.0)
...		...		deppad	−224.9 (0.2)	politik	−33.8(10.9)

Table 1: Extract from lexicon. Average values for the most positve and negative words. We also present the words in the middle of the list, i.e. words deemed neutral, and some more examples. (Standard deviations for 20 repetitions of the method in Figure 1 within parentheses.)

For each word w calculate v:

1. Repeat I number of times:
   - Let $v_+ = 0$ and $v_- = 0$
   - Walk randomly in the graph from w according to $\mathrm{prob}(s, d)$ for a maximum of m steps.
     - The first time we hit a word in S_+ (S_-) calculate the path length l using $\mathrm{len}(i, j)$, let $v_+ = v_+ + m/l$ ($v_- = v_- + m/l$).
2. Let $v_+ = v_+/I$ and $v_- = v_-/I$
3. Let $v = v_+ - v_-$

Figure 1: Random Walk. We use $I = 100$ and $m = 250$ and repeat all the above 20 times to calculate mean values and standard deviations.

The values have very different magnitudes. This may in part stem from that we use the synonymity level to define both transition probability and the length of the edges. The large standard deviations for some words are interesting. Perhaps they indicate that some words that should be connected are not.

If we only consider words with a polarity value bigger than their standard deviation we have 908 positive words and 441 negative words, this starting from only the very small lists of Section 3.

5. Conclusions and Future Work

From a small set of seed words we have constructed a first large, weighted polarity lexicon using the People's Dictionary of Synonyms. The lexicon consists of words from all word classes and different stylistic classes and could be a valuable resource for polarity classification in Swedish.

We will improve this work by considering larger and other sets of seed words. The seed words are not among the highest weighted words. One idea on how to address this is to include edges from each word to itself.

We intend to evaluate the lexicon by presenting positive, negative, and neutral words to human judges. The lexicon will become freely available.

6. References

Ahmed Hassan and Dragomir R. Radev. 2010. Identifying text polarity using random walks. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 395–403, Uppsala, Sweden, July. Association for Computational Linguistics.

V. Kann and M. Rosell. 2005. Free construction of a free Swedish dictionary of synonyms. In *Proc. 15th Nordic Conf. on Comp. Ling. – NODALIDA '05*.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135.

Leonid Velikovich, Sasha Blair-Goldensohn, Kerry Hannan, and Ryan McDonald. 2010. The viability of web-derived polarity lexicons. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 777–785, Los Angeles, California, June. Association for Computational Linguistics.