Distributional Semantics in a Word-Space for Opinions of the Swedish Web

Sarah Albertsson, Per Fallgren, Marcus Liw, Anna Persson, Johan Rittfeldt, and Daniel Toll

Linköpings University (LiU), Linköping, Sweden.

Abstract. The amount of data available on the Internet these days makes it possible to perform intriguing analyses to conclude such things as expressed opinions from looking at text in social networks, such as forums or blogs. We implement the word-space model built upon distributional semantics to be able to perform analyses regarding paradigmatic/syntagmatic relationships between words to reveal semantic relationships in expressed opinions. We also introduce a new form of sentiment analysis that builds upon these relationships. We evaluate the output from our system to determine how well it correlates with the events and trends of the Swedish people and current events. We also explore the future possible capabilities of the system.

Keywords: word-space model, vector space analysis, sentiment analysis, paradigmatic/syntagmatic relationships, public opinion.

1 Introduction

In recent years a new term called "big data" has become more common. Big data refers to large amounts of data, usually found online, that are so large that they cannot be processed by traditional data processing applications due to their size. An example of the applications of big data is that of analysing social networks. Sobokowicz et al.[18] talk about analyses of social networks (such as weblogs and multi-user virtual environments) and how such analyses can shed light on the underlying social structures and dynamic interactions among social actors. Sahlgren[14] propose that something called a "word-space model" is an attractive general framework compared to frameworks such as latent semantic analysis (LSA). The word-space model performs a more complex analysis, rather than just reporting on how many times for example a word representing a brand has occurred. It is also able to report semantic relations to that word.

A big part of this type of analysis is distributional semantics, a branch of linguistics focusing on looking at distributions of words in a data set. Looking at distributions can offer certain types of knowledge. This is something that was discussed as early as in the 1950s. Harris[3] concludes that "If we consider words or morphemes A and B to be or different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C". This means that looking at distributions can lead to some derivation of meaning of textual expressions. Distributional semantics together with the word-space model, which can be used as a framework for distributional semantics and allow analyses to be made in a vector-space, coupled with sentiment analysis (SA), lays the groundwork for what we will attempt to achieve.

1.1 Purpose and Hypothesis

This paper will focus on exploring the possibilities of using the word-space model to analyse the opinions of the Swedish population. The data will consist of Swedish texts from various kinds of Swedish news and social web sites. The question(s) we aim to be able to answer is the following:

Will we be able to monitor the opinions and sentiments of the Swedish population by analysing data crawled from Swedish social web sites using the linguistic theories and see social patterns? This hypothesis can be broken down into several smaller ones.

(1.) Will we be able to say which kinds of words the Swedes use to describe certain words, such as the names of the Swedish political parties, and are these words accurate?
(2.) Which words do the Swedes use in the same way, i.e. which words are they using as if they were synonyms/antonyms and can this be verified?
(3.) Can the output from our system be seen to be correlating with events that are discussed on a national scale?
(4.) How much data is needed for the system to be able to generate accurate and reliable results?

2 Theory

To implement our system a number of background theories have been used within the interdisciplinary field of human languages. Areas being used in this project is natural language processing and computational linguistics. The goal with the different facets of the field is to have computers process natural language pertinently. Most tasks within the field faces problems with language ambiguity. To handle the classification problem of disambiguation[4], different classifiers which derive from artificial intelligence (AI) have been implemented. Computational models such as artificial neural networks, support vector machines and latent semantic analysis have been used with varying results in earlier studies within the field[15][9][6][19][10][1][16]. We have built our system's main understanding upon the word space model which builds upon the distributional hypothesis. The word space model is a spatial representation of a word in its context window of 2+2 that captures its semantic meaning in a vector-space[17]. By looking at the position of a word's vector in a multidimensional space instead of looking at its linguistics properties. Every word has a specific point in the space with a vector allied to it, this vector is a representation of the word and is being used to define its meaning. Within the model a cosine similarity function is used to compare vectors. For opinion mining buzz monitoring can be used to uncover subjects that are especially talked about at a specific time. With the extracted opinions, from social networking sites, sentiment analysis can be applied to unveil user's expressions of sentiment about a subject. Sentiment analysis has many different approaches such as N-gram modelling, machine learning and subjective lexicons[7]. New methods of natural languages need to consider aspects of volume, velocity, and variety of information in order to avoid higher costs and reduced performance.

3 Implementation

Armed with the distributional hypothesis and the know-how of how to utilise a word-space to bring the distributional hypothesis to life we are able to create a system that automatically gathers text and feeds it into a word-space which can handle inquiries. We have some different methods to present. Firstly the crawler, a crawler is in simple terms a software that analyses web pages and extracts certain text from it. After the data was collected it was tokenised, i.e. the texts words are separated from all separators. This text is then used to create the word-spaces, also called memories. The word-space is the foundation of this project. Two different types of word-spaces were created, represented as matrices. One for paradigmatic relations and another for syntagmatic relations. We also have four different memories because that is based on texts with different time stamps. The big memory includes data from 2014 while the small memories are from 1-21 May. This so that we can see both trending and stable words for both syntagmatic and paradigmatic relations. Another method used in this system is frequency counting. It keeps track of every unique words appearance at a document level. It can also be used to keep track of subjects, like a political party, rather than words. We also have a sentiment analysis model implemented in the system. Here two vectors are created, one which represent positive words and one which represent negative words[12]. To find a polarity, a words syntagmatic related words and their paradigmatic vectors are compared to an average positive and a negative vector respectively. Finally, the design and possibility for a webpage was explored. The webpage was made with HTML5, CSS3 and JavaScript for visualizing the system. Essentially what it does, is run the Python code analysing the system's matrices and then return desired values on to our predefined HTML.

List of Swedish domains that have been crawled:

- http://www.aftonbladet.se
- http://www.expressen.se
- https://www.flashback.org
- http://www.familjeliv.se
- http://www.bloggportalen.se/BlogPortal /view/Home

4 Testing

The first test concerns the paradigmatic relations that the system generates. Due to the fact that words that have a strong paradigmatic relationship occur in the same contexts and share the same neighbours or co-occurring words they ought to have the same part of speech. To test this the words generated from the system have been manually tagged with their respective part of speech. In table 1 the words tested can be found in the second row, with their part of speech tag above them in the first row. The rest of the column, from the second row and downward, consist of the words generated by the system. The table's rows represent how close the generated relationship is. The closer the word is to the tested word the closer the relationship is. Words that are in bold, are words that have the same part of speech as the word tested. On average the system managed to generate words of

which 96% share the same part of speech as the word tested.

Verb {springa}	Adjective {underbar}	Proper Noun {Per}	Noun {stol}
springer	härlig	Birgit	stolen
sprang	fantastisk	Fritiof	hakan
sprungit	skön	Sven	mulen
knata	fin	Bert	usb
glider	trevlig	Erik	svensson
traska	hemsk	Lars	missuppfattning
kila	obehaglig	Lidén	2m
flyger	ljuvlig	Borg	knegare

Table 1 – Paradigmatically related words and their part of speech

External correlation concerns the question of whether the system can generate output that correlate with external events. To answer this the system has been tested with the Eurovision Song Contest (ESC), which has millions of viewers in Sweden. The chosen event has been of such magnitude that it ought to have generated a lot of discussion online. The output from the system that has been tested is firstly the "trending" syntagmatic relations, meaning syntagmatic relations that exist only during the particular time frame. Trending words, also known as semantic wobble[5], is a measure of a word's cosine before and after an observation. By looking at this it is possible to examine if something has made a semantic impact on a word's meaning. The idea is that if people are debating, discussing or commenting an event or introducing a new term the word will change its semantic meaning over time.

For ESC the first word in each column of in the table was tested. Conchita Wurst was the winner of the contest, her first name and surname have been tested. Also the words "Eurovision", the name most often used to refer to the contest, and "Nielsen" – the surname of the artist representing Sweden in the contest. ESC took place the 10th of May 2014, the data used was from May 7 to May 13. As in the table before, the rows in Table 2 represent how close the generated relationships are.

Conchita	Wurst	Eurovision	Nielsen
Wurst	Conchita	Contest	Sanna
Wursts	Wursts	Song	29,
Österrikes	Österrikes	Nielsen	Eurovision
Neuwirth	Nielsens	song	delagationen
Sanna	Neuwirth	Wursts	Nielsens
Nielsens	Nielsen	Sanna	Köpenhamn
Eurovision	Eurovision	Wurst	seglat
skäggiga	skäggiga	Tittarna	Wursts
Vladimir	sanna	Vladimir	Wurst

Table 2 – Syntagmatic relations to words related to the ESC.

The result are that "Conchita" is strongly related with her surnames, also with "sterrikes", meaning "Austria's", as Wurst comes from Austria. Moreover the name "Neuwirth" also occurs which is Wurst's real first name, as "Conchita" is a stage name. Also "Sanna" and "Nielsens" (which is the first and last name of the Swedish contestant) comes up as relations, meaning that people talk about the Swedish contestant when they talk about "Conchita". The Swedish adjective "skäggiga" also comes up, meaning "the bearded", this due to Conchita being recognized for her beard (as Conchita Wurst is a drag-queen artist). "Vladimir" is probably referring to Russia's president Vladimir Putin. Before the ESC there were political debates in the country about broadcasting a show on national TV with a drag-queen artist, something of relevance due to the fact that Wurst is an openly homosexual man[8]. The second output that has been used is frequency counts of the occurrence, or conversational volume, of words related to the events. If something happens that concerns a topic that usually is not one that is much talked about, the occurrences of words related to that topic can be expected to increase if the topic is something that is discussed. Hence we are able to look at the frequencies of words related to events of interest in the Swedish media and see if the frequencies are stable and then spike up around the same time these events occur. If they spike at the same time as the events take place we can conclude that this is a reflection of people discussing these events. Below Figure 1 shows the collected frequencies of the word "Conchita" and "Wurst" combined.

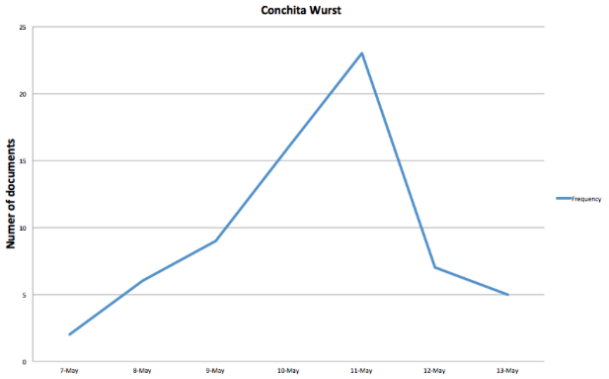

Figure 1 – The combined frequencies of "Conchita" and "Wurst" during the period around ESC (the contest's final took place on 10th of May).

4.1 Datasize

In this section we look at tests that have been performed to determine the amount of that is needed to make the system perform in the manner described in the above tests. The strongest relations exist due to them being the most frequently used ones. Other used less frequently can be skewed due to the small sample lacking enough data to generate how words are used in the majority of cases. It is therefore expected that the larger the amount of data fed into the system is the better the output will be, and the more stable it will be. Six different data sets were used, their sizes were the following:

#1 - 1 day, 7 May, 366,102 words.
#2 - 3 days, 7 May, 1,110,468 words.
#3 - 6 days, 7-12 May, 2,288,856 words.
#4 - 9 days, 7-15 May, 3,287,856 words.
#5 - 21 days, 1-21 May, 8,622,416 words.
#6 - 5 months, 1 Jan-21 May, 59,449,284 words.

4.2 Paradigmatic Tests

The data amount test evaluates the paradigmatic relations. It is expected that the most data that is fed into the system the most commonly used paradigmatic relations should emerge. Since paradigmatic relations state that words that have a strong paradigmatic relationship have similar meaning and are used in a similar manner they also should have the same part of speech, as discussed and tested above. The test therefore consists of checking how many words share the same part of speech as the word tested. It is expected that the more data the better the performance will be, and at a certain point the performance will start leveling out.

The first word tested was "Norge" ("Norway") which is a proper noun. Words generated that are also proper nouns are in bold in Table 4 below. The higher up the word is in the list the closer the relationship is to "Norge".

Norge	#1 (366,102)	#2 (1,110,468)	#3 (2,288,856)
	Sverige **USA** **Europa** kväll livet verkligheten styrelsen **stockholm**	**Sverige** idag detta **USA** ... polisen min också han	**Tjeckien** premiären **Ryssland** **Danmark** **Gefle** **Örebro** **Finland** match öppningsmatchen
	#4 (3,287,630)	#5 (8,622,416)	#6 (59,449,284)
	Ryssland **Danmark** **Finland** **Tjeckien** premiären match antikroppar kritiken öppningsmatchen	**Tjeckien** **Slovakien** **Kanada** **Finland** **Ryssland** **Frankrike** **Danmark** bryter ställs	**Finland** **Belgien** landskamperna **Tyskland** **Holland** **Legoland** **Schweiz** **Polen** grannlandet

Table 3 – Paradigmatically related to the words to the word Norge with differently sized data.

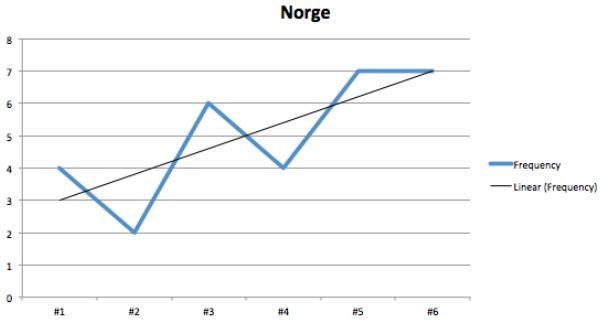

Figure 2 – How many paradigmatically generated words to the word Norge that had the same part of speech in different sized data sets.

5 Analysis

One of the fundamental problems of computational linguistics and the meaning of words, could be expressed by its lack of being able to explain the actual meaning of words[2]. Other studies within the area of SA has focused on the author's intended meaning and emotions in texts[11]. We have chosen to overlook intended meaning as a problem by focusing on the words that the authors actually express.

5.1 Syntagmatic and Paradigmatic Relationships

One of the most interesting things one could use syntagmatic relations for is to understand what people

talk about when they talk about a specific subject. This differs somewhat from the paradigmatic relations that instead look at semantic similarity between words. The syntagmatic relations could thus be an attractive ability for e.g. political parties and businesses. The paradigmatic relations can be used for finding explanations for words by looking at other words that are distributed similarly in text, or what people actually mean when they talk about a specific word.

5.2 Sentiment Analysis

The implementation of SA in our system was completely experimental and does not stem from any previous research which has used the same method. The idea for the system we implemented was quite promising, however, since the data we gathered is limited, along with limited background on objectively positive/negative words, the system did not work optimally. The system looks at a word's syntagmatic neighbours to decide the words polarisation and one can tell what people think about the things that are most closely connected to a specific word. However, this system has some flaws. It is pretty good at polarising a word correctly, but the system cannot decide if a word is more positive respectively negative than another word.

5.3 External Correlation

The already existing functionality of the system lets us use the output from the system to check how this correlates with external events. The tests performed in the previous chapter generated very interesting results, however they rely on certain assumptions. For example when testing related words to the ECS, the assumption was that when "Conchita" shows up as a related word this is due to the fact that the winning contestant of the contest first name is Conchita. However, there still exists a possibility that the relation exists due to random factors or even that the Conchita that is talked about is not the Conchita that took part in the ESC. We have made the assumption that the probability of "Conchita" showing up as a related word to "Eurovision" and this relation existing due to other factors than that "conchita" was a contestant is negligible. From these assumptions it can be concluded that the system performs very well.

5.4 Datasize Analysis

The last couple of tests that were performed in section 4 aimed to conclude if there is a certain amount of data needed to make the system perform as it did in the previous tests. This was done by creating data sets of six different data sizes which were tested with different methods. The first test looked at paradigmatic relations and at how much data was needed to generate words that had the same part of speech. It can clearly be seen in Figure 2 that the amount of generated paradigmatic relations that have the same part of speech as the word tested increases the larger the data set. This can also be seen for the syntagmatic tests. There does exist a bit of turbulence with the smaller data sets so that in some cases the results are high with the smallest data set and then lowers for the larger data sets and then goes up again. This can have several causes, such as something being normally only discussed during the weekend causes the data to be skewed towards a temporal factor. The trend lines in the graphs of the section does point toward a performance increased with larger data sets. There exists a question of whether or not there is a limit at which when the data set becomes stagnant and does not get any better, as have been found in the experiments of Sahlgren[13]. This is not something that can be seen in our system, which causes us to believe that more data is needed to get up to this point.

6 Conclusion

It can be concluded that the system needs a lot of data to present reasonable results. However, with room for improvement, the results were overall positive and the system shows promising results for our hypotheses. We were indeed able to generate words that Swedes use in context of other terms, such as ESC. Also, with the functions for paradigmatic relations and nearest neighbours the system partly generated synonyms or terms that can be used similarly to the analysed word.

Correlation with external events, which are discussed nationally or internationally, was discovered. Although the data was limited, it was enough to represent peaks and valleys of the trends in discussions. The sentiment analysis model we introduce shows promising results, albeit with definite need for improvement.

Concerning the question whether or not there is a limit of how much data is needed to generate accurate results, we conclude that it gets better with more and more data, but we could not find a concrete line that separated bad results from good results.

The system and its different functions could have several interesting applications and uses in the future, given that a few of the issues of its implementation are solved. The system could show great potential in the field of language research, as it shows very clearly

how words and languages are used today. The system could also have an entirely different kind of function if developed enough to make language filters that would detect for example online bullying or to detect if a person is deemed to be in the risk zone of committing suicide. Or even scanning social network feeds in real-time to be able to detect accidents or emergencies. However, such a sophisticated system would need a lot more work and development.

We can conclude that the system shows promising results for our hypotheses. We were indeed able to generate words that Swedes use in context of other terms, such as political parties. This was tested and analysed and while the result had its flaws, the words were accurate and got better and better with more data. With the functions for paradigmatic relations and nearest neighbours the system partly generated synonyms or terms that can be used similarly to the analysed word, this positively answers one of our hypotheses.

Correlation with external events which are discussed nationally or internationally was discovered. The examples used were related to specific events that took place during the month of May. Although the data was limited, it was enough to represent peaks and valleys of the trends in discussions. The sentiment analysis model we introduce shows promising results, albeit with definite need for improvement, we can see from the results of human opinions that the model is pretty good at deciding a polarity for a particular word. However, if asked which of the two words that are more positive than another, the system does not perform well. SA is also, as many other parts of the system, dependent on more data.

Concerning the hypothesis whether or not there is a limit of how much data is needed to generate accurate results, we conclude that there is not a simple answer. However, it gets better with more and more data, but we could not find a concrete line that separated bad results from good results.

References

1. Anjaria, M., and Guddeti, R. M. R. Influence factor based opinion mining of twitter data using supervised learning. *2014 Sixth International Conference on Communication Systems & Networks (COMSNETS)* (2014), 1.
2. Clarke, D. A contexttheoretic framework for compsitionality in distributional semantics. *Computational Linguistics 38*, 1 (2012), 4171.
3. Harris, Z. Distributional structure. In *Papers in Structural and Transformational Linguistics*, Formal Linguistics Series. Springer Netherlands, 1970, pp. 775–794.
4. Jurafsky, D., and Martin, J. H. *Speech and language processing: Pearson New International Edition. Second Edition / Daniel Jurafsky, James H. Martin.* Prentice Hall series in artificial intelligence. Edinburgh Gate, Harlow : Pearson Education Limited, cop. 2014, 2014.
5. Karlgren, J. New measures to investigate term typology by distributional data. In *Proceedings of the 19th Nordic Conference of Computational Linguistics, May 22–24, 2013, Oslo University, Norway. NEALT Proceedings Series 16* (2013).
6. Koppel, M., and Schler, J. The importance of neutral examples for learning sentiment. *Computational Intelligence 22*, 2 (2006), 100–109.
7. Liu, B. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies 5*, 1 (2012), 1–167.
8. Luhn, A. Russian politician condemns eurovision as europe-wide gay parade, May 2014.
9. Pang, B., Lee, L., and Vaithyanathan, S. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* (2002), pp. 79–86.
10. Paradis, R. D., Guo, J. K., Moulton, J., Cameron, D., and Kanerva, P. Finding semantic equivalence of text using random index vectors. *Procedia Computer Science 20*, Complex Adaptive Systems (2013), 454 – 459.
11. Pettit, A. Identifying the real differences of opinion in social media sentiment. *International Journal of Market Research 55*, 6 (2013), 757 – 767.
12. Rosell, M., and Kann, V. Constructing a swedish general purpose polarity lexicon random walks in the peoples dictionary of synonyms. *SLTC 2010* (2010), 19.
13. Sahlgren, M. The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.
14. Sahlgren, M. The distributional hypothesis. *Italian Journal of Linguistics 20*, 1 (2008), 33–54.
15. Saraswathi, K., and Tamilarasi, A. Investigation of support vector machine classifier for opinion mining. *Journal of Theoretical and Applied Information Technology 59*, 2 (2014), 291–296.
16. Sarikaya, R., Hinton, G., and Deoras, A. Application of deep belief networks for natural language understanding. *IEEE/ACM Transactions on Audio, Speech and Language Processing 22*, 4 (2014), 778 – 784.
17. Smith, C., and Jönsson, A. *Automatic Summarization As Means Of Simplifying Texts, An Evaluation For Swedish.* 2011.
18. Sobkowicz, P., Kaschesky, M., and Bouchard, G. Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. *Government Information Quarterly 29*, 4 (2012), 470–479.
19. Turney, P. D. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews.