Creating an Annotated Corpus for Aspect-Based Sentiment Analysis in Swedish

Jacobo Rouces[1], Lars Borin[1], and Nina Tahmasebi[1]

Språkbanken, Department of Swedish, University of Gothenburg
{jacobo.rouces,lars.borin,nina.tahmasebi}@gu.se

Abstract. Aspect-Based Sentiment Analysis constitutes a more fine-grained alternative to traditional sentiment analysis at sentence level. In addition to a sentiment value denoting how positive or negative a particular opinion or sentiment expression is, it identifies additional aspects or 'slots' that characterize the opinion. Some typical aspects are target and source, i.e. who holds the opinion and about which entity or aspect is the opinion. We present a large Swedish corpus annotated for Aspect-Based Sentiment Analysis. Each sentiment expression is annotated as a tuple that contains the following fields: one among 5 possible sentiment values, the target, the source, and whether the sentiment expressed is ironic. In addition, the linguistic element that conveys the sentiment is identified too. Sentiment for a particular topic is also annotated at title, paragraph and document level. The documents are articles obtained from two Swedish media (Svenska Dagbladet and Aftonbladet) and one online forum (Flashback), totalling around 4000 documents. The corpus is freely available and we plan to use it for training and testing an Aspect-Based Sentiment Analysis system.

Keywords: Aspect-Based Sentiment Analysis, Swedish, Sentiment Gold Standard, Sentiment Annotation.

1 Introduction

The field of sentiment analysis, also known as opinion mining, has been very popular during the last decade, and it has offered a large array of applications, from customer service and marketing to medicine and political polling [6].

However, up to recently, most research has been somewhat constrained on two different dimensions.

First, the focus has been for the most part on document or sentence-level sentiment analysis, which assigns a (generally positive or negative) sentiment to specific units of text. While this is useful in certain contexts, it neglects information such as what the sentiment or sentiment is *about* (the target), or who holds the sentiment (the source), which is not necessarily always the author of the text, or a constant entity either. For example, across the same text, the author can express his support or liking for something and at the same time report on someone's different opinion on the same thing or something else.

For instance, a review about a hotel may contain a sentence like "The apartment is well located but it's too small". Another example is the following excerpt from a

debate between Governor George W. Bush and Vice President Al Gore during October 2000. “I have actually not questioned Governor Bush’s experience. I have questioned his proposals.”

This is especially true for texts outside the few-hundred-characters limits found in microblogging, for instance traditional news or less-constrained blogs and fora.

Furthermore, most research has focused on extracting sentiment from texts in English. However, nearly half of natural language in the Internet is not in English [1]. Therefore, there is an increasing need of natural language processing tools in languages other than English.

An increasing amount of research has recently been focusing on aspect-based sentiment analysis [5], whose goal is more fine-grained than sentence-level sentiment analysis. Instead of providing a simple pair (text, sentiment value), it produces tuples with additional elements, such as the target of the sentiment, the source, the time, etc. The most relevant one is usually the target, and it can be identified as a text span within the text (ideally a syntactic constituent), such as “The railroad infrastructure” in “The railroad infrastructure is in a deplorable state”, or as an entity, PRICE in “The hotel was too expensive”.

In this paper, we describe the creation of an annotated corpus for aspect-based sentiment analysis in the Swedish language, combining targets and sources. Targets are identified in two different ways: as text spans and as a particular entity. The entity is the concept of immigration in Sweden, which is a source of widely different political opinion. Irony and anaphora are also identified.

2 State of the Art

There has recently been an increasing interest in aspect-based sentiment analysis because it offers more fine-grained information than just associating a sentiment value to a piece of text.

A relatively recent review of the state of the art in aspect-based sentiment analysis has been made available [5]. Most aspect-based sentiment analysis systems can be regarded as producing a tuple like (text, sentiment value, target expression, target entity, source expression, time, ...), although most systems will produce a subset of these elements. The two first elements are what would produce traditional sentiment analysis, while the existence of at least the third or the fourth is the most common addition in most aspect-based sentiment analysis systems. Sometimes, the text field may still be a sentence, or a larger unit like a whole document, while others it can focus on a sentiment expression.

The SemEval-2016 Task 5 has been dedicated to aspect-based sentiment analysis [4]. It includes two subtasks. One extracts tuples (sentiment expression, sentiment value, target expression, target entity), with the possible entities having a closed set of attributes that act as sub-entities (i.e RESTAURANT+PRICE). The second subtask exchanges sentiment expressions with whole review texts, i.e. aggregates the sentiment value across a whole review.

[1] `https://w3techs.com/technologies/overview/content_language/all`

A specific approach can be found in [1], where Basque and Catalan hotel reviews are annotated for aspect-level sentiment analysis. It focuses on tuples of the form (text, sentiment value, target expression, source expression). Although the target entity is not determined, it focuses on tuples that deal with aspects of hotels. Conditional Random Fields are used for classification.

Several types of aspect-based sentiment analysis tasks has been evaluated against a manually annotated corpus in [3]. The corpus consists of 76,732 tweets (micro-blogging entries) of whom half deal with the aspect of transport within the Île-de-France region in France, and half don't. The best performing methods obtain micro-averaged f-scores of 0.90 and 0.82 for the first and second task respectively.

3 Description

The Swedish-language sources from which we have extracted the text are the following:

- Editorials and opinion pieces from *Svenska Dagbladet* (SvD) (`http://www.svd.se/`), a daily newspaper with a daily circulation of 143,400 (2013).
- Editorials and opinion pieces from *Aftonbladet* (`http://www.aftonbladet.se/`), a daily newspaper with a daily circulation of 154,900 (2014).
- Posts from *Flashback* (`https://www.flashback.org/`), a Swedish-speaking Internet forum, with an Alexa ranking of 9,978, the 42nd in Sweden (2018).[2]

The timestamps of the articles and posts extracted date back to the year 2000. The documents have been sampled from those containing one among a list of 60 terms related to immigration. Otherwise, the frequency in which the chosen target entity (immigration) would have been covered would be too low and would generate too few positive or negative cases.

The corpus contains two types of annotations: tuple-based and entity-based.

- The **tuple-based** annotations are of the type (source, target, sentiment expression, sentiment value, irony) for every time a sentiment is expressed in the text. The source, target and sentiment expressions are annotated as text spans. The three of them are independently optional, since sentiment expressions can lack an explicit linguistic element for any of them (c.f. examples from Figure 1). The sentiment value can be one among 5 possible options: ('very negative', 'negative', 'neutral', 'positive', 'very positive'). Irony is a boolean value. We consider irony to be especially relevant in the context of social media and online forums.
  The text spans in a tuple are expected to be contained in the same sentence. When the source or the target are pronouns, we create another text span where the anaphora is resolved and we connect the two text spans with a 'refers to' link. This allows the annotated corpus to contain sources and targets in the syntactic sense (syntactically connected to the other elements of the tuple), but also in the semantic sense (where in the text the entity is named).
  When the sentiment expression is actually several non-contiguous text spans, a 'part of' link is used.

[2] `https://www.alexa.com/siteinfo/flashback.org`

– The **entity-based** annotations are of the type (text, sentiment value, target entity). They declare when a sentiment is expressed around a specific entity or topic, which is immigration. Namely, it reflects whether there is a positive or negative perception towards immigration as a phenomenon in Sweden, which constitutes the target entity. This is done at document, title and paragraph levels. In order to avoid creating a whole new annotation scheme, we apply the scheme used for tuple-based annotations on "handles" that we insert in the text. Each handle starts with a # character to distinguish it from the real text, and is inserted before the unit of text that it represents: before each paragraph (#P IMM), and also before the title, which we insert at the beginning of the document, (#T IMM) and before the whole document, at the uppermost position (#D IMM). When a unit of text contains a sentiment towards the target entity under consideration, the handle is annotated as a sentiment expression with a sentiment value (the handle is, of course, not a real sentiment expression, but this is done as a convention to assign the value to the handle).

Table 1: Translation of example sentences.

Line	English translation
3	How should immigrants become Swedes - without us?
6	In the debate on integration, a "Swedish factory" is sought, regardless of party color, where we can put in newly arrived immigrants and get employable people with "Swedish values" and perfect Swedish in speech and writing.
8	We are happy to discuss how such a factory should be built but at the same time we forget the most important question: is it possible at all to achieve integration without involving ourselves who are established in society?
10	There are many proposals about how authorities or educational institutions should implement changes and thereby find solutions to integrate the newly arrived both into society and into the labor market.
11	It is, of course, convenient to believe that one of these players will solve this, but there is a great risk that we put resources and energy on systems and solutions that do not lead to any results.
21	There is great potential for improvement among our authorities and the results that the "factory" produces today are not good enough.
22	But based on the current conditions, it is probably the only result that we can expect.
43	In addition to engaging private individuals, an important part of my solution for getting jobs for the newly arrived is also to build personal relationships with local companies.
44	I know many business executives and boards who want and are willing to invest both time and money to be part of the solution.
45	Of course, there is a business interest behind the collaboration, but it is also about leadership and the feeling of making a difference.
46	Because a CEO or a board clearly indicate that when one wants to be a positive force in society, one also creates a spirit of progress within the organization.

Figure 1 show example annotations from Swedish newspaper *Aftonbladet*. Blocks of non-contiguous sentences are separated with dots. The blocks contain whole paragraphs. Sources and targets are denoted by text spans of a generic type 'Something',

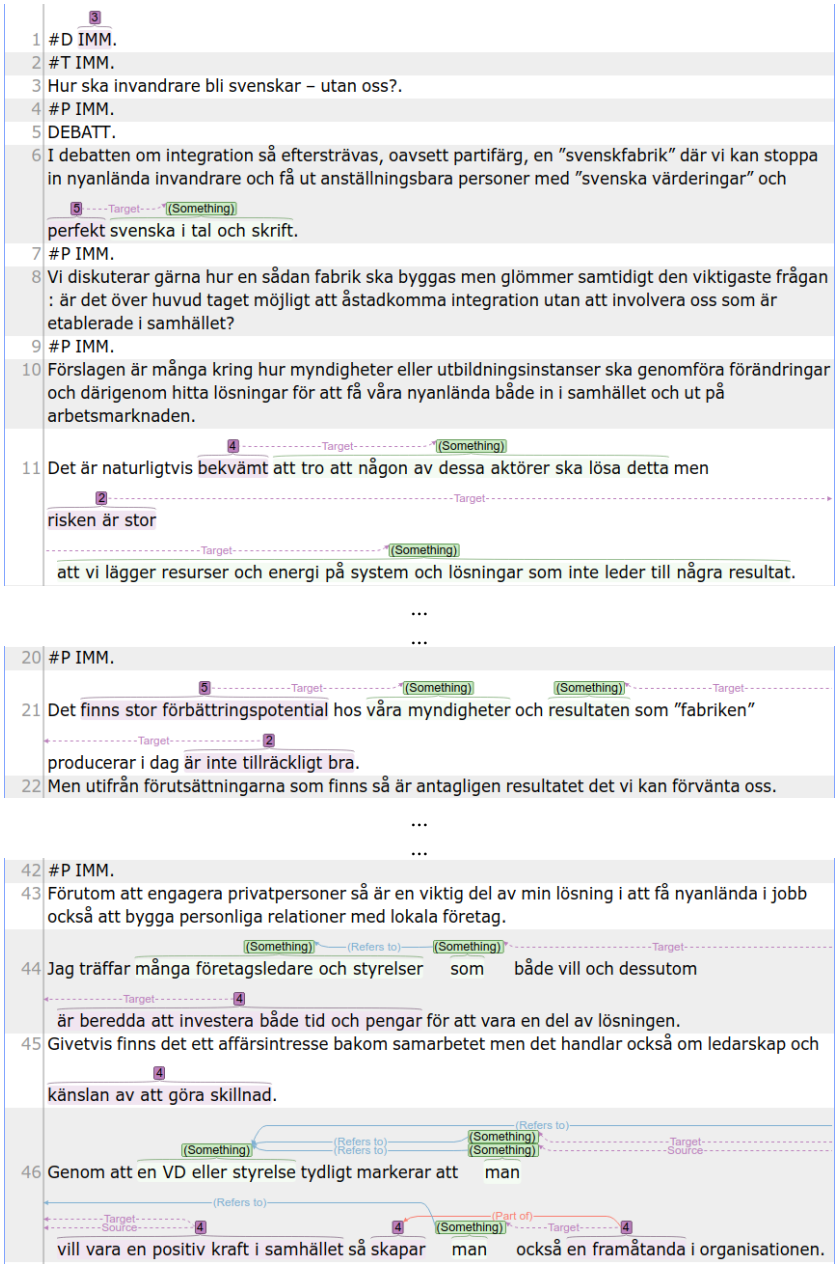

Fig. 1. Example annotations from an article from Swedish newspaper *Aftonbladet*.

so their specific type is denoted by the inbound relation. Sentiment value (non-negative integer) and irony (boolean) are represented as a property of the text span for the sentiment expression. Sentiment values ('very negative', 'negative', 'neutral', 'positive', 'very positive') are represented with numerical values (1,2,3,4,5) respectively because using non-negative integers allowed the annotators to use keyboard shortcuts. Table 1 shows the English translations of each line.

We have employed 9 annotators to annotate 3,870 documents, containing in total 1,574,226 words in 108,132 sentences. All annotators had at least undergraduate background in linguistics. Prior to the annotation tasks, they were provided with an annotation manual and some annotated example documents. Table 2 shows the number of documents and paragraphs annotated for each source.

Table 2: Number of documents and paragraphs in the annotated corpus

source	documents	paragraphs	sentences	words
SvD	852	19360	44884	677385
Aftonbladet	1184	18979	42408	617442
Flashback	1834	12021	20840	279399
Total	3870	50360	108132	1574226

A total of 40 documents were annotated by all annotators, so inter-annotator agree-ments could be calculated. Krippendorff's alpha using an interval metric was 0.34 for document-level annotations and 0.44 for paragraph-level annotations.

The tool employed is Webanno[2, **?**] due to the flexibility of its data model, allowing annotations in the form of labeled text spans as well as the creation of labeled relations or edges connecting them. Webanno also allows several export formats (several CoNLL, several UIMA, Weblicht TFC, and the native WebAnno TSV3 (v3).

4 Conclusions

We have created an annotated 1,574,226-word corpus for aspect-based sentiment analysis in Swedish, using texts obtained from Swedish traditional newspapers and one online forum. The corpus also contains sentiment annotations at document, title and paragraph levels towards a specific semantic topic. This amounts to 3 870 documents annotated (and the same number of titles) and 50,360 paragraphs.

The corpus can be used as gold standard for training and evaluation of an aspect-based sentiment analysis system. It is freely and publicly available under an open-source CC-BY 4.0 license at `https://spraakbanken.gu.se/swe/resurs/swe-absa-bank`. The format of the annotations is WebAnno TSV3. Metadata linking file names with source URLs and publication timestamps is also available in a separate file.

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 Swedish Research Council and the University of Gothenburg.

References

1. Barnes, J., Badia, T., Lambert, P.: Multibooked: A corpus of basque and catalan hotel reviews annotated for aspect-level sentiment classification. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018). European Language Resource Association (2018), http://aclweb.org/anthology/L18-1104
2. Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S.M., Hartmann, S., Gurevych, I., Frank, A., Biemann, C.: A web-based tool for the integrated annotation of semantic and syntactic structures. In: Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH). pp. 76–84. The COLING 2016 Organizing Committee, Osaka, Japan (Dec 2016), https://www.aclweb.org/anthology/W16-4011
3. Paroubek, P., Grouin, C., Bellot, P., Claveau, V., Eshkol-Taravella, I., Fraisse, A., Jackiewicz, A., Karoui, J., Monceaux, L., Torres-Moreno, J.M.: Deft2018 : recherche d'information et analyse de sentiments dans des tweets concernant les transports en Île de france. In: Actes de la conférence TALN 2018 Volume 2 : Démonstrations, articles des Rencontres Jeunes Chercheurs, ateliers DeFT. vol. 2, pp. 219–229. ATALA (May 2018), https://deft.limsi.fr/actes/Actes_TALN2018_vol2.pdf
4. Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Mohammad, A.S., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., et al.: Semeval-2016 task 5: Aspect based sentiment analysis. In: Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016). pp. 19–30 (2016).
5. Schouten, K., Frasincar, F.: Survey on aspect-level sentiment analysis. IEEE Transactions on Knowledge & Data Engineering (1), 1–1 (2016).
6. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting elections with twitter: What 140 characters reveal about political sentiment. Icwsm **10**(1), 178–185 (2010).