Improving the Detection of Comparison Arguments in Product Reviews

Wiltrud Kessler

Institute for Natural Language Processing
Universität Stuttgart
Pfaffenwaldring 5b
70569 Stuttgart
wiltrud.kessler@ims.uni-stuttgart.de

Abstract: A common way to express sentiment about some product in a review is by comparing it to a different product. In order to get meaningful information about the comparison, we need to identify the separate parts: the predicate that expresses the comparison, the two compared entities, and the aspect that they are compared in. In this work, we assume the sentence and the predicate to be given and work on argument identification and classification. We show that syntax is more helpful than context for the task. We also improve performance on argument identification by including information about the type of the comparison. We were not able to prove our hypothesis that including information on the sentiment of the predicate improves performance on argument classification.

1 Introduction

Sentiment analysis is an area in Natural Language Processing that deals with the task of determining the polarity (positive, negative) of an opinionated document or a sentence. Sentiment analysis has attracted much attention in recent year due to the growing importance of social media where users often express their opinions about products or services. In product reviews, sentiment is usually determined with regard to some target product, e.g., the sentence "X has a good lens" expresses positive sentiment towards X. A common way to express sentiment about some product is by comparing it to a different product. Such comparisons cannot be treated the same way as non-comparative sentiment expressions, because they involve more than one target entity and may involve assignment of more than one polarity, e.g., the statement "X has a better lens than Y" expresses positive sentiment towards X and less positive or maybe even negative sentiment towards Y.

For our purposes we define a comparison to be any statement about the similarity or difference of two entities. Besides the linguistic category of comparative sentences, this includes a wide variety of expressions found in user generated texts, such as "X blows away all others", "X and Y have the same sensor", or "X wins over Y". We call the word or phrase that is used to express the comparison the *comparative predicate*. A comparative predicate has three arguments: The two *entities* that are compared and the *aspect* they are compared

in. In our data, most of the entities are products (cameras, cars, phones). The term aspect denotes any attribute or part of the entity that is being compared, e.g., the lens, size or resolution of a camera. In the sentence "X has a better lens than Y" the word "better" is the comparative predicate, "X" and "Y" are the two entities and "lens" is the aspect.

In this short paper, we start with a given comparison sentence and a given comparative predicate. The task we want to solve is to find the corresponding arguments. We build on existing work [KK13] and attempt to improve argument identification and classification based on the following hypotheses: Syntax information is more helpful than context information and including information on the sentiment of the predicate and the type of the comparison improves performance.

2 Related Work

With the growing importance of user generated content in the Web 2.0, sentiment analysis (or opinion mining) which often deals with the opinions expressed in such texts has become an active research area. For an overview of methods and challenges see [Liu12]. The focus has over time shifted from document-level approaches that determine the overall polarity of a whole document to more fine-grained analyses that determine sentiment in smaller units and with respect to some target entity. One of the challenges on sentence level are comparison sentences. Jindal and Liu [JL06a] are the first to identify comparison sentences by using class sequential rules based on keywords as features for a Naive Bayes classifier. In this work we assume that we are given a set of such sentences.

Several approaches have been presented for the detection of comparison predicates and arguments. In follow-up work on their sentence identification, Jindal and Liu [JL06b] detect comparison arguments with label sequential rules and in a second step identify the preferred entity in a ranked comparison [GL08]. Approaches inspired by semantic role labeling that detect predicates and subsequently their arguments have been used for Chinese [HL08] and English [KK13]. Xu et al. [XLLS11] use conditional random fields to extract relations between two entities, an attribute and a predicate phrase.

We build upon the work of Kessler and Kuhn [KK13] who use a three step approach: predicate identification, argument identification and argument classification. In this work, we assume predicates to be given and design features to improve the last two steps.

3 Approach

The input to our system is a sentence where at least one comparative predicate has been identified. The result of our processing are three arguments for each predicate: The two entities that are being compared, and the aspect they are compared in. Any of the arguments may appear more than once or be empty.

We use the MATE Semantic Role Labeling system [BHN09] with default settings and

without the reranker. This is equivalent to the setup in [KK13]. We start with the predicates already identified. Our system performs two steps. In a first binary classification step, all arguments are identified (*argument identification*). In a second step, for each identified argument it is determined whether it is entity 1, entity 2 or aspect (*argument classification*). The system uses regularized linear logistic regression for classification.

We vary feature sets for the experiments, but always use the same feature sets for both steps. For every word in question we extract its form and the part of speech (e.g., noun, verb) as word features. The given predicate and the current argument candidate are always accessible for extracting word features. The easiest way to add more information is to include the context words before and after the candidate argument. Alternatively, the sentence can be parsed and information from the syntactic structure can be used. We use dependency parses that model the structure as a set of relations between each word and its parent. The parser is trained on news texts but applied to texts from blogs and reviews. These texts contain many errors in spelling, grammar and punctuation, so the parses may not always be correct which might lead to confusing information for the classifier. We also introduce two features to model additional sentiment and comparison information.

In our experiments we compare the following feature sets:

Base Word features extracted from current argument and predicate, position of the argument relative to the predicate (before, after or same word).

Context Base features + Word features for the three context words to the left and right of the current argument.

Syn Base features + Features extracted from the output of the MATE dependency parser [Boh10], including word features for the argument children, argument siblings, predicate parent, predicate children and syntactic path between predicate and argument. This corresponds to the setting used in [KK13]. For more detailed descriptions of the features see [BHN09].

Sent Feature to model the polarity of the predicate as defined by the MPQA list of subjectivity clues[1] [WWH05] containing 2304 positive and 4152 negative words. Possible values are positive, negative and neutral.

Type Feature to model the type of the comparison. There are four possible types (not all of them are annotated in all datasets, see data description):

- ranked comparisons (also called non-equal gradable) in which one entity is ranked as better or worse as the other, e.g., "X has higher resolution than Y",
- equative comparisons in which both entities are ranked as being equally good or bad, e.g., "X and Y have the same sensor",
- superlative comparisons in which one entity is ranked as better or worse than all other entities, e.g., "X is the best", and
- differences in which a difference between the two entities is stated, but no ranking is introduced, e.g., "picture quality is different in X and Y".

[1] `http://www.cs.pitt.edu/mpqa/subj_lexicon.html`

4 Experiments

Data. We use four datasets in our experiments: the cameras and cars datasets from the JDPA corpus[2] [KECN10], the J&L dataset[3] [JL06b], and the IMS camera review dataset[4] [KK14]. We process the data in the same way as described in [KK13] except that we add information about the type of the comparison. For J&L and IMS we use the annotations provided in the data (equative, ranked, superlative in J&L; equative, ranked, superlative, difference in IMS). For the JDPA corpus, we use the type equative if the value in the "same" slot is set to "true", otherwise we assume the type ranked. Further, we map the "scale" annotation to "aspect" in the IMS data to make the task identical to the other datasets. In IMS and J&L, the two entities are distinguished by order of appearance in the sentence. In the JDPA datasets, entities are identified as the one evaluated as better (mapped to entity 1), and the one evaluated as worse (mapped to entity 2). We extract all sentences where we find at least one comparative predicate as our data. We use annotated predicates (gold predicates) as a starting point for the experiments, as our features are tailored to arguments and do not improve predicate identification.

Evaluation Setup. We evaluate on each dataset separately using 5-fold cross-validation. We report precision (P), recall (R), and F1-measure (F1). Bold numbers denote the best result in each column and task. We mark a F1-measure result with * if the difference to *Syn* is statistically significant at $p < .05$ using the approximate randomization test [Nor89] with 10000 iterations.

Results. Table 1 shows the results for argument identification, Table 2 those for argument classification. We can see that using information from argument and predicate alone (*Base*) performs very poorly. Using syntactic information (*Syn*) outperforms using only context information (*Context*) by a considerable margin in both tasks in all datasets. The difference is significant (except for JDPA cam, entity 1 where $p = 0.0833$). We therefor use the syntactic features in the following to judge the impact of the two other features. Adding sentiment information (*Syn+Sent*) does not improve performance, it often even hurts. We were hoping for an improvement in distinguishing entity 1 from entity 2 in the JDPA data, where entity 1 (the preferred entity) occurs before the predicate for positive predicates ("A is better than B") and after the predicate for negative predicates ("B is worse than A"). The result may be due to the fact that we only consider the sentiment of the predicate itself without taking into account any context like negation that might reverse the polarity. Adding information about the type of the comparison (*Syn+Type*) improves results in nearly every setting. The difference is significant for argument identification, but only in some cases for argument classification.

[2] http://verbs.colorado.edu/jdpacorpus/

[3] http://www.cs.uic.edu/~liub/FBS/data.tar.gz

[4] http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/reviewcomparisons/

	JDPA cam			JDPA car			J&L			IMS cam		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Base	**69.7**	12.5	21.2*	55.2	9.5	16.2*	57.8	13.1	21.4*	56.3	21.2	30.8*
Context	63.4	26.4	37.3*	60.3	27.0	37.3*	64.0	30.7	41.5*	66.0	43.3	52.3*
Syn	69.0	36.6	47.8	70.2	41.6	52.2	67.4	44.5	53.6	78.1	57.2	66.0
Syn+Sent	69.0	35.9	47.3*	**70.4**	41.3	52.1	**68.5**	44.4	53.9	**78.7**	57.0	66.1
Syn+Type	65.6	**38.8**	**48.7***	68.3	**43.0**	**52.8***	65.0	**48.3**	**55.4***	72.8	**63.0**	**67.6***

Table 1: Results for argument identification on gold predicates (Precision, Recall, F1-measure)

		JDPA cam			JDPA car			J&L			IMS cam		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Entity 1	Base	30.0	1.7	3.3*	19.5	2.4	4.2*	52.6	15.1	23.5*	47.6	20.2	28.3*
	Context	40.0	14.3	21.0	38.1	17.5	24.0*	55.0	34.6	42.5*	60.4	40.8	48.7*
	Syn	**41.5**	17.3	24.5	39.9	22.4	28.7	56.4	46.6	51.0	65.9	49.8	56.7
	Syn+Sent	41.0	16.6	23.6	**40.1**	22.2	28.5	**57.8**	46.6	51.6*	**66.5**	49.4	56.7
	Syn+Type	37.8	**18.3**	**24.7**	39.2	**24.0**	**29.7***	55.5	**49.9**	**52.5***	60.3	**56.0**	**58.1***
Entity 2	Base	29.4	2.9	5.3*	29.2	3.3	5.9*	46.6	8.1	13.8*	48.4	23.6	31.7*
	Context	36.8	13.7	20.0*	47.5	19.6	27.8*	**62.0**	22.5	33.0*	56.5	36.9	44.7*
	Syn	45.7	22.7	30.3	49.7	**30.7**	**37.9**	57.1	32.6	41.5	64.8	49.6	56.2
	Syn+Sent	**46.0**	22.7	30.4	**50.2**	30.5	**37.9**	59.4	34.1	43.3*	**65.6**	49.5	56.4
	Syn+Type	42.9	**24.9**	**31.5**	47.5	30.3	37.0*	56.7	**36.8**	**44.6***	60.4	**54.3**	**57.2***
Aspect	Base	**78.7**	27.3	40.5*	**62.9**	17.6	27.5*	46.9	8.7	14.7*	47.1	11.3	18.2*
	Context	70.7	35.6	47.4*	53.8	25.1	34.2*	52.6	19.3	28.3*	54.5	34.8	42.5*
	Syn	72.2	47.0	56.9	60.0	36.0	45.0	60.7	30.7	40.8	72.5	49.6	58.9
	Syn+Sent	72.4	46.4	56.6	59.7	35.7	44.7	**61.8**	30.7	41.0	**72.8**	49.5	58.9
	Syn+Type	70.0	**48.3**	**57.2**	59.4	**38.1**	**46.4***	55.3	**33.3**	**41.6**	68.0	**52.8**	**59.4**

Table 2: Results for argument classification on gold predicates (Precision, Recall, F1-measure)

5 Conclusions

In this short paper we present experiments on improving the identification and classification of arguments (entity 1, entity 2, aspect) of comparisons in product reviews. We show that syntax information is more helpful than context information, and that including the type of the comparison improves performance. We were not able to prove our hypothesis that including information on the sentiment of the predicate improves performance. This may be due to the fact that our feature does not take into account any context like negation that might act as a polarity reverser. For future work, we want to further investigate the interaction of sentiment and comparison arguments. We also plan to address the inherent diversity of expressions typical for user generated content by using generalization techniques, e.g., to detect product names.

Acknowledgments

The work reported in this paper was supported by a Nuance Foundation grant.

References

[BHN09] Anders Björkelund, Love Hafdell, and Pierre Nugues. Multilingual Semantic Role Labeling. In *Proceedings of CoNLL '09 Shared Task*, pages 43–48, 2009.

[Boh10] Bernd Bohnet. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of COLING '10*, pages 89–97, 2010.

[GL08] Murthy Ganapathibhotla and Bing Liu. Mining opinions in comparative sentences. In *Proceedings of COLING '08*, pages 241–248, 2008.

[HL08] Feng Hou and Guo-hui Li. Mining Chinese comparative sentences by semantic role labeling. In *Proceedings of ICMLC '08*, pages 2563–2568, 2008.

[JL06a] Nitin Jindal and Bing Liu. Identifying comparative sentences in text documents. In *Proceedings of SIGIR '06*, pages 244–251, 2006.

[JL06b] Nitin Jindal and Bing Liu. Mining comparative sentences and relations. In *Proceedings of AAAI '06*, pages 1331–1336, 2006.

[KECN10] Jason S. Kessler, Miriam Eckert, Lyndsay Clark, and Nicolas Nicolov. The 2010 ICWSM JDPA Sentiment Corpus for the Automotive Domain. In *Proceedings of ICWSM-DWC '10*, 2010.

[KK13] Wiltrud Kessler and Jonas Kuhn. Detection of Product Comparisons - How Far Does an Out-of-the-box Semantic Role Labeling System Take You? In *Proceedings of EMNLP '13*, pages 1892–1897, 2013.

[KK14] Wiltrud Kessler and Jonas Kuhn. A Corpus of Comparisons in Product Reviews. In *Proceedings of LREC '14*, pages 2242–2248, 2014.

[Liu12] Bing Liu. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012.

[Nor89] Eric W. Noreen. *Computer-intensive methods for testing hypotheses – an introduction*. Wiley & Sons, 1989.

[WWH05] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT '05*, pages 347–354, 2005.

[XLLS11] Kaiquan Xu, Stephen Shaoyi Liao, Jiexun Li, and Yuxia Song. Mining comparative opinions from customer reviews for Competitive Intelligence. *Decis. Support Syst.*, 50(4):743–754, 2011.