Detection of Product Comparisons – How Far Does an Out-of-the-box Semantic Role Labeling System Take You?

Wiltrud Kessler and **Jonas Kuhn**
Institute for Natural Language Processing
University of Stuttgart
wiltrud.kessler@ims.uni-stuttgart.de

Abstract

This short paper presents a pilot study investigating the training of a standard Semantic Role Labeling (SRL) system on product reviews for the new task of detecting comparisons. An (opinionated) comparison consists of a comparative "predicate" and up to three "arguments": the entity evaluated positively, the entity evaluated negatively, and the aspect under which the comparison is made. In user-generated product reviews, the "predicate" and "arguments" are expressed in highly heterogeneous ways; but since the elements *are* textually annotated in existing datasets, SRL is technically applicable. We address the interesting question how well training an out-of-the-box SRL model works for English data. We observe that even without any feature engineering or other major adaptions to our task, the system outperforms a reasonable heuristic baseline in all steps (predicate identification, argument identification and argument classification) and in three different datasets.

1 Introduction

Sentiment analysis deals with the task of determining the polarity of an opinionated document or a sentence, in product reviews typically with regard to some target product. A common way to express sentiment about some product is by comparing it to a different product. In the corpus data we use, around 10% of sentences contain at least one comparison. Here are some examples of comparison sentences from our corpus:

(1) a. "[This camera]$_{\mathrm{E}+}$ ... its [screen]$_{\mathrm{A}}$ is much **bigger** than the [400D]$_{\mathrm{E}-}$."

b. "[D70]$_{\mathrm{E}+}$ **beats** [EOS 300D]$_{\mathrm{E}-}$ in almost [every category]$_{\mathrm{A}}$, EXCEPT ONE."

c. "[Noise suppression]$_{\mathrm{A}_1\mathrm{A}_2}$ was generally **better**$_1$ than the [D80]$_{\mathrm{E}-_1}$'s and much **better**$_2$ than the [Rebel]$_{\mathrm{E}-_2}$'s."

d. "A striking **difference** between the [EOS 350D]$_{\mathrm{E}-}$ and the new [EOS 400D]$_{\mathrm{E}+}$ concerns the [image sensor]$_{\mathrm{A}}$."

Note that our definition of comparisons is broader than the linguistic category of comparative sentences, which only includes sentences that contain a comparative adjective or adverb. For our work, we consider comparisons expressed by any Part of Speech (POS).

A comparison contains several parts that must be identified in order to get meaningful information. We call the word or phrase that is used to express the comparison ("better", "beats", ...) a *comparative predicate*. A comparison involves two *entities*, one or both of them may be implicit. In our data, most of the entities are products, e.g., the two cameras "D70" and "EOS 300D" in sentence 1b. In graded comparisons, entity+ (E+) is the entity that is being evaluated positively, entity- (E-) the entity evaluated negatively. In many sentences one attribute or part of a product is being compared, like "image sensor" in sentence 1d. We call this the *aspect* (A).

The task we want to solve for a given comparison sentence is to detect the comparative predicate, the entities that are involved and the aspect that is being compared. We borrow our methodology from Semantic Role Labeling (SRL). In SRL, *events* are expressed by predicates and *participants* of these events are expressed by arguments that fill different semantic roles. Adapted to the problem of detecting comparisons, the events we are interested in are comparative predicates and the arguments are the two entities and the aspect that is being compared.

Due to the diversity of possible ways of expressing comparisons, the "predicates" and "arguments"

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1892–1897,
Seattle, Washington, USA, 18-21 October 2013. ©2013 Association for Computational Linguistics

in this task are more heterogeneous categories than in standard SRL based on PropBank and NomBank annotations. Moreoever, the existing labeled datasets are based on an annotation methodology which gave the annotators a lot of freedom in deciding on the linguistic anchoring of the "predicate" and "arguments". This adds to the heterogeneity of the observed constructions and makes it even more interesting to ask the question how far an out-of-the-box SRL model can take you.

In this work, we re-train an existing SRL system (Björkelund et al., 2009) on product review data labeled with comparative predicates and arguments. We show that we can get reasonable results without any feature engineering or other major adaptions. This is an encouraging result for a linguistically grounded modeling approach to comparison detection.

2 Related Work

The syntax and semantics of comparative sentences have been the topic of research in linguistics for a long time (Moltmann, 1992; Kennedy, 1999). However, our focus is on computational methods and we also treat comparisons that are not comparative sentences in a linguistic sense.

In sentiment analysis, some studies have been presented to identify comparison sentences. Jindal and Liu (2006a) report good results on English using class sequential rules based on keywords as features for a Naive Bayes classifier. A similar approach for Korean is presented by Yang and Ko (2009; 2011b; 2011a). In our work, we do not address the task of identifying comparison sentences, we assume that we are given a set of such sentences.

The step we are concerned with is the detection of relevant parts of a comparison. To identify entities and aspect, Jindal and Liu (2006b) use an involved pattern mining process to mine label sequential rules from annotated English sentences. A similar approach is again presented by Yang and Ko (2011a) for Korean. In contrast to their complicated processing, we simply use an existing SRL system out of the box. Both approaches consider only nouns and pronouns for entities and aspects, we use all POS and allow for multi-word arguments. Jindal and Liu (2006b) base the recognition of comparative predicates on a list of manually compiled keywords. We use this as our baseline. Our approach is not dependent on a set of keywords and is therefore more easily adaptable to a new domain.

All works label the entities according to their position with respect to the predicate. This requires the identification of the preferred entity in a non-equal comparison as an additional step. Ganapathibhotla and Liu (2008) use hand-crafted rules based on the polarity of the predicate for this task. As we label the entities with their roles from the start, we solve both problems at the same time.

Xu et al. (2011) cast the task as a relation extraction problem. They present an approach that uses conditional random fields to extract relations (*better*, *worse*, *same* and *no_comparison*) between two entitites, an attribute and a predicate phrase.

The approach of Hou and Li (2008) is most related to our approach. They use SRL with standard SRL features to extract comparative relations from Chinese sentences. We confirm that SRL is a viable method also for English. In their experiments they report good results on gold parses, but observe a drop in performance when they use their method on automatic parses. All our experiments are conducted on automatically obtained parses.

3 Approach

The input to our system is a sentence that we assume to contain at least one comparison. The result of our processing are one or more comparative predicates and for each predicate three arguments: The two entities that are being compared, and the aspect they are compared in. More formally speaking, for every sentence we expect to get one or more 4-tupels (predicate, entity+, entity-, aspect). Entity+ is the entity that is being evaluated as better than entity-. Any of the arguments may be empty. Currently, we treat only single words as comparative predicates. Annotated multi-word predicates are mapped to one word. We allow for multi-word arguments, but annotate only the head word of the phrase and treat it as a one word argument for evaluation. We do not place any restrictions on possible POS.

We use a standard pipeline approach from SRL. As a first step, the comparative predicate is identified. The next step in SRL would be predicate

disambiguation to identify the different frames this predicate can express. As we do not have such frame information, predicate disambiguation is not performed in our pipeline.

After we have identified the predicates, the next step is to identify their arguments. The identification step is a binary classification whether a word in the sentence is some argument of the identified predicate. As a final classification step, it is determined for each found argument whether this argument is entity+, entity- or the aspect.

We use an existing SRL system (Björkelund et al., 2009)[1] and the features developed for SRL, based on the output of the MATE dependency parser (Bohnet, 2010). Features use attributes of the predicate itself, its head or its dependents. Additionally, for argument identification and classification there are features that describe the relation of predicate and argument, the argument itself, its leftmost and rightmost dependent and left and right sibling.

For the classification tasks of the pipeline, the SRL system uses regularized linear logistic regression from the LIBLINEAR package (Fan et al., 2008). We set the SRL system to train separate classifiers for predicates of different POS. In preliminary experiments, we have found this to perform slightly better than training one classifier for all kinds of predicates, although the difference is not significant. We do not use the reranker.

4 Experiments

Data. We use the JDPA corpus[2] by J. Kessler et al. (2010) for our experiments. It contains blog posts about cameras and cars. We use the annotation class "Comparison" that has four annotation slots. We convert the "more" slot to entity+, the "less" slot to entity- and the "dimension" slot to the aspect. For now, we ignore the "same" slot which indicates if the two mentions are ranked as equal.

We have also tested our approach on the dataset used in (Jindal and Liu, 2006b)[3]. We use all comparisons annotated as types 1 to 3 (ignoring type 4, non-gradable comparisons). In this dataset (J&L), entities are annotated as entity 1 or entity 2 depending on their position before or after the predicate. We keep this annotation and train our system to assign these labels.

	JDPA		J&L
	cameras	cars	
all sentences	5230	14003	7986
comparison sentences	505	1094	649
predicates	642	1327	695
distinct predicates	147	252	122
preds. occurring once	87	147	61
Entity+ / 1	517	1091	657
Entity- / 2	511	1068	331
Aspect	623	1107	526

Table 1: Statistics about the datasets

We do sentence segmentation and tokenization with the Stanford Core NLP[4]. Annotations are mapped to the extracted tokens. We ignore annotations that do not correspond to complete tokens. In the JDPA corpus, if an annotated argument is outside the current sentence, we follow the coreference chain to find a coreferent annotation in the same sentence. If this is not successful, the argument is ignored. We extract all sentences where we found at least one comparative predicate as our dataset.

Table 1 shows some statistics of the data.

Evaluation Setup. We evaluate on each dataset separately using 5-fold cross-validation. We report precision (P), recall (R), F1-measure (F1), and for argument classification macro averaged F1-measure ($F1_m$) over the three arguments. Bold numbers denote the best result in each column and dataset. We mark a F1-measure result with * if it is significantly higher than all previous lines.[5]

Results on Predicates. We have implemented two baselines based on previous work. The simplest baseline, *BL POS* classifies all tokens with a comparative POS ('JJR', 'JJS', 'RBR', 'RBS') as predicates. A more sophisticated baseline, *BL Keyphrases*, uses a list of about 80 manually com-

[1] http://code.google.com/p/mate-tools/

[2] Available from http://verbs.colorado.edu/jdpacorpus/ – we ignore cars batch 009 where no arguments of comparative predicates are annotated.

[3] Available from http://www.cs.uic.edu/~liub/FBS/data.tar.gz – although the original paper works on some unknown subset of this data, so our results are not directly comparable to the results reported there.

[4] http://nlp.stanford.edu/software/corenlp.shtml

[5] Statistically significant at $p < .05$ using the approximate randomization test (Noreen, 1989) with 10000 iterations.

		P	R	F1
cams	BL POS	66.6	38.2	48.5
	BL Keyphrases	53.1	**62.8**	57.5*
	SRL	**73.8**	58.7	**65.4***
cars	BL POS	62.5	34.7	44.6
	BL Keyphrases	51.9	**56.5**	54.1*
	SRL	**73.2**	55.5	**63.2***
J&L	BL POS	74.3	52.9	61.8
	BL Keyphrases	61.5	**80.0**	69.5*
	SRL	**77.0**	68.1	**72.3***

Table 2: Results predicate identification

		P	R	F1
cams	BL	49.4	**47.1**	48.2
	SRL	**66.5**	38.0	**48.4**
cars	BL	50.2	**50.1**	50.1
	SRL	**68.7**	42.2	**52.3***
J&L	BL	38.7	44.6	41.5
	SRL	**68.5**	**45.2**	**54.5***

Table 3: Results argument identification (gold predicates)

		Entity+ / 1			Entity- / 2			Aspect			$F1_m$
		P	R	F1	P	R	F1	P	R	F1	
cams	BL	30.1	**31.7**	**30.9**	21.2	21.3	21.3	61.8	**51.2**	56.0	36.1
	SRL	**38.6**	17.4	24.0	**43.7**	**24.5**	**31.4**	**69.9**	47.7	**56.7**	**37.3**
cars	BL	31.1	**32.7**	**31.9**	23.0	24.0	23.5	49.3	**44.5**	**46.8**	34.0
	SRL	**39.5**	22.9	29.0	**48.1**	**31.0**	**37.7**	**58.4**	36.2	44.7	**37.1***
J&L	BL	43.2	39.4	41.2	19.0	31.1	23.6	15.0	17.1	16.0	26.9
	SRL	**58.3**	**47.2**	**52.1**	**60.8**	**35.6**	**45.0**	**58.8**	**30.6**	**40.3**	**45.8***

Table 4: Results argument classification (gold predicates)

piled comparative keyphrases from (Jindal and Liu, 2006a) in addition to the POS tags.

Table 2 shows the result of our experiments. Our method significantly outperforms both baselines in all datasets. The generally low recall values are mainly a result of the wide variety of predicates that are used to express comparisons (see Discussion).

Results on Arguments. To get results indepent of the errors introduced by the relatively low performance on predicate identification, we use annotated predicates (gold predicates) as a starting point for the argument experiments. All results drop about 10% when system predicates are used.

As a *baseline (BL)* for argument identification and classification, we use some heuristics based on the characteristics of our data. Most entities are (pro)nouns and most predicates are positive, so we classify the first noun or pronoun before the predicate as entity+ (entity 1 for J&L) and the first noun or pronoun after the predicate as a entity- (entity 2). If the predicate is a comparative adjective, we classify the predicate itself as aspect, because this type of annotation is very frequent in the JDPA data. For other predicates except nouns and verbs, we classify the direct head of the predicate as aspect.

Table 3 shows the results for argument identification, the results for argument classification can be seen in Table 4. Our system outperforms the baseline for all datasets. The differences are significant except for the cameras dataset. In general, the numbers are low. We will discuss some reasons for this in the next section.

5 Discussion

Sparseness. There are many ways to express a comparison and the size of the available training data is relatively small. This strongly influences the recall of our system as many predicates and arguments occur only once. As we can see in Table 1, 60% of the predicates in the cameras dataset occur only once. In contrast, only 12 predicates occur ten times or more. The trends are similar in the other datasets. This particularly affects verbs and nouns, where many colloquial expressions are used ("hammers", "pwns", "go head to head with", "put X to the sword", ...).

Argument identification and classification would benefit from generalizing over the many different product identifiers like "EOS 5D" or "D200". We want to try to use a Named Entity Recognition system trained on this type of entities for this purpose.

Sentiment Relevance. The following examples show a problem that is typical for sentiment analysis and responsible for many false positive predicates:

(2) a. "Relatively [**lower**]$_{\mathrm{A}}$ noise at higher ISO ..."
b. "... but [**higher**]$_{\mathrm{A}}$ then [Sony]$_{\mathrm{E+}}$"

Although "higher" often expresses a comparison like in sentence 2b, in sentence 2a it only describes a camera setting and should not be extracted as a comparative predicate. There has been considerable work in the areas of subjectivity classification (Wilson and Wiebe, 2003) and the related sentiment relevance (Scheible and Schütze, 2013) which we will try to use to detect such irrelevant, "descriptive" uses of comparative words.

Linguistic anchoring. In contrast to SRL, the task of comparison detection in reviews is a relatively new task without universally recognized definitions and annotation schemes. The annotators of the corpora had a lot of freedom in their choice of linguistic anchoring of the predicates and arguments. Consider these examples from the cameras dataset:

(3) a. "[**Lighter**]$_{\mathrm{A}}$ in weight compared to the [others]$_{\mathrm{E-}}$."
b. "... [its]$_{\mathrm{E+}}$ [better]$_{\mathrm{A}}$ and faster **compared** vs the [SB800 flash]$_{\mathrm{E-}}$ as well."
c. "... this camera's [screen]$_{\mathrm{E+}}$ is [**smaller**]$_{\mathrm{A}}$ than the [ones]$_{\mathrm{E-}}$ on some competing models ..."

Sentences 3a and 3b show a situation where two words are used to express the same comparison and it is unclear which one to chose as a predicate. The decision is left to the individual annotators.

There is some variety of annotations on arguments as well. In the JDPA data, a comparative adjective is often annotated as aspect, sometimes even when there is an alternative, e.g., "weight" in sentence 3a. Also, for a phrase like "its screen", we find "screen" annotated as the aspect (sentence 1a) or an entity (sentence 3c) – and both have their merit. We want to further study how different linguistic anchorings of comparisons effect classification performance.

Equative comparisons. As we can see from the confusion matrix of our system, the distinction between entity+ and entity- is very difficult to learn. In graded comparisons, the distinction is informative, but sentiment information would be needed for the correct assignment. There are also some problematic cases where the ranking cannot be inferred without the broader context, e.g., sentence 1d.

A more annotation-related problem concerns equative comparisons, i.e., both entities are rated as equal. The difference between entity+ and entity- is meaningless in this case. In the JDPA corpus, entities still have to be annotated as either entity+ or entity- and the annotation guidelines allow the annotator to choose freely. As a result, the data is noisy, for the same predicate sometimes entity- is before the predicate, sometimes entity+. If we eliminate this noise by always assigning the entities in order of surface position, we see a gain in macro averaged F1-measure for all systems of about 2% (cameras) to 4% (cars).

6 Conclusions

We presented a pilot experiment on using an SRL-inspired approach to detect comparisons (comparative predicate, entity+, entity-, aspect) in user generated content. We re-trained an existing SRL system on data that is labeled with comparative predicates and arguments. Even without feature engineering or major adaptions, our approach outperforms the baselines in three datasets in every task. This is an encouraging result for a linguistically grounded modeling approach to comparison detection.

For future work, we plan to include features that have been tailored specifically to the task of detecting product comparisons. To address the inherent diversity of expressions typical for user generated content, we want to employ generalization techniques, e.g., to detect product names. We also want to further study the different possible linguistic anchorings of comparisons and their effect on classification performance. Studies of this kind may also inform future data annotation efforts in that certain ways of anchoring the elements of a comparison linguistically may be more helpful than others. We also believe that the explicit modeling of different types (equative, superlative, non-equal gradable) of comparisons will have a positive effect on performance.

Acknowledgments

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

References

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

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

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June.

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

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

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

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

Christopher Kennedy. 1999. *Projecting the Adjective: The Syntax and Semantics of Gradability and Comparison*. Outstanding Dissertations in Linguistics. Garland Pub.

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

Friederike Moltmann. 1992. *Coordination and Comparatives*. Ph.D. thesis, Massachusetts Institute of Technology.

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

Christian Scheible and Hinrich Schütze. 2013. Sentiment relevance. In *Proceedings of ACL '13*, pages 954–963.

Theresa Wilson and Janyce Wiebe. 2003. Annotating opinions in the world press. In *Proceedings of SIGdial '03*, pages 13–22.

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

Seon Yang and Youngjoong Ko. 2009. Extracting comparative sentences from Korean text documents using comparative lexical patterns and machine learning techniques. In *Proceedings of the ACL-IJCNLP '09*, pages 153–156.

Seon Yang and Youngjoong Ko. 2011a. Extracting comparative entities and predicates from texts using comparative type classification. In *Proceedings of HLT '11*, pages 1636–1644.

Seon Yang and Youngjoong Ko. 2011b. Finding relevant features for Korean comparative sentence extraction. *Pattern Recogn. Lett.*, 32(2):293–296, January.