Detecting Comparative Sentiment Expressions – A Case Study in Annotation Design Decisions

Wiltrud Kessler and **Jonas Kuhn**
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 is by comparing it to a different product. The anchor for the comparison is a comparative predicate like "better". In this work we concentrate on the annotation of multiword predicates like "more powerful". In the single-token-based approaches which are mostly used for the automatic detection of comparisons, one of the words has to be selected as the comparative predicate. In our first experiment, we investigate the influence of this decision on the classification performance of a machine learning system and show that annotating the modifier gives better results. In the annotation conventions adopted in standard datasets for sentiment analysis, the modified adjective is annotated as the aspect of the comparison. We discuss problems with this type of annotation and propose the introduction of an additional argument type which solves the problems. In our second experiment we show that there is only a small drop in performance when adding this new argument type. [1]

1 Introduction

Sentiment analysis is an area in Natural Language Processing that deals with the task of determining the polarity (positive, negative, neutral) of an opinionated document, sentence or other text unit. 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. As many standard approaches assume one polarity to be assigned to one target entity, they cannot deal with comparisons which involve more than one target entity and may involve more than one polarity. It is thus necessary to analyze comparisons separately.

For our purposes we define a comparison to be any statement about the similarity or difference of two entities. Comparative sentences in the linguistic sense ("X is better than Y" or "X is the best") are included in this definition and indeed many comparisons are of this form, but user generated texts also contain many more diverse statements, e.g., "X blows away all others".

In most popular sentiment corpora to date, comparisons are anchored on one word that "expresses the comparison" (*comparative predicate*) which has three arguments: the two *entities* that are compared and the *aspect* they are compared in (Jindal and Liu, 2006b; Kessler et al., 2010). Most comparative predicates are single words like "better" or "best", but English grammar rules systematically introduce multiword predicates. Consider the following variations of a sentence (predicates in bold, arguments in brackets):

(1) a. "$[\text{It}]_{\text{entity1}}$ had a **sturdier** $[\text{feel}]_{\text{aspect}}$."
b. "$[\text{It}]_{\text{entity1}}$ had a **less sturdy** $[\text{feel}]_{\text{aspect}}$."

Sentence 1a compares the aspect "feel" of a camera to some other camera with the comparative predicate "sturdier". If we change the direc-

tion of the comparison, we get a multiword predicate with the modifier "less" added (sentence 1b). In the following we will refer to all such modifiers as *function words* and to the modified adjective heads as *content words*.

In the literature to date, most approaches to automatically detect comparative predicates are single-token-based. For multiword predicates these approaches require one word to be chosen as the comparative predicate (either the function word or the content word, respectively). The first question we want to address in this case study is how this design decision influences the classification performance of a machine learning system trained to detect comparisons.

In many available corpora, the function word is annotated as the comparative predicate and the content word is annotated as the aspect. This creates the counterintuitive situation that changing the direction of the comparison may introduce a new aspect. Also, annotation schemes that define only one aspect slot force a decision whenever a content word and a real aspect are present. This may lead to loss of information or annotation inconsistencies. We propose to solve these problems by introducing a new argument type and as a second question in this case study investigate the effect on performance.

2 Related Work

The syntax and semantics of comparatives have been the topic of research in linguistics for quite some time (Bresnan, 1973; von Stechow, 1984). In the context of sentiment analysis, Jindal and Liu (2006a) are the first to propose an approach for the identification of sentences that contain comparisons. Their system uses 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 and aim at identifying the components of the comparisons.

Several approaches have been presented for the detection of comparative predicates and arguments. In follow-up work on their sentence identification, Jindal and Liu (2006b) detect comparison arguments with label sequential rules and in a second step identify the preferred entity in a ranked comparison (Ganapathibhotla and Liu, 2008). Semantic Role Labeling has inspired approaches that detect predicates and subsequently their arguments, those have been applied to Chinese (Hou and Li, 2008) and English (Kessler and Kuhn, 2013). Xu et al. (2011) use Conditional Random Fields to extract relations between two entities, an attribute and a predicate phrase.

All these studies assume a specific way of annotating comparative predicates and arguments and do not investigate the impact this design decision has on actual classification results.

3 Multiword Predicates

Multiword predicates account for about 10-20% of comparative predicates in our data. Some are expressions like "X has the edge over Y" or "X is on par with Y" which we will not discuss in this work. The focus of this study are multiword predicates like "less sturdy" which are systematically introduced by English grammar rules for expressing comparisons. These constitute the majority of multiword predicates and are composed of a modifying function word and a content word. Besides the modifiers "less" / "more" for comparative forms, and "most" / "least" for the superlative, the list of function words includes "as" which is used to introduce an equative comparison like "X is as good as Y".[2]

In the literature to date, single-token-based approaches are mostly used for the automatic detection of comparative predicates. A strong argument can be made to select the function word as the token anchor for the comparative predicate. There will be more training instances to use in machine learning for a given function word than for the individual content words, so sparseness is reduced. On the other hand, choosing the content word may be more informative for end users.

The first question we want to investigate in this study is whether the different annotation decisions translate into a difference in classification performance. In our first experiment we identify all occurrences of multiword predicates. In one setting (*function predicates*), we annotate the modifying function word as the comparative predicate. In the second setting (*content predicates*), we annotate the modified content word.

[2]Note that not all occurrences of the keywords indicate multiword predicates, e.g., in "X has less noise" the word "noise" is not part of the predicate but the compared aspect.

The following illustrates the different annotations for an example sentence:

(2) a. "...had a **less** [sturdy]$_{\text{aspect}}$ [feel]$_{\text{aspect}}$ ..." *(function predicates)*

b. "...had a less [**sturdy**]$_{\text{aspect}}$ [feel]$_{\text{aspect}}$ ..." *(content predicates)*

In both cases we have the same number of comparative predicates, only the annotations differ. Argument annotations are identical.

The second question deals with the annotation of the content word when we use function predicates. Most corpora annotate the content word as an aspect. We will illustrate some problems with this approach in the following examples:

(3) a. "...a **sturdier** [feel]$_{\text{aspect}}$ ..."

b. "...a **less** [sturdy]$_{\text{aspect}}$ [feel]$_{\text{aspect}}$ ..."

c. "...a **less** [sturdy]$_{\text{aspect}}$ feel ..."

d. "...a **less** sturdy [feel]$_{\text{aspect}}$ ..."

e. "...a **less** flimsy [feel]$_{\text{aspect}}$ ..."

If we compare sentences 3a and 3b we see that changing the direction of the comparison introduces a new aspect. This is counterintuitive because what is compared (i.e., the aspect) should not depend on the introduced ranking. Additionally, if there is only one slot for the aspect, as is the case in one of the corpora we use, annotators will need to decide between annotations 3c and 3d. Annotation 3c is inconsistent when compared to annotation 3a as both compare the same property of the product but have different annotations for aspect. With annotation 3d we lose information about the actual sentiment polarity that is expressed as we are not able to distinguish it from the annotation in sentence 3e.

To solve these issues, we propose to introduce a separate argument with the sole purpose of modeling the content word in a multiword predicate. In our second experiment we use function words as predicates and change the label of the content word from aspect (used in *function predicates*) to this new argument we will call scale (*function preds. w. scale*) to determine the influence on argument classification. This results in the following annotations being compared:

(4) a. "...had a **less** [sturdy]$_{\text{aspect}}$ [feel]$_{\text{aspect}}$ ..." *(function predicates)*

b. "...had a **less** [sturdy]$_{\text{scale}}$ [feel]$_{\text{aspect}}$ ..." *(function predicates with scale)*

	J&L	J-C	J-A	IMS
total preds.	668	642	1327	2108
multiword preds.	36	71	127	245
– more	*13*	*26*	*68*	*123*
– less	*4*	*6*	*12*	*18*
– most	*2*	*1*	*4*	*12*
– least	*0*	*0*	*1*	*1*
– as	*17*	*38*	*42*	*91*

Table 1: Multiword predicates in the data.

The tasks of predicate and argument identification are independent of argument labels, so the only change will be in argument classification. We expect a drop in classification performance due to the increased number of classes, but hope that the drop is not significant as the new argument class is well-defined and should be relatively easy to distinguish from real aspects.

4 Data

We use four datasets in our experiments: the J&L data[3] (Jindal and Liu, 2006b), the camera (J-C) and car (J-A) parts of the JDPA corpus[4] (Kessler et al., 2010), and our own set of camera reviews (IMS)[5] (Kessler and Kuhn, 2014).

We extract all sentences where we find at least one comparative predicate. Table 1 contains some statistics about the number of multiword predicates in these datasets.

In the JDPA data the function word is annotated as the comparative predicate and the content word as the aspect. For every annotated predicate that matches our function word keywords, we check if the token directly following the predicate is annotated as the aspect. If the predicate is "as", we take the aspect as the content word. For the other function words we use the word only if it is an adjective (as determined by the Stanford POS Tagger). This serves to distinguish "less sturdy" which we want to include in our experiments from "less noise" where the noun "noise" should be the aspect, not part of the predicate.

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

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

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

		P	R	F1	Δ
J&L	function preds.	**76.4**	**66.8**	**71.3**	
	content preds.	75.2	60.9	67.3	-4.0
J-C	function preds.	74.3	**59.3**	**66.0**	
	content preds.	**75.6**	55.0	63.7	-2.3
J-A	function preds.	**74.6**	**59.5**	**66.2**	
	content preds.	74.5	53.2	62.1	-4.1
IMS	function preds.	84.4	**76.4**	**80.2**	
	content preds.	**84.5**	72.6	78.1	-2.1

Table 2: Results predicate identification.

In the J&L data the complete phrase "as X as" is annotated as the predicate. We check if the first and last word of a predicate is "as", and take the words in between as content words. For the other function words annotation is like in the JDPA corpus, so we proceed the same way.

In our IMS data, the function word is always annotated as the predicate. The content word is annotated as a separate argument scale which we can use directly. For the first experiment we map the scale annotations to aspect.

The resulting annotations for JDPA and J&L are a bit noisy, but manual inspection shows that nearly all of the content words are correctly identified. We miss some occurrences of multiword predicates in cases where some other aspect is present and has been annotated instead of the content word (cf. example 3d).

5 Experiments

Setup. We use the MATE Semantic Role Labeling system (Björkelund et al., 2009)[6] with default settings and without the re-ranker. We re-train the system on our datasets to identify comparative predicates and arguments. We perform three classification steps: predicate identification, argument identification and argument classification. The classification uses features based on the output of a dependency parser. Features are extracted for predicates and arguments as well as the predicate head, predicate dependents and the path between argument and predicate. We use the same features for all experiments and identify predicates and arguments of all parts of speech. This setup is equivalent to (Kessler and Kuhn, 2013).

We evaluate on each dataset separately using 10-fold cross-validation. We report precision (P), recall (R), and F1-measure (F1). All results are calculated on all predicates and arguments annotated in the data. Bold numbers denote the best result in each column and dataset.

We cannot calculate significance because annotations change between experiments, but we report the absolute differences in F1-measure to the function predicate setting (Δ).

Function predicates vs. content predicates. Table 2 shows the results for predicate identification. We can see that annotating the content word decreases performance in all datasets. This fits our expectation as lexical features have a big weight in the model and by choosing a number of different adjectives over few function words we make the data more sparse. The decrease is quite large compared to the relatively small number of changes we are making.

Table 3 and the first two lines for each dataset in Table 4 show the results of argument identification resp. classification. With system predicates, due to the decreased performance in predicate identification, performance on arguments suffers to a similar degree. With gold (annotated) predicates, performance still suffers for J&L and IMS, but the JDPA datasets are not as much affected or even gain. Part of this is due to the fact that *content predicates* over-generates aspects that are the same token as the predicate even for single word predicates like "faster". Such annotations never occur in the other datasets but are common in the JDPA datasets. The increased recall for aspects balances the loss on the other arguments.

Aspect annotations vs. scale annotations. The second experiment influences only argument classification, compare lines 1 and 3 for every dataset in Table 4. As we introduce more classes, we expect overall performance to drop. Indeed there is a drop, but the difference between the two configurations is small. When we look at the confusion matrices for all datasets, we see that there are nearly no confusions of the scale with an entity and only few of scale and aspect.

We have analyzed some cases where the scale has been confused with the aspect in the IMS data.

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

		with system predicates				with gold predicates			
		P	R	F1	Δ	P	R	F1	Δ
J&L	function predicates	**57.6**	**31.9**	**41.1**		69.4	**46.3**	**55.5**	
	content predicates	56.8	28.1	37.6	-3.5	**69.6**	45.2	54.9	-0.6
J-C	function predicates	**57.4**	**25.7**	**35.5**		**67.9**	37.2	**48.1**	
	content predicates	56.7	24.9	34.6	-0.9	67.6	**37.3**	**48.1**	-0.0
J-A	function predicates	**57.2**	**27.5**	**37.2**		**70.4**	41.7	52.4	
	content predicates	56.8	25.8	35.5	-1.7	**70.4**	**42.1**	**52.7**	+0.3
IMS	function predicates	**70.7**	**44.1**	**54.3**		**78.9**	**57.4**	**66.4**	
	content predicates	70.4	41.9	52.5	-1.8	77.9	56.5	65.5	-0.9

Table 3: Results argument identification.

		with system predicates				with gold predicates			
		P	R	F1	Δ	P	R	F1	Δ
J&L	function predicates	**50.2**	**27.8**	**35.8**		**59.9**	**40.0**	**48.0**	
	content predicates	48.9	24.2	32.4	-3.4	59.7	38.8	47.0	-1.0
	function preds. w. scale	49.6	27.5	35.4	*-0.4*	59.2	39.5	47.4	*-0.6*
J-C	function predicates	**49.5**	**22.2**	**30.7**		**55.5**	**30.4**	**39.3**	
	content predicates	47.0	20.6	28.7	-2.0	54.5	30.1	38.8	-0.5
	function preds. w. scale	49.4	22.1	30.6	*-0.1*	55.2	30.2	39.1	*-0.2*
J-A	function predicates	**43.8**	**21.1**	**28.5**		50.2	29.7	37.3	
	content predicates	**43.8**	20.0	27.4	-1.1	**51.0**	**30.5**	**38.2**	+0.9
	function preds. w. scale	43.3	20.8	28.1	*-0.4*	49.7	29.4	37.0	*-0.3*
IMS	function predicates	**63.0**	**39.3**	**48.4**		**69.0**	**50.2**	**58.1**	
	content predicates	62.4	37.1	46.5	-1.9	67.7	49.1	56.9	-1.2
	function preds. w. scale	62.4	38.9	47.9	*-0.5*	68.4	49.8	57.6	*-0.5*

Table 4: Results argument classification (micro-average over all classes).

Confusions occur mostly with untypical scale arguments like "more feature rich" or "more pro" where the system predicts an aspect because the content word is tagged as a noun. We have also found a few annotation errors where annotators mistakenly annotated an aspect instead of a scale.

6 Conclusions

In this short paper we present experiments on how different annotations of multiword comparative predicates ("more powerful", "as good as", ...) affect the classification performance of a machine learning system that identifies comparative predicates and arguments. Our experiments indicate that it is more helpful to annotate function words than content words as predicates. In the annotation conventions adopted in standard datasets for sentiment analysis, the modified adjective is annotated as the aspect of the comparison. We discuss problems with this type of annotation and propose the introduction of an additional argument type which solves the problems. In our second experiment we show that there is only a small drop in performance when adding this new argument type. For future work we plan to look more closely at the annotation of other (non-systematic) multiword predicates such as "on par with".

Acknowledgments

We thank the reviewers for their helpful comments. 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.

Joan W. Bresnan. 1973. Syntax of the comparative clause construction in English. *Linguistic Inquiry*, 4(3):275–343.

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.

Wiltrud Kessler and Jonas Kuhn. 2013. 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.

Wiltrud Kessler and Jonas Kuhn. 2014. A corpus of comparisons in product reviews. In *Proceedings of LREC '14*.

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*.

Arnim von Stechow. 1984. Comparing semantic theories of comparison. *Journal of semantics*, 3:1–77.

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.