A Corpus of Comparisons in Product Reviews

Wiltrud Kessler, Jonas Kuhn

Institute for Natural Language Processing
University of Stuttgart
wiltrud.kessler@ims.uni-stuttgart.de

Abstract

Sentiment analysis (or opinion mining) deals with the task of determining the polarity of an opinionated document or sentence. Users often express sentiment about one product by comparing it to a different product. In this work, we present a corpus of comparison sentences from English camera reviews. For our purposes we define a comparison to be any statement about the similarity or difference of two entities. For each sentence we have annotated detailed information about the comparisons it contains: The *comparative predicate* that expresses the comparison, the *type* of the comparison, the two *entities* that are being compared, and the *aspect* they are compared in. The results of our agreement study show that the decision whether a sentence contains a comparison is difficult to make even for trained human annotators. Once that decision is made, we can achieve consistent results for the very detailed annotations. In total, we have annotated 2108 comparisons in 1707 sentences from camera reviews which makes our corpus the largest resource currently available. The corpus and the annotation guidelines are publicly available on our website.

Keywords: Sentiment analysis, comparisons, product comparisons corpus

1. Motivation

Sentiment analysis (or opinion mining) deals with the task of determining the polarity of an opinionated document or sentence (Liu, 2012). In product reviews, sentiment is typically assigned with regard to some target product. This assignment is difficult when more than one product is discussed in a sentence, i.e, when some product is compared to a different product. Such comparisons are a common way of expressing sentiment and they are of interest for companies that do not only want to know what aspects of their product users like or dislike, but also where they stand in relation to their competitors.

For our purposes we define a comparison to be any statement about the similarity or difference of two entities (Liu, 2012). Comparative sentences in the linguistic sense ("A is better than B" or "A is the best") are included in this definition and indeed many comparisons are of this form. But comparisons in user generated texts also contain many more diverse statements that we also include in our definition, e.g., "A blows away all others", "A and B have the same X", "A wins over B".

The analysis of such comparisons has not received a lot of attention in the sentiment analysis community. While there is a wealth of corpora annotated with sentiment information, to our knowledge only two corpora with annotations for comparisons in English reviews are publicly available. Each of them individually is a relatively small source of comparison sentences. Combining them leads to a high degree of heterogeneity since not only the domains but also the annotation schemes vary.

In this work we present a dedicated gold standard corpus of comparison sentences. Our annotation guidelines combine the strengths of the existing corpora while eliminating possible sources of inconsistency. We have limited ourselves to one domain of texts (camera reviews) in order to get more homogeneous data. To our knowledge, our corpus is the largest source of comparison sentences in reviews to date.

In the following, we will present our annotation scheme and annotation process. We will present results of an agreement study and give some statistics about the data we have annotated. The corpus and the annotation guidelines are publicly available on our website[1].

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; Staab and Hahn, 1997). In most of these works, expressions are considered as a comparative only when they contain an adjective or adverb in comparative or superlative form. However, for our work we also want to include comparisons that are not comparatives in a linguistic sense.

The motivation for our work on comparisons comes from the analysis of user generated content, especially sentiment analysis. Sentiment analysis is a challenging task that has received a lot of attention in recent years. Liu (2012) gives an overview of the general tasks and methods that are involved. Work in sentiment analysis has gradually moved from determining the overall opinion of a document down to more fine-grained levels. In non-comparative sentiment expressions, one polarity is generally assigned to one sentiment target. A comparison has to be treated differently, as it involves more than one target entity and may involve assignment of more than one polarity, e.g., the statement "A is better than B" expresses positive sentiment toward A and negative sentiment toward B.

Previous work has tried to identify and further analyze comparison sentences in reviews. Jindal and Liu (2006a) identify sentences that contain a comparison by using class sequential rules based on keywords as features for a Naive Bayes classifier. In follow-up work, they also further analyze the entities and aspects in these sentences (Jindal and

[1] www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/reviewcomparisons/

Liu, 2006b) and identify the preferred entity in a non-equal comparison (Ganapathibhotla and Liu, 2008). Similar approaches have been used to identify and analyze comparisons in other languages, e.g., Korean (Yang and Ko, 2011) or Chinese (Huang et al., 2008).

More recently, Xu et al. (2011) cast the task of analyzing comparisons as a relation extraction problem and use conditional random fields to extract relations between two entities, an attribute and a predicate phrase. Hou and Li (2008) use a semantic role labeling approach for Chinese, Kessler and Kuhn (2013) for English.

As we are presenting a new resource in this work, we will focus on the data used in previous approaches. Of the approaches working on English, only two corpora in the domain of reviews are publicly available. The first one[2] called J&L in the following is from (Jindal and Liu, 2006b) and contains approximately 650 comparison sentences from reviews, blog posts and forum discussions about various topics ranging from digital cameras to soccer. This corpus was designed specifically for the task of identifying comparisons. The second one is the JDPA corpus[3] by J. Kessler et al. (2010) and consists of blog posts about cameras (506 sentences) and cars (1100 sentences). Comparisons are annotated, but were not the focus of the annotation.

One problem for using these existing datasets as training data for machine learning is that each of them individually is a relatively small source of comparison sentences. Combining the three datasets leads to a high degree of heterogeneity in the data since not only the domains but also the annotation schemes vary.

Both corpora identify a comparative predicate as the anchor of a comparison, and for each predicate two entities and an aspect. There are two main differences between the annotation schemes which concern comparison types and the annotation of entities. In the J&L corpus every comparison is assigned one out of four types of comparisons. The JDPA corpus only distinguishes between equative and non-equative comparisons. Entities in the J&L corpus are annotated as entity 1 or 2 based on the order of appearance in the sentence, the preferred entity is not marked. In the JDPA corpus entities are annotated explicitly as the preferred or non-preferred entity, even if the comparison does not introduce an ordering. In the corpus presented here, we annotate entities by order of appearance which can be done for all annotations. Additionally, for ranked comparisons where there is a preferred entity, we annotate the direction of the comparison (entity 1 or 2 preferred).

Apart from structural differences, the corpora also differ in what they regard as a comparison and include in their data. Consider the following examples:

(1) a. "this is my **first** digital camera."
b. "The DMC-FZ18 is my first DSLR experience."
c. "The 2008 Subaru Impreza WRX STI is **based** on the Impreza WRX hatchback . . . "
d. "This engine is heavily based on the older CHV designs . . . "

Sentence 1a is annotated as a comparison in the J&L data. Comparable sentences in the JDPA data (e.g., sentence 1b) are never annotated as comparisons. Even inside the JDPA corpus we note heterogeneity between batches. Sentence 1c from JDPA cars batch 5 has "based" annotated as the comparative predicate (marked in bold). Similar sentences in batch 1 (e.g., sentence 1d) contain no annotation for "based". This is a general development, in the later batches many more diverse predicates are annotated. In general, as comparisons have not been the focus of annotation for the JDPA corpus, the annotation guidelines leave many decisions to the annotators which results in considerable inconsistency among annotations.

To conclude, the existing corpora differ in what is included as a comparison and lack important information (comparison type in JDPA, preferred entity in J&L). Many decisions were left to the annotators which results in a number of inconsistencies in the annotations. In our work, we have designed our annotation guidelines to address these issues with the goal to produce a more comprehensive and consistent gold standard corpus of comparison sentences.

3. Data sources and annotation procedure

To get data from the same domain as the existing corpora, we have chosen to annotate camera reviews. In this way the JDPA camera dataset with some adaptations could be added to our new dataset to form a larger (albeit noisier) training set for machine learning.

As only about 10% of sentences contain a comparison in a typical review, we use a two-stage annotation process. First, we use crowd-sourcing to identify sentences that with high probability contain a comparison. Only these sentences were then passed on to the second stage and annotated in more detail.

3.1. Data sources and preprocessing

We base our corpus of comparison sentences on the camera data set provided by Branavan et al. (2008)[4]. They downloaded a set of camera reviews from `epinions.com` and separated the review text from the other information. From the 12586 reviews in total we extracted the first 608 reviews for annotation of comparisons which gave us 11232 sentences. Reviews that were duplicates or off-topic were manually removed when the annotators detected them (44 reviews in total). We removed HTML tags from the review texts and used the Stanford Core NLP[5] to automatically do sentence segmentation and tokenization.

3.2. Identifying sentences that contain a comparison

We decided to use crowd-sourcing for the task of deciding whether a sentence contains a comparison. We designed a HIT (Human Intelligence Task) and uploaded it on Amazon Mechanical Turk (AMT). The workers were given short instructions about their task and a few examples of comparisons and non-comparisons. Every sentence was annotated by two AMT workers. Possible labels were "comparison", "no comparison" or "not sure". To discourage the use of

[2] `www.cs.uic.edu/~liub/FBS/data.tar.gz`
[3] `verbs.colorado.edu/jdpacorpus/`
[4] `groups.csail.mit.edu/rbg/code/precis/`
[5] `nlp.stanford.edu/software/corenlp.shtml`

"not sure", workers were asked to provide feedback on why they were not sure when they chose this value.
A batch of 500 sentences was annotated on AMT in about one hour. The results imply that the task is difficult for AMT workers. The workers agreed on "no comparison" for 40% of sentences and on "comparison" for 20% of sentences. Compared to the 10% of comparison sentences we would expect to find, this corresponds to high recall. If both AMT workers chose "comparison", the sentence was passed on to the second stage, if both chose "no comparison" the sentence was discarded. In the rest of the cases (40% of sentences), the AMT workers did not agree or both workers chose "unsure". These sentences were roughly checked for obvious non-comparisons by one of the authors and then passed on to the second stage.

3.3. Fine-grained comparison annotation

The more fine-grained annotation was carried out by three students hired and trained specifically for the task. One is a graduate student of computational linguistics, one an undergraduate student of media and the last one a graduate student of computer science. All have an advanced level of English, but none is a native speaker.
The annotators were given annotation guidelines with detailed instructions. We chose 30 sentences from our data as a training set. This set contained all types of comparisons and was annotated by one of the authors as well. Each annotator had to first annotate this training set and any disagreements with the author were discussed.
After training, the actual annotation was carried out by the annotators independently. The annotators only saw the sentence they were currently annotating, without the context of the review. Annotators had the possibility to decide that the sentence is not a comparative sentence, this is necessary as the first stage was designed to give high recall, not precision. Apart from the annotations, annotators were encouraged to provide additional feedback in hard cases. These cases were discussed with one of the authors and the feedback was used to refine the annotation guidelines.

4. Annotation scheme

The data to be annotated consists of sentences from user generated content, namely reviews of digital cameras. The goal of the annotation is to provide fine-grained information on comparisons. We did not annotate any other information besides comparisons.
We define a comparison as any statement about the similarity or difference of two entitites. This definition covers a wide variety of expressions, also beyond the linguistic category of comparative sentences. A sentence may contain none, one or several comparisons.
The main anchor for a comparison is the comparative predicate. It has the following arguments (relations): The two entities that are being compared (E1, E2), and the aspect they are compared in (A). In some special cases a fourth argument scale for modified adjectives/adverbs (S) is added. Predicates and arguments are annotated as token spans in the same sentence, we do not annotate parts of tokens or cross-sentence relations. In addition to the arguments, for each predicate we annotate the type of the comparison and the types of the involved entities. The parts of comparison to be annotated are discussed in more detail below. As an illustration, consider the following example sentences from our data (all examples are presented with original spelling and punctuation):

(2) a. "But [this new XT]$_{\mathrm{E1}}$ compared to [the old rebel]$_{\mathrm{E2}}$ has MUCH **better** [picture quality]$_{\mathrm{A}}$."
b. "[The XT]$_{\mathrm{E1}}$ **beat** [the 300D]$_{\mathrm{E2}}$ in the [file writing]$_{\mathrm{A}}$ department as well . . . "
c. "The biggest **difference** is the [quality of the jpegs]$_{\mathrm{A}}$."
d. "By the way, [Nikon]$_{\mathrm{E1}}$ is at the **top of the line** in [flashes]$_{\mathrm{A}}$."
e. "[It]$_{\mathrm{E1}}$ shared the **same** [sensor]$_{\mathrm{A}}$ as [Nikon D200]$_{\mathrm{E2}}$, and [the latest D80]$_{\mathrm{E2}}$."
f. "In fact I think [the D80]$_{\mathrm{E1}}$ is **better** at [handling noise]$_{\mathrm{A}}$ and [suppressing banding artifacts]$_{\mathrm{A}}$ at higher ISO's."

4.1. Comparative predicate

The central part of any comparison is the *comparative predicate*. The comparative predicate is the syntactic marker that introduces a comparison. Predicates can be of any part of speech, e.g., adjectives (sentence 2a), verbs (sentence 2b), nouns (sentence 2c). We allow annotators to annotate multi-word expressions. This mainly concerns expressions such as "top of the line" in sentence 2d.
Comparisons can express some personal opinion or belief (subjective, sentence 2a), or state verifiable facts (objective, sentence 2e). We annotate both subjective and objective comparisons. We do not include expressions that on the surface look like comparisons, but are used as descriptions of environments or states, i.e., "at higher ISOs" or "I am more than happy".

4.2. Scale (S)

There is a limited number of predicates that are function words and do not by themselves contain the information about what distinguishes the entities. Consider these sentences from our data:

(3) a. "Only [the S3]$_{\mathrm{E1}}$ has a potentially **better** [movie record mode]$_{\mathrm{A}}$."
b. ". . . [the SD800]$_{\mathrm{E1}}$ has a **more** [powerful]$_{\mathrm{S}}$ and overall flexible [movie capture mode]$_{\mathrm{A}}$."
c. "That's a pretty [good]$_{\mathrm{S}}$ [price]$_{\mathrm{A}}$ **compared** to [everything else that I've seen]$_{\mathrm{E2}}$ . . . "
d. "[it]$_{\mathrm{E1}}$'s just as **as** [capable]$_{\mathrm{S}}$ as [the D200]$_{\mathrm{E2}}$."

The annotation for the predicate "better" in sentence 3a is straightforward and does not require an additional argument. We would like to have the annotations for the comparison "more powerful" in sentence 3b parallel the annotations for the comparison "better". One possible way to go would be to annotate "more powerful" as a multi-word predicate. A second possible way is to split the predicate into two parts, annotate the function word "more" as the predicate and the modified adjective "powerful" as an argument which we call scale (S). We chose to split the predicate as this allows us to also capture cases where the two parts are not adjacent (sentence 3c). There is a limited number of predicates that allow the annotation of a scale argument. Besides ranked comparisons with "less" and "more",

another frequent predicate is "as" when used to introduce an equative comparison (sentence 3d).

4.3. Entities (E1, E2)

A comparison involves two *entities* that are compared. One or both of the entities may be implicit. Most often the entities are products, e.g., the two cameras "this new XT" and "the old rebel" in sentence 2a.

We decided to annotate entities based on the order of appearance in the sentence. The first entity is annotated as entity 1, the second as entity 2. The information which entity is preferred if the comparison introduces a ranking between the entities is included in the comparison type. This style of annotation is much easier for annotators, as they are not forced to chose a "better" and "worse" entity if there is no obvious ranking.

An entity can consist of a group of products that are listed individually. As an example consider sentence 2e, where entity 1 "it" (presumably the camera under review) is compared to two other cameras, "Nikon D200", and "the latest D80". Both cameras together make up entity 2. We annotate each listed item individually.

4.4. Aspects (A)

In most sentences one attribute or part of a product is being compared, like "picture quality" in sentence 2a. We follow the terminology of aspect-oriented sentiment analysis and call this the *aspect*[6]. The notion of an aspect includes parts, properties or attributes of the product, results of using the product, and actions performed with it.

Like with entities, there can be more than one aspect compared at the same time, e.g., "handling noise" and "suppressing banding artifacts" in sentence 2f. Aspects may not always be explicit, e.g., "expensive" implies the aspect price. We only annotate explicit aspects.

4.5. Entity type

For each entity in a comparison we annotate its type. Most of the entities are products (cameras in our case), but we distinguished whether a single product (E1 in sentence 2e) or a set of products (E2 in sentence 2e) is compared. Entities can also be a reference to a general standard (E2 in sentence 3c) or a company (E1 in sentence 2d).

Entities can also be of the type "aspect". While this may sound confusing at first, this occurs is because aspects form a hierarchy. It is possible to talk about an aspect of an aspect of the product. Consider the following sentences:

(4) a. "Performance: [The D80]$_{\mathrm{E1-product}}$ uses essentially the **same** [sensor]$_{\mathrm{A}}$ as [the D200]$_{\mathrm{E2-product}}$."

b. "[This full-sized framed sensor]$_{\mathrm{E1-aspect}}$ contains the exact **same** [megapixel density]$_{\mathrm{A}}$ as [the one in the Rebel XT and the 20D]$_{\mathrm{E2-aspect}}$ ..."

In sentence 4a, "sensor" is the aspect that is compared in a comparison between two cameras. In sentence 4b, two sensors are compared in their aspect "megapixel density". The two entities should get the type "aspect" in this sentence. The distinction between entity and aspect for the purpose of our annotation is not only dependend on semantic class (e.g., camera or camera part), but also involves the function in the sentence.

This distinction may be very relevant for a system that lists all differences between two products A and B. For sentences like sentence 4b we would want to list the comparison of an aspect of A under the product A itself. Whenever an entity has the type "aspect", it would serve as an indicator that some linking of the aspect back to the product it belongs to has to be done.

In total, we distinguish six possible types for entities: product, set of products, standard, company, aspect or other. This list of types is tailored to our domain of product reviews and will need adaptation for other domains.

4.6. Comparison type

Comparisons have two main types, gradable and non-gradable (Liu, 2012). A gradable comparison expresses an ordering relationship of the entities being compared. This ordering relationship can have three forms, it can set one entity over the other (ranked comparison / non-equal gradable, sentence 2a), one entity above/below all others (superlative comparison, sentence 2d), or rate all entities as equal (equative comparison, sentence 2e). The first two relations also have two subtypes that clarify the direction of the relation (E1 is superior/inferior to E2), i.e., the order of preference for the entities.

Non-gradable comparisons express a difference between two entities, but do not rank the entities. We annotate non-gradable comparisons only if there is a direct comparison between two entities in an aspect they share (sentence 2c). We do not annotate statements that list existing or missing aspects, even if two entities occur in the sentence, e.g. "X has A, but Y not" or "X has A, but Y has B".

As a result, like in the J&L corpus, we distinguish four types of comparisons that are annotated in our corpus: ranked, superlative, equative comparisons, and non-graded differences. In addition to the J&L annotation of comparison types, we annotate the direction (superior/inferior) for ranked and superlative comparisons.

The type of comparison cannot be determined solely on the basis of the predicate. Syntactic context, especially negation, changes the type of comparison. As an example take sentence 3d which is an equative comparison. If we use "not as capable" instead of "as capable", the result would be a ranked comparison where entity 2 is preferred.

Sometimes the direction of a ranked comparison is unclear, especially with predicates like "smaller" or "higher". In such cases, the annotators are asked to rely on their world knowledge (e.g., high resolution is good, high price is bad) or any context available in the sentence.

5. Corpus Analysis

5.1. Inter-annotator agreement

All annotators had to annotate a set of 100 sentences for the purpose of calculating agreement between annotators as a measure of consistency. These sentences do not include the annotator training set. All of the sentences were judged to be comparison sentences in the first annotation stage.

For categorical label assignments we measure observed agreement A and chance-corrected agreement with Cohen's

[6] Other terms are "feature", "attribute" or "dimension".

Categorical:	A	κ	# agreed
Sentence level	0.8800	0.4448	82
# comparisons	0.8327	0.6314	58
Comparison type	0.7872	0.6065	57
Type entity 1	0.8442	0.6512	45
Type entity 2	0.8102	0.6478	29
Text spans:	agr_s	agr_l	# agreed
Predicate	0.7987	0.8303	80/84
Scale	0.8943	0.8943	14/14
Entity 1	0.7551	0.8555	48/58
Entity 2	0.7280	0.8254	33/40
Aspect	0.5587	0.7547	22/31

Table 1: Inter-annotator agreement. All agreement numbers are averages over pairwise measures.

κ for each pair of annotators. We report the averages of pairwise A and κ. All metrics are calculated only on items that have been annotated by both annotators, i.e., for the comparison type the predicates, for entity types the entity spans have to match. For the comparison of the number of comparisons a sentence contained we put all values over 4 into the same bucket, as this is very rare.

For the annotation of text spans where each annotator individually picks some words from the sentence instead of assigning a label from a predefined set, we follow J. Kessler et al. (2010) and use text span agreement. For sets of annotations X and Y by annotators x and y, the agreement of x to y is calculated as

$$agr(x||y) = \frac{|X \text{ matches } Y|}{|X|}$$

We consider two varieties of matching. In the strict version agr_s, two spans are considered to match only if they are exactly the same. In the lenient version agr_l, two spans are considered to match if they have at least one overlapping token. Text span agreement is calculated for each pair of annotators, we report the average over all pairs of annotators. We only compare spans of the same type, if one annotator annotates some span as entity 1 and the other as aspect, this is not a match.

Table 1 shows the results of the agreement study. The last column shows the number of items on which all annotators agreed (lenient match/strict match). Only matches of non-empty text spans are counted.

Overall, agreement is in the range of values reported for other sentiment annotation tasks. To our knowledge, there are no values reported for the annotation of comparisons in previous work. J. Kessler et al. (2010) report text span agreement for some annotations, but not for comparisons. For the task of identifying sentiment text spans which might be comparable to identifying predicates, they report a lenient agreement of 0.75. For the identification of the sentiment-target relation which may roughly be comparable to identifying the entities in a comparison, they report 0.66. The values are not directly comparable across corpora and annotation schemes, but may give a general idea about the range of expected agreement values.

In our data, agreement on sentence-level is close to 90%, but κ is low (a value between 0.4 and 0.6 is considered moderate agreement). One reason is the very skewed distribution as all of the sentences were selected as containing comparisons in the first annotation stage. Of the 82 sentences that all annotators agreed on, 77 were agreed to contain at least one comparison. But this also confirms the results we got from the AMT experiments that the decision whether a sentence contains a comparison is sometimes not as easy at it may seem. There are of course many sentences that are very obviously a comparison (like those from example 2) or non-comparisons (like "I bought A last week", "B is fantastic"). But here are some examples of sentences from our data where the annotators did not agree:

(5) a. "The images were great both indoors and out."
b. "...athough I would like to see it a little faster."
c. "It mirror flip doesn't sound like a mechanical camera like my 7D ..."
d. "I had to compare this camera with the Nikon D80."

Sentence 5a is a comparison between different usages of the same product, sentence 5b is a wish, sentence 5c is a description of the sound the camera makes. Sentence 5d states that there is going to be a comparison, but this is not (yet) it. We updated the annotation guidelines to explicitly exclude these categories of sentences and had the annotators check their annotation again to eliminate these categories of sentences.

For the assignment of comparison type, we have an agreement value of 79% and $\kappa = 0.61$. In general, κ values between 0.6 and 0.8 are considered substantial agreement. The main source of disagreements about the type of comparison is that one annotator tended to annotate comparisons as type difference if they were ranked but the direction was not clear to the annotator.

Once it is established that there is a comparison, text span agreements are high. Disagreements mainly come from determining the exact boundary, e.g., one annotator annotated "RAW photo quality" as the aspect, another "photo quality". Entities are easier to identify than aspects, probably because aspects require more domain knowledge.

For the assignment of entity types we again have substantial agreement. Disagreements about the entity type come mainly from two sources. The main source of disagreements is missing domain knowledge or missing context. In many sentences entities are referred to by only a model number. When an annotator was lacking domain knowledge, such occurrences were mislabeled as 'other'. It is also common to refer to a product by the company name only, some such cases were mislabeled as 'company'. Most of these errors can be spotted by comparing the types assigned to the two entities. Usually the two compared entities should have the same type. Missing context or domain knowledge is responsible for almost all the cases where entities 1 and 2 have different types. In a few cases the annotation of different types reflects an error of the sentence author, e.g., in sentence 6a were the correct reference would actually be "the one on the 600".

The second source of disagreements is when entity spans match leniently, but this changes the type of the entity. Consider the annotations for the following sentence:

(6) a. "...[the 630]$_{\text{E1}}$'s [lcd]$_{\text{A}}$ seemed **less** [sharp]$_{\text{M}}$ compared to [the 600]$_{\text{E2}}$." (annotator 1)

b. "...[the 630's lcd]$_{E1}$ seemed **less** [sharp]$_{M}$ compared to [the 600]$_{E2}$." (annotator 2)

The text spans of entity 1 overlap, so entity types are compared. But "the 630" refers to a product and "the 630's lcd" to an aspect, so the types do not match. This example also illustrates that the differentiation between entity and aspect is sometimes not that clear-cut.

The results from this agreement study have been used to refine the annotation guidelines. Annotators have been asked to review their annotations for these errors.

5.2. Statistics about the data

Statistics about our complete corpus can be found in Table 2. These numbers do not include the annotator training set or the agreement set. In total we collected 1707 sentences that contain at least one comparison. This is 15% of the sentences extracted from the reviews.

The average number of comparisons per sentence is 1.23. The overwhelming majority of sentences (over 80%) contains only one comparison, nearly all of the remainder contain 2 or 3 comparisons. The maximum number of comparisons is 17 (!), this is a sentence that consists of an enumeration of basically every aspect of a camera[7]. If we exclude this sentence, the maximum is 5.

The overwhelming majority of comparisons is ranked. The most frequent predicates in ranked comparisons are "better", "more", "as" (in sentences with "A is not as X as B") and "less". The most frequent predicates for equative comparisons are "same", "as", "similar" and "like". Statements of difference have "difference", "compared" and "different" as the most frequent predicates. Nearly all superlative comparisons contain an adjective in superlative form, the others are statements like "nothing beats A".

There are far more ranked and superlative comparisons with direction superior than inferior (78% vs. 22%). This is consistent with the bias on positive statements repeatedly reported in sentiment analysis.

Most of the comparisons have one entity 1 and one entity 2, but is also common to drop entity 2, especially for superlatives. Usually at least one aspect is present as well, except for superlatives and differences where more than half of the instances do not contain an aspect.

There are 337 annotations for the argument scale. Apart from the expected predicates "more", "less", "as", we have several occurrences of expressions like "compared". There are a few instances where the argument contains a multiword predicate that had to be split up, e.g., in the sentence "it puts the others to shame." the annotator wanted to chose "put to shame" as the predicate. As we do not allow gaps in predicates and she did not want to include the entity into the predicate, she chose "put" as the predicate and annotated "to shame" as the modified "adjective".

Most of the compared entities are products (one product or a set of products). A number of entities are of the type "aspect", this mainly includes parts of cameras like the flash or the sensor, and the pictures produced with the camera.

[7] An excerpt: "...is **easier** to [navigate]$_{A}$, has **more** [features]$_{A}$, **more** [auto-focus points]$_{A}$ (9 vs 7), **better** [quality]$_{A}$ and **faster** [auto-focus]$_{A}$, ..."

		# items
Sentences	All sentences	11232
	Comparison sentences	1707
	with 1 comparison	1406
	with 2 comparisons	238
	with 3 comparisons	45
	with ≥ 4 comparisons	18
Comparison types	Predicates	2108
	Ranked	1389
	Equative	357
	Difference	216
	Superlative	146
Arguments	Scale	337
	Entity 1	1907
	Entity 2	1360
	Aspect	1407
Entity types	Product	1906
	Set of products	450
	Standard	124
	Company	45
	Aspect	470
	Other/unknown	98

Table 2: Statistics about the data

In total, our corpus contains more than twice as many sentences as the J&L corpus, more than three times as many as the camera part from the JDPA corpus and 1.5 times as many as the car part. This makes it the largest resource dedicated to comparisons in reviews currently available.

6. Conclusion

In this work, we presented a dedicated gold standard corpus of comparison sentences from English camera reviews. For each sentence we have annotated detailed information about the comparisons it contains: The *comparative predicate* that expresses the comparison, the *type* of the comparison, the two *entities* that are being compared, and the *aspect* they are compared in. We have described our annotation process and given an overview of our annotation guidelines. The results of our agreement study show that the decision whether a sentence contains a comparison is difficult to make even for trained human annotators. Once that decision is made, we can achieve consistent results for the very detailed annotations. In total, we have annotated 2108 comparisons in 1707 sentences from camera reviews which makes our data the largest resource of comparisons in reviews currently available. The corpus and our annotation guidelines are publicly available on our website.

7. Acknowledgements

We would like to thank the AMT workers and our student annotators for their work. The work reported in this paper was supported by a Nuance Foundation grant.

8. References

Branavan, S. R. K., Chen, Harr, Eisenstein, Jacob, and Barzilay, Regina. (2008). Learning document-level semantic properties from free-text annotations. In *Proceedings of ACL '08*, pages 263–271.

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

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

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

Huang, Xiaojiang, Wan, Xiaojun, Yang, Jianwu, and Xiao, Jianguo. (2008). Learning to identify comparative sentences in chinese text. In Ho, Tu-Bao and Zhou, Zhi-Hua, editors, *PRICAI 2008: Trends in Artificial Intelligence*, volume 5351 of *Lecture Notes in Computer Science*, pages 187–198. Springer Berlin Heidelberg.

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

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

Kessler, Wiltrud and Kuhn, Jonas. (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.

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

Liu, Bing. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Staab, Steffen and Hahn, Udo. (1997). Comparatives in context. In *Proceedings of AAAI '97 / IAAI '97*, pages 616–621.

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

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

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