Multilingual Sentiment based Review Summarizer

Siddharth Dasani
8209990799
sdasani@usc.edu

Piyush Gupta
2344499379
piyush@usc.edu

Susanth Dahal
7553849598
dahal@usc.edu

Dishant Shahani
4798718230
dshahani@usc.edu

I. Introduction

A. *Description of the problem & proposed solution*

We, as humans, have a natural inclination towards opinions. While purchasing a new product or looking for a new restaurant, we always seek and consider others opinion. With the rapid growth of e-commerce and review websites, there is a huge size of review corpus the users are exposed to. Also the websites currently provide only an absolute rating for a product or service without describing its individual features. For a user to identify the positive and negative features of a product, one would have to do the arduous task of going over all the reviews. Our project aims to solve this problem by generating a concise sentiment based summary which would provide the users with a holistic idea of a product′s features.

In this project, given a large review corpus in French or German, we aim to identify the significant features and respective descriptors from a product′s reviews. Furthermore, we use a sentiment classifier to classify the pairs as positives or negatives of a product and finally use semantic analysis to display the pairs in the order of relevancy. This provides a user with a comprehensive summary regarding a products features which could prove helpful in making a buying decision.

B. *Why is it interesting & our contribution*

This project is interesting because nowadays as the amount of data available is enormous, the task of managing and interpreting the information is gaining significance. This process of sentiment based summarization of a data corpus can be used not only for product reviews but also for summarizing books, news articles, magazines etc. and in domains like feature based information retrival. Moreover we decided to design it to work for multilingual corpora(currently French & German). This would enable us to provide a clear summarization of a given subject or product irrespective of the data source language.

The task was technically challenging because while extracting the relevant feature-descriptor pairs, identifying only the nouns-adjective phrases using POS tagging was not sufficient. We had to perform dependency parsing to identify which noun a given adjective describes. Finally as a contribution to current work, most existing summarizers take only the frequency of features into account, but in our implementation we consider semantic similarity as well to identify relevancy. We also perform sentiment based classification which helps to classify the feature-descriptors pairs as positive or negative.

II. Data Source & Annotations

We used Webis-CLS-10 corpora available through The Bauhaus-University′s website. The Cross-Lingual Sentiment(CLS) dataset was first used in Prettenhofer and Stein (2010). It comprises of Amazon reviews for Books, Dvd′s and Music in four languages (French, German, Japanese, English). It consisted of around 70000 reviews in each language.We currently implemented a summarizer for French & German.

The data was available as a raw XML file. Each review was represented by an <item> element. Each <item> element has following child elements:

1) category: The product category.
2) rating: The number of stars (1-5).
3) url: The url of the webpage from which the review was fetched.
4) text: The review text.
5) title: The name of the product.
6) reviewer: The name of the author.
7) location: The location of the author (may be empty).
8) date: The date when the review was written.

The above structure was same for both the languages, French & German. As an input to our implementation we give a folder having reviews in both languages. We use a python language detection library to identify the language. Out of the above given tags, we used the content under the <Text> and <Title> tags for the purpose of this project. For a given specific <Title> we used a dictionary to aggregate all the reviews written by all reviewers into one list. This made a specific products reviews easily available to work with. Although the available review dataset was pre-annotated with a sentiment label classifying the whole review as positive or negative, it was not useful in our sentiment analysis task because there can be a case when a review consists of many positive features of a product but the overall sentiment of the review is negative. eg: ”It has a great camera but not worth the money”.

Our task was to identify feature-descriptor pairs and classify them as positives or negatives for a given product. For the purpose of evaluation, as the pre-annotated review sentiment was not helpful, we filtered out a subset of 2000 reviews, 1000 in french & 1000 in German. We translated them to english using Google translate. We then manually identified the positive and negative feature-descriptor pairs from the translated reviews. We use this as a reference solution and compare our output with this reference using the F1 score measure.

III. Procedure

A. Data Preprocessing

Initially as we were dealing with two languages (French & German), we used the python package **langdetect** to identify the language of a given review text. This helped us segregate the reviews based on language. We then need to identify the significant features of a given product from reviews. Most of the features of a particular product generally occur in the reviews as Nouns and Noun Phrases. The opinion about these features are adjectives that describe these features. We use **pattern.fr** and **pattern.de** python libraries developed by CLiPS(Computational Linguistics And Psycholinguistics) to perform POS tagging for french and German reviews respectively. This helped us to identify the Nouns and Adjectives in the sentence.

B. Dependency Parsing

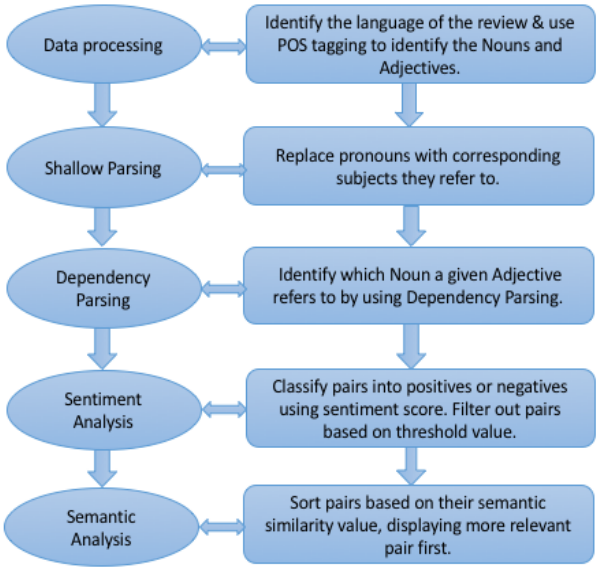
We noticed that in many reviews the product name was replaced by a pronoun. So in this part, we replaced the pronoun with the Noun it refers to. For this purpose we used the **parsetree** function available in pattern library. It generates the parse tree of a sentence and extract typed dependencies showing the grammatical relationships among the words. This helped to identify which noun a given pronoun is referring to. Furthermore, for a sentence having multiple feature-descriptor pairs, in order to identify which Noun a given Adjective modifies, we used the same dependency parsing using the the parsetree function of the pattern library.

C. Sentiment based classification & Filtering

Now that we have extracted the related feature-descriptor pairs, the next part is getting the sentiment of each pair of phrases.This would help us to classify the specific feature as positive or negative. For this task, we used the sentiment function of the pattern.fr and pattern.de library for french and german respectivelty. The function returns a value between [-1, 1] where values less than 0 represents negative sentiment and value more than 0 represent positive sentiment. In order to filter out less relevant pairs, we defined the threshold value of [-0.2, 0.2] and disregarded the pairs having sentiment value in this range.

D. Sorting features using Semantic Similarity

Finally having obtained the positive and negative features-descriptor pairs for a product, we need to sort them based on the relevancy of the subject. To do that we take into account their semantic similarity. For this task, we use a Java API called **DISCO** that is available as a precompiled jar and can be queried through the terminal. We have to give the string as a parameter along with the similarity measure to be used. We used the COSINE similarity measure as it provided most relevant results for the sample reviews. Having obtained the semantic similarity value for each pair, we use it as a measure to sort the positive and negative features in a descending order of their similarity value. These sorted pairs of phrases within positives and negatives are then presented to the user if form of a word cloud.



IV. Evaluation

As a final output we get feature-descriptor pairs classified into positives and negatives group and sorted according to relevance based on the semantic similarity value. In order to compare our results with the program output, we took a sample of 2000 reviews and manually created a summary in form of product features. For the sample dataset, we measure it′s performance by taking into account the following factors:

1) The number of features correctly identified by the model.
2) The correctness in terms of identifying the sentiment of a feature by the model.

A. Evaluation - Feature Identification

We have Used F-Score to measure the performance of our model in terms of feature detection. Let the number of correct features identified by the model be N_{cf} and the total number of features identified be N_{tf} . Also, let N_f be the correct number of features in the overall reviews of a product which we identified manually. Now we define the Precision of identified features, P_f as the ratio of the number of correct features identified by the model and the total number of features identified by the model.

$$P_f = \frac{N_{cf}}{N_{tf}}$$

We have defined the Recall of identified features, R_f as the ratio of the number of correct features identified by the model and the actual number of correct features in all the reviews for the product.(annotated manually).

$$R_f = \frac{N_{cf}}{N_f}$$

Hence the F-Score of identified features, F_s can hence be calculated as:

$$F_s = \frac{2P_f R_f}{P_f + R_f}$$

B. *Evaluation - Sentiment Analysis*

Let N_{tp} be the number of correctly identified positive features, N_{fp} be the number of negative features classified as positive, N_{tn} be the number of correctly identified negative features and N_{fn} be the number of positive features classified as negative. As in the previous section, we define the precision (P_p), recall (R_p) and F-Score (F_p) to be,

$$P_p = \frac{N_{tp}}{N_{tp} + N_{fp}}$$

$$R_p = \frac{N_{tp}}{N_{tp} + N_{fn}}$$

$$F_p = \frac{2P_pR_p}{P_p + R_p}$$

Similarly, for negatively classified subjects, we define the precision (P_n), recall (R_n) and F-score (F_n) to be,

$$P_n = \frac{N_{tn}}{N_{tn} + N_{fn}}$$

$$R_n = \frac{N_{tn}}{N_{tn} + N_{fp}}$$

$$F_n = \frac{2P_nR_n}{P_n + R_n}$$

V. Results

A. *Performance Measure*

In order to evaluate the performance, we have manually annotated data for both the languages (French & German). We used 2000 manually annotated reviews, 1000 in each language, and used it as a baseline to find the F-score for our implementation. The precision, recall and F-Score values for the French data set of Books, Dvd′s and Music combined are given as:

	Features	Positives	Negatives
Precision	0.75	0.77	0.75
Recall	0.73	0.76	0.72
F-Score	0.74	0.76	0.73

The precision, recall and F-Score values for the German data set of Books, Dvd′s and Music combined are given as:

	Features	Positives	Negatives
Precision	0.71	0.73	0.70
Recall	0.69	0.74	0.67
F-Score	0.69	0.74	0.68

Sample wordclouds showing positive and negative features of a specific DVD are shown below. Their corresponding english translation is also shown for reference.

Fig. 1. Sample wordcloud for Positive Reviews (French)

Fig. 2. Sample wordcloud for Positive Reviews (English)

Fig. 3. Sample wordcloud for Negative Reviews (French)

Fig. 4. Sample wordcloud for Negative Reviews (English)

B. *What it gets Right or Wrong?*

If a sentence has a simple structure and is not contextually dependent on the previous statements, our model performs fairly good. For example,

1) **The original text in French:** Ceci est un bon film avec des acteurs impressionnants.
2) **A word-by-word gloss:** This East a good movie with of the cast impressive.
3) **A translation into English:** This is a good movie with awesome actors.

For sentences having a negation about an opinion or the ones in which the opinion contextually depends on the previous statements, the implementation does not give a correct output. For the below given example, our implementation identifies feature as "very good movie" but this is not contextually correct as per the overall review.

1) **The original text in French:** Comme il tait la suggestion de mon ami, je pensais que ce film serait trs bon , mais il tait pas la hauteur de mes attentes.
2) **A word-by-word gloss:** As the was the suggestion of my friend , l thought this movie would be very good, but its was not at the height of my expectations.
3) **A translation into English:** As it was the suggestion of my friend , I thought this film would be very good, but it was not upto my expectations.

VI. Discussion

We have summarized the reviews in French and German by identifying the relevant feature-descriptor pairs and classifying them as positive or negative. Evaluating the results we saw that our summarizer was able to identify most of the positive and negative features of a product for which the reviews statements had a simple parse tree. With more semantic analysis and greater contextual reference we can make our summarizer to work properly for reviews having contradicting sentiments. This can be done by finding contrasting conjunctions in sentences and figuring out if the noun they are referring to is the feature being analyzed. Along with that, we can identify sentence relationships showing contradiction in opinion about the feature and remove the feature-descriptor pair from consideration. This will improve our summarizer significantly.

For future work, we plan to make it compatible for few other languages like Japanese and Hindi. This will help us to identify the positive and negative features of a product on a regional basis. As we know that an opinion may differ from person to person and from region to region, our summarizer can be of a great help while identifying which features are preferred in a place or a region and which are not. This could prove helpful for a manufacturer to decide which features to improve in the upcoming product versions in a geographically targeted manner. Furthermore, we also plan to extend our summarizer to be capable of summarizing journals and newspapers. In this way we can even categorize the trending news and extract the good and bad happenings on a regional basis. It can be an exciting experiment to see what is the good work being done by a government/organization in a region and the decisions with which the citizens/employees are not satisfied.

References

- Minqing Hu and Bing Liu. "Mining and summarizing customer reviews". Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004, full paper), Seattle, Washington, USA, Aug 22-25, 2004.
- Peter Prettenhofer and Benno Stein. "Cross-Language Text Classification using Structural Correspondence Learning". In 48th Annual Meeting of the Association of Computational Linguistics (ACL 10), pages 1118-1127, July 2010. Association for Computational Linguistics.
- Amazon review dataset: http://www.uni-weimar.de/en/media/chairs/webis/corpora/corpus-webis-cls-10/
- CLiPS pattern library for text analysis: http://www.clips.ua.ac.be/pattern
- DISCO API to retrieve the semantic similarity between arbitrary words and phrases. http://www.linguatools.de/disco/disco_en.html

Division of labour

- **Siddharth:** Sentiment & Semantic Analysis, Performance Evaluation, Data Annotation.
- **Piyush:** Dependency Parsing, Semantic Analysis, Performance Evaluation, Data Annotation.
- **Susanth:** Data collection & Preprocessing, Shallow Parsing, Data Annotation.
- **Dishant:** Data Preprocessing, POS Tagging, Frequent Itemset mining, Data Annotation.

Word Count: 2094