Joshua Pham (jhp2352), Dune Blum (djb3657)
LIN 350: Computational Semantics
Katrin Erk
11 October 2016
Initial Project Description

What are the main questions that you want to answer, the main language phenomena you want to address, or the main ideas you want to explore?

We essentially want to look at feature-specific sentiment analysis on forum post based reviews. Our plan is to first scrape data from forum posts in order to create a corpus. We will attempt to filter said data by trying to figure out the topic of the forum post (viz. is it a review?) and then removing stopwords and punctuation in order to create a set of words that fit our original goal: to attempt to figure out which features are well regarded, and which are not so much so. Once we have our filtered corpus, we plan to do some sort of topic modeling on the data in an attempt to relate feelings to features, and essentially map opinions about features to the features themselves.

Our first question is whether it is possible to organize information from forum posts using semantic parsing tools. Forum posts are very broad sets of data, and neither is it that every forum post is a review, nor is it that every part of a forum post that is considered a review is related to the review itself. We are going to try to see if we can filter this data to the point where it is obvious that the poster, at said point in the post, is giving an opinion on the product's features.

Next, we are going to see if we need to break down sentences within forum replies to determine the product features they describe. Is there a way we can relate a sentence to a certain feature? How big does our context window need to be in order to understand what a sentence is referring to? These are all questions we need to answer when building our corpus.

Further, we want to understand if it is possible to differentiate between a question about a feature and an opinion about a feature. Do we need to develop a way to analyze each sentence, or post, and see if said sentence/post is providing an opinion or is there an easier way to do it? Is there a keyword or tag on a specific forum we should be looking for that helps us pinpoint whether the user is providing an opinion on a feature, or merely asking a question? If we cannot distinguish between said posts, will our goal still be achievable?

Finally, will the corpus we create in the end be detailed enough in order for us to notice meaningful patterns inside said corpus? Will the general opinion about a feature be obvious, or will there be too much statistical noise to deduce any true trends? If we are unsuccessful in filtering out enough meaningless data through the previously discussed methods, then it may be impossible to draw any conclusions. Even if we do filter out enough of said meaningless data, will we have enough actual useful data to accurately map said opinions?

What distributional model will you use, or what kinds of rules are you planning to state? Be as detailed as you can. (Yes, I know you will not have worked out every detail at this point, but strive to work out as many as you can.)

Upon first glance, Latent Semantic Analysis seems to fit our needs well. We want our model to define features, and LSA reduces documents to low-dimensional spaces that could be useful for studying synonymy between words used to describe features. However, our past assignment, which used LDA, showed that topics might not be able to be described using a single word, but rather a collection. Topics that emerge from our product review analysis might be described using fewer words, since they describe certain features. This may make it easier to label topics as correct and complete.

Selectional preferences might also be useful. For example, assuming we are able to discern what the features are, we can then look at the adjectives used to describe each feature in order to understand what the user is saying about the feature. If a poster were to say "The sound quality is great," then the use of the word 'great' implies that he or she likes the feature.

If you do a distributional project, it is vital that you figure out as early as possible what data you can use to learn your model. Is there enough data? Is it freely available? Do you have to contact someone to get it?

We will not be analyzing a pre-assembled corpus, with uniform length, formality of language, or author. Instead, we will obtain our corpus by scraping product review forums. For headphones, the Head-Fi community has a lively forum with an abundance of expert reviews of products, opinionated responses, and discussions about headphone purchases. Similar communities exist for flashlights, with the Candle Power Forums, and for cameras, with the Digital Photography Review forums. Not having the uniformity of a pre-assembled corpus may raise challenges. We plan to address this by developing a general method for scraping, then choosing the forum for which these methods work best.

Since our objective is to group posts by their commentary on specific product features, we have to develop a way to first learn these features. Usually, the original post of a review thread breaks up its commentary by feature. We could train our model on a collection of these original posts in order to determine the topics that we will analyze in the replies. However, threads vary between product reviews and general discussion. We will look into taking advantage of "review" tags, subforums, and other categorical structures to ensure that our data is as clean as possible.

Gensim will be invaluable to us for obtaining distributional models from our corpora, but additional tools are required to build this corpus. Since individual posts will be relatively short compared to the bodies of text that we have been analyzing with Gensim in class, using NTLK to tag parts of speech and break down sentences might be helpful to make the most of the data. In order to scrape forums for their posts and replies, we intend to use a Python library called Beautiful Soup to parse HTML.