Semantic Change of Words Across Different Corpora

Victoria Huang
University of Pennsylvania
vichuang@seas.upenn.edu

1 Introduction

The meaning of a word or the way it is used can differ according to context. For example, words can have informal and formal meanings. The meaning being used depends on the level of formality of the context. For this project, we want to analyze how word semantics change according to the level of formality of the corpus and what the main differences in word meaning are. In particular, we are interested in how the formality of the platform affects word meaning. That is, we want to analyze how words shift meanings when used in informal (Twitter) versus formal (Google Newswire) environments. We describe the methods used in Section 2. We describe how to set up the conducted experiment in Section 3. We analyze the results and how the word semantics change across corpora in Section 4. We present our conclusions in Section 5.

2 Approach

The main issue with using two different word2vec models is that the word vectors for our two corpora live in different vector spaces, meaning that we can't directly compare the word vectors of the two corpora. Instead, we used the two following methods to compare the semantic features of the words across Google Newswire and Twitter.

2.1 K-Means Clustering

Our first approach was to use K-Means clustering with on both corpora separately. We can compare word-pairs to see if they end up in the same cluster for both corpora. If they do not, that indicates that there was a change in semantic features across the two corpora.

2.2 English-English Translation

Our next approach was to project the word vectors from one space to the other. Similarly to how words in one language can be mapped to its translation in another language via a translation matrix, we will map the words in the "Twitter language" onto words in the "Google Newswire" language. By using an English-English translation matrix, we can project the Twitter vectors into the same space as the Google Newswire vectors and map it onto the closest Google Newswire word vector. Therefore, if all word vectors now live in the same vector space, we can directly compare their cosine similarities.
The smaller the cosine similarity between two word vectors, the bigger the difference in the words' semantic features. So we are interested in the words that have low cosine similarities with their translations.

3 Experiment Setup

3.1 Dataset

The two different corpora used are: Google Newswire and Twitter. In particular, we used the pre-trained word2vec model for Google Newswire available at https://code.google.com/archive/p/word2vec/, and the pre-trained Twitter word2vec model available at https://www.fredericgodin.com/software/.

3.2 Projecting Twitter Vectors into Google Newswire Vector Space

Note that the goal of the project is to analyze the change of word semantics across different corpora. Therefore, we will only be projecting and comparing the vectors of words that exist in both corpora.
After this initial filtering, we will train an English-English translation matrix on the common vocab-

ulary. We exclude some words that we are interested in or are unsure of as to what their translation should be from the training process (i.e. 'hot', 'dope', 'cool', etc.). We can get an idea of the words that will have diverging semantic features across the different corpora from the K-Means clustering approach as described above. We don't want our model to fit on these examples as we don't know what their translation should be, so we can these words from training so that the model does not fit on them. For the rest of the words, we can label their translations as themselves. The assumption we make here is that almost all words have the same semantic meaning in both "Twitter English" and "Google Newswire English". While there will be noise (for words whose semantic features change across these two corpora), it should be insignificant and should not have much impact on the translations / projections.

Once the translation matrix is trained, we can project all the word vectors in the Twitter model into the Google Newswire space. For each <original word, translation> pair, we compute the cosine similarity.

4 Results

We ranked and sorted all of the computed <original word, translation> word-pairs by their cosine similarity scores. For the word-pairs with diverging semantic features, we observed several patterns.

4.1 Google Newswire Jargon

Google Newswire is more formal and employs some jargon terms. Consider the following word-pairs:

Twitter	Translation	Cosine Similarity
midterm	exam	0.25396731652290466
chip	flake	0.26952469113041666
gross	disgusting	0.27578299487161095

Twitter is a social media platform and is less likely to contain technical terms then Google Newswire. Therefore, a word in Google Newswire may carry additional jargon / more technical meanings compared to the Twitter version of the word. For example, a 'chip' in Twitter refers to a snack, while in Google Newswire, it can also refer to a microprocessor chip. While you can see news articles on microprocessor chips, it is unlikely that Twitter users will post about them. Twitter is a platform where its users mainly post updates on their everyday lives, which generally does not require the use of technical terms.

4.2 Popular Culture and Informal Language

Overwhelmingly, however, the changes in semantics are due to the use of informal language or to references to popular culture (or a combination of both) on Twitter. Informal language is frequently used on Twitter. It is not uncommon to see words written in non-standard ways. For example, the first letter of proper nouns is often not capitalized.

Twitter	Translation	Cosine Similarity
keystone	beer	0.009829291967978692
conjuring	movie	0.060863178876562125
divergent	Watchmen	0.07687172191081817
sonic	ice_cream	0.0981947546533584
anchorman	movie	0.11500306277711157
converse	jeans	0.12478039515327173
vans	jeans	0.1444160812833883
jets	Jets	0.2090948263253235
patriots	Pats	0.2575398724404167
eagles	Eagles	0.26790506816277715
titanic	Titanic	0.30913109928491067
nirvana	Nirvana	0.34803883577795375

As seen in the table above, names of movies, sports teams, bands, and brands have improper capitalization on Twitter. These names also happen to be common nouns. Since the proper nouns have a different meaning than the common nouns, the improperly-cased words in Twitter gain an additional meaning. Since words generally have the correct capitalization in Google Newswire, the common nouns do not carry the semantic features of the proper nouns. This is not the case for Twitter, which creates a divergence in word semantics. Likewise, Twitter contains abbreviations that are also real English words with alternate meanings:

Twitter	Translation	Cosine Similarity
apt	house	0.06170169034192631
tics	tickets	0.08039011235548371
bud	fellas	0.1911652810106025
nips	nipples	0.19959199939057065
tat	tattoo	0.19988062952034433
cause	anyway	0.21251091330647834
gorge	gorgeous	0.22521648429516022

Since words in Google Newswire are more likely

to be fully spelled out, they will not have the semantic features of the words they can be abbreviations for. For example, 'apt' is short for 'apartment' on Twitter (which is semantically close to 'house'), but 'apt' is used as an adjective in Google Newswire.
Words also tend to take on their slang meaning on Twitter and their more formal meaning in Google Newswire:

Twitter	Translation	Cosine Similarity
ratchet	trashy	0.05511561200786945
chilling	lounging	0.14716498679894663
dope	awesome	0.17426874577229667
Blunt	Crack	0.18268346552441278
Curves	Sexy	0.23456158420126624
hoe	bitch	0.27551218031664076
crib	house	0.2861536968167244
weed	booze	0.30338048160816644
Chick	Girl	0.30469176907734163
buff	sexy	0.3081253347936383
sketchy	weird	0.32090988531885906
hot	sexy	0.3644036397889235
trolls	haters	0.45507806734231077
chill	cool	0.48756233547397204
shady	sleazy	0.583640349396314

Twitter is mainly used by a younger demographic, which uses a lot of slang and informal language on a daily basis. On the other hand, Google Newswire uses a more formal language, so its words would not contain these semantic features.

5 Conclusion

We can treat "Twitter English" and "Google Newswire English" as two distinct languages. The differences in the formality of the two corpora lead to differences in word semantics. Twitter contains a lot of informal language, and its words can be written in non-standard ways. This creates ambiguity and adds meanings to the improperly-written words. Even if a word is correctly spelled on Twitter, its semantics can diverge from the same word in Google Newswire. We are more likely to see slang and popular culture references on Twitter than on Google Newswire. Google Newswire also uses a more formal and standard English language, so its words lack these additional meanings. Therefore, its words lack the informal semantic features of the Twitter ones. On the other hand, Google Newswire words can have additional jargon or more technical meanings that aren't present in the Twitter words. All of these differences lead to semantic differences between the two corpora.