[Insert Title Here]: Compound word classifier

Main idea: Creating a compound word classifier.

-Given an unseen compound word, our model tries to predict whether a compound word is real or not.

Research and Brainstorming:

Our ideas for predicting compound words using its constituents were inspired by this paper.

Reference research paper: https://www.aclweb.org/anthology/W19-5105.pdf

(Published in Aug 2, 2019)

-I can give you a summary of it later (what the paper's about, and what ideas we grabbed from it – i.e. predicting valid compound words, creation of corrupt compound words, part of their dataset)

Data Source and Dataset Generation:

Selecting our real compound words:

We decided to use the LADEC word dataset to generate our English compound word dataset.

Background:

The Large Database of English Compounds (LADEC) consists of over 8,000 English words that can be parsed into two constituents that are free morphemes, making it the largest existing database specifically for use in research on compound words. Both monomorphemic (e.g., *wheel*) and multimorphemic (e.g., *teacher*) constituents were used. The items were selected from a range of sources, including CELEX, the English Lexicon Project, the British Lexicon Project, the British National Corpus, and Wordnet, and were hand-coded as compounds (e.g., *snowball*).

Other interesting attributes include ratings of how predictable a compound's meaning is from its parts and linguistic characteristics (i.e. frequency, family size, and bigram frequency).

LADEC Research article and Database Download page:

https://www.ncbi.nlm.nih.gov/pubmed/31347038

https://link.springer.com/article/10.3758/s13428-019-01282-6#Sec2

LADEC Database Download page:
https://era.library.ualberta.ca/search?search=LADEC&tab=item

From the LADEC dataset, we further specified to only use the compound words that came from the English Lexicon Project (potentially dismissing words from other sources, including CELEX (from Germany), the British Lexicon Project, the British National Corpus, and Wordnet (from Princeton); we wanted to focus on compound words used in the US). This is identified by values of 1 in the "inELP" column of our dataset.

We also selected the compound words that are common, according to the list of common words identified by Mathematica's built-in WordData package. This is identified by a "1" in the "isCommonstim" column of the dataset.

(for reference, the database labels can be found in the "Appendix", pg 24-25)

Slide information:

To facilitate the selection of items that are in common usage or contain constituents that are in common usage, we included indicator variables denoting whether the compound and first and second constituents appear on a list of 40,127 common English words. This list is part of Mathematica's built-in WordData package. Of the full set of 8,961 items (including incorrectly parsed items), 3,664 appear in the list of common words, 8,310 have first constituents appearing in the list of common words, and 6,599 have second constituents appearing in the list of common words. In terms of unique items (i.e., counting each parse only once, rather than counting all possible parses), 3,414 compounds appear on the list of common words.

After carefully selecting our dataset, we found that we had

700 uncommon compound words

2449 common compound words

"Real" Compound Word Dataset: 3149 compound words in the English corpus belonging to the English Lexicon Project dataset, which were chosen by creators of the LADEC database.

According to Mathematica's WordData package, 700 of these words are uncommon and 2449 of these are common.

Just a side-note quote (optional):

“Importantly, our algorithm and subsequent coding identified 4,605 compounds that were not present in either the ELP or BLP corpora, and thus extends the number of compounds available to researchers. Furthermore, our list of items indicates that ELP and BLP contain largely nonoverlapping sets of compounds.”

Generating our “corrupt” compound words:

So, previously we have created a set of existing valid compound words in the English language. We must also provide examples of non-existing, invalid compound words in the English language.

For every word, we randomly replaced constituents of existing compound words with corresponding constituents of other existing compound words. We made a set of constituent 1 words, a set of constituent 2 words, and took the Cartesian product of these two sets to generate a very large set of new invalid, non-existing compound words.

This set of “corrupt” compound words consists of 1363417 words in total. We also verified that these compound words do not exist by checking if it is listed in the English Lexicon project dataset. (Disclaimer: We want to check for compound words used in the US, not used in the English language worldwide.)

Our entire dataset now consists of 3149 real compound words in the English corpus (700 of which are uncommon and 2449 of which are common) and 1363417 corrupt compound words, generated by replacing constituents of real compound words with corresponding constituents of other real compound words.

Then we took a random sample of the entire dataset in order to generate our test dataset; the rest we used for our training dataset. We performed a 70/30 split.

Other steps to do:

Training the model

Getting model accuracy by running on test datasets

Repeat training and testing step on various other classifier models

Then finalize our results and report our findings.

Potential Features to use for our model:

-if c1 is common AND c2 is common - probably indicative that the compound word is valid.

- …

Citations slide to generate:

Inspiration Research paper: https://www.aclweb.org/anthology/W19-5105.pdf

LADEC Research article: https://www.ncbi.nlm.nih.gov/pubmed/31347038 OR https://link.springer.com/article/10.3758/s13428-019-01282-6#Sec2

Finished:

Formulated our entire project idea, from start to finish (J D)

Preprocessed the data (J)

Steps to do:

Remake the PPT slides (T)

Train the model and test it (D)

Create other classifier models to compare accuracy (D)

Finalize our results and finish slides on Saturday 3-5pm (J D T)

Send the email w explanation about changed project idea, late apology, and our finished project (ipynb notebook and slides) (T D)