Part of Speech and Wordnet

Until now we have considered only bag of words implementations. We have been basing our results on counts or the presence of unique words as features, but we haven't incorporated any information about the meaning of those words (except for removing certain 'stopwords). 'Semantic' methods take different types of meaning into account.

These methods are rarely used by political scientists and they represent tremendous opportunity, if a researcher can demonstrate that incorporating semantics adds significant value to an analysis. The exception is international events data research (discussed on Monday – see the work of Phil Schrodt and more recently Mike Ward). As discussed, they take subject-verb-object relationships in news stories.

This homework just provides a few examples.

For example, using **POS tagging**, you can refine the features of an analysis by: combining words, excluding certain types of words, focusing attention on certain types of words etc.

- Bigrams: not, happy vs. not happy
- Proper nouns (NNP): Find and extract just the names cited in documents
- Verbs and adjectives (VB, VBZ etc): Focus attention on the words most likely to capture sentiment
- Pronouns (PRP, PRPS): Study gender references, or a politician's emphasis on me versus us versus them

The options are endless: https://sites.google.com/site/partofspeechhelp/home/prp_prp

Wordnet relates words based on their meaning. This information can be used (for example) to identify similar words in order to treat them as a similar feature, and to learn about different usages that might prove problematic for an analysis. The example is 'dog.' There are lots of kinds of dogs. In addition, someone might use dog in a sentence to refer to something other than a canine.

Where does this stuff come from? Before machine learning, people thought that the path to artificial intelligence was to create a complete mapping of language. Tons of money went into this effort, but it was not very successful. Machine learning revived the quest and today, NLP (natural language processing) methods effectively incorporate the things developed in the first phase with machine learning.