Instructions for COLING-2014 Proceedings

First Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Second Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Abstract

In this paper we present a feasibility study for rewriting multiword expressions as single words, which NLP systems could potentially process more easily than the original phrases. Here we investigate PPDB: The Paraphrase Database to get a mapping from multiword expressions onto single words, using the MWE categorization system as described in Baldwin, et al.

1 Introduction

Multiword expressions (MWEs) are phrases whose meanings are different than the literal interpretation of the words in the phrase. MWEs include verb-particle constructions, fixed expressions, compound nominals, and decomposable idioms, to name a few.

MWEs are difficult for non-native speakers of English to understand, and also for NLP systems to identify.

2 Experimental Design

The Penn Paraphrases Database (PPDB) contains English paraphrases. We have characterized a subset of the paraphrases found in the PPDB, according to categories of multi-word expressions (MWEs), syntactic changes in the expansion from a word to its paraphrase, and what parts of speech appear in the corpus. We also looked at how many of the paraphrases in the PPDB appear to be spurious.

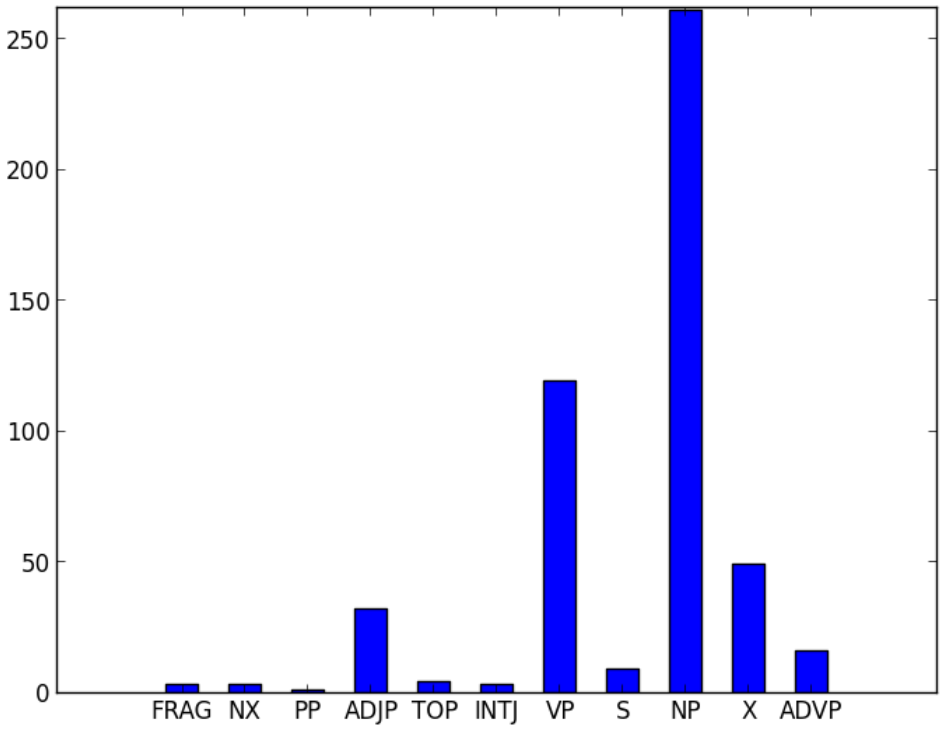
The categories of MWEs we looked at were light verbs, verb-particle constructions, negation, and superlatives. We also included Tim Baldwin's categories for MWEs: fixed expressions, non-decomposable idioms, compound nominals, proper names, and decomposable idioms.

In addition to MWE categories, we also included categories for syntactic changes from a word to its paraphrase: change of tense followed by a paraphrase, nominalizations, infinitival to, adverbial modifier, one or more words the same as part of the original word, determiner followed by a one-word paraphrase, determiner followed by the plural form, and change of tense. Finally, we included acronyms, hypernym-hyponym pairs, times, extra punctuation marks, and numbers as categories, as well as unspecified expansions and bad paraphrases.

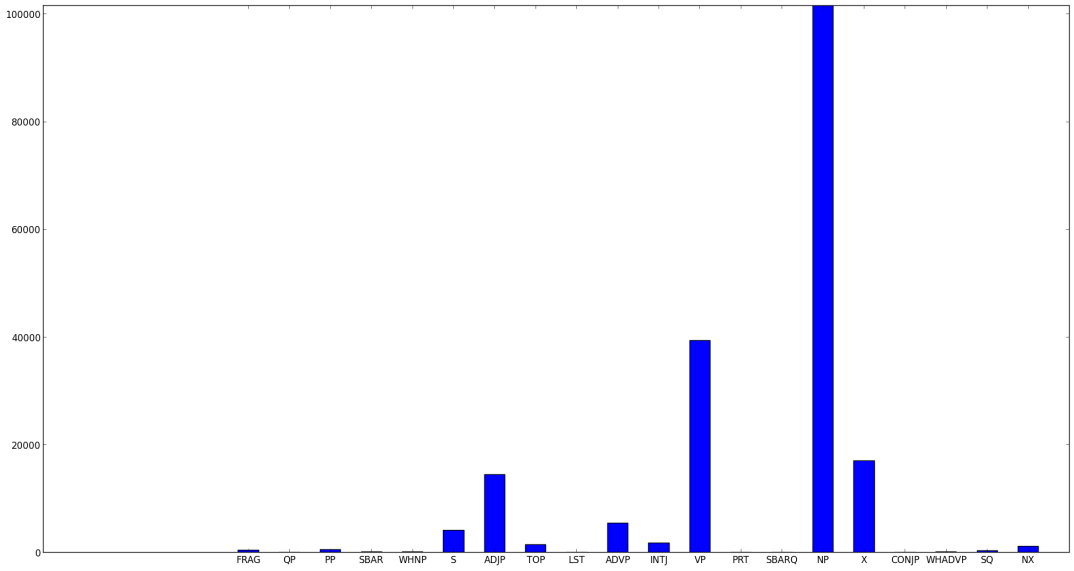
3 Results

Of a random sample of 500 paraphrases from the L one-to-many paraphrase file, the most common types of paraphrase were expansions using the same morphological form (117 instances, or 23.4

The distribution of the parts of speech from this random sample is depicted in the histogram below.



The distribution of all of the parts of the speech from the L one-to-many paraphrase file is depicted in the following histogram:



In both samples, the most common part of speech is NP, followed by VP.

Below are illustrative examples of all categories of multiword expressions:

In addition to categorizing a random sample of paraphrases, I searched for instances of light verbs, verb-particle constructions, negation, comparatives, and superlatives. The light verbs were those with the verb have, take, make, hold or give, followed by a noun phrase. The verb-particle constructions were

MWE Category	Example
verb-particle	torched, burnt down
fixed expression	applied, put into effect
non-decomposable idiom	furious, as mad as hell
proper noun	markov, mr markov
decomposable idiom	nuts, out of your mind
light verb	issued, made available
superlative	notably, most particularly
negation	unused, not utilized

Table 1: Examples from each category.

any verbs followed by the particles down, up, on, out, over or upon. Negation instances had the word not either in the original or the expanded paraphrase. Comparatives had the word more, and superlatives had the word most.

The results from these searches are summarized in the table below.

4 Analysis

Acknowledgements

The acknowledgements should go immediately before the references. Do not number the acknowledgements section. Do not include this section when submitting your paper for review.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Association for Computing Machinery. 1983. *Computing Reviews*, 24(11):503–512.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.