A Mechanism for Analyzing Compounds in the Penn Treebank

Mary Dungan, Sandiway Fong, Josh Harrison
University of Arizona
sandiway@email.arizona.edu

Noun compounds require both linguistic and world knowledge to interpret and parse correctly according to their modifier-modifiee relationships. Automatic parsing systems have only a narrow amount of information available to them. Given that the interpretation of a compound is based upon its syntactic structure, deciding the correct parse goes a long way in deciding a compound's meaning. Our system uses a variety of information sources, including corpus statistics, syntactic information, a word list, and a list of proper names. Compound types are analyzed in terms of Di Sciullo's (2005) F-tree proposal.

The Penn Treebank (Marcus et al. 1993) is a large-scale syntactically annotated corpus that is widely used as training data for stochastic parsing systems. Although detailed syntactic information is supplied for all major phrases, this does not extend to noun compounds. In the Penn Treebank, compounds are neither properly identified or analyzed. The goal of our system is to extend the Penn Treebank by automatically identifying, categorizing and parsing all noun compounds in the corpus. It is anticipated that the resulting augmented corpus will be of considerable added value as a training data set.

The first step for our parser is to identify compounds based on syntactic information, e.g. noun phrases containing a conjunction or multiple nouns. The system then eliminates syntactically uninteresting compounds such as proper names based on a dictionary of names collected from LSA and government census sources. Compound words are split up into their constituents based upon a dictionary. Then to analyze the compounds correctly, the parser must decide whether the compound is right or left branching. At this point in the system, the parser has no syntactic or lexical information to make its decision, so it calculates statistical information about the compounds constituents.

An example of a compound that must be parsed with either left or right branching is "wooden soup bowl." While it is obvious to an English speaker that "wooden" modifies "bowl" rather than "soup," an automatic parser does not have such knowledge immediately available to it. Rather than default to either right or left branching of such phrases, we propose that a modification of the techniques presented in Lauer (1995) will result in more accurate parses.

Lauer's (1995) dependency model can be contrasted with the adjacency model as outlined by Pustejovsky *et al* (1993). We tested both adjacency and dependency models, using bigrams derived in two ways from the Wall Street Journal section of the Penn Treebank. Potential compound NPs were extracted and used as the basis for one set of bigrams, while the other set was pulled from the whole sentence. Bigrams were calculated from each data set with each of the following methods, using a window of two, using a window of four, and using a weighted window of four. While adjacency models mark quantitatively more compounds as right bracketed than the corresponding dependency models in tests, the accuracy of those marked compounds is not as high as those marked using the dependency model.

Lauer, Mark. 1995. Corpus statistics meet the noun compound: some empirical results. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*. 47-54.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19:313—330.

Pustejovsky, J., Bergler, S. and Anick, P. 1993. Lexical Semantic Techniques for Corpus Analysis. In *Computational Linguistics* Vol 19(2), Special Issue on Using Large Corpora II, pp331-58.

Di Sciullo, Anna Maria. 2005. Decomposing Compounds. *Skase Journal of Theoretical Linguistic*s 2 (2005), 14–33.