Edmond Lau May 12, 2004
6.xxx – Human Intelligence Enterprise Term Project – Executive Summary

Learning Lexical Clusters in Children's Books

Children demonstrate an amazing ability to learn the meanings and correct usages of many new vocabulary words every day. Intuitively, from hearing a phrase such as "Cinderella's glass slippers," a child may also strengthen her confidence to use a related phrase such as "Barbie's plastic slippers." If we are to understand how human beings acquire and use knowledge about language, we need to examine the role of statistical regularity in language learning; in particular, we need to determine whether the statistical frequency with which a child experiences certain word patterns also contributes significantly to her ability to learn language. As a first step toward tackling this problem, I proposed the concept of *lexical clusters* – collections of words with statistically similar linkage structures with other words – and implemented a clustering algorithm in Java to explore the role that lexical clusters may play in language learning.

The core of my project centered around the integration of Deniz Yuret and Steven Larson's two powerful ideas for exploiting statistical regularity. Yuret demonstrated that lexical attraction models of language can correctly identify word links by simply using the likelihood of pairwise relations between words; Larson demonstrated that a system can develop class definitions by clustering together statistically similar information vectors. Combining these two ideas, I proposed the concept of *lexical clusters* as a mechanism for defining word classes based on subsymbolic descriptions acquired by a lexical attraction parser. For example, from parsing the phrase "the city mouse and the country mouse," the words *city* and *country* may form a lexical cluster because they both modify the word *mouse* in the same manner.

To illustrate the concept of lexical clusters, I designed and implemented an algorithm for comparing word similarity using the frequency tables constructed by Yuret's lexical attraction parser. The algorithm associates each of N unique parsed words to a distinct number from 1 to N. For each word, the algorithm then builds a length-N linkage vector, in which the jth term of the vector denotes the number of links that the parser has ever assigned to words i and j. Two words belong to the same lexical cluster if they link to all other words in the similar proportions; in vector language, the similarity between each pair of words is determined to be the cosine of the angle between the two words' vectors and assumes a value ranging from 0 (not similar) to 1 (very similar).

Using the clustering algorithm, I conducted an experiment in which I trained a lexical attraction parser on ten children's short stories and fables, including *Cinderella*, *Jack and the Beanstalk*, and six chapters of *Alice in Wonderland*, for a total of 20,663 words; I then used the clustering algorithm to build lexical clusters based on the linkage frequencies accumulated during training. Though imperfect, the system identified, among other less inspirational results, the following lexical clusters:

- country, trap, little, city
- advised, forgave, addressed, embraced, swept, desired, taught, asked, laughing
- hasted, run, hurried
- knight, gentleman, lived, person

Because the majority of the lexical clustering results contained unrelated words, however, I concluded that insufficient statistical regularity exists in children's books to offer a satisfactory explanation how children can learn to use language correctly with relative ease. By analyzing these results, I was surprised to discover that the direction of word links, which might not have mattered to Yuret in determining the dependency structures of sentences, plays a significant role in probabilistic experiments that deal with word semantics. In particular, the direction of the word link determines which word functions as the modifier and which word functions as the modified, such as in the phrases *city mouse* versus *mouse trap*.

Using the clustering algorithm developed for the experiment, I also explored a possible way in which lexical clusters may help to bootstrap the language learning process. In particular, one potentially interesting experiment would be to run Yuret's lexical attraction parser on some fraction of the Penn Treebank corpus, execute the clustering algorithm, and then for each subsequent word link parsed, increase the link frequency of the words associated with the word pair in addition to the word pair itself. The clustering algorithm may bootstrap the learning process and increase the resulting accuracy.