Lexsub Experiments

Diana McCarthy and Roberto Navigli

University of Sussex, University of Rome "La Sapienza"

dianam@sussex.ac.uk, navigli@di.uniroma1.it

July 3rd 2007

This is a brief document which explains some of the outcomes of our lexical substitution experiment. There is a more detailed paper at:

ftp://ftp.informatics.susx.ac.uk/pub/users/dianam/semevaltask10.pdf

and a presentation at

http://www.informatics.sussex.ac.uk/research/nlp/mccarthy/files/McCarthyNodalida07.pdf

The Data:

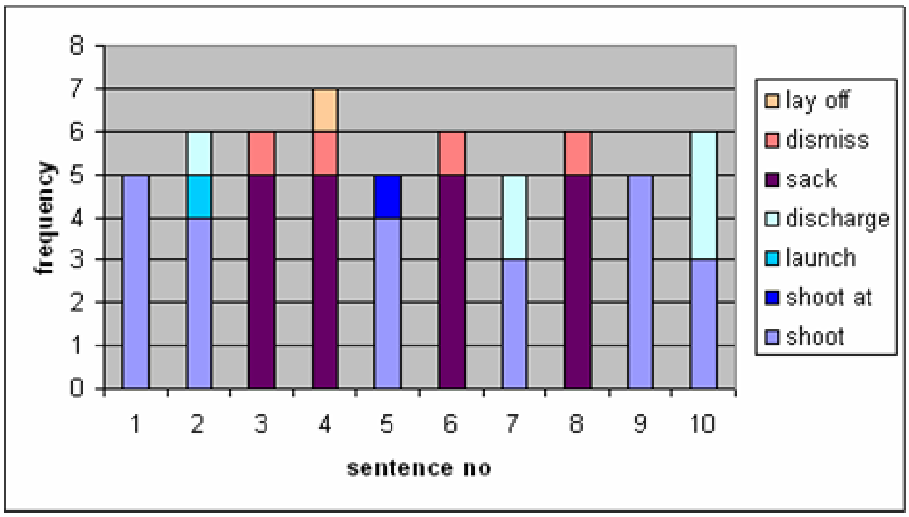
The experiment involved 5 annotators (people) finding words or phrases that mean the same as target words in the context of a sentence. For example, given the sentence:

"The ideal preparation would be a light meal about 2-2 1/2 hours pre-match , followed by a warm-up hit and perhaps a top-up with extra fluid before the ***match****."*

An annotator might replace the second "***match***" with "***game***".

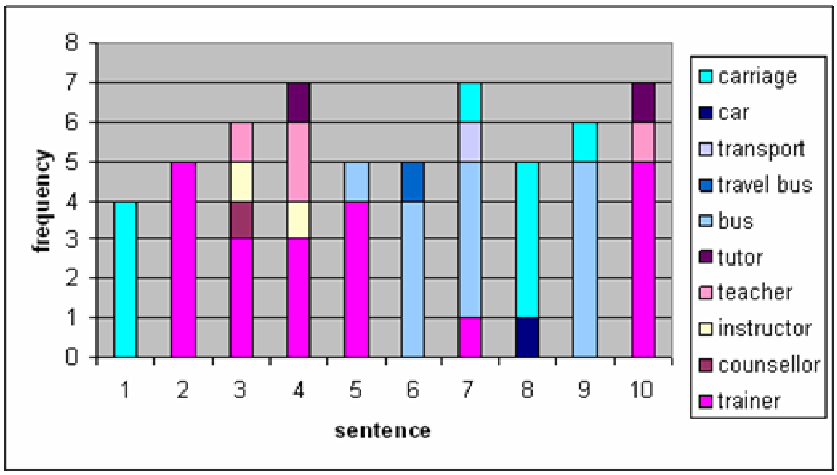
The sentences used were obtained from sampling data from the World Wide Web. We used 10 sentences for every word and a total of 201 words. Annotators were allowed to provide more than one substitute (up to 3) if they all fitted the context equally well and they were also allowed to provide a NIL response if they could not think of a reasonable substitute. Below we provide graphs displaying the substitutes for a small sample of our words:

1) the verb "*fire*"



Notice that the substitutes tend to fall into two broad meanings "*sack*" and "*shoot*" but that there is further variation because you can "*shoot a gun*" or "*shoot a person*" and some substitutes won't work for the latter. The colours were chosen by us to highlight relationships between the substitutes.

2) The noun "*coach*"

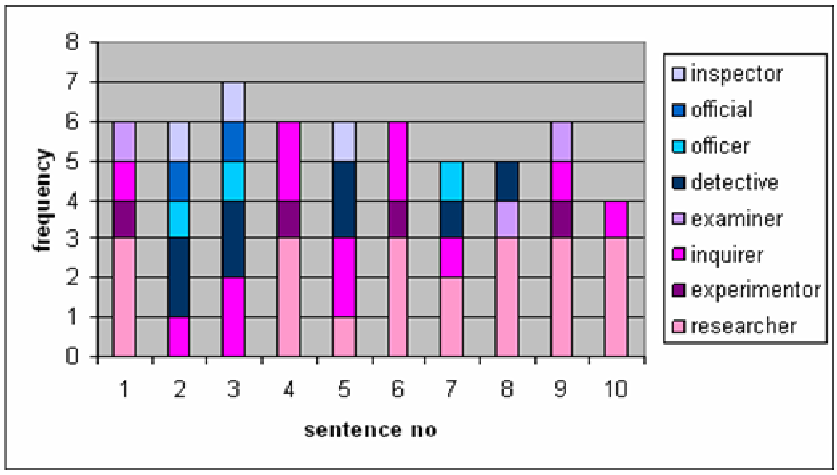


In this example we see that whilst there are two main meanings related to "*bus*" and "*instructor*" there is some cross over at sentences 5 and 7. This occurs because the context of one sentence is not enough to be sure of the meaning. For example sentence 5 is

"*The Championship by-law states that the SA will pay 100 % of travel, accommodation , uniform for* ***coaches*** *and airfare too ?*"

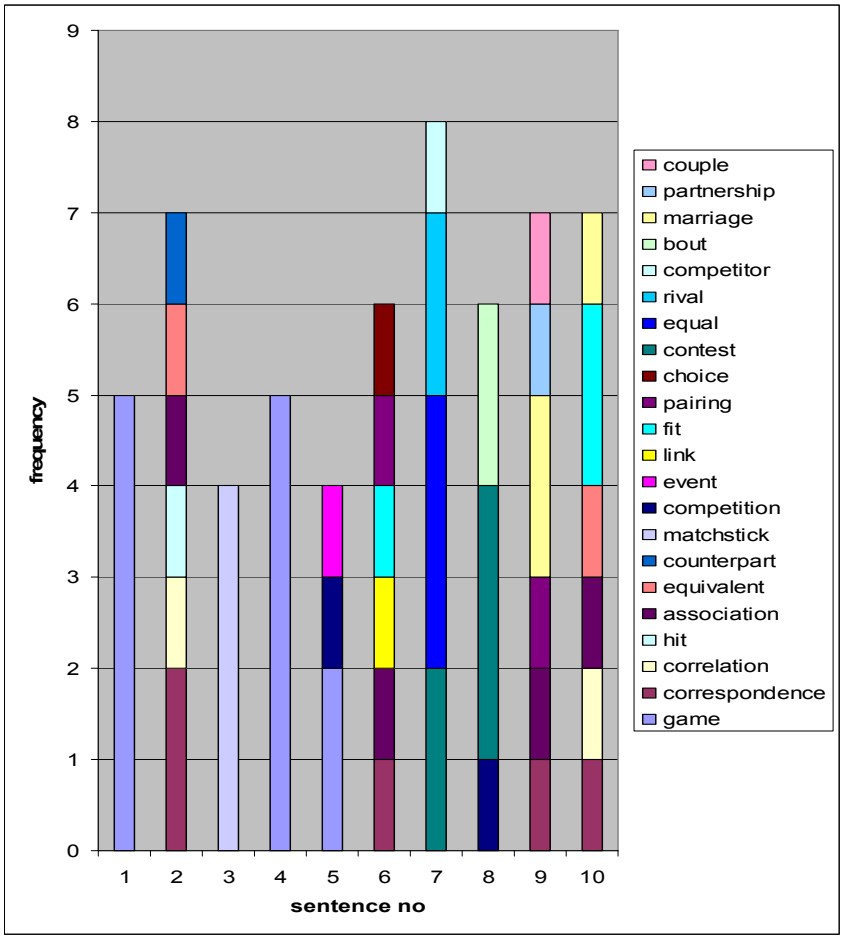
So whilst the majority verdict is that the meaning is "*trainer*", one person thought "*bus*" might also fit; that is the uniform might be required for wearing on the bus.

3) The noun "*investigator*"



Here the meanings of the target word (*investigator*), substitutes and contexts merge.

4) The noun "*match*"



N.B. Here we haven't changed the automatically generated colour scheme. This example shows that some words have many substitutes and the context dependencies of the substitutes show us that there are many meanings of the word. There are also relationships between the different meanings made evident by the partial overlap of substitutes for different sentences.

Experiments:

Whilst this data is useful for investigating word meanings, our primary goal is to evaluate computer systems that might be able to understand and summarise human language and answer questions using the vast amount of information available on the world wide web in a variety of languages.

Our data was used for an International Competition in which research teams tried to get computers to do the same task of selecting substitutes. Details of the competition are at:

http://www.informatics.sussex.ac.uk/research/groups/nlp/mccarthy/task10index.html

We used various measures to check there was a reasonable amount of overlap in the annotators' responses and then various scores to see how well the systems do:

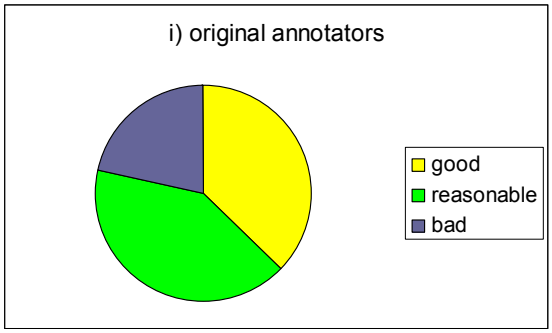
i) finding the best substitute
ii) finding substitutes when they can have 10 attempts
iii) finding when the candidate word is part of a phrase which has a special meaning

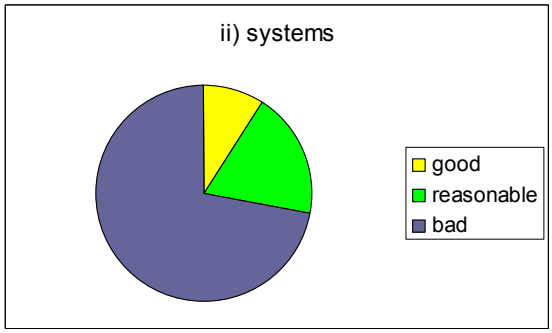
The competition culminated in a meeting in Prague (23rd/24th June 2007) to discuss results. Most systems used on-line thesauruses created by humans to identify candidate substitutes (synonyms) and then exploited a large sample of language from the web to make predictions of the most likely synonyms in a given sentence. There were some other more sophisticated techniques and we still have further analysis to work out which approach works best in what circumstances.

Whilst the participating systems all used electronic dictionaries built by humans along with automatic methods of deciding which word should go in which context, we are interested in possibilities for systems that learn the candidate substitutes automatically. They can do this by comparing the contexts of words in a large sample of text, for example they might discover that "*sack*" can mean the same thing as "*fire*" because these words both co-occur with words such as "*boss*" "*company*" and "*employee*". They might also find that "*fire*" bears some resemblance to "*shoot*" because of contexts such as "*gun*" "*intruder*" "*bullet*".
We have tried some of these automatic methods of finding synonyms and they work nearly as well on the lexsub data as using one of the manually produced thesauruses. We hope to combine an automatic method of finding the synonyms with an automatic method of finding the right synonym in the right context using web data.

The decision on what is a good substitute is not a cut and dried one and there is naturally a good deal of variation, just as there is a lot of variation in the ways that people express themselves. As part of our experiments we had to be sure that the substitutes provided by our annotators were a useful 'gold standard'. Humans often cannot think of all possibilities and we wanted to be sure that on the whole, the human substitutes were better than the system responses. We conducted a further experiment on 100 of the sentences where there were more than 2 answers provided. We mixed all the responses from both the original annotators and the participating systems and then asked 3 new annotators to grade all responses as good, reasonable or bad. In the pie charts below we show the proportion of responses from i) the 5

original annotators and ii) computer systems for each grade (good, reasonable or bad) where the grade was that selected by the majority vote of the 3 annotators.





This is just the beginning. We have a great deal more we wish to do on the analysis of our data. There is also the possibility that we want to annotate more data. If you are interested in hearing more about our work, or annotating some more data please do get in touch with us. We need annotators who are over the age of 25 and are native English speakers.