On Ambiguity in Internet Searches

Gordana Ilic Holen, Janne von Koss Torkildsen and Janne Bondi Johannessen

Tekstlaboratoriet, University of Oslo
P.b. 1102 Blindern
N-0317 Oslo
Norway
jannebj@ilf.uio.no

1 Introduction

1. 1 Prospect

The aim of this project is to decide to what extent the results of Internet searches contain irrelevant information because of ambiguous search words. If the amount of irrelevant information is great, users will be discouraged from employing the Internet for information retrieval. There are several different types of ambiguity. Some of these seem easy to filter out. Above all, we are interested in finding ambiguity which coincides with membership in different
grammatical categories, because this type of ambiguity could easily be reduced by using a tagger for grammatical disambiguation.

1.2 About the project

We have studied log files containing the search words in Fast's search engine and reconstructed the searches for the most frequent words. If a search word were ambiguous, we sorted its meanings according to several different criteria. In deciding whether a word was ambiguous or not, we employed a number of different sources, among these encyclopaedias and dictionaries. The results of this inquiry show that almost one fourth of the most frequent search words were ambiguous, and in about 90% of these cases there was a correlation between meaning and grammatical category. (This number means that whenever a search word was multiply ambiguous, at least two grammatical categories could be found to match at least two of the different meanings.)

The paper is organized as follows: Section 2 explains the research methods and the material we used. Section 3 contains examples. In section 4 we systematize the results by providing some statistics, and in section 5 we present some concrete proposals which all make use of tagging. Section 6 is the conclusion.

2.0 Methods and research material

2.1 Period picture

The project took place in the summer of 2000, employed two students and lasted two months, thus constituting four months of labor.

2.2 The log files of search words

The research material consists of about 900 000 search words form Fast Search and Transfer's log files coming from several different search engines. We sorted the material by frequency, and grouped together identical word forms. From the sorted word list, we picked the 5500 most frequent words for further investigation. In this way we made sure to work on only very common search words.

Even if the material contained no less than 5500 search words, the number of lexemes is much smaller. There are two reasons for this: a) the files distinguish between capital and non-capital letters (e.g. Liv and liv), b) in several cases many different spellings are used for the same word (e.g. pokemon, Pokemon, pokèmon, pokémon, Pokèmon)

We have not gathered together all the words that belong to the same lexeme, but simply used the word forms as they appeared in the log files.

Some users have searched for more than one word at a time. We chose to split up all these complex searches. The result is that function words such as *i* 'in', *på* 'on' appear as search words in our files, even though the chances are that no one actually searched only for these words. In any case, this choice does not seem to have affected the general result.

The most frequent content words were *sex* and *chat* having 52 000 and 36 000 searches respectively. (We keep the truly most frequent function words like *i* ('in') and *på* ('on') out of the discussion.) The least frequent of the 5500 search words were *hyttetomter* and *a.s*, 275 and 210 searches respectively.

2.3 Information about the search words

A lot of the words in the Norwegian language are ambiguous, and the task of seeing all the different meanings a word can have is not an easy one. Usually the ambiguous words appear in a grammatical context which rules out most of the possible interpretations and in a pragmatic context which gives a clue as to which meanings are interesting to consider.
Take the sentence *Per var høy* 'Per was tall'. Each of the words has at least two meanings:

(1)

```
"<Per>"
        "*per" subst mask prop          (egennavn)
        "per" prep                      (= pr.)
 "<var>"
        "var" adj pos m/f ub ent        (= følsom)
        "var" subst nøyt appell ent ub  (= et putevar)
        "var" subst nøyt appell fl ub   (= flere putevar)
        "vare" verb imp                 (= ikke slutt!)
        "være" verb pret                (= hadde egenskapen)
 "<høy>"
        "høy" adj pos m/f ub ent        (= ikke lav)
        "høy" subst nøyt appell ent ub  (= gress)
```

As language users, we are usually unaware of such ambiguities, and therefore we do not think about possible ambiguity in the search criterions we use in Internet searching. The students working on this project were more aware of such problems than the average user, and did also use dictionaries and encyclopedia.

In order to find out if certain words were ambiguous, we used the Internet edition of *Multitaggeren* (developed by Tekstlaboratoriet and Dokumentasjonsprosjektet at the University of Oslo, and which is a further development of, among others, Bokmålsordboka). The main difference between *Multitaggeren* and standard dictionaries is that it also provides information on inflected words, not only the word forms. In other words, it provides information not only on a word form such as *bilde* 'picture', but also *bilder* 'pictures'. *Multitaggeren* provides information about the grammatical properties of a certain word, but not about its meaning.

Regarding the meaning(s) of the words, we used the web edition of *Bokmålsordboka* (Landrø and Wangensteen 1986), and the encyclopedia *Store Norske Leksikon* by Kunnskapsforlaget.

2. 4. The Work Process

All together, we examined the 5500 most frequent search words in the log files. For each search word, we checked whether it was ambiguous by looking it up in *Multitaggeren*, and when ambiguity was found we investigated what kind of ambiguity was involved, and how grammatical categories and properties correlated with the different meanings. Note that we have only considered those differences in meaning that can be gathered under the term homonymy, and not those which are counted as polysemy. (We have for example considered the “putevar” and “følsom” meanings of *var* as relevant, but not the-lowest-part-of-a-leg and the-lowest-part-of-a-mountain meanings of *fot* 'foot'.)

For each word we conducted a search on the Internet using Fast’s search engine Alltheweb.com. In this way we found out what kind of results the search word led to, whether the hits varied depending on the ambiguity of the words, and whether the hits was relevant according to what we considered to be the preferred meaning of these words. We only looked at the first 100 hits for each word. This number should probably have been higher, as the hits are often grouped in such a way that the hits involving the same meaning of a search word appear together. However, we had limited time available, and it was extremely time-

consuming to go through the search results, since it was often impossible to see which meaning of the search word was used from the few lines that accompanied each search result. In such cases, it was necessary to check on the web-page itself, and the web-link could be broken, or the page could be under construction or simply take time to download.

For this reason, the data presented should not be regarded as conclusive. Nevertheless, the material provides a good indication of what is to be found.

We completed a little form for each search word.

(2)

1. Ambiguity:
    - Important for search?
2. Inflected form:
3. Name:
    - Ambiguity:
        - Person:
        - Place:
        - Firm/org:
        - Other:
4. Note

In the talk we will say more about how we filled in these forms, and about the results, illustrating with examples.