Lexical Disambiguation: Sources of Information and their Statistical Realization

Ido Dagan *
Computer Science Department, Technion, Haifa, Israel
and
IBM Scientific Center, Technion City, Haifa, Israel

Abstract

Lexical disambiguation can be achieved using different sources of information. Aiming at high performance of automatic disambiguation it is important to know the relative importance and applicability of the various sources. In this paper we classify several sources of information and show how some of them can be achieved using statistical data. First evaluations indicate the extreme importance of local information, which mainly represents lexical associations and selectional restrictions for syntactically related words.

1 Disambiguation Sources

The resolution of lexical ambiguities in unrestricted text is one of the most difficult tasks of natural language processing. In machine translation we are confronted with the related task of ***target word selection*** – the task of deciding which target language word is the most appropriate equivalent of a source language word in context. In contrast to computational systems, humans seem to select the correct sense of an ambiguous word without much effort and usually without even being aware to the existence of an ambiguous situation. This fact naturally led researches to point out various sources of information which may provide the necessary cues for disambiguation, either for humans or machines. The following paragraphs classify these sources into two major types, based on either ***understanding*** of the text or ***frequency*** characteristics of it.

One kind of information relates to the understanding of the meaning of the text, using semantic and pragmatic knowledge and applying reasoning mechanisms. The following sentences, taken from foreign news sections in the Israeli Hebrew press, demonstrate how different levels of understanding can provide the disambiguating information.

(1) ẖaver ha-bayit ha-'elyon shel ha-parlament ha-sovieti zaka be-monitin ke-ẖoker shel ha-shẖitut be-kazaẖstan.

This sentence translates into English as:

(2) The member of the upper house of the soviet parliament acquired a reputation as an investigator of the corruption in Kazakhstan.

The two most frequent senses of the ambiguous noun 'ẖaver' correspond to the English words 'friend' and 'member'. In the above example, the information for selecting the correct sense is provided by the semantic knowledge that 'a house of parliament' typically has members but not friends. Computationally this kind of information is usually captured by a shallow semantic model of selectional restrictions. In other cases, such as example (3), it is necessary to use deeper understanding of the text, which involves some level of reasoning:

(3) be-het'em le-ẖoq ha-hagira ha-ẖadash tihye le-kol ezraẖ sovieti ha-zkut ha-otomatit lekabel darkon bar tokef le-ẖamesh shanim.

This sentence translates into English as:

(4) According to the new emigration bill every soviet citizen will have the automatic right to receive a passport valid for five years.

The Hebrew word 'hagira' is used for the two subsenses 'emigration' and 'immigration'. In order to make the correct selection it is necessary to reason that since the soviet bill relates to soviet citizens then it concerns with leaving the country rather than entering it.

Another kind of information source, which was originally raised in the psycholinguistic literature, relates to the relative frequencies of word senses and associations between word senses. These factors were shown to play an important role in lexical retrieval, and were suggested as relevant for lexical disambiguation [4, 3]. Hanks [1], for example, lists different words associated with the two senses of the word 'bank', such as ***money, notes, account, investment*** etc. versus ***river, swim, boat*** etc.

[*] This research was partially supported by grant number 120-741 of the Israel Council for Research and Development

Aiming for high performance in automatic disambiguation, it is important to know (a) what is the portion of ambiguous cases in running text which can be resolved by each source of information and (b) how to set preferences among these sources when they provide contradicting evidence.

2 Statistical Information

A tempting starting point for answering the above questions is to use various types of statistical data about word senses and evaluate their contribution to disambiguation. In recent years, statistical data were used successfully for other linguistic tasks. The process of acquiring statistical data is usually faster and more standard and objective than manual construction of knowledge. This makes such data suitable for the evaluation task we are confronted with. The following paragraphs describe the kinds of statistics we use and explain how they reflect different types of disambiguating information.

In another paper [2] we describe a new multilingual approach in which we gather statistics about senses of ambiguous words of one language using a corpus of a different language. For example, the different word associations for the two senses of 'bank' will be identified in a Hebrew corpus, where a distinct word is used for each of the senses. This method enabled us to collect statistics from very large corpora without manually tagging the occurrences of the ambiguous words with their word senses. In our first experiment we have examined about one hundred examples of ambiguous Hebrew words which were selected randomly from foreign news sections in the Israeli press. For each sense of a Hebrew word we have collected statistics (in an English corpus) on its absolute frequency and its cooccurrences with other words that were syntactically related with it in the example sentence.

Two kinds of statistics were maintained. One statistic was the number of times in which the related words were identified in the corpus having the same syntactic relation as in the example sentence. This kind of statistic reflects both selectional restrictions, like the relation between 'member' (versus 'friend') and 'a house of parliament', and also word associations, like the association between 'member' and 'reputation', which is stronger than the association between 'friend' and 'reputation'. In the first case we expect a null frequency for the semantically illegal alternative, while in the second case we expect the difference in frequencies to represent the different degrees of association between the competing alternatives and their surrounding context. In getting this syntactically based statistic we are of course limited by the coverage and the accuracy of the parser, thus getting smaller and somewhat noisy counts relative to the real counts in the corpus.

A second and more robust statistic is obtained by counting the number of times in which the two words cooccurred within a limited distance [1]. For instance, the words 'member' and 'acquire' cooccurred 81 times in the corpus within a maximal distance of 7 words. This statistic is partly correlated with the first statistic, capturing also cases that were missed by the parser, but it also reflects lexical associations between words that tend to cooccur adjacently without having a specific syntactic relation between them. For instance, in one of our examples the word 'hatsba'ah', which means either 'voting' or 'indication', cooccurred in the same sentence with the word 'b<u>h</u>irot' (elections). We expect that the adjacency statistic will indicate the strong association between 'voting' and 'elections', and thus would prefer 'voting' as the appropriate sense.

The results reported in [2] together with further examination of our data have clearly indicated some interesting facts. In the vast majority of cases enough disambiguating information is provided by the immediate context, especially by syntactically related words. The absolute frequency of a word sense does not seem very useful, since it usually can be overridden successfully by the local context. An encouraging fact is that deep understanding of the text is rarely necessary, and seems to be required only for very delicate distinctions such as in example (3). In future work we intend to further analyze our data and test more examples, so that we can reach more decisive and quantitive conclusions. We believe that such conclusions will contribute to improve lexical disambiguation in broad coverage systems.

References

[1] Church, K. W., and Hanks, P., Word association norms, mutual information, and Lexicography, *Computational Linguistics*, vol. 16(1), 22-29 (1990).

[2] Dagan, Ido, Alon Itai and Ulrike Schwall, Two languages are more informative than one, submitted to ACL-91.

[3] Meyer, D., Schvaneveldt, R. and Ruddy, M., Loci of contextual effects on visual word-recognition, in P. Rabbitt and S. Dornic (eds.), *Attention and Performance V*, Academic Press, New-York, 1975.

[4] Simpson, Greg B. and Curt Burgess, Implications of lexical ambiguity resolution for word recognition, in Small, S. L., G. W. Cotrell and M. K. Tanenhaus, (eds.) *Lexical Ambiguity Resolution*, Morgan Kaufman Publishers, 1988.