Using a Semantic Tagger as a Dictionary Search Tool

Laura Löfberg[1], Jukka-Pekka Juntunen[2], Asko Nykänen[2], Krista Varantola[1], Paul Rayson[3] and Dawn Archer[4]
[1]School of Modern Languages and Translation Studies
University of Tampere,33014 Tampereen yliopisto, Finland
laura.lofberg@uta.fi, krista.varantola@uta.fi
[2]Kielikone Ltd, PL 126, 00211 Helsinki, Finland
jpj@kielikone.fi, asko.nykanen@kielikone.fi
[3]Computing Department and [4]Department of Linguistics and MEL
Lancaster University, Lancaster LA1 4YR/T, UK
paul@comp.lancs.ac.uk, d.archer@lancaster.ac.uk

Abstract

The USAS semantic tagger is a powerful language technology tool that has proven to be very effective in various applications such as content analysis, discourse analysis and information extraction. In the Benedict project, we intend to use the semantic taggers for the English and Finnish languages as search tools in electronic dictionaries, thereby enabling users to carry out context-sensitive dictionary searches.
The aim of this paper is to envisage ways in which a semantic tagger can help users find the "right" answer from a dictionary (i.e. the answer that the user needs). We begin with a brief introduction of the semantic taggers for the English and Finnish languages. Thereafter, we focus on the presentation of the context-sensitive dictionary look-up, and show how the dictionary software will be able (i) to determine the correct sense in the context at hand, and (ii) to highlight that sense for the user. The new search tool will be commercially available in the Benedict software.

1 Introduction

The electronic format of dictionaries has revolutionized the dictionary user's search process. The dictionary is no longer merely an alphabetical list of entries, but can be searched in a number of ways and from different angles depending on the kind of information that the user requires. For example, one can now carry out a full text search, a phonetic search, a morphological search, or extend the search to other information categories such as corpora, grammar books and Web-sources. The possibility of carrying out the kinds of searches that (i) can be adapted according to search requirements, and (ii) will answer the users' context-sensitive needs is only one of the factors that constitute the intelligence of an electronic dictionary. Other such factors include the possibility of customising dictionary content, making suggestions for correct spelling, and carrying out morphological analysis in case the user is trying to look up an inflected item etc. These factors are examples of "shallow intelligence" – that is, intelligence that is characterized by straightforward, deterministic algorithms.

In contrast to "shallow intelligence" applications, "deep intelligence" applications assist the dictionary user in the same way a knowledgeable human expert might (typical

examples include user profiles and context sensitive dictionary applications). In the Benedict project[1] we have been developing a new type of context-sensitive dictionary search tool that will not only lead users to the correct main entry but will also highlight the relevant sense of the looked-up item. This, in turn, has led to our further developing the English semantic tagger, and creating a parallel semantic tagger for the Finnish language. The following section describes the English and Finnish semantic taggers in some detail.

2 Semantic Tagger

The English semantic tagger (henceforth EST) is a program developed by Lancaster University's UCREL team (University Centre for Computer Corpus Research on Language). It includes an English lexicon, and software that automatically links words in a text to one or more semantic categories. To date, the EST has been used for market research, content analysis and information extraction. The EST's new role as a dictionary search tool has necessitated further development of its disambiguation methods and software systems, and the creation of a tool to link EST output to dictionary entries.

The EST uses a set of semantic tags that, at its conception were loosely based on Tom McArthur's (1981) Longman Lexicon of Contemporary English, but which have now been considerably revised in the light of ongoing research. A semantic tag indicates the conceptual field of the word. It groups together word senses that are related at some level of generality with the same mental concept indicating not only synonymy and antonymy but also hypernymy and hyponymy. The tagset is arranged in a hierarchy with 21 major discourse fields expanding into 232 category labels. The following figure shows the 21 labels at the top level of the hierarchy.

A general and abstract terms	B the body and the individual	C arts and crafts	E Emotion
F food and farming	G government and public	H architecture, housing and the home	I money and commerce
K entertainment, sports and games	L life and living things	M movement, location, travel and transport	N numbers and measurement
O substances, materials, objects and equipment	P education	Q language and communication	S social actions, states and processes
T Time	W world and environment	X psychological actions, states and processes	Y science and technology
Z names and grammar			

Figure 1: The top level of the UCREL Semantic Analysis System (USAS)

The EST combines a lexicon containing a set of possible tags for each word and various disambiguation template rules that deduce which of the tags is the correct one in the text at hand. Currently, the English lexicon contains around 43,400 words. Additionally, there's a template list containing over 18,500 multi-word units. The EST is reported to have obtained 92% accuracy on general English texts (Rayson and Wilson 1996). A key disambiguation technique relies on part-of-speech tags provided by CLAWS (Garside and Smith 1997) to distinguish semantic tags by major word class.

In the Benedict project, we are working on improving the EST and constructing a parallel tool for Finnish, a synthetic and non-Indo-European language that is very different from English. The aim is to avoid building a completely new system but to use the existing software created for the English language as far as possible. Our experiments so far have shown that the semantic categories developed for the EST are for the most part compatible with the semantic categorizations of objects and phenomena in Finnish. However, the different grammatical features between these two very different languages have proved to be extremely challenging. By way of illustration, unlike English, Finnish is highly inflectional, uses a lot of compounds, and has a relatively free word order (Löfberg et al. 2003). To overcome these and similar issues, a new software module has been appended to the semantic tagger. TextMorfo, as it is called, is a Finnish syntactical and morphological analysis tool, which will mirror the part-of-speech tagger for English (CLAWS). In addition, we are building a compound engine to process rarer compounds that are not yet included in the Finnish lexicon.

At present, the Finnish lexicon contains around 30,000 words. In compiling the Finnish lexicon we have followed the same principles as the UCREL team followed when compiling the English lexicon. This means that the lexicons are theoretically comparable in terms of sense. However, their structures differ greatly. For example, because of the limited number of inflected forms in English, the English lexicon contains both base forms and inflected forms. Due to the highly inflectional and agglutinative nature of Finnish, the same is not possible for the Finnish lexicon. Consequently, it contains only the base forms. That said, the morhological analysis tool (TextMorfo) enables the system to recognize any inflected forms that occur in text[s].

The initial test results have been promising. The program still needs refining and the coverage of the Finnish lexicon needs to be expanded and supplemented by a template list of multi-word units, but it seems clear that the program can indeed process the Finnish language.

3. Introducing the Context-Sensitive Dictionary Look-Up

Let us imagine a typical dictionary usage situation: The user is reading a text on a computer screen and discovers a word that he does not know. He looks up the word in a dictionary. Let us further suppose a typical scenario where the word has many senses (see figure 2 below). How can the user in such a situation determine which sense is the right one in a given context?

arm[1] (ɑ:m) *n* **1** (in man) either of the upper limbs from the shoulder to the wrist. Related adj: brachial. **2** the part of either of the upper limbs from the elbow to the wrist; forearm. **3a** the corresponding limb of any other vertebrate. **3b** an armlike appendage of some invertebrates. **4** an object that covers or supports the human arm, esp. the sleeve of a garment or the side of a chair, sofa, etc. **5** anything considered to resemble an arm in appearance, position, or function, esp. something that branches out from a central support or larger mass: *an arm of the sea; the arm of a record player.* **6** an administrative subdivision of an organization: *an arm of the government.* **7** power; authority: *the arm of the law.* **8** any of the specialist combatant sections of a military force, such as cavalry, infantry, etc. **9** *Nautical.* See yardarm. **10** *Sport, esp. ball games.* ability to throw or pitch: *he has a good arm.*

Figure 2. Noun senses of the word *arm* from the *Collins English Dictionary*[2]

This is the problem that the context-sensitive search tool of the Benedict software will solve. Unlike electronic dictionaries in general, the Benedict software scans the entire context of the word. The context is then fed into a semantic tagger and the tagger output is used to filter out incorrect senses.

Such disambiguation is perhaps not necessary for a text that is in the user's native language, but it is particularly helpful for a non-native language user potentially confused by cultural and lexical variations. This particular feature will therefore be optional in the actual dictionary software. With the aid of this disambiguation, further information such as collocation information and usage examples for the selected sense can also be presented.

4. Mapping to a Dictionary

How does the context determine the sense? For a human reader, disambiguation is an automatic, natural process, but for a computer it can be a daunting task. The Benedict software uses the semantic tagger to categorize the looked-up item and its context into the semantic categories described earlier. Suppose, for example, that we are looking for the translation of the word *arms* in the sentence of the figure 3.

Figure 3. Sample context

The sentence is first syntactically analysed, and the word *arms* receives a noun tag. The base form (*arm*) and other syntactic attributes of the word are also marked. The sentence is then tagged semantically, resulting in the following output:

He_Z8m was_Z5 sewing_C1 ***arms_B5*** *onto_Z5 the_Z5 jacket_B5.*

Figure 4. Tagged sample context

As Figure 4 makes clear, the plural noun *arms* receives the tag *B5*, which stands for the subcategory "Clothes and personal belongings" of the top-level domain "B: the body and the individual" (see figure 1). The remaining words of the sentence also receive their respective semantic tags.

Although a seemingly straightforward procedure, the system has to be able to select and highlight the correct sense(s) among all the senses given in the dictionary. Initially, we disambiguate with the syntactic information in the same manner in which many existing implementations do (e.g. Prózéky & Kis 2002). Thus, in our example the noun entry (figure 2) for the word is selected. Nevertheless there are still ten senses to select from. We therefore need an algorithm that effectively chooses the best one(s) for our context.

4.1 Mapping semantic categories to dictionary domain markers

Most dictionaries mark the senses of a word with domain tags or subject field markers (e.g. *Nautical, Sport* in figure 2). This is the single most valuable source of information for our sense selection problem: the dictionary domain markers in many ways resemble the semantic categories used in the semantic tagger. A mapping from the semantic categories to the dictionary domains of a specific dictionary can be compiled and has actually already been done by Archer et al (2003) for the Collins Dictionary domain system. Thus, the B5 tag of our example can be mapped to the dictionary domains *Clothing, Personal Arts & Crafts, Hairdressing & Grooming, Clothing & Fashion, Jewellery* and *Tanning* in the *Collins English Dictionary.* The mapping is certainly not a one-to-one mapping. However, it is still an important input for our algorithm.

4.2 Domain Detection System (DDS)

Unfortunately, not all of the senses in the dictionaries are marked with dictionary domain tags. In addition, it is often the case that the context does not provide enough information for the semantic tagger to unequivocally select the correct sense. Even so, we still want to be able to highlight the most probable one. We are in the process of developing a software solution to this problem - the *Domain Detection System (DDS).* The DDS relies on the following two assumptions:

- *The domain specific words tend to co-occur in text. (Assumption 1, Domain Boundness Assumption)*
- *At least one of the dictionary senses is always the correct one for the given context. (Assumption 2, Dictionary Completeness Assumption)*

Basically, the DDS uses all the information available to it to select the best sense(s) in a dictionary entry for the given context. Based on Assumption 1, all the words and phrases in the context of the looked-up item, that is, the words *He, was, sewing, arms, onto, the* and *jacket* in our example, may give some hints respecting the domain of the word *arm*. Likewise, all the words in the usage examples, collocations, synonym lists etc., of a single dictionary sense can assist in finding the correct tag. The DDS is also flexible enough to cope with entries that are missing some of this information. Hence the DDS semantically tags both the context of the looked-up item and the dictionary entry, and then calculates the "domain distances", that is, the semantic distance between the context of the looked-up item and the dictionary senses. Based on Assumption 2, the sense with the smallest distance is selected and highlighted.

4.3 Semantic distance calculation in DDS

The implementation of the distance calculation can be carried out in a variety of ways. In our algorithm, it relies on vector arithmetic, that is, each text passage receives a feature vector that effectively gives a score to each semantic category. Thus, our sample sentence in figure 4 receives the vector:

(Z8:1/7, Z5:3/7, C1:1/7, B5:2/7)
=
(Z8:0.14, Z5:0.43, C1:0.14, B5:0.29)

Figure 5. Semantic vector of the sample sentence.

Likewise, all the dictionary senses receive similar vectors. These vectors can now be drawn in a multi-dimensional Euclidian space where each vector is presented by a point. We can easily calculate the distance between any two points as Euclidian distance in the space or as an angle between the two vectors.

4.4 Additional features

The simple vector calculation algorithm described above illustrates the core idea of the DDS. Still, it is an oversimplification of the system and would not suffice for real data. The final DDS is a hybrid system that uses many statistical and rule-based components to solve various sub-problems. For example, semantic categories like Z4, Z5, and Z8 capture function words (as opposed to content words) and therefore tend to occur frequently in a majority of text types. If they were not weighted properly, they would become too dominant. Consequently, different dictionary fields are also prioritised and weighted using a balanced corpus like the BNC to ensure that we capture the most valuable domain information.

In addition, the DDS is designed so as to identify – and make use of - the syntactic-semantic patterns that appear near the search word. In the small context that we have, the syntax and syntactic relations of the words become a hugely important source of information. Thus, we are able to select the correct sense of figure 2 above, because the definition - "***4** an object that covers or supports the human arm, esp. the sleeve of a garment or the side of a chair, sofa, etc.*" – contains a number of words that, like *arm* of Figure 3, are

categorised as belonging to the B5 semantic field, 'Clothes and personal belongings' (e.g. *sleeve, garment*).

There are still other corrections to be made to cope with various sub-problems like context length, variations between dictionaries etc. However, preliminary tests of the DDS have produced encouraging results.

5. Conclusion and Further Implications

In this paper we have been envisaging the development and the functionality of a context-sensitive dictionary search tool enabled by the use of a semantic tagger. With the aid of this tool, the user will not only be guided to the entry of the item that s/he is looking up but the dictionary software will also highlight for him/her the sense within that entry that is the mostly likely, given the context. The modules needed for the dictionary look-up algorithms (syntactic and semantic taggers, morphological analysis tools, lexicon, etc.) in the English language already exist in the main, and the development of the modules for the Finnish language will be completed soon.

We have briefly presented the highlights of our Domain Detection System, which calculates semantic distances between two text passages. Of course, our system still needs some fine-tuning. We have made two rather strong assumptions that should certainly be questioned. Assumption 1 (Domain Boundness) is mostly true for technical documents but not always valid for newspaper and more casual texts where the domain is often a moving target. That said, the system can be very useful for a user who is totally confused as to the mostly likely context and/or is not able to understand any of the key words in the text.

Assumption 2 should generally hold for large quality dictionaries. However, natural language is constantly evolving and no dictionary is ever complete. The system is nevertheless restricted to the dictionary content and can only give answers that are as good as the data source itself.

Despite these hindrances, the system will certainly shed some light on the problems with ambiguous word senses. It will be the task of the user interface to display this information in a feasible way.

Endnotes

1 The Benedict Project is funded by the European Union, and it belongs to the Information Society Technologies programme of the Fifth Framework Programme (Action Line: Multilingual Web). The consortium is made up of the University of Tampere, Kielikone Ltd., Gummerus Kustannus and Nokia from Finland, and Lancaster University and HarperCollins Publishers from UK. The project started 1 March 2002 and will end 28 February 2005.

2 Collins English Dictionary 5th Edition first published in 2000 © HarperCollins Publishers 1979, 1986, 1991, 1994, 1998, 2000

References

Archer, D., Rayson, P., Piao, S. and McEnery, T. (forthcoming). Comparing the UCREL Semantic Annotation Scheme with Lexicographical Taxonomies. Accepted to EURALEX 2004 conference.

Garside, R. and Smith, N. 1997. 'A Hybrid Grammatical Tagger: CLAWS4' in Garside, R., Leech, G., and McEnery, A. (eds.), Corpus Annotation: Linguistic Information from Computer Text Corpora. Longman: London.

Löfberg, L., Archer D., Piao, S., Rayson, P., McEnery, T., Varantola, K. and Juntunen, J-P. 2003. 'Porting An English Semantic Tagger To The Finnish Language' in Archer, D., Rayson, P., Wilson, A. and McEnery, T. (eds.), Proceedings of the Corpus Linguistics 2003 conference (UCREL technical paper number 16). UCREL: Lancaster University: Lancaster.

Manning, C.D. and Schütze, H. 2003. Foundations of statistical natural language processing. The MIT Press. Cambridge, Massachusetts. London, England.

McArthur, T., 1981. Longman Lexicon of Contemporary English. Longman: London.

Prózéky, G. and Kis, B. 2002. 'Development of a Context-Sensitive Electronic Dictionary' in Braasch, A. and Povlsen C. (eds.), Proceedings of the Tenth EURALEX International Conference (EURALEX 2002). Center for Sprogteknologi: Denmark.

Rayson, P. and Wilson, A. 1996. 'The ACAMRIT Semantic Tagging System: Progress Report' in Evett, L.J. and Rose, T.G. (eds.), Language Engineering for Document Analysis and Recognition, LEDAR (AISB96 Workshop proceedings). Brighton: England: Faculty of Engineering and Computing, Nottingham Trent University, UK.