Word lookup as an ongoing dialogue between a user and a lexicon

Michael Zock

LIMSI-CNRS, **currently** Department of Computer Science, Tokyo Institute of Technology, Tokyo

zock@limsi.fr or zock@cl.cs.titech.ac.jp

Abstract : We all experience now and then the problem of being unable to find the word expressing adequately the idea we have in our mind. Instinctively we reach for a dictionary. Yet, dictionaries may be of little help, if they expect from us precisely what we are looking for : a perfectly spelled word, expressing the message we try to convey. While perfect input may be reasonable in the case of analysis (comprehension), it certainly isn't in the case of synthesis (generation) where the starting point is conceptual in nature: a message, the (partial) definition of a word, or a word related to the target word. The language producer needs a dictionary allowing for reverse access. A thesaurus does that, but only in a very limited way (the entry points are basically topical).

People use various methods to initiate search in their mind : words, concepts, partial descriptions, etc. If we want to mimic these functionalities by a computer, we must build the resource accordingly. Let's assume that the text producer is looking for a word that he cannot access, instead he comes up with another word (or concept) somehow related to the former. He may not know precisely how the two relate, but he knows that they are related. He may also know how close their relationship is, and whether a given link is relevant or not, that is, whether it will lead directly (synonym, antonym, hyperonym) or indirectly to the target word. Since the relationship between the source- and the target word is often indirect, several lookups may be necessary : each one of them having the potential to contain either the target word (direct lookup), or a word leading towards it (indirect lookup).

In order to allow for this kind of search or access, we propose to build an associative network. Our idea is to index an existing electronic dictionary (or, thesaurus) by extracting associations from an encyclopedia. Since all words (or concepts they express) are somehow connected, one can enter the network at any point (by giving the source word : word or concept coming to your mind) and follow the links (associations) to reach the word one is looking for (target word). While the notion of association is very old, its use to support *word access* via computer is new. In other words, the resource still needs to be built. We discuss here some of the problems involved in accomplishing this by using computers.

I. Introduction

Every text producer (speaker/writer) encounters now and then the following problem: given some conceptual input (meaning or message) he fails to find the corresponding word(s). The typical reaction in a case like this is to resort to a dictionary (someone else's mental dictionary, some traditional paper dictionary or its electronic equivalent) and to look for the corresponding word. The question is how to find the target word in a dictionary that is word based, alphabetically ordered and doesn't allow for search on the basis of (partial) input. While being fairly adequate for the language receiver, such kind of dictionaries are much less useful for the language producer whose starting point are meanings and not words.

II. How reasonable is it to expect perfect input ?

The expectation of perfect input is unrealistic even in analysis, but even more so in generation. The user may well be unable to provide the required information: be it because he cannot access in time the word he is looking for, even though he knows it,[1] or be it because he does not know the word yet expressing the idea he wants to convey. This latter case typically occurs when learning or using a foreign language. Yet, not being able to find the right word, does not imply that one does not know anything concerning the target word. Actually, quite often the contrary is the case.

Suppose, you were looking for a word expressing the following ideas : *domesticated animal*,

[1] Temporary amnesia, known as the TOT, or tip-of-the-tongue problem (Brown & Mc Neill, 1966 ; Zock & Fournier, 2001).

producing milk, *suitable for making cheese*. Suppose further that you knew that the target word was neither *cow* nor *sheep*. While none of this information is sufficient to guarantee the access of the intended word *goat*, the information at hand (part of the definition) could certainly be used (For a concrete proposal going in this direction, see Bilac et al. 2004). Next to definition information, people have other kind of knowledge concerning the target word. For example, they know how the latter relates to other words. For example, they know that *goats* and *sheep* are somehow connected, that both of them are animals, that *sheep* are appreciated for their wool and meet, that sheep tend to follow each other blindly, while *goats* manage to survive while hardly eating anything, etc. In sum, people have in their mind lexical networks : all words, concepts or ideas they express are highly interconnected. As a result, any one of them has the potential to evoke the others. The likelihood for this to happen depends, among other things, on factors such as frequency (associative strength) and distance (direct vs. indirect access). As one can see, association is a very general and powerful mechanism. No matter what we hear, read or say, any idea is likely to remind us of something else. This being so, we should make use of it.

III. Search based on the relations between concepts and words

If one agrees with what we've just said, one could view the *mental dictionary as a huge semantic network composed of nodes (words and concepts) and links (associations),with either being able to activate the other*. [2] Finding a word amounts thus to entering the network and following the links leading from the source node (the first word that comes to your mind) to the target word (the one you are looking for). Suppose you wanted to find the word 'nurse' (*target word*), yet the only token coming to your mind were 'hospital'. In this case the system would display by *category* (chunks) all the words linked to it (figure 1b). Put differently, the system would build internally a small semantic network with 'hospital' in the center and all the words having a direct link with it (figure 1a) as immediate satellites. If the candidate is in any of these lists, search stops – otherwise it goes on. The user could choose a word occuring in any of the lists or a new token (relation).

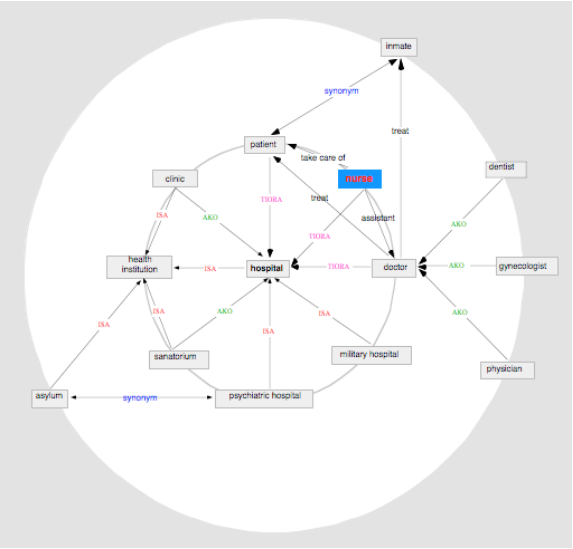
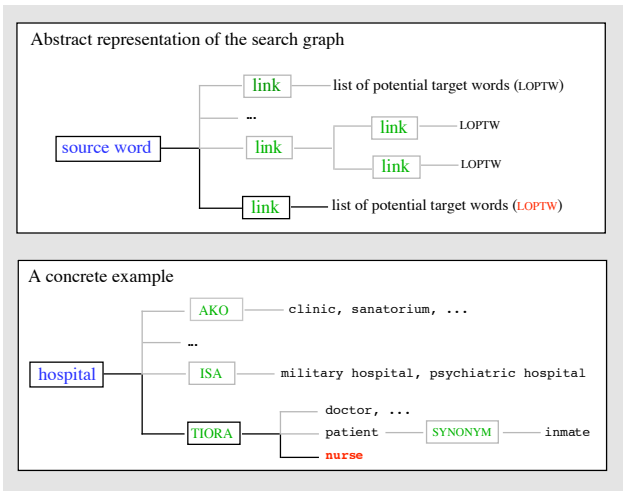


Figure 1a : Search based on propagation in a network (*internal* representation)



Figures 1b and 1c : proposed candidates, grouped according to the *nature* of the link

AKO : a kind of ; **ISA** : subtype ; **TIORA** : typically involved object, relation or actor

[2] Actually, one could question the very notion of « mental dictionary », which is convenient but misleading in as it supposes a dedicated part or locus for this task in our memory. A multiply indexed mental encyclopedia, composed of polymorph information (concepts, words, metalinguistic information) seems much more plausible to me.

The fact that the links are labelled has some very important consequences : while maintaining the power of a graph (possible cyclic navigation) it has at the interface level the simplicity, hence reassuring effect of a tree. Each node points only to data of the same category, whereas the graph may point to heterogeneous elements. Using at the interface level a tree rather than a graph, words are presented now by clusters, hence navigation can be done by category. The assumption being that the user generally knows to which category the target word belongs, and that categorical search is in principle faster than search in a huge list of unordered (or, alphabetically ordered) words.

The distance between the source and target word is certainly variable. If the source word (given input) is directly related, revealing the target word is trivial and straightforward, otherwise search requires several steps. Finding the wanted word requires thus several lookups. Ideally, each one of them brings the user closer to the target word.

IV. A resource still to be built

Word access, as described here, amounts to navigating in a huge associative network. Of course, such a network has to be built. The question is how. Our proposal is to build it automatically by parsing an existing corpus (ideally an encyclopedia). This would yield a set of associations,[3] which still need to be labelled. This could be done via an ontology.

Unlike private information,[4] which by definition cannot and should not be put into a public dictionary,[5] encyclopedic knowledge can be added in terms of associations, as this information expresses commonly shared knowledge, that is, the kind of associations most people have when encountering a given word. Take for example the word *elephant*. An electronic dictionary like WordNet associates the following gloss with the headword: large, grey, four-legged mammal, while Webster gives the following information:

> A mammal of the order Proboscidia, of which two living species, *Elephas Indicus* and *E. Africanus*, and several fossil species, are known. They have a proboscis or trunk, and two large ivory tusks proceeding from the extremity of the upper jaw, and curving upwards. The molar teeth are large and have transverse folds. Elephants are the largest land animals now existing.

While this latter entry is already quite rich (trunk, ivory tusk, size), an encyclopedia contains even more information.[6] If all this information were added to an electronic resource, it would enable us to access the same word (e.g. *elephant*) via many more associations than ever before. By looking at the definition here above, one will notice that many associations are quite straightforward (color, size, origin, etc.), and most of them appear presumably quite frequently. If one agrees with these views, the remaining question is how to extract this encyclopedic information and to add it to an existing electronic resource. The answer is fairly straightforward: run a parser on an existing encyclopedia, replace (parts of) the syntactic information with ontological labels (the noun: *africa* yielding the category, continent: Africa) and integrate the words and links into an existing electronic resource. While reaching this goal is not trivial, [7] achieving it would extend considerably the number of words from which a target word can be accessed, which is precisely our goal.

V. Discussion and conclusion

I have raised and partially answered the question of how a dictionary should be indexed in order to

---

[3] The assumption being that every word co-occurring with another word in the same sentence is a candidate of an association. The more frequently two words co-occur in a given corpus, the greater their associative strength.

[4] For example, the word *elephant* may remind you of a specific trip and a specific country in Africa.

[5] This does not (and should not) preclude the possibility to add it to one's personal dictionary.

[6] You may consider taking a look at http://en.wikipedia.org/wiki/ which is free.

[7] There are at least two problems : the problem of *ambiguity*, and the problem of classifying information not available in the ontology (eg. dogs love to play with children). One might also want to check to what extent a parser is necessary, or, whether a bag of words wouldn't do the job.

support word access. I was particularly concerned with the language producer, as his needs (and knowledge at the outset) are quite different from the ones of the language receiver (listener/reader). It seems that, in order to achieve our goal, we need to do two things: add to an existing electronic dictionary information that people tend to associate with a word, that is, build and enrich a semantic network, and provide a tool to navigate in it. To this end I've suggested to label the links and to convert the graph into a tree-like structure. Actually my basic proposal is to extend a resource like WordNet by adding certain links, in particular on the horizontal axis (syntagmatic relations). These links are associations, and their role consists in helping the encoder to find ideas (concepts/words) related to a given stimulus (brainstorming), or to find the word he is thinking of (word access).

One problem that we are confronted with is to identify possible associations. Ideally we would need a complete list, but unfortunately, this does not exist. Yet, there is a lot of highly relevant information out there. For example, Mel'cuk's lexical functions (Mel'cuk, 1992), Fillmore's FRAMENET (http://www.icsi.berkeley.edu/~framenet/), work on ontologies (Cyc), thesaurus (Roget), WordNets (the Princeton and Eurowordnets), and, last but not least, the FACTOTUM project (http://humanities.uchicago.edu/homes/MICRA/). Of course, one would need to make choices here and probably add links. Another problem is to identify *useful* associations. Not every possible association is necessarily plausible. Hence, the idea to take as corpus something that expresses shared knowledge, for example, an encyclopedia. The associations it contains can be considered as being plausible. We could also collect data by watching people using a dictionary and identify search patterns.[8] Next, we could run psycholinguistic experiments. [9] While the typical paradigm has been to ask people to produce a response (red) to some stimulus (rose), we could ask them to identify or label the links between words (e.g. *apple-fruit*, *lemon-yellow*, etc.). The ease of labeling will probably depend upon the origin of the words (the person asked to label the link or somebody else).

Another approach would be to extract collocations from a corpus and label them automatically. There are tools for extracting co-occurrences (Ferret, 2002), and ontologies could be used to qualify some of the links between collocational elements. While this approach might work fine for couples like *coffee-strong*, or *wine-red* (since an ontology would reveal that *red* is a kind of *color*, which is precisely the link type : i.e. association), one may doubt that it could reveal the nature of the link between *smoke* and *fire*. Yet, most humans would immediately recognize this as a causal link. As one can see, there are still a few major problems to be solved. Nevertheless, I do believe that these obstacles can be removed, and that the approach presented here has the potential to improve word acces, making the whole process more powerful, natural and intuitive, hence efficient.

VI. References :

Bilac, S., Watanabe W., Hashimoto T., Tokunaga T. & H. Tanaka (2004). Dictionary search based on the target word description. This volume

Brown, R. and Mc Neill, D. (1966). The *tip of the tongue* phenomenon. Journal of Verbal Learning and Verbal Behavior, 5, 325-337

Ferret O. (2002) Using collocations for topic segmentation and link detection, *COLING 2002*, p. 260-266, Taipei.

Mel'cuk, I. et al. (1992) Dictionnaire Explicatif et Combinatoire du français contemporain. Recherche lexico-sémantique III. Les presses de l'université de Montréal.

Zock, M. & J.-P. Fournier (2001). How can computers help the writer/speaker experiencing the Tip-of-the-Tongue Problem ?, Proc. of RANLP, Tzigov Chark, pp. 300-302

---

[8] One such pattern could be: give me the word for a *bird with yellow feet and a long beak, that can swim* . Actually, word access problems frequently come under the form of questions like: *What is the word for **X** that **Y**?*, where **X** is usually a hypernym and **Y** a stereotypical, possibly partial functional/relational/case description of the target word.

[9] Actually, this has been done for decades, but with a different goal in mind. (see D. Nelson, C. McEvoy & T. Schreiber : http://cyber.acomp.usf.edu/FreeAssociation/).