Enhancing electronic dictionaries with an index based on associations

Olivier Ferret
CEA –LIST/LIC2M
18 Route du Panorama
F-92265 Fontenay-aux-Roses
ferreto@zoe.cea.fr

Michael Zock[1]
LIF-CNRS
163 Avenue de Luminy
F-13288 Marseille Cedex 9
michael.zock@lif.univ-mrs.fr

Abstract

A good dictionary contains not only many entries and a lot of information concerning each one of them, but also adequate means to reveal the stored information. Information access depends crucially on the quality of the index. We will present here some ideas of how a dictionary could be enhanced to support a speaker/writer to find the word s/he is looking for. To this end we suggest to add to an existing electronic resource an index based on the notion of association. We will also present preliminary work of how a subset of such associations, for example, topical associations, can be acquired by filtering a network of lexical co-occurrences extracted from a corpus.

1 Introduction

A dictionary user typically pursues one of two goals (Humble, 2001): as a *decoder* (reading, listening), he may look for the definition or the translation of a specific target word, while as an *encoder* (speaker, writer) he may want to find a word that expresses well not only a given concept, but is also appropriate in a given context.

Obviously, readers and writers come to the dictionary with different mindsets, information and expectations concerning input and output. While the decoder can provide the word he wants additional information for, the encoder (language producer) provides the meaning of a word for which he lacks the corresponding form. In sum, users with different goals need access to different indexes, one that is based on form (decoding), the other being based on meaning or meaning relations (encoding).

Our concern here is more with the encoder, *i.e.* lexical access in language production, a feature largely neglected in lexicographical work. Yet, a good dictionary contains not only many entries and a lot of information concerning each one of them, but also efficient means to reveal the stored information. Because, what is a huge dictionary good for, if one cannot access the information it contains?

2 Lexical access on the basis of what: *concepts* (*i.e.* meanings) or *words*?

Broadly speaking, there are two views concerning lexicalization: the process is **conceptually-driven** (meaning, or parts of it are the starting point) or **lexically-driven**[2]: the target word is accessed via a source word. This is typically the case when we are looking for a *synonym*, *antonym*, *hypernym* (paradigmatic associations), or any of its syntagmatic associates (red-rose, coffee-black), the kind of association we will be concerned with here.

Yet, besides conceptual knowledge, people seem also to know a lot of things concerning the lexical form (Brown and Mc Neill, 1966): number of *syllables*, beginning/ending of the target word, *part of speech* (noun, verb, adjective, etc.), *origin* (Greek or Latin), *gender* (Vigliocco et al.,

[1] In alphabetical order

[2] Of course, the input can also be hybrid, that is, it can be composed of a conceptual and a linguistic component. For example, in order to express the notion of *intensity*, MAGN in Mel'čuk's theory (Mel'čuk *et al.*, 1995), a speaker or writer has to use different words (very, seriously, high) depending on the form of the argument (ill, wounded, price), as he says *very* ill, *seriously* wounded, *high* price. In each case he expresses the very same notion, but by using a different word. While he could use the adverb *very* for qualifying the state of somebody's health (he is *ill*), he cannot do so when qualifying the words *injury* or *price*. Likewise, he cannot use this specific adverb to qualify the noun *illness*.

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 281–288,
Sydney, July 2006. ©2006 Association for Computational Linguistics

1997). While in principle, all this information could be used to constrain the search space, we will deal here only with one aspect, the words' relations to other concepts or words (associative knowledge).

Suppose, you were looking for a word expressing the following ideas: *domesticated animal*, *producing milk suitable for making cheese*. Suppose further that you knew that the target word was neither *cow, buffalo* 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[3]. Besides this type of information, people often have other kinds of knowledge concerning the target word. In particular, they know how the latter relates to other words. For example, they know that *goats* and *sheep* are somehow connected, sharing a great number of features, that both are *animals* (hypernym), that *sheep* are appreciated for their wool and meat, that they tend to follow each other blindly, etc., while *goats* manage to survive, while hardly eating anything, etc. In sum, people have in their mind a huge lexico-conceptual network, with words[4], concepts or ideas being highly interconnected. Hence, any one of them can evoke the other. The likelihood for this to happen depends on such factors as *frequency* (associative strength), *saliency* and *distance* (direct vs. indirect access). As one can see, associations are a very general and powerful mechanism. No matter what we hear, read or say, anything is likely to remind us of something else. This being so, we should make use of it.

3 Accessing the target word by navigating in a huge associative network

If one agrees with what we have just said, one could view the *mental lexicon* as a huge semantic network composed of *nodes* (words and concepts) and *links* (associations), with either being able to activate the other[5]. Finding a word involves 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 is *hospital*. In this case the system would generate internally a graph with the *source word* at the center and all the associated words at the periphery. Put differently, the system would build internally a semantic network with *hospital* in the center and all its associated words as satellites (see Figure 1, next page).

Obviously, the greater the number of associations, the more complex the graph. Given the diversity of situations in which a given object may occur we are likely to build many associations. In other words, lexical graphs tend to become complex, too complex to be a good representation to support navigation. Readability is hampered by at least two factors: *high connectivity* (the great number of links or associations emanating from each word), and *distribution*: conceptually related nodes, that is, nodes activated by the same kind of association are scattered around, that is, they do not necessarily occur next to each other, which is quite confusing for the user. In order to solve this problem, we suggest to display by category (chunks) all the words linked by the same kind of association to the source word (see Figure 2). Hence, rather than displaying all the connected words as a flat list, we suggest to present them in chunks to allow for categorial search. Having chosen a category, the user will be presented a list of words or categories from which he must choose. If the target word is in the category chosen by the user (suppose he looked for a hypernym, hence he checked the ISA-bag), search stops, otherwise it continues. The user could choose either another category (e.g. AKO or TIORA), or a word in the current list, which would then become the new starting point.

[3] For some concrete proposals going in this direction, see dictionaries offering reverse lookup: http://www.ultralingua.net/ ,http://www.onelook.com/reverse-dictionary.shtml.

[4] Of course, one can question the very fact that people store words in their mind. Rather than considering the human mind as a *wordstore* one might consider it as a *wordfactory*. Indeed, by looking at some of the work done by psychologists who try to emulate the mental lexicon (Levelt et al., 1999) one gets the impression that words are synthesized rather than located and call up. In this case one might conclude that rather than having words in our mind we have a set of highly distributed, more or less abstract information. By propagating energy rather than data —(as there is no message passing, transformation or cumulation of information, there is only activation spreading, that is, changes of energy levels, call it weights, electronic impulses, or whatever),— that we propagate signals, activating ultimately certain peripherical organs (larynx, tongue, mouth, lips, hands) in such a way as to produce movements or sounds, that, not knowing better, we call words.

[5] While the links in our brain may only be weighted, they need to be labelled to become interpretable for human beings using them for navigational purposes in a lexicon.

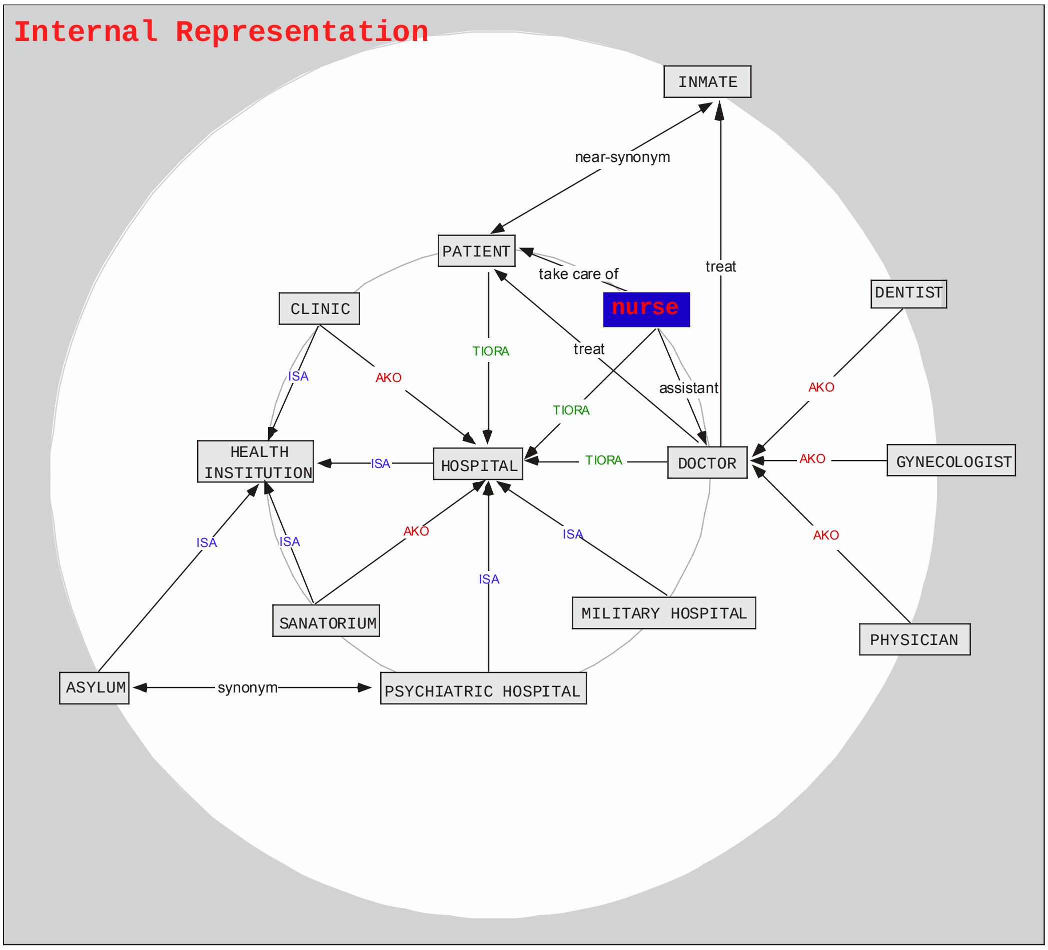

Figure 1*:* Search based on navigating in a network (internal representation)
AKO: a kind of; **ISA**: subtype; **TIORA**: **T**ypically **I**nvolved **O**bject, **R**elation or **A**ctor.

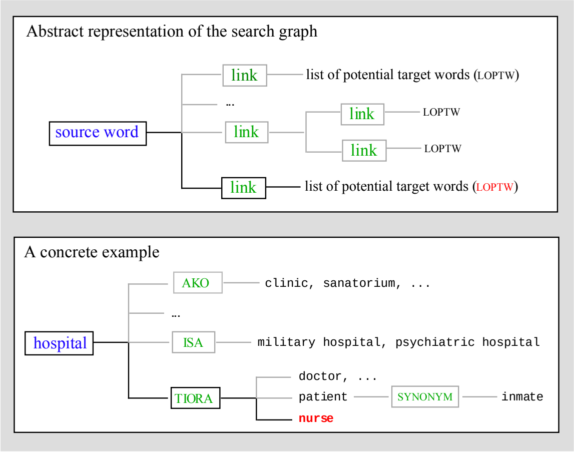
Abstract representation of the search graph
link
list of potential target words (LOPTW)
...
source word
link
link
LOPTW
link
LOPTW
link
list of potential target words (LOPTW)
A concrete example
AKO
clinic, sanatorium, ...
...
hospital
ISA
military hospital, psychiatric hospital
TIORA
doctor, ...
patient
SYNONYM
inmate
nurse

Figure 2: Proposed candidates, grouped by family, *i.e.* according to the nature of the link

As one can see, the fact that the links are labeled has some very important consequences:

(a) While maintaining the power of a highly connected graph (possible cyclic navigation), it has at the interface level the simplicity of a tree: each node points only to data of the same type, *i.e.* to the same kind of association.

(b) With words being presented in clusters, navigation can be accomplished by clicking on the appropriate category.

The assumption being that the user generally knows to which category the target word belongs (or at least, he can recognize within which of the listed categories it falls), and that categorical search is in principle faster than search in a huge list of unordered (or, alphabetically ordered) words[6].

Obviously, in order to allow for this kind of access, the resource has to be built accordingly. This requires at least two things: (a) indexing words by the associations they evoke, (b) identi-

[6] Even though very important, at this stage we shall not worry too much for the names given to the links. Indeed, one might question nearly all of them. What is important is the underlying rational: help users to navigate on the basis of symbolically qualified links. In reality a whole set of words (synonyms, of course, but not only) could amount to a link, *i.e.* be its conceptual equivalent.

fying and labelling the most frequent/useful associations. This is precisely our goal. Actually, we propose to build an associative network by enriching an existing electronic dictionary (essentially) with (syntagmatic) associations coming from a corpus, representing the average citizen's shared, basic knowledge of the world (encyclopaedia). While some associations are too complex to be extracted automatically by machine, others are clearly within reach. We will illustrate in the next section how this can be achieved.

4 Automatic extraction of *topical* relations

4.1 Definition of the problem

We have argued in the previous sections that dictionaries must contain many kinds of relations on the syntagmatic and paradigmatic axis to allow for natural and flexible access of words. Synonymy, hypernymy or meronymy fall clearly in this latter category, and well known resources like WordNet (Miller, 1995), EuroWordNet (Vossen, 1998) or MindNet (Richardson *et al.*, 1998) contain them. However, as various researchers have pointed out (Harabagiu *et al.*, 1999), these networks lack information, in particular with regard to syntagmatic associations, which are generally unsystematic. These latter, called TIORA (Zock and Bilac, 2004) or *topical relations* (Ferret, 2002) account for the fact that two words refer to the same topic, or take part in the same situation or scenario. Word-pairs like *doctor–hospital*, *burglar–policeman* or *plane–airport*, are examples in case. The lack of such topical relations in resources like WordNet has been dubbed as the *tennis problem* (Roger Chaffin, cited in Fellbaum, 1998). Some of these links have been introduced more recently in WordNet via the *domain* relation. Yet their number remains still very small. For instance, WordNet 2.1 does not contain any of the three associations mentioned here above, despite their high frequency.

The lack of systematicity of these topical relations makes their extraction and typing very difficult on a large scale. This is why some researchers have proposed to use automatic learning techniques to extend lexical networks like WordNet. In (Harabagiu & Moldovan, 1998), this was done by extracting topical relations from the glosses associated to the synsets. Other researchers used external sources: Mandala *et al.* (1999) integrated co-occurrences and a thesaurus to WordNet for query expansion; Agirre *et al.* (2001) built topic signatures from texts in relation to synsets; Magnini and Cavagliá (2000) annotated the synsets with Subject Field Codes. This last idea has been taken up and extended by (Avancini *et al.*, 2003) who expanded the domains built from this annotation.

Despite the improvements, all these approaches are limited by the fact that they rely too heavily on WordNet and some of its more sophisticated features (such as the definitions associated with the synsets). While often being exploited by acquisition methods, these features are generally lacking in similar lexico-semantic networks. Moreover, these methods attempt to learn *topical knowledge* from a lexical network rather than *topical relations*. Since our goal is different, we have chosen not to rely on any significant resource, all the more as we would like our method to be applicable to a wide array of languages. In consequence, we took an incremental approach (Ferret, 2006): starting from a network of lexical co-occurrences[7] collected from a large corpus, we used these latter to select potential topical relations by using a topical analyzer.

4.2 From a network of co-occurrences to a set of Topical Units

We start by extracting lexical co-occurrences from a corpus to build a network. To this end we follow the method introduced by (Church and Hanks, 1990), *i.e.* by sliding a window of a given size over some texts. The parameters of this extraction were set in such a way as to catch the most obvious topical relations: the window was fairly large (20-words wide), and while it took text boundaries into account, it ignored the order of the co-occurrences. Like (Church and Hanks, 1990), we used mutual information to measure the cohesion between two words. The finite size of the corpus allows us to normalize this measure in line with the maximal mutual information relative to the corpus.

This network is used by TOPICOLL (Ferret, 2002), a topic analyzer, which performs simultaneously three tasks, relevant for this goal:

- it segments texts into topically homogeneous segments;
- it selects in each segment the most representative words of its topic;

[7] Such a network is only another view of a set of co-occurrences: its nodes are the co-occurrent words and its edges are the co-occurrence relations.

- it proposes a restricted set of words from the co-occurrence network to expand the selected words of the segment.

These three tasks rely on a common mechanism: a window is moved over the text to be analyzed in order to limit the focus space of the analysis. This latter contains a lemmatized version of the text's plain words. For each position of this window, we select only words of the co-occurrence network that are linked to at least three other words of the window (see Figure 3). This leads to select both words that are in the window (first order co-occurrents) and words coming from the network (second order co-occurrents). The number of links between the selected words of the network, called *expansion words*, and those of the window is a good indicator of the topical coherence of the window's content. Hence, when their number is small, a segment boundary can be assumed. This is the basic principle underlying our topic analyzer.

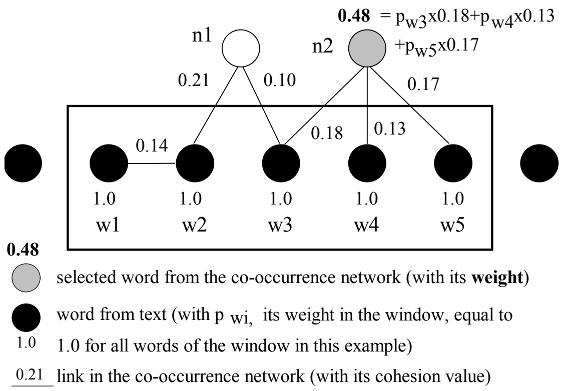
0.48 = p_w3x0.18+p_w4x0.13
+p_w5x0.17
n1
n2
0.21
0.10
0.17
0.14
0.18
0.13
1.0
1.0
1.0
1.0
1.0
w1
w2
w3
w4
w5
0.48
selected word from the co-occurrence network (with its **weight**)
word from text (with p wi, its weight in the window, equal to
1.0
1.0 for all words of the window in this example)
0.21
link in the co-occurrence network (with its cohesion value)

Figure 3: Selection and weighting of words from the co-occurrence network

The words selected for each position of the window are summed, to keep only those occurring in 75% of the positions of the segment. This allows reducing the number of words selected from non-topical co-occurrences. Once a corpus has been processed by TOPICOLL, we obtain a set of segments and a set of expansion words for each one of them. The association of the selected words of a segment and its expansion words is called a Topical Unit. Since both sets of words are selected for reasons of topical homogeneity, their co-occurrence is more likely to be a topical relation than in our initial network.

4.3 Filtering of Topical Units

Before recording the co-occurrences in the Topical Units built in this way, the units are filtered twice. The first filter aims at discarding heterogeneous Topical Units, which can arise as a side effect of a document whose topics are so intermingled that it is impossible to get a reliable linear segmentation of the text. We consider that this occurs when for a given text segment, no word can be selected as a representative of the topic of the segment. Moreover, we only keep the Topical Units that contain at least two words from their original segment. A topic is defined here as a configuration of words. Note that the identification of such a configuration cannot be based solely on a single word.

Text words	Expansion words
surveillance (*watch*)	police_judiciaire (*judiciary police*)
téléphonique (*telephone*)	écrouer (*to imprison*)
juge (*judge*)	garde_à_vue (*police custody*)
policier (*policeman*)	écoute_téléphonique (*phone tapping*)
brigade (*squad*)	juge_d'instruction (*examining judge*)
enquête (*investigation*)	contrôle_judiciaire (*judicial review*)
placer (*to put*)	

Table 1: Content of a filtered Topical Unit

The second filter is applied to the expansion words of each Topical Unit to increase their topical homogeneity. The principle of the filtering of these words is the same as the principle of their selection described in Section 4.2: an expansion word is kept if it is linked in the co-occurrence network to at least three text words of the Topical Unit. Moreover, a selective threshold is applied to the frequency and the cohesion of the co-occurrences supporting these links: only co-occurrences whose frequency and cohesion are respectively higher or equal to 15 and 0.15 are used. For instance in Table 1, which shows an example of a Topical Unit after its filtering, *écrouer* (*to imprison*) is selected, because it is linked in the co-occurrence network to the following words of the text:

juge (judge): 52 (frequency) – 0.17 (cohesion)
policier (policeman): 56 – 0.17
enquête (investigation): 42 – 0.16

word	freq.	word	freq.	word	freq.	word	freq.
scène (*stage*)	884	théâtral (*dramatic*)	62	cynique (*cynical*)	26	scénique (*theatrical*)	14
théâtre (*theater*)	679	scénariste (*scriptwriter*)	51	miss (*miss*)	20	Chabol (*Chabol*)	13
réalisateur (*director*)	220	comique (*comic*)	51	parti_pris (*bias*)	16	Tchekov (*Tchekov*)	13
cinéaste (*film-marker*)	135	oscar (*oscar*)	40	monologue (*monolog*)	15	allocataire (*beneficiary*)	13
comédie (*comedy*)	104	film_américain (*american film*)	38	revisiter (*to revisit*)	14	satirique (*satirical*)	13
costumer (*to dress up*)	63	hollywoodien (*Hollywood*)	30	gros_plan (*close-up*)	14		

Table 2: Co-occurrents of the word *acteur* (*actor*) with a cohesion of 0.16
(the co-occurrents removed by our filtering method are underlined)

4.4 From Topical Units to a network of topical relations

After the filtering, a Topical Unit gathers a set of words supposed to be strongly coherent from the topical point of view. Next, we record the co-occurrences between these words for all the Topical Units remaining after filtering. Hence, we get a large set of topical co-occurrences, despite the fact that a significant number of non-topical co-occurrences remains, the filtering of Topical Units being an unsupervised process. The frequency of a co-occurrence in this case is given by the number of Topical Units containing both words simultaneously. No distinction concerning the origin of the words of the Topical Units is made.

The network of topical co-occurrences built from Topical Units is a subset of the initial network. However, it also contains co-occurrences that are not part of it, *i.e.* co-occurrences that were not extracted from the corpus used for setting the initial network or co-occurrences whose frequency in this corpus was too low. Only some of these "new" co-occurrences are topical. Since it is difficult to estimate globally which ones are interesting, we have decided to focus our attention only on the co-occurrences of the topical network already present in the initial network.

Thus, we only use the network of topical co-occurrences as a filter for the initial co-occurrence network. Before doing so, we filter the topical network in order to discard co-occurrences whose frequency is too low, that is, co-occurrences that are unstable and not representative. From the use of the final network by TOPICOLL (see Section 4.5), we set the threshold experimentally to 5. Finally, the initial network is filtered by keeping only co-occurrences present in the topical network. Their frequency and cohesion are taken from the initial network. While the frequencies given by the topical network are potentially interesting for their topical significance, we do not use them because the results of the filtering of Topical Units are too hard to evaluate.

4.5 Results and evaluation

We applied the method described here to an initial co-occurrence network extracted from a corpus of 24 months of *Le Monde*, a major French newspaper. The size of the corpus was around 39 million words. The initial network contained 18,958 words and 341,549 relations. The first run produced 382,208 Topical Units. After filtering, we kept 59% of them. The network built from these Topical Units was made of 11,674 words and 2,864,473 co-occurrences. 70% of these co-occurrences were new with regard to the initial network and were discarded. Finally, we got a filtered network of 7,160 words and 183,074 relations, which represents a cut of 46% of the initial network. A qualitative study showed that most of the discarded relations are non-topical. This is illustrated by Table 2, which gives the co-occurrents of the word *acteur (actor)* that are filtered by our method among its co-occurrents with a high cohesion (equal to 0.16). For instance, the words *cynique (cynical)* or *allocataire (beneficiary)* are cohesive co-occurrents of the word *actor*, even though they are not topically linked to it. These words are filtered out, while we keep words like *gros_plan (close-up)* or *scénique (theatrical)*, which topically cohere with *acteur (actor)* despite their lower frequency than the discarded words.

	Recall[8]	**Precision**	**F1-measure**	**Error (P_k)**[9]
initial (I)	0.85	0.79	0.82	0.20
topical filtering (T)	0.85	0.79	0.82	0.21
frequency filtering (F)	0.83	0.71	0.77	0.25

Table 3: TOPICOLL's results with different networks

In order to evaluate more objectively our work, we compared the quantitative results of TOPICOLL with the initial network and its filtered version. The evaluation showed that the performance of the segmenter remains stable, even if we use a topically filtered network (see Table 3). Moreover, it became obvious that a network filtered only by frequency and cohesion performs significantly less well, even with a comparable size. For testing the statistical significance of these results, we applied to the P_k values a one-side t-test with a null hypothesis of equal means. Levels lower or equal to 0.05 are considered as statistically significant:

p_{val} (I-T): 0.08
p_{val} (I-F): 0.02
p_{val} (T-F): 0.05

These values confirm that the difference between the initial network (I) and the topically filtered one (T) is actually not significant, whereas the filtering based on co-occurrence frequencies leads to significantly lower results, both compared to the initial network and the topically filtered one. Hence, one may conclude that our method is an effective way of selecting topical relations by preference.

5 Discussion and conclusion

We have raised and partially answered the question of how a dictionary should be indexed in order to support word access, a question initially addressed in (Zock, 2002) and (Zock and Bilac, 2004). We were particularly concerned with the language producer, as his needs (and knowledge at the onset) 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 we have suggested to label the links, as this would reduce the graph complexity and allow for type-based navigation. Actually our basic proposal is to extend a resource like WordNet by adding certain links, in particular on the syntagmatic axis. 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, 1995), Fillmore's FRAMENET[10], work on ontologies (CYC), thesaurus (Roget), WordNets (the original version from Princeton, various Euro-WordNets, BalkaNet), HowNet[11], the work done by MICRA, the FACTOTUM project [12], or the Wordsmyth dictionary/thesaurus[13].

Since words are linked via associations, it is important to reveal these links. Once this is done, words can be accessed by following these links. We have presented here some preliminary work for extracting an important subset of such links from texts, topical associations, which are generally absent from dictionaries or resources like WordNet. An evaluation of the topic segmentation has shown that the relations extracted are sound from the topical point of view, and that they can be extracted automatically. However,

[8] Precision is given by N_t / N_b and recall by N_t / D, with D being the number of document breaks, N_b the number of boundaries found by TOPICOLL and N_t the number of boundaries that are document breaks (the boundary should not be farther than 9 plain words from the document break).

[9] P_k (Beeferman et al., 1999) evaluates the probability that a randomly chosen pair of words, separated by k words, is wrongly classified, *i.e.* they are found in the same segment by TOPICOLL, while they are actually in different ones (miss of a document break), or they are found in different segments, while they are actually in the same one (false alarm).

[10] http://www.icsi.berkeley.edu/~framenet/

[11] http://www.keenage.com/html/e_index.html

[12] http://humanities.uchicago.edu/homes/MICRA/

[13] http://www.wordsmyth.com/

they still contain too much noise to be directly exploitable by an end user for accessing a word in a dictionary. One way of reducing the noise of the extracted relations would be to build from each text a representation of its topics and to record the co-occurrences in these representations rather than in the segments delimited by a topic segmenter. This is a hypothesis we are currently exploring. While we have focused here only on word access on the basis of (other) words, one should not forget that most of the time speakers or writers start from meanings. Hence, we shall consider this point more carefully in our future work, by taking a serious look at the proposals made by Bilac et al. (2004); Durgar and Oflazer (2004), or Dutoit and Nugues (2002).

References

Eneko Agirre, Olatz Ansa, David Martinez and Eduard Hovy. 2001. Enriching WordNet concepts with topic signatures. In *NAACL'01 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*.

Henri Avancini, Alberto Lavelli, Bernardo Magnini, Fabrizio Sebastiani and Roberto Zanoli. 2003. Expanding Domain-Specific Lexicons by Term Categorization. In *18th ACM Symposium on Applied Computing (SAC-03)*.

Doug Beeferman, Adam Berger and Lafferty. 1999. Statistical Models for Text Segmentation. *Machine Learning*, 34(1): 177-210.

Slaven Bilac, Wataru Watanabe, Taiichi Hashimoto, Takenobu Tokunaga and Hozumi Tanaka. 2004. Dictionary search based on the target word description. In *Tenth Annual Meeting of The Association for Natural Language Processing (NLP2004)*, pages 556-559.

Roger Brown and David McNeill. 1996. The tip of the tongue phenomenon. *Journal of Verbal Learning and Verbal Behaviour*, 5: 325-337.

Kenneth Church and Patrick Hanks. 1990. Word Association Norms, Mutual Information, And Lexicography. *Computational Linguistics*, 16(1): 177-210.

Ilknur Durgar El-Kahlout and Kemal Oflazer. 2004. Use of Wordnet for Retrieving Words from Their Meanings, In *2nd Global WordNet Conference*, Brno

Dominique Dutoit and Pierre Nugues. 2002. A lexical network and an algorithm to find words from definitions. In *15th European Conference on Artificial Intelligence (ECAI 2002)*, Lyon, pages 450-454, IOS Press.

Christiane Fellbaum. 1998. *WordNet - An Electronic Lexical Database*, MIT Press.

Olivier Ferret. 2006. Building a network of topical relations from a corpus. In *LREC 2006*.

Olivier Ferret. 2002. Using collocations for topic segmentation and link detection. In *COLING 2002*, pages 260-266.

Sanda M. Harabagiu, George A. Miller and Dan I. Moldovan. 1999. WordNet 2 - A Morphologically and Semantically Enhanced Resource. In *ACL-SIGLEX99: Standardizing Lexical Resources*, pages 1-8.

Sanda M. Harabagiu and Dan I. Moldovan. 1998. Knowledge Processing on an Extended WordNet. *In WordNet - An Electronic Lexical Database*, pages 379-405.

Philip Humble. 2001. *Dictionaries and Language Learners*, Haag and Herchen.

William Levelt, Ardi Roelofs and Antje Meyer. 1999. A theory of lexical access in speech production, *Behavioral and Brain Sciences*, 22: 1-75.

Bernardo Magnini and Gabriela Cavagliá. 2000. Integrating Subject Field Codes into WordNet. In *LREC 2000*.

Rila Mandala, Takenobu Tokunaga and Hozumi Tanaka. 1999. Complementing WordNet with Roget's and Corpus-based Thesauri for Information Retrieval. *In EACL 99*.

Igor Mel'čuk, Arno Clas and Alain Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*, Louvain, Duculot.

George A. Miller. 1995. WordNet: A lexical Database, *Communications of the ACM*. 38(11): 39-41.

Stephen D. Richardson, William B. Dolan and Lucy Vanderwende. 1998. MindNet: Acquiring and Structuring Semantic Information from Text. In *ACL-COLING'98*, pages 1098-1102.

Piek Vossen. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publisher.

Gabriella Vigliocco, Antonini, T., and Merryl Garrett. 1997. Grammatical gender is on the tip of Italian tongues. *Psychological Science*, 8: 314-317.

Michael Zock. 2002. Sorry, what was your name again, or how to overcome the tip-of-the tongue problem with the help of a computer? In *SemaNet workshop, COLING 2002*, Taipei. http://acl.ldc.upenn.edu /W/W02/W02-1118.pdf

Michael Zock and Slaven Bilac. 2004. Word lookup on the basis of associations: from an idea to a roadmap. In *COLING 2004 workshop: Enhancing and using dictionaries*, Geneva. http://acl.ldc.upenn.edu/ coling2004/W10/pdf/5.pdf