Evolutionary basic notions for a thematic representation of general knowledge

Alain Joubert, Mathieu Lafourcade
LIRMM – UM2
Laboratoire d'Informatique, Robotique et Microélectronique de Montpellier
E-mail : {alain.joubert, lafourcade}@lirmm.fr

Abstract

In the field of Natural Language Processing, in order to work out a thematic representation system of general knowledge, methods relying on thesaurus have been used for about twenty years. A thesaurus consists of a set of concepts which define a generating system of a vector space modelling general knowledge. These concepts, often organized in a treelike structure, constitute a fundamental, but completely fixed tool. Even if the concepts evolve (we think for example of the technical fields), a thesaurus as for it can evolve only at the time of a particularly heavy process, because it requires the collaboration of human experts. After detailing the characteristics which a generating system of the vector space of knowledge modelling must have, we define the "basic notions". Basic notions, whose construction is initially based on the concepts of a thesaurus, constitute another generating system of this vector space. We then approach the determination of the acceptions expressing the basic notions. Lastly, we clarify how, being freed from the concepts of the thesaurus, the basic notions evolve progressively with the analysis of new texts by an iterative process.

1. Introduction: thematic representation of general knowledge

(Lafourcade and Sandford 1999) then (Lafourcade 2001) developed a system of thematic representation[1] of general knowledge. This system is based on a vector representation which is initially built from the hierarchy of concepts of the Larousse thesaurus (Larousse 1999). This thesaurus initially defines the extent of general knowledge: we consider that a term belongs to general knowledge when it belongs to one of the non-specialized dictionaries (ex: Larousse or Robert for French language). The use of a vector space for modelling has existed for a long time in Information Retrieval, for example (Salton and MacGill 1983). The thesaurus we used features a tree-like structure; it contains 873 leaf concepts. These concepts form a generative system of the vector space named C_{873}; this space constitutes a modelling of our thematic representation system of general knowledge. This vector approach, based on a set of predetermined concepts, is the one that (Chauché 1990) recommended.

The concepts defining the generative system constitute the fundamental elements in representing knowledge. The purpose of a generative system is thus to be able to represent the maximum of general knowledge in a minimum number of vectors. Do the 873 concepts of the thesaurus used, given a priori, establish the "best" generative system? This question is all the more legitimate that other general thesauruses were published, for example the Roget Thesaurus (Kipfer 2001) .

The 873 concepts used do not constitute a base of the C_{873} vector space. Indeed, we can notice that it is possible to find relations between some of these 873 generative vectors, and thus it is obvious that the dimension of the C_{873} space is smaller than 873. However, for reasons of simplicity, the decomposition of an acception remains unique in the considered thesaurus. This interdependence between concepts has sometimes been processed, as for example in the LSA model (Deerwester and al. 1990). What is the dimension of the vector space of thematic modelling of general knowledge? It is impossible to answer this question. It depends in particular on the degree of precision wanted in this modelling: the more precision we want, the more concepts we need to distinguish similar (and thus nearby) notions.

Every acception, that is, every meaning of a term, is thus represented on C_{873} by a unique vector, called abstract vector (Schwab 2005): each of the 873 components of an abstract vector represents the intensity of one of the 873 generative ideas. The abstract vectors are at the moment built from the hierarchy of a thesaurus, thus totally fixed. This construction is achieved by automatic learning from various sources: dictionaries, lists of synonyms or antonyms … For example, in the case of a definition of an acception, we combine the abstract vectors of the various acceptions that we meet in the text of this definition to form the abstract vector of the defined acception. Figure 1 shows the 873 components of the abstract vector for the term *'Peace'* .

The 873 concepts of the Larousse thesaurus establish a generative system. This system is not unique! We can consider that the generative system of the Larousse thesaurus corresponds in fact to one of a typical individual. Can we find a "better" generative system of the vector space modelling the thematic representation of knowledge than the generative system defined by the concepts of the used thesaurus? What is the meaning of "better"?

[1] We only consider thematic relations; "transverse" relations between objects not belonging to the same sub-tree of the hierarchy are not taken into account.

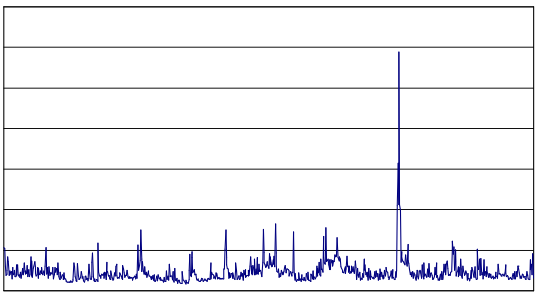

Figure 1: The 873 components of the abstract vector for the term *'Peace'*.

In the thesaurus (Larousse 1999), the concept *652_PAIX* exists: so, this component has here a dominating value.

2. What are the criteria of a good generative system?

It could be pretentious to assert what a good generative system of a vector space modelling our thematic representation of general knowledge would be. It is however possible to clarify several minimal characteristics that such a system must have. This is close to the work of (Wilks 1977).

2.1 Extent: it is a generative system

A generative system has to cover the whole non-specialized domain (because it is the domain considered here). What do we mean by "the whole non-specialized domain"? Even if it is impossible to define the limits of the domain in a clear way, we can consider that the terms belonging to common non-specialized dictionaries give us a reasonable idea of what these limits are. We cannot however conclude that any new acception that could be expressed according to already existing acceptions can be considered as being part of the domain because a lot of specialized acceptions can be expressed with non-specialized terms.

2.2 Representativeness: it is a generative system close to a base

A generative system should distinguish, in a minimum number of concepts, a maximum of "common" notions: our experiments show that a thousand of concepts seem sufficient to be such a system. The generative system, even if it is not a base of the representation space (can we really verify it?), must be relatively close to it.

2.3 Evolution capacities: it is not a fixed system

A thesaurus is mainly criticized for its fixed character. A generative system must be able to take into account the evolution of notions easily. Furthermore, a generative system, like a thesaurus, cannot a priori be exhaustive: it seems difficult, even impossible, to define the limits of the considered domain. This shows that the generative system has to be evolutionary. It will be necessary however to make sure that these abilities to evolve do not work against the representativeness.

3. Method: definition and evolution of the basic notions

3.1 Definition of the basic notions

In an incremental way, as texts are analysed, we meet words. Let k be the number of different terms met. Most of the terms being polysemous, these different k terms have k' acceptions and thus generate k' vectors on the C_{873} vector space. Furthermore, it is necessary to balance these k' vectors according to how often they appear in the texts studied. This weighting function is inevitably an increasing function: the more frequent an acception is, the more important it is for our general knowledge. But connecting proportionally this importance with the frequency of appearance seems too strict for relatively infrequent terms. Numerous phenomena in physics are measured according to logarithmic scales (for example: decibels in acoustics, Richter scale for earthquakes, brightness of stars in the Greek antiquity…). We decided to use a logarithmic function.

It also seems reasonable to weight every occurrence of these k' vectors according to the depth of the corresponding acception in the syntactic analysis tree of the text studied; indeed, it is logical to think that the more a word is "lost" in the depths of a sentence, the less important it is for the global meaning of this sentence[2]. It seems natural that this weighting function, inevitably decreasing, should be a negative exponential function of depth.

In a more formal way, if we call p the depth of a term t in the tree of syntactic analysis of the studied text, the norm of the vector v representing the corresponding acception (after disambiguation between the possible various acceptions of t) is:

$$\|v\| = e^{-\alpha p}$$ where α is a weighting coefficient.

The ith occurrence of this acception a of the term t will be represented on C_{873} by a vector $v_{a,i}$ the norm of which will be:

$$\|v_{a,i}\| = e^{-\alpha p_{a,i}}$$ where $p_{a,i}$ indicates the depth in the analysis tree of the ith occurrence of the acception a.

If we want to take into account the occurrences of the acceptions, we should think about a (logarithmic) summation of the different vectors representing every occurrence of a same acception. As a result, we find that the vector v_a representing the acception a on all the texts processed will have as norm:

$$\|v_a\| = \mathrm{Log}\left(\Sigma_i \, e^{\|v_{a,i}\|} \right).$$

In order to simplify, and if we want to take into account the thematic closeness[3], the k' vectors v_a obtained can be

[2] Of course, the importance of a word also depends on its function in the sentence, and its depth in the tree of syntactic analysis is only a criterion of the importance of the word in question.

[3] It is possible to measure this thematic closeness by means of the angular distance between vectors: this one is based on the

grouped together in n clusters, with n << k'. The n barycentric vectors of these n clusters form a generative system of a vector space B_n. So, we group together the terms thematically close to a same concept. Although the method is different, our aim resembles the one developed by (Landauer and Dumais 1997) for the Latent Semantic Analysis (LSA) method. By considering a relatively large number n, probably from several hundreds to a few thousands, this B_n space can almost be confused with C_{873}, provided that the texts studied sweep all the general knowledge.

What do these n vectors correspond to? They are the "basic notions" deducted from the analysis of texts. Figure 2 illustrates the principle of our methodology.

3.2 Labelling the basic notions

For each of these n basic notions, called b_i, it is possible to find the closest term by using the notion of thematic distance defined on C_{873}. Providing that a term close enough of b_i exists, it will express this basic notion. In fact, it is not a question of finding the closest term, but the closest acception of b_i. The question of what we should name the cluster has to be asked to avoid any risk of ambiguity. Furthermore, it is essential to take into account the frequency of the candidate acception to name the basic notion: can a rare acception be a good concept? In the opposite, the most frequent acception, or more exactly the one the vector of which has the biggest norm in the cluster (see fig.2), is not necessarily the 'best' concept, if it is relatively far from the barycenter of the cluster of vectors which it is supposed to represent.

3.3 Evolution capacities of the system

As new texts are analysed, the clusters of vectors evolve: it is "the evolution of the notions", with a possible phenomenon of differentiating clusters or grouping them together. This is to be set against the definition of the 873 basic concepts which are unchangeable. We can notice that these evolutionary basic notions are initially expressed according to the 873 fixed concepts. From the generative system of the fixed concepts, we built the generative system of the basic notions. Any acception can now be expressed in the space of the basic notions; therefore, any new acception expressed according to an already existing acception (which is the case of the definitions from dictionaries) can be expressed in the space of the basic notions. So, the thesaurus concepts are used for initializing the process. After a first definition of the basic notions, it is possible to break free from concepts, because any new acception can be decomposed on the space generated by the basic notions.

The process of evolution includes two stages:

1°/ evolution of the basic notions, that is, evolution of the generative system:

Every introduction of a new acception modifies the cluster to which it belongs. The barycenter of this cluster is thus moved, which leads to a modification of the vector of the corresponding basic notion. In some cases, the introduction of a new acception may have such an influence on the cluster that it can split in more restricted clusters, or on the contrary regroup with a nearby cluster.

2°/ modification of the coordinates of the acception vectors for which the components on the vectors of the modified basic notions is not equal to zero:

After modifying the vector of a basic notion, it is essential to go through the existing acception vectors and to modify the coordinates of those for which the projection on the modified basic notion vector is not equal to zero: it is a classical change of mark. Indeed, during the revision of a vector of acception, we revise not only the intensity of its components, but also its components themselves, namely the basic notions on which it leans.

Figure 3 shows this iterative principle: every introduction of a new acception (or every new definition of an already existing acception) generates an iteration leading to an evolution of the generative system of the space of the basic notions. This updating, even if it can be relatively long, is totally automatic. It is not the case with a fixed thesaurus defined a priori that can be modified only by a human expert. In this last case, the first stage, completely 'manual', is actually much heavier; the second stage, which is automatic, is similar to the one realized in the space of the basic notions.

4. Conclusion

The modelling by a vector space for representating thematically general knowledge has been used for about twenty years. The definition of the generative system of this space usually rests on the hierarchy of a thesaurus: it is a fundamental, but totally fixed structure. Evolution of knowledge requires a generative system which can evolve. That is why we defined the basic notions. Their construction initially relies on the skeleton which constitutes the concepts of a thesaurus. The appearance of acceptions allows the evolution of the generative system of the space of knowledge representation, as new texts are analyzed. This evolutionary definition of the basic notions totally frees itself from the fixed hierarchy of the concepts. It allows to contemplate applications generated by the evolution of notions.

---

notion of similarity between vectors, often used in information retrieval, for example (Baeza-Yates and Ribeiro-Neto 1999).

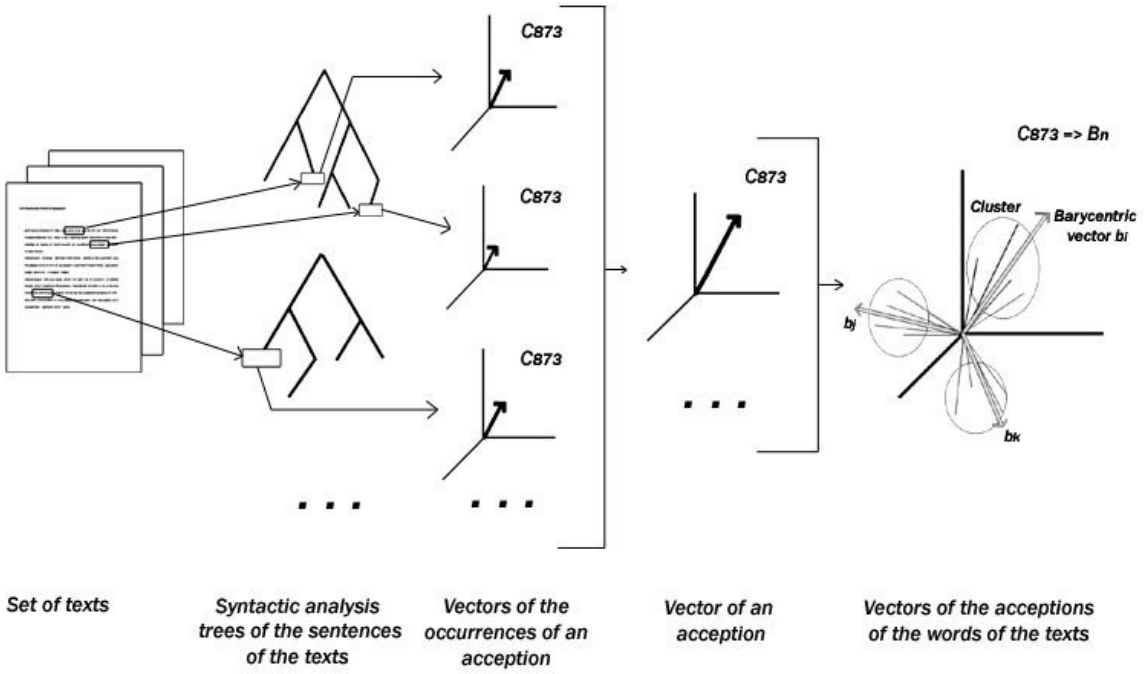

Fig.2: this diagram illustrates the method we used to define the basic notions.

Studying the texts allows to get the trees of morpho-syntactic analysis of the sentences constituting these texts. This analysis is conducted thanks to the tool SYGFRAN (for the French language) developed with SYGMART (Chauché 1984). A semantic analysis is then carried out and every occurrence of an acception gives a vector on C_{873} whose norm is function of its depth in the tree. By adding (logarithmically) the vectors of the various occurrences of a same acception, we get the vector corresponding to this acception on C_{873}. The vectors of various acceptions can be grouped together in clusters the barycenters of which define the basic notions.

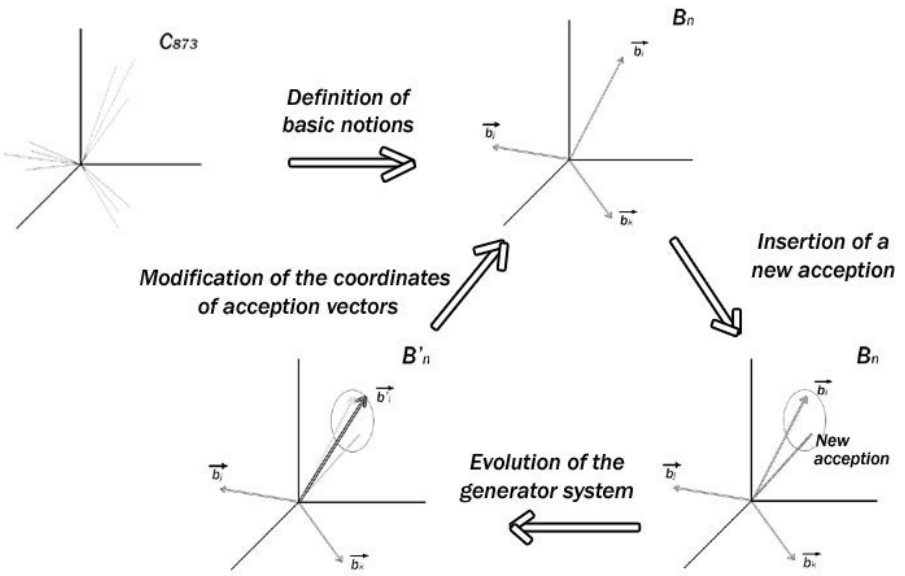

Fig.3: this diagram which recapitulates the definition of the basic notions, illustrates the iterative structure allowing their evolution after introducing a new acception.

5. References

Baeza-Yates R., Ribeiro-Neto B. (1999). *Modern Information Retrieval*, Addison Wesley Longman, 1999, 514 p.

Chauché J. (1984) "Un outil multidimensionnel de l'analyse du discours", *Proceedings of the 22nd conference on Association for Computational Linguistics*, Stanford California, 11-15.

Chauché J. (1990) "Détermination sémantique en analyse structurelle : une expérience basée sur une définition de distance", *TA Information*, vol. 31, n°1, 17-24.

Deerwester S., Dumais S., Landauer T., Fumas G., Harshman R. (1990). Indexing by latent semantic analysis, *Journal of the American Society of Information Science*, 416 (6), 391-407.

Kipfer B.A. (2001) *Roget's International Thesaurus*, sixth edition, Harper Resource (First Edition : 1852).

Lafourcade M., Sandford E. (1999) "Analyse et désambiguïsation lexicale par les vecteurs sémantiques", *TALN'1999*, Cargèse, France, 351-356.

Lafourcade M. (2001) "Lexical sorting and lexical transfer by conceptual vectors", *Proc. of the First International Workshop on Multimedia Annotation (MMA'2001)*, Tokyo.

Landauer T., Dumais S. (1997) "A solution to Plato's problem : The latent semantic analysis theory of acquisition, induction and representation of knowledge", *Psychological Review*, 104(2), 211-240.

Larousse (1999) *Thésaurus Larousse – des idées aux mots, des mots aux idées*, Larousse.

Salton G., MacGill M.J. (1983) *Introduction to Modern Information Retrieval*, McGraw-Hill, New York.

Schwab D. (2005) "Approche hybride -lexicale et thématique- pour la modélisation, la détection et l'exploitation des fonctions lexicales en vue de l'analyse sémantique de textes", Thèse de doctorat, Université de Montpellier II.

Wilks Y. (1977). Good and bad arguments about semantic primitives, *Communication and Cognition,* 181-221.