Conceptual Vector Learning
Comparing Bootstrapping from a Thesaurus or Induction by Emergence

Mathieu Lafourcade

LIRMM
161, rue ADA – 34392 MONTPELLIER Cedex 5
FRANCE.
lafourca@lirmm.fr

Abstract

In the framework of the Word Sense Disambiguation (WSD) and lexical transfer in Machine Translation (MT), the representation of word meanings is one critical issue. The conceptual vector model aims at representing thematic activations for chunks of text, lexical entries, up to whole documents. Roughly speaking, vectors are supposed to encode *ideas* associated to words or expressions. In this paper, we first expose the conceptual vectors model and the notions of semantic distance and contextualization between terms. Then, we present in details the text analysis process coupled with conceptual vectors, which is used in text classification, thematic analysis and vector learning. The question we focus on is whether a thesaurus is really needed and desirable for bootstrapping the learning. We conducted two experiments with and without a thesaurus and are exposing here some comparative results. Our contribution is that dimension distribution is done more regularly by an emergent procedure. In other words, the resources are more efficiently exploited with an emergent procedure than with a thesaurus terms (*concepts*) as listed in a thesaurus somehow relate to their importance in the language but not to their frequency in usage nor to their power of discrimination or *representativeness*.

1. Introduction

In the framework of the Word Sense Disambiguation (WSD) and lexical transfer in Machine Translation (MT), the representation of word meanings is one critical issue. The conceptual vector model aims at representing thematic activations for chunks of text, lexical entries, locutions up to whole documents. Roughly speaking, vectors are supposed to encode *ideas* associated to words or expressions. The main applications of the model are thematic text analysis and lexical disambiguation [Lafourcade 2001]. Such thematic representation is to be used together with more associative information like lexical networks. Conceptual vectors are more on the verge of improving *recall* than lexical networks which focus more on *precision*.

Practically, we have built a system, with automated learning capabilities, based on conceptual vectors and exploiting monolingual dictionaries (available on the web). So far, from French, the system learned around 145000 lexical entries corresponding to roughly 560000 vectors (the average meaning number for polysemous words being 5.3). We are conducting the same experiment for English. The issue of dimensionality in semantic space has been quite debated (see [Lowe 2000] for a theorization of those subjects), but some questions about the qualitative nature of the produced vector space are still largely untackled.

In this paper, we first expose the conceptual vectors model and the notions of semantic distance and contextualization between terms. Then, we present in details the text analysis process coupled with conceptual vectors, which is used (with very small adjustments) in text classification, thematic analysis and vector learning. The question we focus on is whether a thesaurus is really needed and desirable for bootstrapping the learning. We conducted two experiments with and without a thesaurus and are exposing here some comparative results. Our contribution is that dimension distribution is done more regularly by an emergent procedure. In other words, the resources (the vector components) are more efficiently exploited with an emergent procedure than with a thesaurus (this property seems to be independent of the thesaurus structure or concepts set). Key terms (*concepts*) as listed in a thesaurus somehow relate to their importance in the language (either general or of a specific domain), but not to their frequency in usage nor to their power of discrimination or *representativeness*. Corpora based approaches behave the other way, but do not explicitly point out semantic relations between word meanings.

2. Conceptual Vectors

We represent thematic aspects of textual segments (documents, paragraph, syntagms, etc) by conceptual vectors. Vectors have been used in information retrieval for long [Salton and MacGill 1983] and for meaning representation by the LSI model [Deerwester *et al.* 1990] from latent semantic analysis (LSA) studies in psycholinguistics. In computational linguistics, [Chauché 1990] proposes a formalism for the projection of the linguistic notion of semantic field in a vector space, from which our model is inspired.

From a set of elementary concepts, it is possible to build vectors (conceptual vectors) and to associate them to lexical items. Lexical items are words or expressions, which constitute lexical entries. For instance, *car* or *white ant* are lexical items. The hypothesis, we call *thesaurus hypothesis*, that considers a set of concepts as a generator to language has been long described in [Roget, 1852].

Polysemic words combine different vectors corresponding to different meanings. This vector approach is based on known mathematical properties, thus it is possible to undertake well founded formal manipulations attached to reasonable linguistic interpretations.

Concepts are defined from a thesaurus (in our prototype applied to French, we have chosen [Larousse 2001] where 873 concepts are identified to be compared with the thousand defined in [Roget, 1852]).

To be consistent with the thesaurus hypothesis, we consider that this set constitutes a generator family for the words and their meanings. This family is probably not free (no proper vector base) and as such, any word would project its meaning on it according to the following principle.

Let be C a finite set of n concepts, a conceptual vector V is a linear combination of elements c_i of C. For a meaning A, a vector V(A) is the description (in extension) of activations of all concepts of C. For example, the different meanings of *door* could be projected on the following concepts (the concept of INTENSITY are ordered by decreasing values):

V(door) = OPENING {0.8}, BARRIER {0.7}, LIMIT {0.65}, PROXIMITY {0.6}, EXTERIOR {0.4}, INTERIOR {0.39}, ...

In practice, the larger C is, the finer the meaning descriptions are. In return, the computing is less easy: for dense vectors (which are those which have very few null coordinates - in practice, by construction, all vectors are dense) the enumeration of activated concepts is long and difficult to evaluate. We prefer to select the thematically closest terms, i.e., the neighborhood. For instance, the closest terms ordered by increasing distance to door are: V$(door) = portal, opening, gate, barrier, …

2.1. Angular Distance

Let us define *Sim(A,B)* as one of the *similarity* measures between two vectors A et B, often used in information retrieval . We can express this function as below with the " . " as the scalar product. We suppose here that vector components are positive or null. Then, we define an angular *distance* D_A between two vectors.

$$Sim(A,B) = \cos(\widehat{A,B}) = \frac{A \cdot B}{\|A\| \times \|B\|}$$

$$D_A(A,B) = \arccos(Sim(A,B))$$

Intuitively, this function constitutes an evaluation of the *thematic proximity* and measures the angle between the two vectors. We would naively consider that, for a distance $D_A(A,B)$ < pi/4 (45 degrees) A and B are thematically close and share many concepts. For $D_A(A,B)$ > pi/4, the thematic proximity between A and B would be considered as loose. Around pi/2, they have no relation. D_A is a real distance function and it verifies the properties of reflexivity, symmetry and triangular inequality. In the following, we will speak of distance} only when these last properties will be verified, otherwise we will speak of measure. We have, for example, the following angles (values are in radian and degrees).

D_A(V(tit), V(tit))=0 (0)
D_A(V(tit), V(bird))=0.55 (31)
D_A(V(tit), V(sparrow))=0.35 (20)
D_A(V(tit), V(rain))=1.28 (73)
D_A(V(tit), V(insect))=0.57 (32)

The first one has a straightforward interpretation, as a *tit* cannot be closer to anything else than itself. The second and the third are not very surprising since a *tit* is a kind of *sparrow*, which is a kind of *bird*. A *tit* has not much in common with a *train*, which explains a large angle between them.

One can wonder why there is 32 degrees angle between *tit* and insect, which makes them rather close. If we scrutinize the definition of *tit* from which its vector is computed (*Insectivorous passerine bird with colorful feather.*) Perhaps the interpretation of these values seems clearer. In effect, the thematic is by no way an ontological distance.

A less naïve approach is to compare the actual angular distance to the mean distance over the vector space. This is a more practical comparison that is relative to the actual vector population. Anyway, the comparison function by itself has no influence on the conceptual vector construction.

2.2. Conceptual Vector Construction

The conceptual vector construction is based on definitions from different sources (dictionaries, synonym lists, manual indexations, etc). Definitions are parsed and the corresponding conceptual vector is computed. This analysis method shapes, from existing conceptual vectors and definitions, new vectors.

It requires a bootstrap with a kernel composed of pre-computed vectors. This reduced set of initial vectors is manually indexed for the most frequent or difficult terms. It constitutes a relevant lexical items basis on which the learning can start and rely. One way to build a coherent learning system is to take care of the semantic relations between items. Then, after some fine and cyclic computation, we obtain a relevant conceptual vector basis. After 2 and half years (after starting in mid 1999), our system counted more than 71000 items for French and more than 288000 vectors, in which more 20000 items are concerned by relations, like antonymy, for example. These items are either defined through negative sentences, or because antonyms are directly in the dictionary. Example of a negative definition: *non-existence: property of what does not exist.* Example of a definition stating antonym: *love: antonyms: disgust, aversion.*

3. Semantic Text Analysis

The text analysis procedure based on conceptual vectors is independent of the underlying vector space. From a morphosyntactic analysis tree of the text, for each term sense (acception) we associate a vector. If the term is not present in the vector database, then the null vector is used instead. Vectors are then propagated upward and downward on the tree. The upward propagation produces merged vectors on the inner nodes of the tree. The downward propagation adjusts the vector of each node according by the context provided by the vectors of the other nodes. This *weak contextualizatio* is by itself an exploitation of the mutual-information contained in vectors. When reaching a term node, each acception node is weighted non-linearly according to the context. The process is globally convergent, although in ambiguous text with several possible interpretation some vectors may oscillates between several states. *For* example, this is the case with typical sentences like *L'avocat est véreux* (Eng. *the lawyer is corrupted* or *The avocado is worm-eaten*) where both interpretations are equally reasonable (without further context). When applied to term definitions as found in dictionaries, this analysis leads to vector learning. The overall learning process is continuously iterated, each term definition and acception vector being automatically revised periodically. The learning process converges globally in less than 10 cycles.

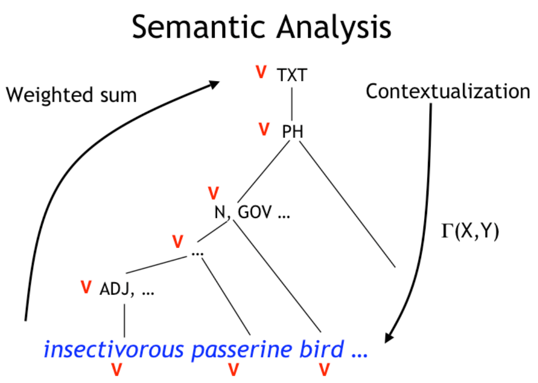


Fig 1 : Semantic analysis with a typical definition of tit as Insectivorous passerine bird with colorful feather

We have undertaken three main experiments. The first one (TH873) is based on the vector space defined in the french thesaurus Larousse, where 873 basic concepts are defined. For bootstrapping the learning process a kernel of roughly 1000 acceptions has been manually indexed on the basis of the thesaurus concepts. The second one (EM873) is done by emergence. No kernel, neither initial concept sets are then needed, but only the dimension of the vectors space is required (the dimension here has been set to 873 for having vector results directly comparable with the TH873 experiment). As no kernel is required, the bootstrapping is induced by randomly generating vectors for unknown terms. These vectors are going to be revised afterward. To keep computed vectors different (and not all converging to a common mean vector), we terminate the computation process of a given vector by an operation, called *amplification* that enhances the *contrast* of the vector. Basically, if the variation coefficient of a vector is extreme (too low or too high), then each component is non-linearly augmented (to a power value over 1) and the vector is then normalized. This process is applied repeatedly until the coefficient variation has a middle value. This process is directly inspired by what is done in photography to augment the contrast of dull pictures. We recognized here, that the most important properties to be achieved when learning vectors is both the coherence between acception vectors (and not their actual component activation) but also the discrimination between them. The third experiment (EM5000) is done by emergence with vector size of 5000.

4. Experiments and results

We found the following comparative results. In TH873 experiment, there is a strong precision induced by the finely crafted concept set issued from the thesaurus, but at the cost of a lack of information sharing. On the other hand, EM873 more evenly distributes the 873 vector components to represent very subtle meaning differences, especially in the vector space region where the lexical density is high. By emergence, the lexical density tends to be more uniform as more components tend to participate (than in TH873). In EM5000, being of a much larger size, vectors describe meanings with much more finesse but at a cost (in space and time). The increase in description is basically logarithmic with the increase in vector size. Globally, from a vector size of 873 to a vector size of 5000, the vector description is increased of (roughly) 33 percent. But, this gain is very significant to discriminate terms than where considered as (quasi) synonymous. For instance, in TH873, the *dragonfly* and *cockroach* have almost identical vectors, although in EM873, they remain quite close although being separable. In the EM5000 experiment, there are quite different and cannot anymore be considered as synonymous. Some assessment with vectors of size 10000 showed that the increase in description quality is negligible (less than 1% percent from EM5000) especially compared to the amount of size occupied.

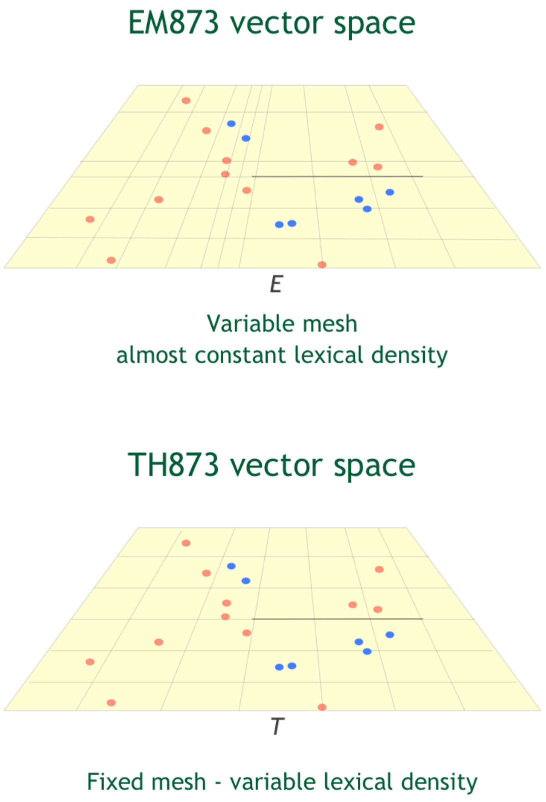
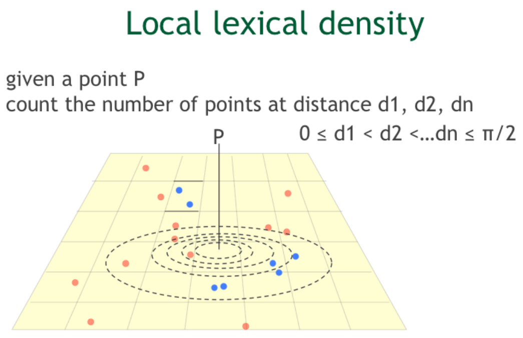

EM873 vector space
E
Variable mesh
almost constant lexical density
TH873 vector space
T
Fixed mesh - variable lexical density

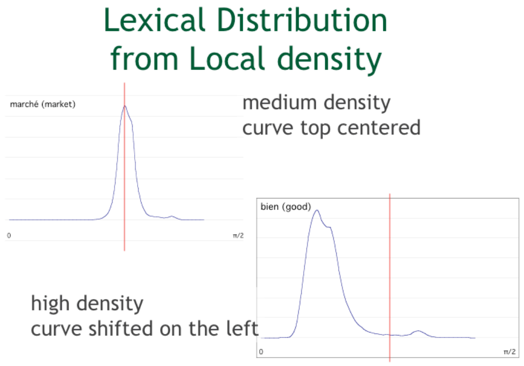
Fig 2 : Schema of a vector space with a fixed set of concepts and with a fixed number of concepts which are computed by emergence



Local lexical density

given a point P
count the number of points at distance d1, d2, dn

Beside a manual evaluation of vectors, done by enumerating and assessing term neighborhood, some functions can globally assess vectors. For example, the evaluation of the lexical concentration gives clues about in increase of vector representation power (in the full paper those functions are detailed with equations). Our conclusion all in all, is that the higher the dimension the better the description both for separating terms that belong to close semantic fields but also for *relating* terms of different semantic fields but might share some relations that could prove being critical for semantic analysis. However, the best ratio between quality and vector size has to be precisely determined and, of course, may depend on application. Our experiments strongly suggest that a vector size around 5000 seems to be a good trade-off between finesse and space for word sense disambiguation and indexation of general texts (like those found in newspapers). Results and lexical data (vectors) of some of our experiments are freely accessible at `<http://www.lirmm.fr/˜lafourcade>` .



Lexical Distribution from Local density

medium density
curve top centered

high density
curve shifted on the left

5. References

[Barrière and Copeck 2001] Barrière C., Copeck T., “Building Domain Knowledge from Specialized Texts”, *TIA 2001*, Nancy, 2001.

[Chauché 1990] Chauché J., “Détermination sémantique en analyse structurelle : une expérience basée sur une définition de distance”, *TA Information*, vol. 31, n_ 1, p. 17-24, 1990.

[Deerwester *et al.* 1990] Deerwester S., Dumais S., Landauer T., Furnas G., Harshman R., “Indexing by latent semantic analysis”, *Journal of the American Society of Information Science*, 416(6), p. 391-407, 1990.

[Hearst 1998] Hearst M.A., “Automated discovery of Wordnet relations”, In C. Fellbaum ed. *Wordnet An Electronic Lexical Database*, MIT Press, Cambridge, MA, p. 131-151, 1998.

[Lafourcade et al. 2001] Lafourcade M., Prince V. and D. Schwab, “Vecteurs conceptuels et structuration émergente de terminologie”, *TAL*, vol 43 - n_ 1, p. 43-72, 2002.

[Lafourcade 2001] Lafourcade M., “Lexical sorting and lexical transfer by conceptual vectors”, *First International Workshop on MultiMedia Annotation* (MMA’2001), Tokyo, 6 p, January 2001.

[Larousse 2001] Larousse, *Thesaurus Larousse - des idées aux mots - des mots aux idées.* Larousse, 1992.

[Lowe 2000] Lowe, W., “Towards a theory of semantic space”, *Proceedings of the 23rd Annual Meeting of the Cognitive Science Society*, Lawrence Erlbaum Associates. pp.576-581.

[Lowe 2000] Lowe,W., “Topographic Maps of Semantic Space”, *PhD Thesis*, institute for Adaptive and Neural Computation, Division of Informatics, Edinburgh University, 2000.

[Resnik 1995] Resnik P., “Using Information contents to evaluate semantic similarity in a taxonomy”, *IJCAI-95*, 1995.

[Riloff and Shepherd 1995] Riloff E., Shepherd J., “A corpus-based bootstrapping algorithm for Semi-Automated semantic lexicon construction”, *Natural Language Engineering*, vol. 5, part. 2, p. 147-156, 1995.

[Roget, 1852] *Thesaurus of English Words and Phrases.* Longman, London, 1852.

[Salton and MacGill 1983] Salton G., MacGill M.J., *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, 1983.

[Salton 1988] Salton G., *Term-Weighting Approaches in Automatic Text Retrieval*, *McGraw-Hill computer science series*, McGraw-Hill, vol. 24, 1988.

[Schwab *et al.* 2002] Schwab D., Lafourcade M., V. Prince, “Antonymy and conceptual vectors.”, In proc. of *COLING’2002*, Taipei, Taiwan, August 2002.

[Sparck Jones 1986] Sparck Jones K., *Synonymy and Semantic Classification*, Edinburgh Information Technology Serie, 1986.

[Ploux et Victorri 1998] Ploux S., Victorri B., “Construction d’espaces sémantiques à l’aide de dictionnaires de synonymes.” *TAL*, vol. 39, n_ 1, p. 161-182, 1998.

[Yarowsky1992] Yarowsky D., “Word-Sense Disambiguation Using Statistical Models of Roget’s Categories Trained on Large Corpora”, *COLING’92*, Nantes, p. 454-460, 1992.