From word sense naming to vocabulary augmentation in Papillon

Fabien JALABERT , Mathieu LAFOURCADE

LIRMM
Laboratoire d'informatique, de Robotique
et de Microélectronique de Montpellier
MONTPELLIER - FRANCE.
{jalabert,lafourca}@lirmm.fr
http://www.lirmm.fr/~{jalabert,lafourca}

Abstract. In the framework of the Papillon project, there are acceptions that are not lexicalized in a given language. They correspond at best to some hypernyms. Moreover, in order to easily supervise a translation process, we would like to be able to name meanings instead of referring to definitions by numbers. Dictionaries define words using a *"genus + differentia"* approach and can be exploited for new compound extraction. This approach is relates for various research efforts in sense naming. The conceptual vector model (CVM) aim to represent meanings in a non lexical way and vectors are calculated through the analyses of multiple dictionary resources. The following article describes how our work on sense naming can be easily applied to the Papillon project and could offer simultaneously a lexical augmentation approach, a disambiguation checking process and a new lexical resource. Using this information, an automatic process helps building a mixed lexical and acception network.

Keywords

Conceptual vectors, meaning representation, vocabulary augmentation, sense naming, Papillon project.

Introduction

In the framework of the Papillon project, the acception base contains items that are not lexicalized in French. For instance, the English words *'giblets'* and *'offal'* have only a common hyperonym in French: *'abats'*. When we want to translate these words in French and distinguish the two meanings, a locution like *'abats de volaille'* (that means *'∼ of fowls'*) and *'abats de porc/boeuf'* (*'∼ of pork/beef'*) would be desirable. So, we would like to recover an associated term that allows generating a new compound entry. Moreover, it would permit an easier checking of a disambiguation process. A word sense tagging system that gives an annotation to each disambiguated word seems highly desirable. For instance *'free'* means *'cost nothing'*, *'not occupied'*, *'familiar'*, *'not captive or tied'*... These senses could be tagged as *free*⟨money⟩, *free*⟨busy⟩, *free*⟨familiar⟩, *free*⟨liberty⟩. Similarly, this checking process is useful during translation from one language into another to identify the designated meaning in the multilingual database. For example, if the process returns *free*⟨money⟩, a language understanding agent can deduce the meaning chosen by the translator, without having to know the proper word in the target language.

In this paper, we describe our work on meaning representation through conceptual vectors and on sense naming. We present some measure that allows to filter and organize possible name candidate for each meaning of polysemeous words. Then, some insights are given toward vocabulary augmentation, i.e. how to create acceptable vocables for unlexicalized acception in a given language (in our case French).

1 Conceptual Vectors

We represent thematic aspects of textual segments (documents, paragraphs, syntagms, etc.) with conceptual vectors. Vectors have been used in information retrieval for long [*Salton & MacGill*, 1983] and for meaning representation by the LSI (Latent Semantic Indexing) model [*Deerwester et al.*, 1990] from latent semantic analysis (LSA) studies in psycholinguistics. In computational linguistics, [*Chauché*, 90] proposed a formalism for the projection of the linguistic notion of semantic field in a vectorial space, from which our model is inspired [*Lafourcade et al.*, 2002]. From a set of elementary notions, dubbed as *concepts*, it is possible to build vectors (conceptual vectors) and to associate them to lexical items. The hypothesis that considers a set of concepts as a generator to language has been long described in [*Rodget*, 1852] (*thesaurus hypothesis*). Polysemous words combine the different vectors corresponding to the different meanings considering several criteria as weights: semantic context, usage frequency, language level, etc. This vector approach, being based on well known and simple mathematical properties, allows 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*, 1992] where 873 concepts are identified to compare with the thousand defined in [*Rodget*, 1852]). To be consistent with the thesaurus hypothesis, we consider that this set constitutes a generator space for the words and their meanings. This space is probably not free (no proper vectorial base) and as such, any word would project its meaning(s) on this space.

1.1 Thematic Projection Principle

Let be $\mathcal{C}$ a finite set of n concepts, a conceptual vector V is a linear combination of elements c_i of $\mathcal{C}$. For a meaning A, a vector $V(A)$ is the description (in extension) of activations of all concepts of $\mathcal{C}$. For example, the different meanings of ‹*quotation*› could be projected on the following concepts (the *CONCEPT*[intensity] are ordered by decreasing values):

V(‹*quotation*›) =
STOCK EXCHANGE[0.7], *LANGUAGE*[0.6], *CLASSIFICATION*[0.52], *SYSTEM*[0.33], *GROUPING*[0.32], *ORGANIZATION*[0.30], *RANK*[0.330], *ABSTRACT*[0.25], ...

In practice, the largest $\mathcal{C}$ is, the finer the meaning descriptions are. In return, the computer manipulation is less easy. It is clear, that for dense vectors the enumeration of the activated concepts is long and difficult to evaluate. We would generally prefer to select the thematically closest terms, i.e., the *neighborhood.* For instance, the closest terms ordered by increasing distance of ‹*quotation*› are:

$\mathcal{V}$(‹*quotation*›) = ‹*management*›, ‹*stock*›, ‹*cash*›, ‹*coupon*›, ‹*investment*›, ‹*admission*›, ‹*index*›, ‹*abstract*›, ‹*stock-option*›, ‹*dilution*›, ...

1.2 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 [*Morin*, 1999] as their normed scalar product. We suppose here that vector components are positive or null. We, then, define an *angular distance* D_A between two vectors A and B as their angle.

$$\begin{aligned} Sim(A,B) = \cos(\widehat{A,B}) &= \frac{A \cdot B}{\|A\| \times \|B\|} \\ D_A(A,B) &= \arccos(Sim(A,B)) \end{aligned} \tag{1}$$

Intuitively, this function constitutes an evaluation of the *thematic proximity* and is the measure of the angle between the two vectors. We would generally consider that, for a distance $D_A(A,B) \leq \frac{\pi}{4}$, (i.e. less than 45 degrees) A and B are thematically close and share many concepts. For $D_A(A,B) \geq \frac{\pi}{4}$, the thematic proximity between A and B would be considered as loose. Around $\frac{\pi}{2}$, they have no relation. D_A is a real distance function. It verifies the properties of reflexivity, symmetry and triangular inequality. We can have, for example, the following angles (values are in degrees):

D_A(‹*profit*›, ‹*profit*›)=0°
D_A(‹*profit*›, ‹*benefit*›)=10°
D_A(‹*profit*›, ‹*finance*›)=19°
D_A(‹*profit*›, ‹*market*›)=28°
D_A(‹*profit*›, ‹*product*›)=32°
D_A(‹*profit*›, ‹*goods*›)=31°
D_A(‹*profit*›, ‹*sadness*›)=65°
D_A(‹*profit*›, ‹*joy*›)=39°

Examples are extracted from http://www.lirmm.fr/˜lafourca

2 Sense Naming

As our environment is multisource to statistically counterbalance imprecision or errors in definitions, a meaning is represented by a global vector calculated from a cluster of definitions. So, we would like to be able to name such a cluster, i.e to name a sense Naming is interresting to supervise the learning process, but also as a new lexical source to be (re)injected in the clusters. For example, consider *’key’* and the following meanings: *’locking device’*, *’winding device’*, *’keyboard component’*, *’vital clue’*, *’explanatory list’*. The neanings are enumerated through the use of *tags* that are contextualized by *key*. Hence, we would consider that the first meaning is named as *key*⟨`locking device`⟩.

For a translation from English into French, the supervisor has to check that each output word corresponds to the proper meaning of the source items. Instead of referring to a particular lexicon with a specific meaning numbering, we would like to induce to meaning by the use of a tag associated to the (polysemous) term. We should recall here that we are strictly motivated by the task at hand, and consider in the context of Papillon that the task is translation. It leads to the fact that of polysemous word where each meaning has the same translation poses no problem. For example, the French word ‹*fenêtre*› can (most of the time) be translated by ‹*window*› whatever the meaning.

Formally, a sense naming process is a function that associates a word T (for *tag*) with a sense S of a word W. The objective is to be able to recover S (supposedly known) with only W and T. For instance, if we choose to name *’free’* using *’liberty’* (resp. *’money’*), then the main objective is to get the proper meaning of ‹*free*› only with *free*⟨`liberty`⟩ (resp. *free*⟨`money`⟩). Our study divides the construction of this function in three steps: (1) extraction of candidate tags, (2) desambiguisation ability evaluation, and (3) association ability evaluation. These processes are focused primarily on precision (opposed to recall) as the main point is to identify as strictly as possible a given word sense.

2.1 Tags Extraction

We first try to extract from lexical resources some candidates as names for a cluster. Enumerating the whole dictionaries is definitively not practical as being time and resource consuming and moreover causes interferences in the result (especially between close co-hyponyms). The difficult part of this process step is to highlight relevant words. The techniques used here are quite classical.

For each definition of a given cluster, we create a set of candidates that are extracted by using a SYGMART analyzer (for identification of syntactic dependancies) and statistical information (term frequency and inverted term frequency). The kept candidates are mostly governors and adjuncts of nominal and verbal groups. Auxiliaries, pronouns, determiners are removed unless they take part of an already identified locution (as *signe du zodiaque* for *poisson*). From another source, we get a synonym list that refers globally to the vocable (for example for *bank* we get *border*, *institution*, etc.). We do also use other sources as the Larousse Thesaurus, etc. Finally, by using anti-dictionaries and frequency lists we filter out the results removing parasitic words like *’be’*, *’have’*, *’do’*, *’action’*, *’noun’*, *’mean’*, etc

For each candidate, a mark is given depending on criteria like the position of the candidate in the definition, its distribution (in dictionaries or in larger corpora), its language level (technical, slang, etc.). As a definition is generally composed by *”genus + differentia”*, the first terms of this definition are mostly best candidates, unless its surface form is *non standard* (passive voice, GNP put in front of sentence etc.).

2.2 Disambiguation Ability Evaluation

In the following formalization, we note the word and the clusters as written in their usage $w = \{w_1, w_2, \cdots, w_p\}$ (for example, $bank = \{bank_1, bank_2, \cdots, bank_p\}$. We note the corresponding conceptual vectors with an arrow like $\overrightarrow{bank_1}, \overrightarrow{bank_2}, \cdots, \overrightarrow{bank_p}$. The global vector of the word *bank* is noted $\overrightarrow{bank}$. We note, a set of the extracted tag candidates as $T = \{T_1, T_2, \cdots, T_p\}$ and $\{t_1, t_2, \cdots t_p\}$.

The disambiguation ability evaluation starts by filtering out tags that may be a source of confusion between two clusters. A strict approach consists in eliminating a tag that appears in more than one set. However, in rare cases, we may end up with empty candidates set. A softer approach consists in keeping the tag only in the set where the conceptual vectors are the closest.

To achieve this goal, we chose a numerical approach based on conceptual vectors. Three new measures have been crafted to identify the proper tags : (1) the *absolute disambiguation margin*, (2) the *relative disambiguation margin* and (3) the *non-sense risk*.

Absolute Margin This measure computes the gap in which the reciprocal function does not risk to associate this tag with another meaning of the polysemous lexical item. Let d_1 be the minimal distance between a meaning t_i of T and meaning w_j of W to be annotated. The d_2 value is minimal distance between a meaning t_i and any meaning of W but with w_j excluded. The absolute margin is defined as:

$$\begin{aligned} d_1 = \min(D_A(\overrightarrow{t_i}, \overrightarrow{w_j})) \quad d_2 = \min(D_A(\overrightarrow{t_l}, \overrightarrow{w_k})) \quad \mathit{and} \quad j \neq k \\ MARGIN_A(W,T) = |d_2 - d_1| \end{aligned} \tag{2}$$

This higher this margin is, the better the probability to find the same association in other lexical resources. The absolute margin does not take into account the distance between a tag and the term. For instance, with two tag meanings t_1, t_2 and their the following values, the absolute margin select t_2 is:

$$d_{1,t_1} = 0.21 \qquad d_{2,t_1} = 0.3 \qquad\qquad d_{1,t_2} = 0.3 \qquad d_{2,t_2} = 0.4$$

$$MARGIN_{A_1} = 0.09 \qquad\qquad MARGIN_{A_2} = 0.10$$

See figure 1 as an example with polysemous word ‹*frégate*›.

Relative Margin We define the relative margin as the ratio between the absolute margin and d_1:

$$MARGIN_R = \frac{MARGIN_A}{d1} \tag{3}$$

With the previous example, the results are:

$$MARGIN_{R_1} = \frac{0.09}{0.21} = 0.428 \qquad MARGIN_{R_2} = \frac{0.1}{0.3} = 0.333$$

The margin R_1 is better than R_2, which means that the tag t_1 disambiguate better than t_2.

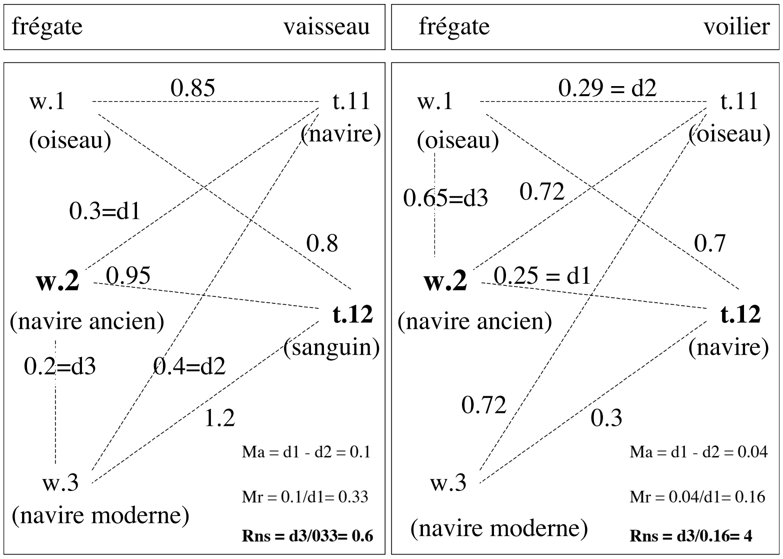

Fig. 1. Example of margin and risk calculation for the word ‹*frégate*› and two tags: ‹*vaisseau*› and ‹*voilier*›. The tag ‹*vaisseau*› is less risky than ‹*voilier*› and would be choosen as annotator for the meaning ‹*frégate.2*›.

Risk of Non Sense Again, the previous measure does not take into account the distance between the various meanings of the word to be anotated and indded it could be a choice factor. The risk of non sense if defined as the ratio between the relative margin and the distance between the meaning to be annotated and the next closest meaning (designated by distance d_2):

$$\begin{aligned} d_3 &= D_A(\overrightarrow{w_j}, \overrightarrow{w_k}) \\ R_{NS} &= \frac{d3}{MARGIN_R} \end{aligned} \tag{4}$$

For instance, for ‹*bar*› the Oxford-Hachette Dictionary defines 3 meanings among others:

- (1) strip of metal or wood
- (2) rod (or pole) used to confine or obstruct in a cell/cage/window
- (3) profession in law context

Suppose that we want to tag the meaning (2) of *'bar'* with *'rod'* or *'cell'* and that we get the following margin results:

$MARGIN_R(\text{‹}\mathit{bar.2}\text{›}, \text{‹}\mathit{rod}\text{›}) = 0.3$ where the first associated meaning is (2) and the second is (1).

$MARGIN_R(\text{‹}\mathit{bar.2}\text{›}, \text{‹}\mathit{cell}\text{›}) = 0.3$ where the first associated meaning is (2) and the second is (3).

We should consider that the first tag is as good as the second. But the meanings may not have the same closeness: $d(1, 2) = 0.3 \qquad d(2, 3) = 0.5$

So, if the disambiguation process makes a mistake, then *'cell'* would be worse than *'strip'* because the last one encompasses more appropriatly both meanings. This last risk measure is used to express the seriousness of an error.

In the context of sense naming, as we focus on precision, we do order all candidates for a given sense by risk decreasing order. The association ability evaluation may alter (or reinforce) this order according to several other criteria.

2.3 Association Ability Evaluation

Several other functions evaluate the multiple candidates and give a corresponding mark. They take into account:

- *The frequency of a candidate in corpora:* a very common term may not characterize and refine the meaning of a term. Otherwise, an unusual word may not be known by the supervisor. The frequency must be adjusted with the word to tag and its context. In a formal document, tag choice should not be the same as in a home page.
- *The co-occurrence between candidate and word is useful:* if this method does not help to disambiguate two senses of a same word, it gives a very good scalable value of the idea associations between a word and its candidates. The system can take into account the context of the co-occurrence in the same principle of the above frequency evaluation. The co-occurrence must consider the distance between to words and the document type and thematic where they appears together.
- *The grammatical form:* to get a more intuitive result, we give priority to terms that have the same morphosyntactic category. For example, *farm:N*$\langle\texttt{building}\rangle$ would be better than *farm:N*$\langle\texttt{build}\rangle$.
- *Occurrence of source word into its tag definition:* The tag has been extracted from definitions of the associated meaning. So, if the annotated item is present in a candidate definition, we cannot consider that the disambiguation is better or not, but the link between the two words is reinforced.

The previous evaluation does not express a disambiguation measure or association between a tag and its associated meaning. It just takes into account the relationship between the two polysemous words. The following evaluator considers the right association between the target and the annotation.

- *Occurrence of tag into rival definition:* If the tag occurs in another meaning definition, then it may not characterize only the associated term. Therefore, the candidate gets a negative mark.

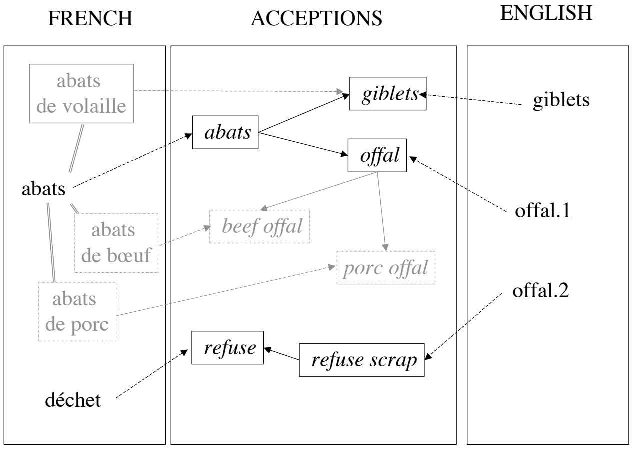

Fig. 2. Example lexical augmentation (items in grey) for english acceptions unlexicalized in french.

3 Lexical augmentation

3.1 Common patterns

In the central acception base, there are meanings that do not correspond to any word in some languages. For example, in French *'rivière'* means a river that does not flow into an ocean or a sea. *'fleuve'* is a river that flow into an ocean or a sea. There is no explicit lexical distinction between both in English. Reciprocally, *'giblets'* and *'offal'* have just a hyperonym in French. During the translation of a document, sometimes, a hyperonym is sufficient, but in some cases, we need a more accurate process [*Mangeot-Lerebours*, 2001]. To compensate for the inexistent word we suggest a method to generate locutions. Indeed, we can note that few patterns can help to generate most of the new terms and depends on grammatical category of the main word and its tag. [*Lehmann, Martin-Berthet*, 1998] contains an accurate study of French compositional patterns. For example, when both are nouns, we can just put the *'de'* preposition.

Giblets = Abats de volaille (volaille = fowl)
Offal =Abats de porc (porc = pork) / Abats de boeuf (bœuf = beef)

Associating an adjective or another noun with another one or a noun or adverb with a verb is very easy. For instance, in French, most of animals have different name depending on the gender.

Chienne = Female dog., *Jument = Female Horse.*, *Laie = Female Boar.*, etc.

This analysis depends on the language and patterns are easy to find using automated process. Some languages have same compositional method that make simpler its translation. [*Takenobu, Yosiyuki*, 1995] highlight the similarity between English and Japanese that are head-final languages with regard to noun compounding while [*Otoguro*, 1995] studies the verb-verb compounding. Bilingual dictionary scanning can highlight lexical items that correspond to two different entries in the target language, and multiple items that have same translation. So, the difficulty lies in getting the correct associated word and pattern. With few patterns it is easy to enumerate and find the correct word with result nearly equivalent to the lexical annotation. The only difference is that this new problem needs stronger constraints that can be summarized in grammatical type and position. The quality of results indeed depends on the resource. The difficulty is higher but there are many bilingual resources that are mostly structured. The position in the definition and a previous analysis of the dictionary structure can therefore produce good candidates. The evaluation in such process is not really essential. The objective is not to extract many words and order them but only find one word that satisfies all constraints many resources. The evaluation process better corresponds to a selection process, a filter that removes words that have not a usable grammatical category, that is thematically far, and then that do not fit in the resource structure.

3.2 Getting association

To extract candidates, many lexical resources and especially dictionaries are available. One difficulty of this project is that we do not have necessarily a bilingual dictionary linking the two given languages. In the above example, the easiest way to find candidates is by using a bilingual English-French dictionary. But if the acception source comes from another language, this dictionary can be inexistent. The objective is to use several lexical resources to find one or more ways that lead to a common associated word. The following enumeration describes the different lexical resources and their properties.

The bilingual dictionary These resources are very useful because most of them have a structured presentation. Indeed, bilingual dictionaries are often more adapted to automatic processed than monolingual ones because their principle is to give corresponding terms to entries. Monolingual must create definitions and paraphrases. The following example describes the definition of *’giblets’*, *’offal’* and *’abats’* in a French-English dictionary:

abats	:	*nmpl [volaille] giblets ; [bœuf,proc] offal*
giblets	:	*npl abbatis mpl or abats mpl (de volaille)*
offal	:	*n (U)(Culin) abats mpl (de boucherie) ;(garbage) déchets mpl, ordures fpl détritus mpl*

We can note that the occurrence of *’volaille’* is very simple to extract but no occurrence of *’beef’* or *’pork’* is present but only *’boucherie’* that means *’butchery’*. The extraction of hyperonymy from other resource could suggest that *’offal’* would be a default term when *’giblets’* can not be used. So, the bilingual dictionaries that are today implemented in a semi-structured language (XML) that makes extraction a fast easy and reliable process.

The monolingual dictionary

A monolingual dictionary mostly defines a word using *”genus + differentia”* approach. Therefore its definitions contain candidates that can be often translated. The following examples are definitions extracted from online dictionaries:

Dictionary.com:

Offal:	(1) Waste parts, especially of a butchered animal (2) Refuse; rubbish
Giblets:	The edible heart, liver, or gizzard of a fowl

Oxford paperback dictionary

Offal:	(1) edible organs of animal, esp. heart, liver,etc (2) refuse; scraps
Giblets:	plural noun liver, gizzard, etc. of bird removed and usually cooked separately

In these two examples, *’bird’* and *’animal’* can be extracted as candidates. *’Offal’* is polysemous but such word is quickly and reliably disambiguated (using angular distance in the vectorial concept database) and translated since the meanings are very different. So, the use of monolingual dictionary is more difficult and needs intermediate steps. Moreover the results depend on dictionary quality. English-French linking remains reliable due to the many existing lexical resources [1]. But they are less structured than bilingual ones and there are languages where we can not find direct translation often do not present many dictionaries.

The synonym dictionaries give good candidates that just need disambiguation and translation. But, these words replace the terms and do not create a lexical augmentation. If the acceptions in the Papillon dictionary do not exist in French, then it may not appear in synonym dictionaries. just can get hyperonyms and often polysemous words.

Papillon and Ontology If the number of intermediate steps is too high, then we need to create a new ontology that is a new set of words gathering common candidates. When a noun complement is needed to characterize a specific meaning, generally, the searched item is not unusual. So, we suggest creating a new small database that keeps only the acceptions that exist in most of or even all languages. Afterwards we keep only the most frequent words that occur in definitions and then remove all too polysemous words. Next a supervisor can scan the results and improve it by adding and removing items. The size of this set would be about several thousand items.

The difficulty is now to get a relevant lexicon without any lexical resource. A multilingual dictionary can be modelled as a semantic graph. Therefore, we can link an acception we want to complement by choosing a term in the neighborhood. We can furthermore know if it is a hyperonym by using miscellaneous resources like WordNet or Lafourcade’s hyperonym extractor [2] [*Lafourcade*, 2002] that gives good results. By choosing a pivot language we can get easily many lexical resources that contain hyperonym relationships. Finally, we should be able to get one of the closest relevant

[1] more than 20 on-line dictionaries are scanned by *http://www.onelook.com*

[2] *http://www.lirmm.fr/ lafourca/SERVICES/semvec-docs/hyperhypo-docs.html*

acceptions and translate it from a language that contains no ambiguities. For instance, remember the case of *'giblets'*, *'offal'* and *'abats'* and consider that we do not have got a bilingual or monolingual dictionary. Then a hyperonym must emerge of the main acception database. Indeed there are languages containing hyponyms that no other language expresses in a unique entry. But most of languages contain common hyperonyms. So, we can get the multilingual base will allow to get easily the hyperonym relationship between *'abats'*, *'offal'* and *'giblets'*. The vectorial model will have next to choose a noun that highlight the secondary thematic in *'giblet'* that is fowl or bird. These nouns activate concepts that are contained in giblet but not in *'abats'*. This contrast is obtained using the strong contextualisation studied by [*Schwab et al.*, 2002] in the framework of antonymy.

3.3 A new lexical resource

By getting strong associations between words using these multiple methods, binding links can be inferred between acceptions. This allows the creation of a set of words and idea associations between words. This new lexical base can become a new resource for meaning base by reinforcing and contrasting relationships. The improvement of the lexical base implies best results and vice versa. Moreover, the appearance of common ideas in most of languages generates a meanings graph that contains candidates for a new vectorial base: the calculation time and space can be adjusted by choosing more or less ideas in a vector. Speciality needs can be satisfied by zooming on thematic of the main graph and choosing more items [*Lafourcade et al.*, 2002]. The projection methods to convert from a base to another can also be automatically deduced. All the conversion agent has to know is the meaning graph and which components are in each vector. The use of a main base allows projection of a concept space onto others. The selection of ideas can be automatic by maximizing the distance average and minimizing the standard deviation or preferably by adapting it in proportion to lexical density of the graph. Some clustering methods studied entropy and other statistical methods to treat such meanings graph.

Conclusion

The NLP team of the LIRMM currently works on thematic aspect of meaning representations and the possible impact on the Papillon acception base. This paper describes an approach based on meaning associations recovering. Moreover, there are terms that are not lexicalized in another language. lexicalizing these terms may be usefull for improving translations. So this problem that is similar to compound nouns translation one lead to generate new complex words to express the proper meaning. The analysis of lexical resources split this experiment in two steps that are the associations recovering and the pattern matching. In the first stage, we consider that a lexie is defined using "genus + differentia" method. Therefore we try to identify these two components using lexical resources. Finally, depending on the part-of-speech category, pattern recognition allows the compound word generation. In fact, in the same principle of our previous work, this study leads to create a new idea database. The multilingual approach of Papillon is the best resource to recover common ideas that exist in most of languages. Such project offers a semantic network that allows the unsupervised generation of a new ontology. This last new resource supply many benefits to conceptual vector model which components can be dynamically selected. The system can zoom in or out on several part of the graph to improve accuracy of meaning representation in speciality domains. Projections between vectors produced in distinct bases can be automatically calculated. This dynamic adaptation to the context and domain is important considering that cognitive interests of several languages are different. This flexibility in a vectorial base and meaning representation is also an important feature in such a large-scale project. A universal multilingual project must unify the different languages but retain uniqueness of each one.

References

[*Chauché*, 90] Chauché J., *Détermination sémantique en analyse structurelle : une expérience basée sur une définition de distance.* TAL Information, 31/1, pp 17-24, 1990.

[*Deerwester et al.*, 1990] Deerwester S., S. Dumais, T. Landauer, G. Furnas, and R. Harshman, *Indexing by latent semantic anlysis.* In Journal of the American Society of Information science, 1990, 416(6), pp 391-407.

[*Lafourcade et al.*, 2002] Lafourcade M., D. Schwab, et V. Prince *Vecteurs conceptuels et structuration émergent de terminologies* In Traitement Automatiques des Langues (TAL), Vol 43, n1/2002, pp. 43-72.

[*Lafourcade*, 2002] Lafourcade M. *Guessing Hierarchies and Symbols for Word Meanings through Hyperonyms and Conceptual Vectors* In Procs of OOIS 2002 Workshop, Montpellier, France, September 2002.

[*Larousse*, 1992] Larousse. *Thésaurus Larousse - des idées aux mots, des mots aux idées.* Larousse, ISBN 2-03-320-148-1, 1992.

[*Lehmann, Martin-Berthet*, 1998] Lehmann A. et F. Martin-Berthet *Introduction à la Lexicologie - Sémantique et morphologie.* Dunod 1998

[*Mangeot-Lerebours*, 2001] Mangeot-Lerebours M. *Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue* Thse 2001.

[*Morin*, 1999] Morin E. *Extraction de liens sémantiques entre termes à partir de corpus techniques.* Thèse de doctorat de l'Université de Nantes, 1999.

[*Otoguro*, 1995] Otoguro R. *Word formation in syntax: Japanese verb-verb compounding and grammatical information spreading.* MA dissertation, University of Essex, 2002

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

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

[*Schwab et al.*, 2002] Schwab D. , M. Lafourcade et V. Prince *Antonymy and conceptual vectors* Coling 2002.

[*Takenobu, Yosiyuki*, 1995] Takenobu T., K. Yosiyuki *Analysis of syntactic structure of Japanese compound nouns* 1995.