Low Cost Automated Conceptual Vector Generation from Mono and Bilingual Ressources

Mathieu Lafourcade, Frédéric Rodrigo, and Didier Schwab
LIRMM - Laboratoire d'informatique, de Robotique
et de Microélectronique de Montpellier
MONTPELLIER - FRANCE.
{lafourcade, rodrigo, schwab}@lirmm.fr
http://www.lirmm.fr/~{lafourcade, schwab}

Abstract

This paper assesses the possibilities of constructing a multilingual lexicon by propagating conceptual vectors through several monolingual and bilingual resources. The system is based on a vector model in order to learn meanings to potentially select and classify meanings. Bilingual resources ensure the possibility to project vectors on the target lexicon and semantic space.

Keywords: conceptual vectors, bilingual dictionaries, translation, vector construction.

1 Introduction

With the advent of the Web, documents are available in a variety of languages, increasing dramatically the need for machine translation and multilingual lexical databases. One of the crucial aspect of the machine translation process, is the choice of a target term from a source language. An original approach to lexical transfert can be sketched with conceptual vectors model (Lafourcade and Schwab, 2002). However, conceptual vector databases still have to be built, which is quite time and resource consumming. In this paper, we elaborate a simple method to automatically create a conceptual vector acception database for a target language, starting from an already existing source of conceptual vectors in a given source language. This process has to be cheap and fast. The result on the proposed method can be used as a first lexical material for lexicographer work in projects like Papillon (Papillon, 2001 2003).

Formally, this research aims at easily populating a target space with the knowledge of the source space and bilingual dictionaries. In (Lafourcade, 2002), a study was already done on building a multilanguages acceptions database. This database was built in two steps: at the beginning, bootstrapping from monolingual databases and linking acceptions between languages with bilingual dictionaries. Starting with a source term, a grammatical morphology and a context (some gloses), this 3-uple can be put in correspondence in one or several equivalents in the target language. The acception must be correctly linked between monolingual and bilingual dictionary in order to merge close meanings. Some problems have been encountered with terms with contrastive meanings (Lafourcade, 2002). For example, *cahier* in French is the equivalent of *exercise book* or *note-book* in English (the appropriate choice depends on context). In the same project (Papillon, 2001 2003), (Schwab and Lafourcade, 2002) have studied semantic relations in acceptions database. These relations are synonymy, antonymy, hyperonyny and holonymy. The goal of this work is to increase the database integrity by respecting some contraints.

The current study differs from previous works by the methodology of the database building. Here, we try to construct a first mockup of the target lexical material from a source database, instead of building the database by linking directly accceptions. The proposed approach focuses more on recall than precision.

2 Conceptual Vectors

The conceptual vector model (Lafourcade and Schwab, 2002) is a formalism which projects semantic fields into a vectorial space. The thesaurus hypothesis is to consider a set of notions that can generate the language lexicon. The leaves of a thesaurus hierarchy tree are used as generator vectors of a space where each meaning is represented by a vector. The space defined in this way is probably not free (no proper vectorial base) and indeed, we take advantage of the implicit redundancy of information to ensure coherency among vectors.

For (an oversimplified) example, if we consider ANIMAL and GRAY as concepts (i.e entries of thesaurus associated to vectors), then *mouse* can be defined with the vectorial sum of ANI-

MAL and GRAY. Monosemous terms have only one vector, but polysemous terms have several vectors, one for each meaning.

Conceptual vectors A and B are comparable with some similarity measures. Let us define *sim* as a possible such function:

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

We then can define an angular distance considered as the thematic proximity between two terms:

$$D_A(A,B) = arccos(sim(A,B))$$

Terms with an angular distance lower than 45° can be considered as close and terms with an angular distance around 90° are considered as loosely related. Here, follow some examples:

D_A(coal, coal)	=	0°
D_A(coal, ore)	=	12°
D_A(coal, mine)	=	18°
D_A(coal, coalmine)	=	25°
D_A(coal, miner)	=	35°
D_A(coal, pit)	=	32°
D_A(coal, energy)	=	64°
D_A(coal, electricity)	=	38°

coal cannot be closer to anything else than *coal*, then angular distance is 0°. The angular distance between *coal* and *ore* and *mine* seem to be reasonnable. Other terms are more less related, the angular distances between them are higher.

Operations on conceptual vectors are also useful in our context. The normed sum merges the concepts of two vectors into a new one:

$$C = A \oplus B \quad | \quad C_i = \frac{A_i + B_i}{\|C\|}$$

The normalized term to term product raises shared information between two vectors:

$$C = A \otimes B \quad | \quad C_i = \sqrt{A_i \times B_i}$$

The weak contextualisation between two terms, concepts of A are reinforced by concepts of B:

$$\gamma(A,B) = A \oplus (A \otimes B)$$

For example, *stack* in context of *banknote* can be a *bundle*; *bay* in context of *Norway* can be a *fjord*.

3 Conceptual Vectors Building through Bilingual Dictionaries

Bilinguals dictionaries translate terms from one language to an other. However even for monosemic terms, several equivalents are generaly provided (at least because of quasi-synonymy) and at a lexical selection should be done by the human reader. For our concern, the goal is to compute a conceptual vector. Precisely, it is equivalent to identifying an appropriate location in the vectorial space, for the target term.

The input data of the process is a named vector ($term_S$ with $vector_S$) in the source language S with its morphological information: $(term_S, vector_S, morph_S)$. The expected output is a close vector associated with a target language T term ($term_T$): $(term_T, vector_S, morph_T)$. For our purpose, we consider translation dictionaries having the following structure:

$$Bd_w \equiv < morph^*, glose^*, equiv^+ >$$

where *glose* stands for an optional context denoting usage (or domain) of the current meaning entry and *equiv* is the list of possible corresponding terms. A simple example in figure 1 about the English term *mouse* shows the morphology *Noun*, the gloses *ZOOLOGY* and *COMPUTER SCIENCE* and the French translation *souris*.

```
mouse:
    (Noun) [ZOOLOGY] souris
    (Noun) [COMPUTER SCIENCE] souris
```

Figure 1: Entry of *mouse* in the English to French (LOGOS, 2003) multilingual dictionary. The glosses are used both to select the appropriate acception in the target language but also to help constructing the target vector.

As illustrated with figure 2, the process consists in taking all possibles translations of one term $term_S$ and choose the appropriate one according to contextual information. The selected translation $term_T$ can be associated with this morphology $morph_T$ and the original vector $vector_S$.

4 Realisation

This correspondence is realised in two steps. First, get the possible translations and compute

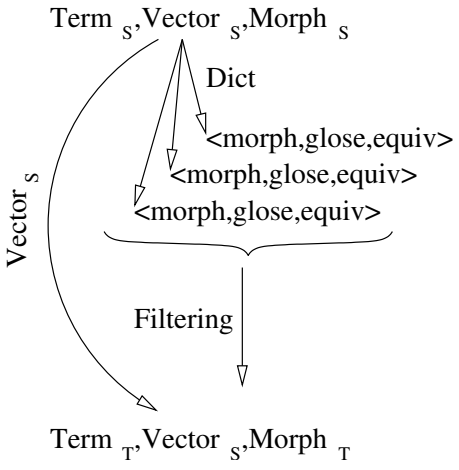

Figure 2: The correspondence process

their conceptual vectors representations. Secondly, choose the best one with a multicriteria filter.

4.1 Accessing Possibles Translations

All entries of translations dictionaries are fetched for a term $term_S$. In order to be integrated in the conceptual vector model, the gloses phrases (in French for a French to English dictionary) must be syntactically analysed (at least to determinate the phrase head). The analysis step is done with (SYGMART, 1990), a French morpho-syntactic analyser, developed by (Chauché, 1990). In order to build a conceptual vector of a text, we combined the produced morpho-syntactic tree and a French source language vector database previsouly built. At the end of this process we have some triplets, structured as:

$$(morph_i, vector(glose_i), equiv_{i,j})$$

4.2 Multicriteria Filters

Filtering triplet translation consists to find the closer $term_T$ of $term_S$ in the context of the glose. In order to do this, a scoring is done. Each filtering step scores the possibles translations. The most highly scored is selected. Lack of data in dictionaries can appear (glosses or morphologic informations may be absent). Consequently, not all filters can be used.

Morphology Filter: as a first naive approach, filter on morphology can be done, and take only translations with the same morphology than $morph_S$. But this choice may *over filter* leading to an empty result. But the adopted strategy is different, in order not to limit the choice on identical morphology, which may far too restrictive. Still, different morphological information will induce less probability for the terms to match. This method allows to find pairing solutions in some more flexible ways. The selected choice relies of one factor, where only a meaning of the terms is reached. Another concern is to avoid a literal translation.

$$score_{i,morph} = \begin{cases} 1 & \text{if } morph_S = morph_i \\ 0 & \text{otherwise} \end{cases}$$

Filtering on Glosse Vectors: distance between $vector(term_S)$ and glose vectors are computed. The measure used is the relinearized angular distance. The closer two vectors are, the closer to 1 is the score:

$$d \leftarrow D_A(vector(term_S), vector(glose_i))$$

$$score_{i,glose} \leftarrow 1 - \frac{2d}{\pi}$$

Moreover, some other filters can be applied. This filters are quite specific to the dictionary used (LOGOS, 2003). This is mainly due to formatting constraint (more than structural constraints). This particular dictionary (Logos) provides a *thema*[1] for meaning and synonyms and antonyms in relation with context of current acception. Bascially a thema is a restricted glosse.

Filtering on Thema: the same kind of process as for the glosses is undertaken:

$$score_{i,subject} \leftarrow sim(vector(term_S), vector(subject_i))$$

Filtering with Semantic Relations: we compute the distance between the vector of term source $vector(term_S)$ and the normalised sum of synonyms vectors. This sum allows the create of a vector with the meanings of all synonyms, consequently this vector *stresses* on common meaning.

[1] a glosse that denotes the thematic field of the acception. For example, *line* can be associated to *railroad transportation*, *fashion*, *geometry*, ...

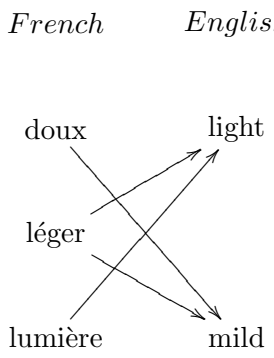

Figure 3: Meanings and translations as provided in a typical bilingual dictionary. Vectors are available beforehand on the French side and *propagated* to the English terms.

Formally, we have:

$$score_{i,lexfunc} \leftarrow sim\big(vector(term_S), vector(\bigoplus_j synonym_{i,j})\big)$$

The global $score_i$ is computed with the score average of the filters.

$$score_i = \frac{score_{i,morph} + score_{i,glose}}{\text{number of used filters}} + \frac{score_{i,subject} + score_{i,lexfunc}}{\text{number of used filters}}$$

On the end, the highest scoring target term $term_T$ is chosen. Then, it is possible to insert $(term_T, vector_S, morph_T)$ to the target conceptual vector database. On the basic process there is not revision or iterated computation.

5 Experience, Results and Evaluations

Terms from French database are put in correspondance with English ones. For each English terms, all meanings are extracted and translated through the process explain in this paper. English terms are sets of English meanings (see figure 3). In order to access the contents of the English database, we calculate lists of neighbour terms around some target terms and evaluate their relevance. Figures 6, 5, 7, and 8 show a small set of the more thematically similar terms for French and English. Figure 9 shows displayed the closed terms of term *stack* in context of one other. They are computed with the weak contextualisation and select the neighbour terms into the English database.

The French database (the source) used in this experiment contains around 96000 terms built by monolingual learning and 160000 vectors. It leads to an average of 1.7 meanings by French term. Only 14500 terms are taken for the translation process (those of the list from (ABU, 1999)). At the end as the translation process, the English database contains 21000 terms. With translation dictionaries, we were able to make the 14500 more common French terms refer to 21000 English terms. Finally, the English database contains 63500 vectors, for an average of 3 meanings by English term.

In order to evaluate the English database both by itself and against the French database, we have made some preliminary tests. We asked a set of persons to evaluate the quality of neighborhood terms of a target term. To facilitate the evaluation, the tester is asked to evaluate two lists: the one produced by the system and one extracted from the the Rodget thesaurus (see figure 11). The tester doesn't know the origin of the list. They are not told about how the lists have been produced. The results are displayed on figure 10. For *beer* testers found the quality of neighbour in databases globally better than in the Rodget thesaurus. Our intrepretation is that it is mostly because the semantic field seems to be larger (not as narrow as in the thesaurus list. See annexes). On the overall, with this kind of evaluation, the English database quality reach around 87% of the Roget thesaurus quality. Nevertheless the tests must continue on larger data and more testers.

Conclusion

The method presented in this paper allows to compute conceptual vectors from a source database in the target database. This way, we can populate a semantic space with conceptual vectors of the target language. Bilingual resources must be provided to ensure a proper correspondence between source and target vectors. However, the morphological information may be substantially different between source and target terms. This method has the drawback of not using explicit links between acceptions, and moreover does not maintain the linking constraints. But, the current study allows all conceptual vector operations on database and aims at being computationally cheap.

Still, the quality of created conceptual vectors could be enhanced. Some improvements are still possible on the process itself without

burdening it with complication. For instance, it would be desirable to do a filter on back translations, i.e. to compute a back proximity value of $term_T$, between $terms_S$ and translation of $term_T$ in source language. It is also possible to reinforce the learning on a target monolingual dictionary with the same rules as the source language, or use lexical resources like (WordNet, 2003). Another option is to directly compare target and source database in order to create links of semantic relations between them. However some semantic networks may be necessary.

The database created in this study will be accessible in the (Papillon, 2001 2003) project, and more databases could be built with this method. This preliminary work shows the necessity to go further into the creation, building and refinement of data from multilingual resources, but also shows some practical aspects of this construction.

References

ABU. 1999. Abu : la bibliothèque universelle, http://abu.cnam.fr/.

Chauché. 1990. Détermination sémantique en analyse structurelle : une expérience basée sur une définition de distance. *TAL Information*, 31(1):17–24.

Prince Lafourcade and Schwab. 2002. Vecteurs conceptuels et structuration émergente de terminologies. *TAL*, 43(1):43–72.

Lafourcade. 2002. Automatically populating acception lexical database through bilingual dictionaries and conceptual vectors. *PAPILLON-2002, Tokyo, Japan.*

LOGOS. 2003. http://www.logosdictionary.com/.

Papillon. 2001-2003. http://www.papillon-dictionary.org/.

Schwab and Lafourcade. 2002. Hardening of acception links through vectorized lexical functions. *PAPILLON-2002, Tokyo, Japan.*

SYGMART. 1990. http://www.lirmm.fr/~chauche/présentationsygmart.html.

WordNet. 2003. http://www.cogsci.princeton.edu/~wn/.

6 Annexes

6.1 Neighbourhood of Terms (including distances)

beer: (ale 0.030) (lambic 0.052) (cognac 0.112) (brandy 0.118) (public house 0.136) (kir 0.137) (white spirit 0.138) (alcohol 0.138) (lemonade 0.140) (julep 0.147) (cider 0.149) (ambrosia 0.150) (piccolo 0.150) (alcoholic drink 0.151) (rough brandy 0.152) (aqua vitae 0.152)

ale: (beer 0.020) (lambic 0.052) (cognac 0.112) (brandy 0.118) (Mark 0.120) (public house 0.136) (kir 0.137) (white spirit 0.138) (alcohol 0.138) (lemonade 0.140) (sea bass 0.143) (julep 0.147) (cider 0.149) (ambrosia 0.150) (piccolo 0.150) (mascara 0.151) (alcoholic drink 0.151) (rough brandy 0.152) (aqua vitae 0.152)

For memory, in French we have for **bière**: (gueuze 0.0935) (pale-ale 0.1156) (cervoise 0.123) (lambic 0.1251) (bibine 0.1467) (tourraillage 0.149) (arack 0.1509) (bière brune 0.1568) (bière ambrée 0.157) (Ctes-du-Rhne 0.1584) (halbi 0.1589) (caboulot 0.1595) (kriek 0.1598) (faro 0.1612) (aquavit 0.1613) (saké 0.1615) (Ava 0.164) (absinthe 0.1661) (irish coffee 0.1672) (reginglard 0.1674) (Armagnac 0.1681) (casse-pattes 0.1682) (daiquiri 0.1683) (Porto 0.1694) (rhumerie 0.1706) (alsace 0.1713) (théier 0.1713) (gnle 0.1713) (casse-poitrine 0.1717) (millésimé 0.1735) (buvette 0.1756) (tequila 0.1756) (taniser 0.1757) (Bordeaux 0.1765) (loufiat 0.1765) (bavarois 0.1766) (blinis 0.1773) (rhumé 0.1774) (aviner 0.1776) (estaminet 0.1782) (rinure 0.1785) (résiné 0.1786) (spiritueux 0.1788) (limé 0.1791)

6.2 Example of simple evaluation lists

beer	bière[boisson]
pale ale	kriek
lambic	cervoise
cognac	bière brune
brandy	pale-ale
stout	citronnade
public house	picrate
kir	lambic
lambic	vinasse
alcohol	saké

Figure 5: Neighbour terms around beer_{en} and bière_{fr}

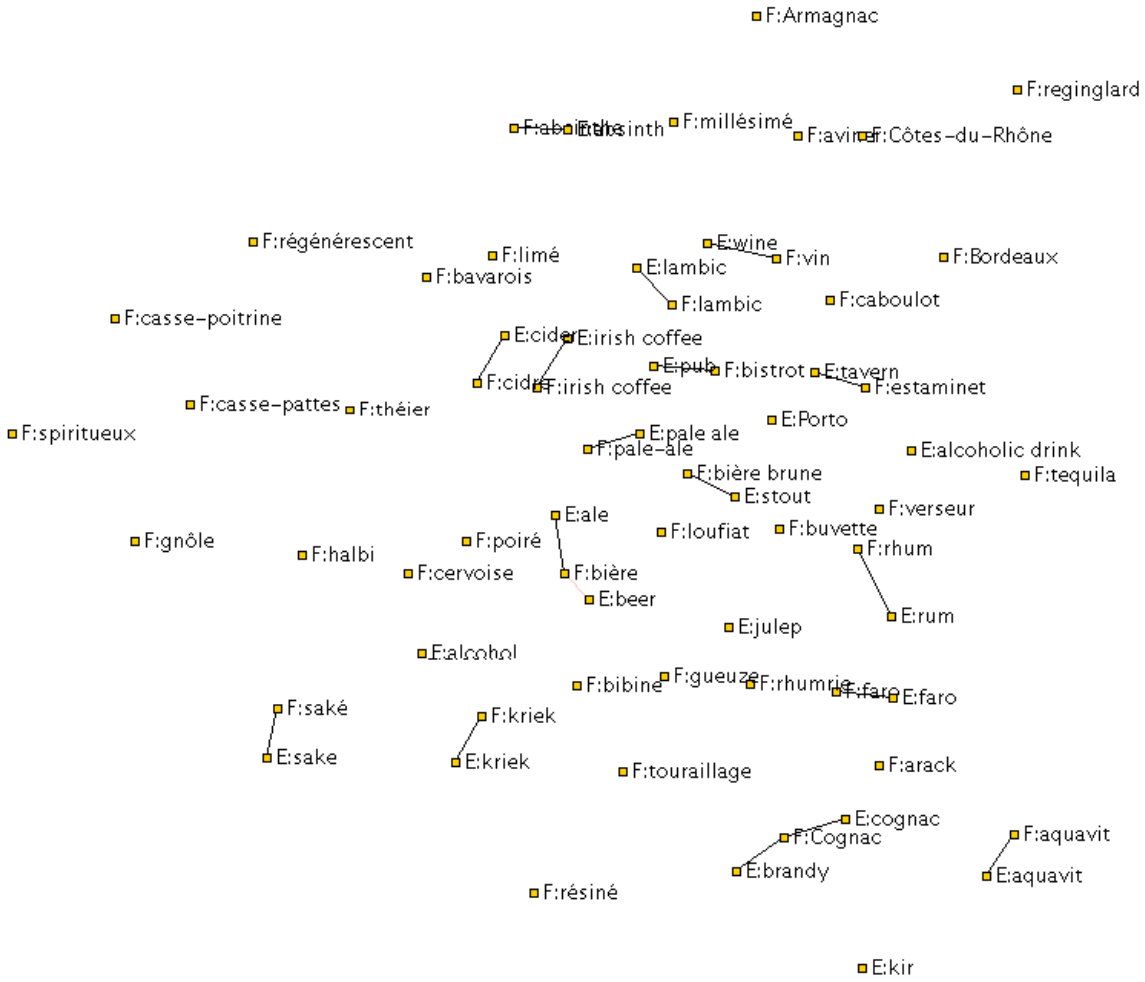

Figure 4: 2D graphical representation, of some neighooring terms of bière and beer. Terms translation of each other are linked.

house	maison
building	chacunière
domicile	hutte
institution	carrée
lodging	building
urban planning	grand ensemble
urbanism	résidence
planology	intérieur
manhole	logement
palace	toit

Figure 6: Neighbour terms around house$_{en}$ and maison$_{fr}$

car[automobile]	voiture[automobile]
motorcar	auto
automobile	automobile
passenger car	tacot
landau	landau
automotive	limousine
Ridge runner	bagnole
limo	parking
jeep	berline
motoring	automobiliste

Figure 7: Neighbour terms around car$_{en}$ and voiture$_{fr}$

commerce
transaction
desktop
agiotage
business
holding company
trade
affiliate
acquisition
firm

commerce
offre et demande
prix marchand
succursale
trafic
agence
holding
étal
caisse de dépôts
caisse de crédit

Figure 8: Neighbour terms around commerce_{en} and commerce_{fr}

6.3 Example of neighbour according to context

stack	money	wood	car	people	food
	stack	stack	stack	stack	stack
	purse	park	park	couple	supply
	denier	odd piece	store	dynastic	provision
	clutch bag	sheaf	lens hood	keep	park
	park	board	railroad station	running	store
	stock exchange	winder	railway station	corner	load
	store	harbor	station	hydrography	omelet
	cash received	course	freightage	ciborium	omelette

Figure 9: The closer terms of *stack* in context of *money*, *wood*, *car*, *people* and *food*

	Beer	Dictionary	Sky	Wood	Moyenne
Rodget	9	16	15	14	14
Voisins	13	16	7	11	12
$\frac{Voisins}{Rodget}$	1.40	1.00	0.45	0.84	0.87

Figure 10: Résultat de l'évaluation, notes entre 0 et 20 au près de sept sondés

Beer dans le thésaurus Rodget	Termes du voisinage thématique de *beer*
ale	alcohol
amber brew	alcoholic drink
barley pop	ale
barley sandwich	ambrosia
belly wash	aqua vitae
bock	brandy
brew	cider
brown bottle	cognac
cold coffee	gin
head	julep
hops	kir
lager	lambic
malt	lemonade
malt liquor	mark
oil	piccolo
porter	public house
slops	rough brandy
stout	strong liquor
suds	white spirit

Figure 11: Data evaluation for ‹*beer*›

Dictionary dans le thésaurus Rodget	Termes du voisinage thématique de *Dictionary*
concordance	alphabet
cyclopedia	article
encyclopedia	designator
glossary	glossary
language	grammar
lexicon	idiom
palaver	lexicography
promptory	lexicon
reference	linguistics
terminology	nomenclature
vocabulary	orthographical
wordbook	vocabulary

Figure 12: Data evaluation for ‹*dictionary*›

Sky dans le thésaurus Rodget	Termes du voisinage thématique de *Sky*
air	attic
azure	banquette
celestial sphere	batten
empyrean	bed curtain
envelope	bezel
firmament	cupola
heavens	eden
lid	esplanade
pressure	lambrequin
substratosphere	mirador
the blue	oasis
troposphere	paradise
upper atmosphere	penthouse
welkin	promenade

Figure 13: Data evaluation for ‹*sky*›

wood dans le thésaurus Rodget	Termes du voisinage thématique de *wood*
copse	antlers
forest	baguette
grove	bush
lumber	fin
thicket	flower bed
timber	forest
timberland	grove
trees	lumber
weald	woodland
woodland	woodwind instruments
woods	woodwinds

Figure 14: Data evaluation for ‹*wood*›