D10 - Technique d'acquisition de connaissances lexicales à partir de corpus analysés en syntaxe

Claire Gardent (CNRS/LORIA)

avec des conributions de:
Claire Mouton (CEA-LIST), Éric de la Clergerie (INRIA)

Abstract.
Ce délivrable D10 présente diverses techniques mises en oeuvre pour l'acquisition de connaissances lexicales à partir de l'analyse syntaxique de gros corpus.
Mots clefs: acquisition, cadre de sous-catégorisation, classes sémantiques, fouille d'erreurs

Document Id	Passage/2009/D10/v1.0
Projet	ANR-06-MDCA-013 PASSAGE
Version	v1.0
Date	30 janvier 2009
État	initial
Distribution	public

Consortium Passage

Ce document fait partie d'un projet de recherche financé par le programme MDCA de l'ANR sous la référence ANR-06-MDCA-013.

Institut National de Recherche en Informatique et en Automatique (INRIA)
Contact: Éric de la Clergerie
E-mail: Eric.de_la_clergerie@inria.fr

Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI)
Contact: Patrick Paroubek
E-mail: pap@limsi.fr

Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA)
Contact: Claire Gardent
E-mail: gardent@loria.fr

CEA-LIST
Contact: Gaël de Chalendar
E-mail: Gael.de-Chalendar@cea.fr

Participants à ce rapport

Les partenaires suivants ont pris une part active au travail conduisant à l'élaboration de ce document, même si ils n'ont pas directement contribué à la rédaction de ce document:

LORIA
CEA-LIST
INRIA

Modifications

Version	Date	Auteur	Modifications
0.1	12.01.09	Claire Gardent	Création
0.2	14.01.09	Claire Gardent	Ajout contribution Claire Mouton (CEA)
1.0	30.01.09	Éric de la Clergerie	Mise en forme et ajout Fouille d'erreurs

Table des matières

Introduction

Dans sa tâche WP5, le projet PASSAGE propose d'utiliser les résultats d'analyse syntaxique d'un large corpus d'au moins 100 millions de mots pour conduire des expériences d'acquisition de connaissances lexicales. Ces connaissances lexicales peuvent porter sur les niveaux morpho-syntaxique, syntaxique ou sémantiques.

La première expérience, relatée par Claire Gardent pour le LORIA, concerne l'acquisition de cadre de sous-catégorisation pour les verbes à partir des diverses occurrences des verbes attestés en corpus. L'importance de tels cadres pour l'analyse syntaxique n'est plus à prouver. Diverses expériences d'acquisition ont été tentées ces dernières années mais, à notre connaissance, jamais à cette échelle pour le français.

La seconde expérience, relatée dans ses grandes lignes par Claire Mouton pour le CEA-LIST, concerne la classification de mots en sens proches en s'appuyant sur la similarité de leurs contextes (syntaxiques et proximité), tels que trouvés au sein d'un grand corpus. Ces informations sont ensuite utiles pour de la désambiguïsation sémantique et pour l'annotation sémantique de documents.

Enfin, la dernière expérience, relatée par Éric de la Clergerie pour l'INRIA, concerne la fouille d'erreurs dans les phrases non analysables d'un grand corpus. Cette fouille permet de faire remonter de l'information sur les mots manquants ou mal renseignés dans un lexique, que ce soit au niveau morphosyntaxique ou syntaxique.

L'ensemble des ces expériences sont préliminaires dans le sens où elles ne s'appuie pas encore sur le format PASSAGE mais sur la version antérieure EASy, trop pauvre en information. Elles soulèvent également des problèmes algorithmiques de passage à l'échelle qui restent à résoudre. Les expériences ne sont pas non plus appuyées sur l'ensemble des analyses de tout les participants sur l'ensemble du corpus Passage de 100 millions de mots.

Chapitre 1

Extracting a syntactic lexicon from a parsed corpus

1.1 Introduction

A syntactic lexicon records for each verb, the nature and the type of its arguments. Additionnally, it may contain information that is specific to the verb such as for instance, the fact that a verb is a control verb or that it accepts the passive transformation. As has been repeatedly argued, a syntactic lexicon [IV94, MGM94] is an important resource for Natural Language Processing in that it provides valuable information e.g., for parsing, for machine translation or for surface realisation [JMdR04, CF04].

Unfortunately, the specification of a large coverage syntactic lexicon is both time consuming and error prone. Moreover, hand written lexicons usually fail to provide probabilistic information about the frequency and relative frequency of the verb/subcategorisation frame pair they contain. Such information is however crucial for many NLP tasks such as for instance, stochastic parsing or verb class acquisition.

To overcome these shortcomings, corpus based, statistical methods have been successfully proposed for English which usually proceed in two steps [BC97, Kor02]. First, a large corpus is parsed and verb dependents are extracted from the available parse trees. Second a statistical filter is applied to determine which of the extracted hypothesis are plausible.

In this deliverable, we report on a first attempt towards automatically extracting a syntactic lexicon for French from a parsed corpus (EASYLEX). The method and results presented are still preliminary. They will be improved and applied to other corpora during the last year of the project. The final version of EASYLEXand a second deliverable will be made available in December 2009.

The structure of the report is as follows. Section 1.2 presents the corpus and the grammatical information used for extraction. Section 1.3 gives an overview of the extraction process that has been carried out. Sections 1.4, 1.5 and 1.6 describe the three main steps of this process namely, the extraction of the verb and its dependents ; a symbolic filter used to normalise certain frames (passive frames, frames with multiple PPs introduced by the same preposition) and the production of the resulting lexicon. Section 1.7 details the work that remains to be done to complete the ex-

traction process and create a lexicon that is as precise as possible. One further aim of the project is to enrich existing syntactic lexicons with information extracted from corpora in particular, frequency information. Under this perspective, section 1.8 reports on the conversion of an existing hand-written syntactic lexicon namely, Dicovalence, to a format compatible with EASYLEX and presents some initial results concerning the overlap between EASYLEX and Dicovalence. Section 1.9 concludes with pointers for further research and sketches a work programme for 2009.

1.2 The CPC-v1_tagparser corpus

The input corpus is a 20 million words corpus parsed with Gil Francopoulo's TAGPARSER. The source text is part of the CPC-v1 (Corpus Passage Court Version v1) corpus and contains articles from Wikipedia FR (`http://fr.wikipedia.org/`)[1] that is, encyclopedia texts. The parsing annotations produced by the TAGPARSER are *easy*-conformant[2]. That is, each sentence parse indicates both the *easy*-constituents found by the parser and the *easy*-relations identified to hold between these constituents. Tables 1.1 and 1.2 summarise the *easy*-constituents and -relations respectively. For a more precise definition, see [GV08].

Additionally, the TAGPARSER provides the following morphosyntactic and semantic information.

- each form is associated with a corresponding lemma (LEMME feature) and a category (POS feature)
- verbs are assigned a MOOD feature whose value can be *conjuguated, participle* or *infinitive*
- Noun and prepositional phrases are assigned a semantic type (the ENTITE feature) with possible value : *dateTime, individual,location,mark,measure, organization,unnamed* or *URLetc*.

Cst.	Gloss	Defn	Example
NV	Noyau verbal	{Clitic*, ?ne}+V	*vient,ne veut,il va*
GN	groupe nominal	?Det+Adj*+N	*Jean, le beau chat, 4*
GP	Groupe prépositionnel	(Prep+ ?(GN|Adv))| *dont*	*avec, avec Jean, en dehors*
GA	Groupe adjectival	Adj|Participle	*roux, retenue, obeissant*
GR	Groupe adverbial	Adv	*aussi, plutôt*
PV	Groupe verbal	Prep+NV_non_finite	*d'ébranler, en reprenant*

FIG. 1.1 – The list of *easy*-constituents

[1] Licence : Licence de documentation libre GNU 1.2 (LGDL 1.2) . Dump used : frwiki-20060402-pages-articles.xml.bz2. Dump access : http ://download.wikipedia.org/backup-index.html

[2] The *easy*syntax for parse annotation is defined by the *easy*-DTD given in annex A.

Reln	Gloss	Example
SUJ_V	verb/subject	**On a** *construit*
AUX_V	verb/auxiliary	*On* **a construit**
COD_V	verb/object	*Jean* **construit une maison**
CPL_V	GP ou PV, circ. ou complt	*Jean* **pense à Marie**
MOD_V	Circ., GN non objet	*Jean* **dort profondément**
ATB_SO	verb/predicate	*Il a* **trouvé** *cette raison* **étrange**

FIG. 1.2 – The list of verb related *easy*-relations. The words in bold face are those involved in the relation under consideration.

1.3 Overview of the extraction process

The extraction of the syntactic lexicon from the Easy corpus is intended to proceed in several steps as follows.

Verb-dependents extraction : All verb/dependents occurrences are extracted from the parsed corpus.

Symbolic filter : A first filter is applied to normalise active/passive variation and "reduce" sequences of dependents with identical function and category.

Lexicon initialisation : For each (verb,list of dependents) pair, a unique summary is created that merges the results of step 1, normalises arguments (e.g., clitics, relative pronoun) and introduces counts and relative frequency information for each (verb,list of dependents) pair.

Frame filter : filters out entries (i.e., (verb,list of dependents) pair) whose frame is unknown. A frame is unknown if it is not part of a predefined set of frames known to be valid for French.

Statistical filter : Filters out lexical entries using e.g., the log likelihood ratio or the binomial hypothesis i.e., improbable (verb,list of dependents) pairs (e.g., list is too long or verb/liste association is too rare).

At the moment of writing, the last two steps are not yet implemented. See Section 1.7 for a detailed discussion of the plans concerning these.

1.4 Extracting the verb and its dependents

In a first step, we extract from the parsed corpus, verb occurrences and their dependents. All dependents are extracted together with some morphosyntactic information about the verb (e.g., mood, auxiliary) and the dependents (head, part of speech of the head, semantic type, etc.). At this stage, the information contained in the parsed corpus about verbs and their dependents is simply recorded. No normalisation, counting or filtering take place.

1.4.1 Identifying verb/dependents occurrences

A verb occurrence is a `verbe` element which is the immediate child of a `relation` element.

```
<relation xlink:type="extended" type="SUJ-V" id="E1R1">
       <sujet xlink:type="locator" xlink:href="E1G1"/>
       <verbe xlink:type="locator" xlink:href="E1G2"/>
</relation>
```

The dependents of a verb occurrence are the left sister element of a verb occurrence. For instance, in the above example the `sujet` element is a dependent.

To extract verb-dependents pattern, we search the parsed corpus for verb occurrences and for their dependents. Information about each verb occurrence and its associated dependents is then recorded in an extraction pattern as defined below.

1.4.2 Extraction patterns

The extraction pattern (EPATTERN) created for each verb/dependents occurrence contained in the parsed corpus has the following form :

```
#S(EPATTERN
:ID verb identifier
:TARGET |verb| ?aux=|aux|
:VFEATS  (mood)
:DEP_i+  a characterisation of each dependent found
)
```

where :

- `?` indicates optionality and `X+` a sequence of one or more `X`
- `|verb|` is the LEMME value extracted from the verb form element
- `verb identifier` is the XLINK value of the verb
- `|aux|` is the LEMME value of the auxiliary form immediately related to the verb form if the verb/subject relation is mediated by an auxiliary (see below)
- `|mood|` is the value of MOOD feature in the verb form and is either *conjuguated, participle* or *infinitive*
- the `DEP_i+` line stands for one or more `DEP_i` line (with `i+`, an integer), each characterising a verb dependent

Example extraction pattern

```
#S(EPATTERN
:ID E2607G3
:TARGET ébaucher
:VFEATS  conjuguated
:SUJ_V E2607G2 Wikipedia properNoun individual
:COD_V E2607G4 objet commonNoun unnamed
:CPL_V E2607G7 ceinture commonNoun de unnamed
)
```

1.4.3 Dependent information

The information extracted for the dependents varies depending on their syntactic category.

For NPs, we extract the identifier and the part of speech, the semantic type and the lemma of the nominal head. For PPs, we additionnally extract the value of the introducing preposition. For sentential complements, we extract the identifier, lemma, part of speech, mood and preposition if any. More precisely, the recorded information follows the following format :

Cat.	Recorded information
GN	`:DEP_i IDENTIFIER (LEMMA POS ?ENAMEX)`
GP	`:DEP_i IDENTIFIER (LEMMA POS PREPOSITION ?ENAMEX)`
PV	`:DEP_i IDENTIFIER (LEMMA POS MOOD ?PREPOSITION)`

where :

- `LEMMA` is the LEMME value of the head. The head of a group is the rightmost form contained in that group which bears the appropriate part of speech for the group namely, *properNoun, commonNoun* for a GN or a GP and *verb* for a PV or an NV.
- `POS` is the POS value of the head.
- `ENAMEX` is the value of the ENTITÉ feature in a GN or a GP. The possible values for `ENAMEX` are : *dateTime, individual,location,mark,measure, organization,unnamed,URLetc*.
- `PREPOSITION` is the value of the LEMME value of the embedded preposition (form with POS value *preposition*).
- `MOOD` is the MOOD value of the verb form.

Dependents and relations. We extract as dependents all items that are related to the verb by one of the following relations : SUJ-V, ATB-SO, COD-V, CPL-V, MOD-V.

ATB-SO is typed by its third argument : s indicates a subject attribute (ATS), o an object one (ATO). Accordingly, an ATB-SO(_, _, s) relation is translated to an ATS dependency and an ATB-SO(_, _, o) to an ATO one.

In the extraction pattern, `DEP_i` , is then instantiated by the extracted dependency relation to one of : SUJ-V, ATB-SO, COD-V, CPL-V, ATS, ATO.

Subjects. The SUJ-V dependency requires special treatment.

1. If the tense is a compound one, the subject relation holds between the subject and the auxiliary while the auxiliary is related to the main verb by an AUX-V relation. Hence we also consider the sequence of relations AUX-V*.SUJ-V to extract subjects. There can be several auxiliaries between the subject and the main verb (*Jean doit avoir été mangé par une souris*.
2. If the verb/subject relation is mediated by an auxiliary, the lemme of the auxiliary immediately related to the verb must be extracted (this will be used to handle passive).
3. If the verb has mood *infinitive* or *participial*, then a subject with category GN and unknown lemma is added

PPs. Some PP dependents either contain no preposition, contain a preposition that needs normalising or have a preposition that is not directly accessible because of coordination. These are handled as follows.

Clitics with GP category (i.e., *lui, en, y*) are recorded as

```
:CPL_V IDENTIFIER LEMMA weakPersonalPronoun noPrep noSem)
```

Similarly, the *dont* relative pronoun is recorded as :

```
:CPL_V IDENTIFIER dont relativePronoun de noSem)
```

duquel is recorded as :

```
:CPL_V IDENTIFIER duquel relativePronoun de noSem)
```

auquel as :

```
:CPL_V IDENTIFIER duquel relativePronoun à noSem)
```

and *où* as :

```
:CPL_V IDENTIFIER où relativePronoun loc location)
```

PPs that do not contain a preposition but a partitive article are recorded as *de*-PPs :

```
:CPL_V IDENTIFIER LEMMA POS de ?ENAMEX)
```

PP that contain a `fusedPreposition` are recorded in such a way that the preposition is normalised i.e., *autour du* is recorded as *autour de*, *aux* is recorded as *à*. More generally, *au(x)* is normalised to *à* and *du* to *de*.

Finally, coordinated PPs are assigned the set of prepositions introducing each of the conjuncts.

1.5 Symbolic filtering

Before grouping together multiple occurrences of verb/dependents pairs, the symbolic filter normalises the extracted patterns in two ways.

Passive/active normalisation. E-PATTERNs which indicate a participial verb form with auxiliary *être* and a CPL_V dependent headed by *par* are converted to an E-PATTERN representing the corresponding active form.

Repetition reduction. This step of the extraction reduces sequences of dependents with identical function and category. The rationale for this is (i) that verbs rarely subcategorise for two or more PPs introduced by the same preposition and (ii) that a given grammatical function can only be realised once for any given verb.

So far, causative constructions with *faire* are not normalised although they should be. We plan to extend the first steps of the extraction process to normalise *X faire VINF ARG0 .. ARGN* sequences into appropriate frames. We also plan to use a list of verbs taking the *être* auxiliary in the passé composé, to appropriately distinguish passive from passsé composé sequences.

Easy relation	Dependent function	Condition
SUJ_V	SUJ	
COD_V	OBJ	
ATB_SO(_,_,s)	ATS	
ATB_SO(_,_,o)	ATO	
CPL_V	AOBJ	if prep = à or lemme = lui|leur|me|te|nous|vous|y
CPL_V	DEOBJ	if prep = de or lemme = en|dont
CPL_V	POBJ	
MOD_V	MOD	

FIG. 1.3 – Grammatical Function Mapping from the *easy*-format

1.6 Lexicon creation

The filtered E-PATTERNs produced are grouped by verb/dependents pairs and associated with frequency and relative frequency information. Specifically, for each Verb/Dependents pair found, a pattern of the following form is created :

```
#S(VDsPATTERN
:IDS verb_identifiers+
:TARGET |verb| ?aux=|AUX|
:DEPENDENTS a summary of the dependents for that entry
:FREQCNT frequency of the dependent list with the verb
:RELFREQ the relative frequency of the dependent list and the verb
:VMOOD (mood of the verb occurrences)
:SUJ (suject head and pos)
:OBJ (object head and pos)
:AOBJ (a-cpl head and pos)
:DEOBJ (de-cpl head and pos)
:POBJ (cpl-v head, pos and prep)
:ATS (ats head and pos)
:ATO (ato head and pos)
:MOD (adverbial head and pos)
)
```

The verb identifiers (`IDS`) list the identifiers of all occurrences of `|verb|` appearing in the corpus together with the dependent types listed in the `DEPENDENTS` line.

The `DEPENDENTS` line specifies the dependent list considered whereby each dependent is described by a `FUNCTION:CATEGORY` pair and where furthermore, functions and categories are derived by the mapping from the *easy* format given in Figures (1.4) and (1.3).

The last 9 fields of the entry records for each type of dependent, the head lemma, the POS tag and if any, the preposition or mood associated with each occurrence of the given dependent.

Easy POS tag	Dependent Category	Condition
GN	NP	
WEAKPERSONALPRONOUN	NP	
GP	PP[Prep]	if Prep $\neq$ à, de, NoPrep
PV	Sinf	if MOOD = *infinitive*
PV	Spart	if MOOD = *participle*
PV	Ssub	if MOOD = *conjuguated*
GA	AP	
GR	AdvP	

FIG. 1.4 – Category Mapping from the *easy*-format

1.7 Discussion

As mention in Section 1.3, the extraction process is still incomplete in that two steps are still missing, namely the frame filter which filters out entries whose frame is unknown and the statistical filter which uses statistical test to filter out lexical entries on the basis of frequency information.

The implementation of the statistical filter involves computing the appropriate counts, defining a gold standard and experimenting with different filters and threshold. This work is relatively well understood and will be carried out once the frame filter has been defined and applied.

For the frame filtering step, a repertory of the possible subcategorisation frames of French is first required. Such a repertory is relatively easy to construct in that it can be extracted from reliable existing lexicons such as, Dicovalence which is hand written, Synlex, which has been manually validated and Treelex, which was extracted from a manually built treebank. A more vexing issue however concerns the mapping between the EASYLEX frames and the frames listed by these resources. More specifically, two issues arise. First the tagsets of these resources are different and need to be aligned. This should be relatively easy. A second more complex issue is that the information referenced by these resources also differ. For instance, while Dicovalence distinguished between pseudo and real arguments, EASYLEX does not. The next important step in improving the lexicon extraction will therefore concentrate on identifying a list of frames that can be used to filter out implausible entries from EASYLEX.

1.8 Converting Dicovalence

One further aim of the project is to obtain a syntactic lexicon for French that contains exact information, includes relative frequency counts and is as complete as possible. To achieve this, we intend to merge the most reliable parts of existing lexicons (Lefff, Synlex, Dicovalence, Treelex, Lexschem) and to enrich the resulting lexicon with the frequency information extracted from corpora.

As a first step towards this aim, we converted Dicovalence to a format that better support a fusion/comparison with EASYLEX. In this format, a *Dicovalence*-lexical entry has the following syntax :

```
#EDV(LEXENTRY
:ID Dicovalence identifier
:TARGET |verb|
:FRAME the DV arguments after unfolding of DV and in the EASY format
:VFEATS syntactic information about the verb
:FREQCNT frequency of the dependent list with the verb
:RELFREQ the relative frequency of the dependent list and the verb
:EXAMPLE the DV example
:ENGLISH the DV english translation
)
```

Furthermore, an argument is represented as a `FUNCTION:CATEGORY` pair where the `FUNCTION` value is defined by the mapping defined in Figure 1.5 and the `CATEGORY` by the mapping defined in Figure 1.6. Similarly, the value of `VFEATS` is defined by the mapping defined in Figure 1.7.

Example *Dicovalence*-entry :

```
VAL$ rétablir: P0 P1
VTYPE$ predicator simple
VERB$ RETABLIR/rétablir
NUM$ 74630
EG$ il a été rétabli dans ses fonctions
TR_DU$ herstellen (in)
TR_EN$ reestablish
P0$ qui, je, nous, elle, il, ils, on, celui-ci, ceux-ci
P1$ qui, te, vous, la, le, les, se réfl., se réc., en Q, celui-ci, ceux-ci,
RP$ passif être, se faire passif
```

Converted *Dicovalence*-entry :

```
#EDV(LEXENTRY
:ID 74630
:TARGET |rétablir|
:FRAME  SUJ:NP OBJ:NP
:VFEATS passivable, se_faire_passive
:FREQCNT frequency found in corpus
:RELFREQ relative frequency
:EXAMPLE il a été rétabli dans ses fonctions
:ENGLISH reestablish
)
```

1.9 Conclusion

Further work on the syntactic lexicon will concentrate on the following points.

P0	SUJ
P1	OBJ
P2	AOBJ
P3	DEOBJ
PL	POBJ
PDL	POBJ
PM	POBJ
PMi	ATB
PP	POBJ
PP⟨PREP*⟩	POBJ
PQ	POBJ
PQ⟨PREP⟩	POBJ
PT	POBJ
PT⟨PREP⟩	POBJ
pseudo_XX	DUMMY
pseudo_XX⟨PREP⟩	DUMMY

FIG. 1.5 – Mapping Dicovalence paradigm to grammatical functions. The grammatical function is extracted either from paradigm description given by Dicovalence (P0, P1, P2, P3, PQ, PM, etc.) or from its category (e.g., pseudo_le). * is the kleene star, XX ranges over strings and PREP over preprositions (e.g., avec, pour, avant)

First, the results produced so far will be improved by applying a frame and a statistical filter as discussed in section 1.7. Depending on the data made available by the Passage participants, we will furthermore adapt and apply our syntactic lexicon acquisition method both to the output of several parsers and to more corpus data. This should in particular, improve the quality of the statistical filtering (“more data is better data”).

Second, the most reliable existing lexicons namely, Dicovalence, Treelex and Synlex will be merged and the resulting lexicon enriched in as far as possible with the frequency information extracted from corpora.

Third, we plan to use the lexicon obtained by lexicon fusion and addition of frequency information to define Levin type verb classes [Lev93]. Following [Sch06, Sch08], we will explore in how far the frame distribution of verbs supports verb clustering in ways that capture their syntactic and hopefully, semantic similarity.

Dicovalence	Easy-format
inf	Sinf
qpsubj	Ssub
qpind#	Sint
qpsubj#	Sint
sipind#	Sint
je,nous,vous,on,elle, ...	NP
réc.	NP[rec]
réfl	NP[refl]
PDL	PP[location]
PL	PP[location]
PL⟨PREP⟩	PP[loc,PREP]
PM	PP[manner]
PMI	PP
PP	PP
PP⟨PREP⟩	PP[PREP]
PQ	PP[quantity]
PQ⟨PREP⟩	PP[quantity,PREP]
pseudo_en	en
pseudo_il	il
pseudo_la	la
pseudo_le	le
pseudo_là	là
pseudo_se	se
pseudo_se_opt	?se
pseudo_soi[PREP]	soi[PREP]
pseudo_soi_opt[avec]	soi[?avec]
pseudo_ça	ça
pseudo_y	y
PT	PP[temporal]
PT⟨PREP⟩	PP[temporal,PREP]

FIG. 1.6 – Mapping Dicovalence categories to the EASYLEX-tagset. PREP stands for a specific preposition.

se faire passif	se_faire_passive
?se faire passif	?se_faire_passive
(se faire passif)	?se_faire_passive
?se passif	?middle
se passif	middle
être passif	passivable
?passif être	?passivable
(passif être)	?passivable
passif être	passivable
P0/P0 [below de_inf in P#]	subj_control
P0/P0 [below inf in P#]	subj_control
P0/P0 [below inf in PP(PREP)]	subj_control
P0/P0 [below inf in PP⟨PREP⟩]	subj_control
P0/P0 [below çà_inf in P#1]	subj_control
P1/P0 [below de_inf in P#0]	obj_control
P1/P0 [below inf in P0]	obj_control
P1/P0 [below inf in PP(PREP)]	obj_control
P1/P0 [below inf in PP⟨PREP⟩]	obj_control
P1/P0 [below ça_inf in P#]	obj_control
P2/P0 [below de_inf in P#]	aobj_control
P2/P0 [below il_de_inf in P#]	aobj_control
P2/P0 [below il_inf in P#]	aobj_control
P2/P0 [below il_çà_inf in P#]	aobj_control
P2/P0 [below inf in P0]	aobj_control
P2/P0 [below inf in PP(PREP)]	aobj_control
P2/P0 [below ça_inf in P#]	aobj_control
P3/P0 [below il_de_inf in P0]	deobj_control

FIG. 1.7 – Converting Dicovalence verb features

Chapitre 2

Induction de sens lexicaux

L'analyse sémantique de textes peut se diviser en deux grandes tâches : la désambiguïsation lexicale (WSD Word Sense Disambiguation) et l'annotation de texte en rôles sémantiques (SRL Semantic Role Labeling). Mon travail de cette année a surtout porté sur la manipulation de mots dans des espaces vectoriels sémantiques issus de cooccurrences syntaxiques. Dans ce type d'espace, un point représente un mot et une coordonnée représente un nombre d'occurences dans un contexte donné (ex : le nombre de fois que le mot chat occure dans le contexte sujet_de_manger).

Les espaces sémantiques utilisés sont ceux de la carte sémantique de [Gre07] construits à l'aide de l'analyseur syntaxique Lima du CEA LIST sur du corpus Web.

Les travaux effectués dans ce paradigme se partagent en 3 principales étapes :

Sylvain Loiseau, Docteur recrute le 01/12/2007 pour 12 mois

2.1 Réduction des espaces

On dispose d'un certain nombre de bases correspondant chacune à une relation syntaxique spécifique. Après restriction du vocabulaire aux 68000 mots les plus fréquents de la langue française, chacune des ces bases contient des matrices creuses à 68000 dimensions. Afin de diminuer les temps de calculs, les besoins en mémoire et de rendre ainsi nos espaces manipulables, on applique des fonctions de hachage spécifiques (Locality Sensitive Hashing) permettant aux vecteurs hashés de conserver une approximation de la distance cosinus [Ravichandran et al., 05]. Cette méthode pourrait être comparée à d'autres méthodes de réductions de dimensionnalité (LSA, SVD, Random Indexing...) mais elle semble plus adaptée à notre probléme.

2.2 Induction des sens de mots

Dans le même ordre d'idées que [Sch98, PL02], ou [Fer04], on cherche à induire des sens de mots par clustering. Contrairement à la perspective de [Sch98] dans laquelle ce sont les instances du mot source dans le texte à traiter qui sont clusterisés, on cherche cette fois à clustériser les termes plus proches voisins que l'on a déterminé au préalable lors de l'analyse des cooccur-

rences d'un corpus à part (notre matrice réduite). Notre méthode est proche de celle de [Fer04], à la différence que nos cooccurrences sont différenciées par la relation syntaxique les unissant. Pour prendre en compte la spécificité sémantique de chacun des espaces, on applique une méthode de clustering inter espaces de représentation (1 espace par relation syntaxique) inspirée de [KKPS04], chaque espace pouvant ainsi voter lors de la constitution des clusters. Les sens induits par ce procédé sont ensuite destinés à être mis en correspondance avec les sens de n'importe quel dictionnaire de référence (ressources manuelles). Ceux-ci étant en effet plus intelligibles pour l'humain et plus adéquats pour les évaluations utilisant des annotations humaines.

2.3 Désambiguïsation lexicale

Lors du traitement d'un texte à désambiguïser, chaque nouvelle instance rencontrée est transcrite à partir de ses contextes syntaxiques en autant de vecteurs-représentations que nous avons d'espaces à notre disposition, puis classifiée par un classifieur knn muti-représentations [KPS05] donnant un vote plus fort aux représentations dans lequelles les k éléments les plus proches sont le plus clairement attribués à un même cluster.

2.4 Implémentation

Les 3 étapes sont implémentées, elles sont en cours d'évaluation et de réglage des paramètres. Un article est en cours de rédaction pour présenter la méthode utilisée.

Ce travail est la première étape d'une expérience itérative visant à construire des espaces sémantiques à partir de corpus lexicalement désambiguisés et de voir si ceux-ci se révèlent d'un usage plus efficace dans des tâches de NLP telles que la désambiguïsation elle-même.

Lorsque le corpus Passage obtenu par ROVER sera disponible, il pourra être utilisé comme données source pour la construction d'une carte sémantique de meilleure qualité et ainsi participer à la constitution de meilleurs sens de mots. L'évaluation actuellement en cours pourra être effectuée sur ces nouvelles données.

Chapitre 3

Fouille d'erreurs

La mise au point de ressources lexicales riches en information et à grande couverture est une tâche complexe et de longue haleine. L'utilisation de techniques d'acquisition est bien sur utile pour acquérir de telles informations. Ce travail peut aussi être complété par de la fouille d'erreurs aidant à l'amélioration des ressources et à leur adaptation pour des nouveaux domaines, comme, par exemple, des sous-domaines techniques.

L'idée très simple est indépendante de la langue (dans une très large mesure) et indépendante des analyseurs. Initiée par des travaux de Van Noord [vN04], elle consiste à identifier les mots (*suspects*) qui apparaissent plus fréquemment que normal dans des phrases non analysables. Avec des corpus suffisamment gros, on obtient ainsi de bons résultats. Dans [SV06a, SV06b], nous avons raffiné cette approche en calculant de manière itérative un poids de suspicion d'autant plus élevé que le suspect apparaît en présence de mots qui ne sont pas suspects. L'algorithme ressemble par certains aspects à un algorithme d'apprentissage non supervisé de type EM (*Expectation Maximization*). L'utilisation d'une interface WEB adaptée permet le dépouillement des suspects et, en général, l'identification rapide des erreurs lexicales à corriger. Il est à noter que si beaucoup d'erreurs détectées sont de nature lexicales, on détecte également des erreurs liées aux particularités du corpus (méta-données, formattage spécifique, ...) et aux manques de la grammaire sous-jacente à l'analyseur.

Le croisement des informations rendues par plusieurs analyseurs permet de mieux identifier les erreurs non liées aux grammaires et donc plus sûrement lexicales, comme expérimenté dans [SV06b]. Ce genre d'approche est potentiellement possible dans le cadre de PASSAGE en exploitant les informations des divers analyseurs impliqués.

De manière plus dépendante des résultats d'un analyseur spécifique, il est possible d'utiliser la détection des suspects pour suggérer des corrections de ceux-ci. Pour un suspect donné, la fouille d'erreurs associe en effet un ensemble de phrases où celui-ci est le suspect principal. En oubliant les informations lexicales connues pour le suspect et en ré-analysant les phrases en question, l'analyseur fournit pour chaque phrase des informations sur le suspect rendant possible l'analyse. Sur l'ensemble des phrases, on peut espérer détecter des régularités dans les informations rendues, ces régularités traduisant l'information lexicale pertinente pour le suspect. Cette approche est en cours d'exploration avec les travaux de Lionel Nicolas [NFV07b, NSÉdLCF08, NFV07a]. Elle a également apporté une confirmation forte de la validité de l'algorithme de fouille d'erreur en montrant

une corrélation forte entre taux de suspicion et taux de ré-analyse réussies.

Enfin, ces techniques de fouilles d'erreurs sont largement généralisables, par exemple sur des séquences de mots, ou séquence de mélange de mots, lemme et catégories syntaxiques pour tenter de détecter des erreurs de nature syntaxiques ou liées à des expressions idiomatiques [SV06a, KVN09].

Plus généralement encore, nous avons montré dans le cadre de corpus botaniques que des extensions de l'algorithme de fouille peut aider à exploiter les régularités pour apprendre à désambiguiser des dépendances entre mots [RGV07, FVV07a, FVV07b, SÉdLC08] en partant de forêts partagées de dépendances. L'idée de base est que des occurrences peu ambiguës de dépendances réduisent progressivement l'ambiguïté d'autres occurrences de ces dépendances. Une telle désambiguisation est importante pour apprendre les rattachements plus plus probables, par exemple pour des cadres de sous-catégorisation ou des restrictions de sélection. Cette approche a été appliquée sur des forêts partagées mais doit pouvoir aussi être testée sur les analyses au format PASSAGE rendues par un ensemble d'analyseurs.

Bibliographie

[BC97] E. Briscoe and J. Carroll. Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, pages 356–363, Washington, DC, 1997.

[CF04] J. Carroll and A. Fang. The automatic acquisition of verb subcategorisations and their impact on the performance of an HPSG parser. In *Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP)*, pages 107–114, Sanya City, China, 2004.

[Fer04] O. Ferret. Discovering word senses from a network of lexical cooccurrences. In *Proceedings of the 20th international conference on Computational Linguistics*, 2004.

[FVV07a] Milagros Fernandez, Éric Villemonte de La Clergerie, and Manuel Vilares. From text to knowledge. In *Proc. of EUROCAST'07 (Eleven international conference on Computer Aided Systems theory)*, 2007.

[FVV07b] Milagros Fernandez, Éric Villemonte de La Clergerie, and Manuel Vilares. Knowledge acquisition through error-mining. In *Proc. of International Conference RANLP'07*, Borovets, Bulgaria, September 2007.

[Gre07] G. Grefenstette. Conquering language : Using nlp on a massive scale to build high dimensional language models from the web. In *Proc of the 8th CICLing Conference*, pages 35–49, Mexico, 2007.

[GV08] V. Gendner and A. Vilnat. Les annotations syntaxiques de référence peas. Technical report, LIMSI, 2008. version 1.11.

[IV94] N. Ide and J. Veronis. Multext : Multilingual text tools and corpora. In *Proceedings of COLING 94*, Kyoto, 1994.

[JMdR04] V. Jijkoun, J. Mur, and M. de Rijke. Information extraction for question answering : Improving recall through syntactic patterns. In *COLING-2004*, 2004.

[KKPS04] K. Kailing, H-P. Kriegel, A. Pryakhin, and M. Schubert. Clustering multi-represented objects with noise. In *Proc. 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2004*, Australia, 2004.

[Kor02] Anna Korhonen. *Subcategorization Acquisition*. PhD thesis, University of Cambridge, 2002.

[KPS05] H.-P Kriegel, A. Pryakhin, and M. Schubert. Multi-represented knn-classification for large class sets. In *10th International Conference on Database Systems for Advanced Applications (DASFAA 2005)*, Beijing, China, 2005.

[KVN09] De Kok and Gert Van Noord. A generalized method for iterative error mining in parsing results. In *Proc. of The 19th Meeting of Computational Linguistics in The Netherlands (CLIN)*, Groningen, NL, January 2009.

[Lev93] B. Levin. *English verb classes and alternations : a preliminary investigation*. Chicago University Press, 1993.

[MGM94] C. Macleod, R. Grishman, and A. Meyers. COMLEX syntax : Building a computational lexicon. In *Proceedings of COLING '94*, pages 268–272, 1994.

[NFV07a] Lionel Nicolas, Jacques Farré, and Éric Villemonte de la Clergerie. Confondre le coupable : Corrections d'un lexique suggérées par une grammaire. In *Proc. of TALN'07*, 2007.

[NFV07b] Lionel Nicolas, Jacques Farré, and Éric Villemonte de la Clergerie. Mining parsing results for lexical corrections. In *Proc. of the 3rd Language & Technology Conference (LTC)*, Poznań, Poland, October 2007.

[NSÉdLCF08] Lionel Nicolas, Benoît Sagot, Éric de La Clergerie, and Jacques Farré. Computer aided correction and extension of a syntactic wide-coverage lexicon. In *Proc. of CoLing 2008*, Manchester, UK, August 2008.

[PL02] P. Pantel and D. Lin. Discovering word senses from text. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2002*, 2002.

[RGV07] François Role, Milagros Fernandez Gavilanes, and Éric Villemonte de la Clergerie. Large-scale knowledge acquisition from botanical texts. In *Proc. of NLDB'07*, 2007.

[Sch98] H. Schütze. Automatic word sense discrimination. *Computational Linguistics*, 24(1) :97–123, 1998.

[Sch06] Sabine Schulte im Walde. Experiments on the automatic induction of german semantic verb classes. *Computational Linguistics*, 32(2) :159–194, 2006.

[Sch08] Sabine Schulte im Walde. The Induction of Verb Frames and Verb Classes from Corpora. In *Corpus Linguistics. An International Handbook.*, chapter 61. 2008.

[SV06a] Benoît Sagot and Éric Villemonte de La Clergerie. Error mining in parsing results. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 329–336, Sydney, Australia, July 2006. Association for Computational Linguistics.

[SV06b] Benoît Sagot and Éric Villemonte de La Clergerie. Trouver le coupable : Fouille d'erreurs sur des sorties d'analyseurs syntaxiques. In *Proc. of TALN'06*, pages 287–296, 2006.

[SÉdLC08] Benoît Sagot and Éric de La Clergerie. Fouille d'erreurs sur des sorties d'analyseurs syntaxiques. *In Traitement Automatique des Langues (T.A.L.)*, 49(1), 2008. à paraître.

[vN04] Gertjan van Noord. Error mining for wide-coverage grammar engineering. In *Proc. of ACL 2004*, Barcelona, Spain, 2004.