Acquisition de collocations verbe + nom à partir de représentations syntaxiques

Atelier sur les corpus spécialisés en terminologie

présenté par Brigitte Orliac
28 avril 2003

Introduction

- Construction d'un modèle d'acquisition automatique de collocations
- Encodage des combinaisons retenues pour des applications en TA
- Approche mixte basée sur des spécifications linguistiques et des méthodes statistiques

La collocation : une expression…

At the United States military's Central Command headquarters in Doha, Qatar, Navy Capt. Frank Thorp told reporters today: "I would say again that we fully expect that there are fierce battles ahead, that there continues to be resistance and that the overall objective of bringing down the regime has not yet been achieved. But it will be."

…non libre

Navy Capt. Frank Thorp spoke to reporters.

Navy Capt. Frank Thorp let out a cry.

Navy Capt. Frank Thorp screamed bloody murder.

Les fonctions lexicales : des outils pour représenter les collocations

- Des fonctions, au sens mathématique, qui définissent, pour chaque UL de la langue, l'ensemble des UL qui lui sont sémantiquement et syntaxiquement associées
- Chaque FL représente une signification de base qui ne peut être appréhendée qu'en pratique, lorsque la fonction est appliquée à une UL particulière
- Pour les FL qui modélisent des collocations, cette signification de base est associée à une relation syntaxique profonde

Exemples de fonctions lexicales

He let out a cry. **Oper**

He took a question from another reporter.

A cry came out. **Func**

A question came up.

Reporters bombarded him with questions. **Labor**

Les collocations verbe + nom spécialisées

- La cooccurrence verbale des termes est encore relativement peu décrite
- Les terminologies privilégient la description de la cooccurrence nominale
  - *cache memory, random access memory, read-only memory, virtual memory, etc.*
  - *batch file, data file, executable file, source file, etc.*
  - *anti-virus program, virus detection program, etc.*

Les collocations verbe + nom spécialisées

- Le discours spécialisé est aussi fait de constructions linguistiques plus larges
  - *the program executes, runs, terminates, works*
  - *[to] configure, close, create, download, install, load, remove, run, start, write a program*
  - *[to] back up, close, create, delete, download, load, name, open, save, upload a file*
  - *[to] keep, load, store something in memory*

Les collocations verbe + nom spécialisées

the program executes, runs, works

the program operates?

[to] close a file

[to] shut a file?

[to] load something into memory

[to] put something into memory?

Les collocations verbe + nom spécialisées

Cache memory stores both instructions or data.

**La mémoire cache emmagasine les instructions aussi bien que les données.*

You can enter and run a request.

**Vous pouvez introduire et diriger une demande.*

Etat de l'art en acquisition automatique

- Les premiers travaux portent sur l'acquisition de chaînes de caractères relativement proches l'une de l'autre (à l'intérieur de fenêtres)
- Développement de mesures d'association pour tester le statut collocationnel de chaque paire extraite (information mutuelle, coefficient de vraisemblance)

Etat de l'art en acquisition automatique

- De bons résultats en ce qui concerne les groupes nominaux anglais
- De moins bons résultats pour les combinaisons verbe + nom et pour les langues autres que l'anglais
- Les collocations manifestent des variations morphologiques et syntaxiques importantes dans les textes

Etat de l'art en acquisition automatique

Word can save files in Word format (i.e. with macros) under any extension.

Your current files are saved in \OS2\ARCHIVES\CURRENT.

By saving recently accessed files from the network, the computer …

This will allow you to save the file to your local hard drive in a place other than the cache directory.

Take a look at the services we explored, try a few out, and start saving and sharing files online.

Etat de l'art en acquisition automatique

- Les nouveaux programmes extraient les collocations de corpus étiquetés (étiquettes de catégories et de liens grammaticaux)
- Les mesures statistiques ne sont plus appliquées à des chaînes de caractères mais à des données linguistiques (séquences morpho-syntaxiques)

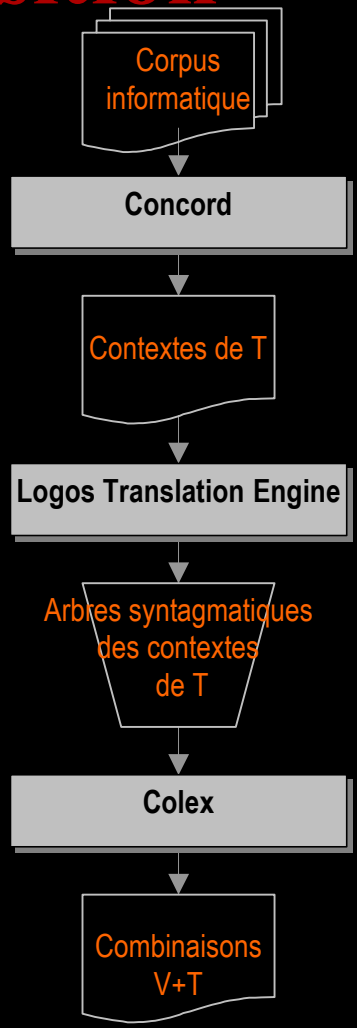
Notre approche

- Acquérir les combinaisons verbe + nom à partir de corpus analysés
  - Analyseur du système de TA Logos
- Valider les combinaisons candidates en utilisant une mesure d'association
  - *[to] boot a computer*
  - *[to] buy a computer*

Notre approche

- Nous acquérons les collocatifs verbaux des termes de l'informatique
- Le terme doit remplir l'un des trois rôles syntaxiques définis pour les bases des FL verbales
  - sujet
  - objet direct
  - objet indirect

Méthode d'acquisition



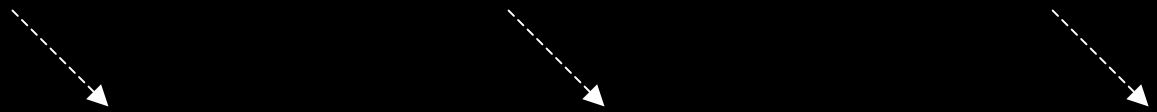
Analyse des contextes de T

All files you add to your Web site should be saved with the suffix .html…

```
[Punc    BOS      bos
]
[NP      file     pl
]
[Aux     should
]
[V       saved    pass
]
[NP      suffix   with
]
[Punc    EOS      un
]
[Punc    CLS-BOS      relCls
]
```

Extraction des combinaisons V + T

```
if(n_string (aux) v_passive (adv) n_obj*
```



NP_file_pl Aux_should_ V_saved_pass NP_suffix_with

```
then {v_passive + n_string + n_obj;}
```

{}→saved→file→with suffix

Résultats pour FILE dans le rôle d'objet direct

f :	22	save	file
f :	18	create	file
f :	18	download	file
f :	17	copy	file
f :	14	locate	file
f :	14	send	file
f :	9	store	file
f :	8	access	file
f :	8	delete	file
f :	8	edit	file
f :	8	find	file
f :	8	load	file
f :	8	read	file
f :	7	attach	file
f :	7	back up	file

Résultats pour OPEN

f :	29	open	file
f :	9	open	it
f :	9	open	window
f :	7	open	icon
f :	6	open	explorer
f :	5	open	program
f :	3	open	browser
f :	3	open	computer
f :	3	open	door
f :	2	open	arm
f :	2	open	copy
f :	2	open	DOS prompt
f :	2	open	page
f :	2	open	site
f :	1	open	address book

Quelques scores pour les collocatifs verbaux de FILE

	LogI	IM	F
open	16.05	3.77	24
copy	14.15	4.44	17
download	13.03	3.98	18
save	11.62	3.19	22
locate	8.87	3.61	14
back up	6.36	4.75	7
edit	5.74	3.94	8
upload	4.99	5.15	5
delete	3.90	3.00	8
attach	3.69	3.16	7
name	3.40	3.79	5
send	3.29	1.91	14
load	2.73	2.38	8
create	2.72	1.49	18
update	2.50	3.05	5

Sous forme graphique