Impact of Collocation Knowledge on Sentence Parsing

Eric Wehrli
LATL-CUI
University of Geneva

16.09.2013

Collocations are useful for NLP

Collocations have potentially a positive impact on NLP, for instance with respect to tasks such as

- word-sense ambiguity
- categorial ambiguity
- attachment (PP-attachment) ambiguity

Disambiguation

- Word sense ambiguity
  - standing room — place debout
  - significant lead — avance considérable
  - loose change — petite monnaie
- Lexical category ambiguity (noun vs. verb)
  - austerity measures — mesures d'austérité
  - labour costs — coût de la main-d'œuvre
  - budget rules — règles budgétaires
- Attachment ambiguity (PP attachment)
  - La ligne de partage des eaux — watershed
  - la force de maintien de la paix — peacekeeping force
  - l'organisation de protection de l'environnement — environment protection agency

Collocation identification with the Fips parser

- collocation identification is best performed on the basis of analyzed data ;
- occurs during parsing, after the application of a right (or left) attachment rule ;
- governing nodes are iteratively considered, halting at the first node of major category (N, V, Adj, Adv) ;
- consider the pair [governing item + governed item] – check whether it constitutes an entry in the collocation database.
- verify the (optional) restrictions associated with the collocation
  to take steps (prendre des mesures) vs. to take a step (faire un pas/avancer)

Collocation database

A collocation database has been added to our monolingual lexical databases, using the collocation extraction system developed by Violeta Seretan and others at LATL.

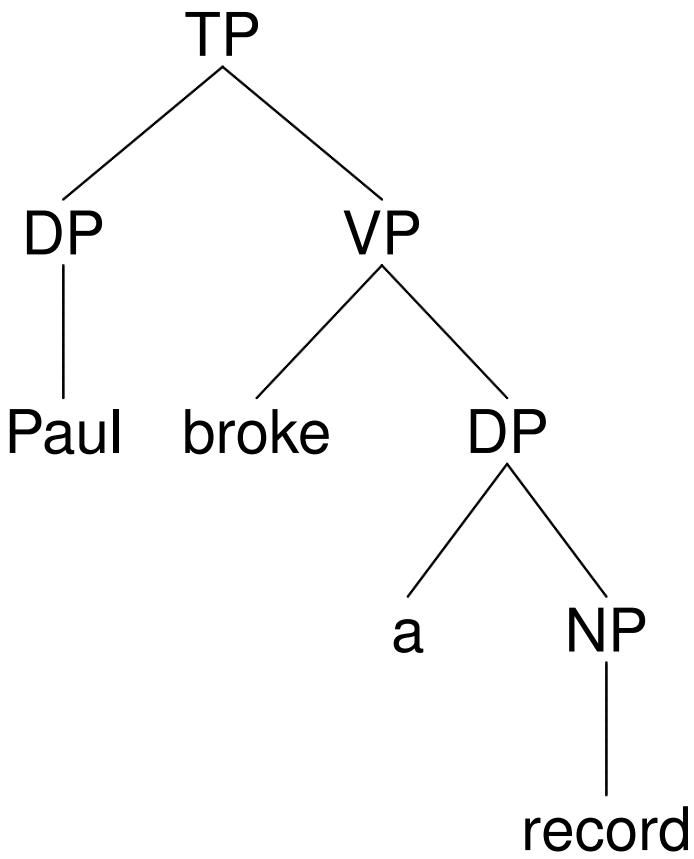
collocation type	English	French	German	Italian
adjective-noun	2,514	4,487	443	1,231
noun-noun	5,004	414	2,002	127
verb-object	554	1,218	168	227
others	1,055	9,495	338	1,418
total	9,149	16,135	3,166	3,061

TABLE: Number and types of collocations in Fips lexical database

Collocation identification - a simple example

(1)a. Paul broke a record

b. [$_{TP}$ [$_{DP}$ Paul] [$_{VP}$ broke [$_{DP}$ a [$_{NP}$ record]]]]



Collocation identification - more complex examples

- *wh*-interrogatives
  Which record did Paul break ?
  $[_{CP}\ [\ _{DP}$ which record$]_i$ did $[_{TP}$ Paul $[_{VP}$ break $[\ _{DP}$ e$]_i$ $]\ ]\ ]$
- relative clauses
  the record that Paul has just broken was very old
- *tough*-movement
  this record seems difficult to break
- *wh*-interrogative + *tough*-movement
  Which record did Paul consider difficult to break ?

another complex example

- left dislocation + small clause + *tough*-movement
  Ce record, Paul le considère difficile à battre
  "this record Paul considers very difficult to break"

  $[_{CP}\ [_{CP}$ ce record $]_i\ [_{TP}$ Paul le$_i$ considère $[_{DP}$ e$]_i$
  $[_{FP}\ [_{DP}$ e$]_i$ difficile $[_{CP}$ à $[_{TP}\ [_{VP}$ battre $[_{DP}$ e$]_i\ ]\ ]\ ]\ ]\ ]$

Research question

- What is the statistical significance of ambiguity resolution based on collocation knowledge ?
- How frequently, in a given corpus, does the detection of a collocation helps the parser make the "right" decision ?

Research program

- Parse a sizeable corpus **with** and **without** collocation detection component
- Compare results of both runs
- Problem : difficult to compare phrase-structure representations

  idea : use Fips as a POS-tagger, much easier to compare POS-tags than structures

Fips output (POS-tag mode) with collocation component

(2) The researchers estimated **the total worldwide labour costs** for the iPad at $33, of which China's share was just $8.

word	tag	position	collocation
the	DET	27	
total	ADJ	31	
worldwide	ADJ	37	
labour	NOUN	47	
costs	NOUN	54	labour costs

TABLE: Parser output with collocation knowledge

Fips output without collocation component

(3) The researchers estimated **the total worldwide labour costs** for the iPad at $33, of which China's share was just $8.

word	tag	position	collocation
the	DET	27	
total	ADJ	31	
worldwide	ADJ	37	
labour	NOUN	47	
costs	VERB	54	

TABLE: Parser output without collocation knowledge

Preliminary results

corpus : 24 articles from *The Economist* – 1672 sentences

- Analyze (tag) corpus with and without collocation component.
- Compare tags (categorial ambiguity).

	with collocations	without collocations
complete analyses	67.94%	67.82%
better tags	86	11
number of collocation	846	-

TABLE: POS-tagging with and without collocation knowledge

Remarks

- Preliminary results are positive and encouraging.
- Must be confirmed on a larger scale and for several languages (German, French, Italian, Spanish).
- Extend evaluation to other types of ambiguities (e.g., PP attachment, word-sense disambiguation)

Next steps...[1]

- Develop software
  - to automatically extract output differences in analyses (with and without collocation component)
  - store them in a database
  - display differences along with context (sentence)
  - let user select the best analysis through GUI and score results
- Enrich collocation database for all languages (en, de, fr, it, es)

Next steps... [2]

- To evaluate collocation contribution with respect to attachment ambiguities
  - modify Fips parser (tagger output) to display attachment type (adjunct vs complement) and lexical head of attachment node (host node)

Fips output with attachment dependencies

(4) Hundreds more are developing software and services to make sense of the sea of data on-line.

...

services	NOM	6058	
to	INF-MKR	6067	
make	VER	6070	DO :sense
sense	NOM	6075	OBJ
of	PRE	6081	nounPrepCompl sense
the	DET	6084	
sea	NOM	6088	
of	PRE	6092	nounPrepCompl sea
data	NOM	6095	
on-line ADJ	6100		
.	PONC	6107	

Word-sense ambiguity [1]

- To evaluate collocation contribution with respect to word-sense disambiguation
  - no significant impact on parsing precision
  - highly relevant for some applications, e.g. translation, IR
  - modify Fips parser (tagger output) to display lexeme information, e.g. lexeme database index (or lexeme correspondence in another language)

English	French
247	296

TABLE: Nouns with more than one lexeme

Word-sense ambiguity [2]

(5) Paul travaille dans un cabinet d'avocats.
"Paul is working in a law firm"

"cabinet" : agency, staff, study, toilet...

il	PRO	211000010	1
travaille	VER	211048661	4
dans	PRE	211045077	14
un	DET	211045922	19
cabinet	NOM	211014638	22
d'	PRE	211047305	30
avocats	NOM	211014481	32
.	PONC	0	39

TABLE: Parser output with collocation knowledge

Word-sense ambiguity [3]

Fips output (tagger mode) without collocation component

(6) Paul travaille dans un cabinet d'avocats.

il	PRO		211000010	1
travaille	VER		211048661	4
dans	PRE		211045077	14
un	DET		211045922	19
cabinet	NOM	211014638	211061132	22
d'	PRE		211047305	30
avocats	NOM	211014481	211057462	32
.	PONC		0	39

TABLE: Parser output without collocation knowledge

Conclusion

- Although the detection of collocations during parsing slightly increases parsing complexity (<10%), preliminary results show a clear and encouraging gain in quality.
- Our research effort within the PARSEME action will bear on an extensive evaluation of the contribution of collocation knowledge on parsing quality, in particular with respect to the cases of ambiguity just discussed.