Syntactic-based Collocation Extraction from Parallel Corpora and from the Web

Luka Nerima, Eric Wehrli, Violeta Seretan
LATL - Language Technology Laboratory
University of Geneva, Switzerland

http://www.latl.unige.ch
{Violeta.Seretan, Luka.Nerima, Eric.Wehrli}@lettres.unige.ch

Framework

- research project

  "**Linguistic Analysis and Collocation Extraction**"

  2002-2003

  Geneva International Academic Network (RUIG-GIAN)

  partner World Trade Organization (WTO)

- main goal
  - automatic extraction of multi-word terminology from texts
  - focus on collocations
  - improve the working environment of terminologists and translators

Outline

- Introduction
  - multi-word expressions classification
  - collocations - definitions, extraction methods
- Syntactic-based collocation extraction
  - syntactic candidate filtering
  - from bigrams to arbitrarily long collocations
- Integrated system:
  - collocation extraction, visualization and validation
- Collocate discovery using Web search

Problem

According ……**to**…… the delegate of the European Communities, it appeared that the measures had been ……**adopted**…… but they would only ……**enter into force**…… later in January 2004.

- collocability/ predictability/ preference
- beyond word level: multi-word expressions
- capture relations between words
  - lexicographic/ automatic means

Multi-Word Expressions

- lexical or syntactical units
- prevalent in language (as many as single words)
- rough classification:
  - compound words
    - service pack, address book, all of a sudden, in front of
  - idioms
    - be up in arms, have a frog in one's throat, be a fifth wheel, *entry into force*
  - collocations
    - massive investment, meet requirement, schedule appointment, depend on, weapons of mass destruction, numerical system, run through, The New York Stock Exchange
- collocations vs. compounds: syntactic flexibility
- collocations vs. idioms: semantic compositionality

Collocation. Two Definitions

- Sinclair, 1991:
  - "Collocation is the occurrence of two or more words within a short space of each other in a text''
  - general, statistical approach
    - words co-occurring more often than by chance
    - "arbitrary and recurrent word combination" (Benson, 1990)
- Manning and Schütze, 1999
  - "an expression consisting of two or more words that correspond to some conventional way of saying things"
  - restrictive, linguistic approach
    - "…each word has a particular and roughly stable likelihood of occurring as argument, or operator, of a given word" (Harris, 1988)

Collocation Acquisition

- lexicography
  - The BBI Dictionary of English Word Combinations (Benson et al., 1986)
  - Collins **COBUILD** English Language Dictionary (Sinclair, 1987)
  - Dictionnaire explicatif et combinatoire du français contemporain (Mel'cuk, 1984)
- automatic extraction
  - Sinclair 1991, Choueka et al. 1983, Church and Hanks 1990, Smadja 1993
  - statistical methods:
    - frequency counts, mobile window, independence hypothesis tests (t, $\chi^{2,}$ log-likelihood ratios), information theoretic measures (mutual information)

Collocation Extraction

1. candidate selection
   - word expressions (usually pairs) that may constitute collocations
   - usually no/very little syntactic processing
2. candidate ranking
   - order according to the collocational strength
   - based on words statistics in the corpus

Syntactic-based Collocation Extraction

- candidate selection
  - syntactic filter
    - candidate: not any pair of words, but only words in a given syntactic relation
  - collocation patterns:

    Adjective - Noun**,** Noun - [Pred] - Adjective, Noun - **Noun,** Verb - Prep, Verb - Prep - Argument, Noun - Prep - Noun, **Noun** - Noun, Adjective - Prep - Noun, …, Subject - Verb, Verb - Object

    nuclear weapon, custom administration, rely on, act of war, share fall, provide supply

- the (filtered) candidates are passed to the statistical test

Advantages

- no textual proximity limitation
  - A proposal for the financing of the variable costs will be made to the Committee…
  - usually, pure statistical methods limit the collocate search space to a window of 5 words (combinatorial explosion)
- distinction among different readings
  - disambiguation during parsing
  - pencher (to lean) vs. se pencher sur une question (to look into an issue)
- afford morpho-syntactic variation
  - words inflection - base word form (lemmatization)
  - inversion - cannonical position
  - extraposition - passivization, relativization, topicalization
    - at a $cost_i$ of $5 billion $that_i$ is chiefly being met e_i by South Korea and Japan

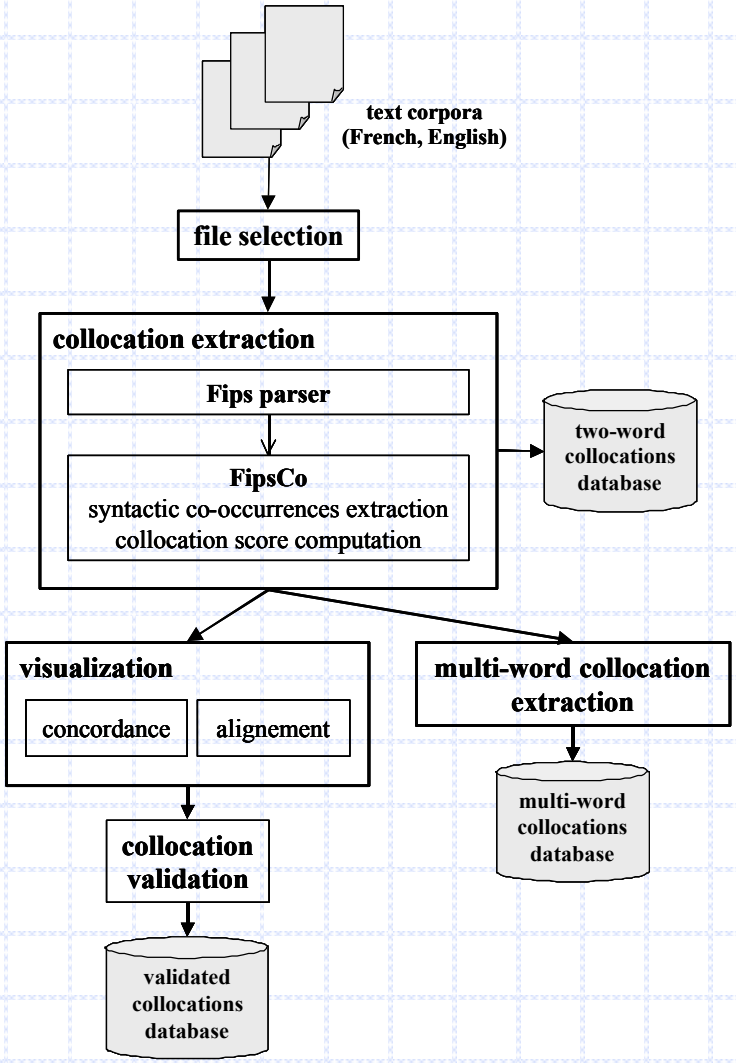
Multi-Word Collocation Discovery

- most concern is for bigrams (word pairs)
  - statistical relevance measures - appropriate for pairs of items only
- many collocations longer than two lexical items
  - round of presidential election, provide a steady supply, financial service supplier, join the euro system, abolish death penalty
- identify chains of bigrams that share common terms
  - round of election
  - presidential election
    - shared word: election
    - multi-word collocation candidate: round of presidential election
- iteratively linking bigrams -> arbitrarily long collocation (candidates)

The System

- Fips parser (Laenzlinger & Wehrli 1991)
  - based on an adaptation of Chomsky's Principles and Parameters theory
  - for English, French
  - FipsCo subsystem - syntactic co-occurrences (bigrams)
- log-likelihood ratios test (Dunning 1993)
  - collocation score
  - ranks bigrams
- visualization tools:
  - concordance
  - alignment in parallel corpora (alignment method)
- collocations validation

System Architecture



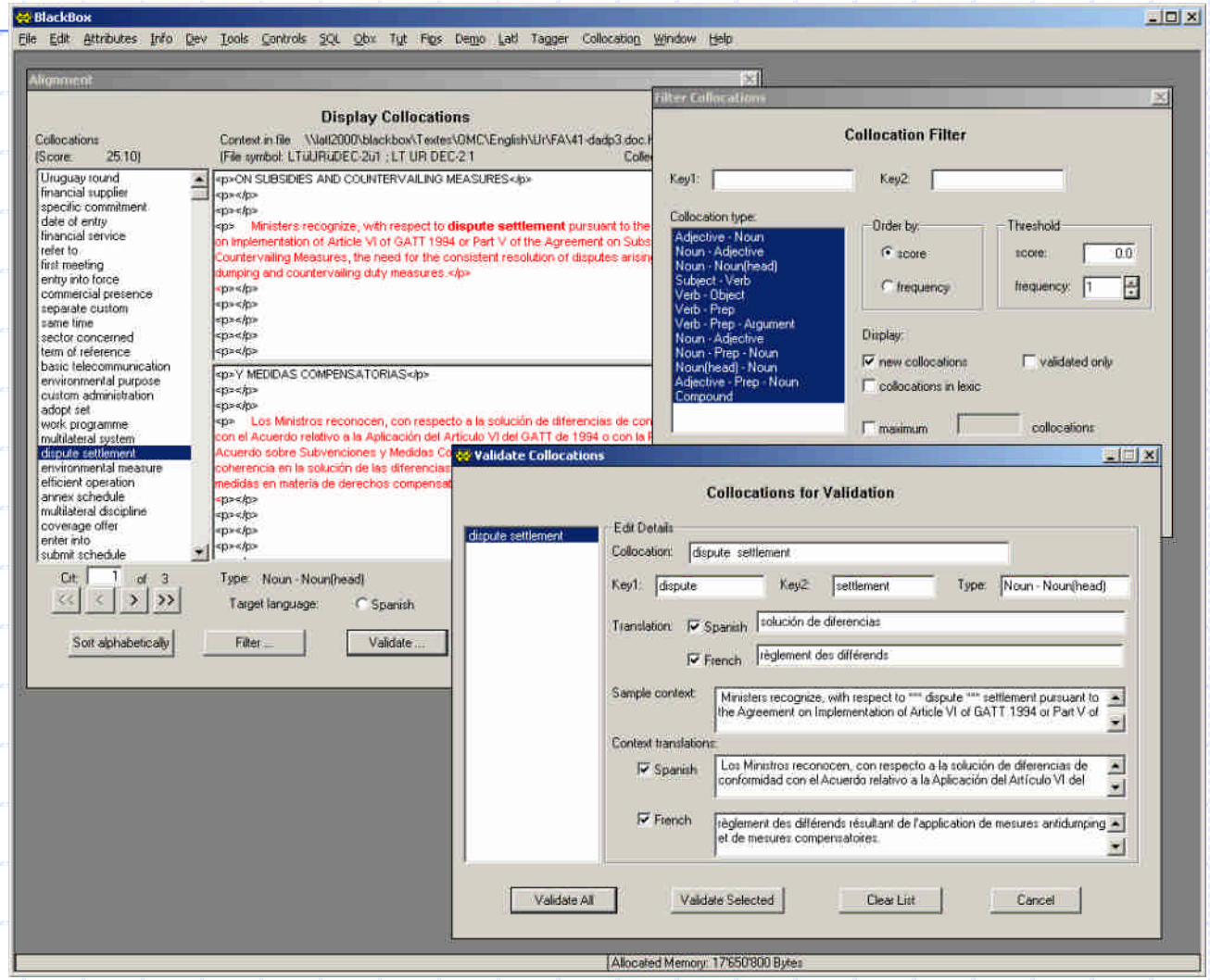
Experimental Results

Corpus	Size	Processing Time	Processing Speed	Bigrams Extracted	Tri-grams Extracted
The Economist	6.20 Mb 879'013 words	7'158 s	0.88 Kb/s 121.5 words/s	161'293 total 106'713 distinct	58'398 total 55'351 distinct
Le Monde	8.88 Mb 1'471'270 words	7'936.2 s	1.14 Kb/s 185.4 words/s	276'932 total 182'298 distinct	119'852 total 113'150 distinct

Bigrams		**Tri-grams**	
The Economist	**Le Monde**	**The Economist**	**Le Monde**
prime minister	milliard de franc	weapon of mass destruction	ministre de affaire étranger
last year	million de franc	have impact on	Front du salut national
mass destruction	premier fois	go out of	ministre de éducation national
interest rate	milliard de dollar	pull out of	tribunal en grande instance
next year	premier ministre	make difference to	président de conseil général
chief executive	Assemblée national	rise in to	membre de comité central
bin laden	Union soviétique	move from to	membre de bureau politique
poor country	million de dollar	rise from in	réaliser chiffre de affaire
central bank	affaire étranger	play role in	franc de chiffre de affaire
see as	fonction public	have interest in	chiffre de affaire de milliard

Top 10 bigrams ordered by the log-likelihood score, and the 10 most frequent tri-grams extracted

Demo


BlackBox
File Edit Attributes Info Dev Tools Controls SQL Obx Tut Figs Demo Lat Tagger Collocation Window Help
Alignment
Display Collocations
Collocations
(Score: 25.10)
Uruguay round
financial supplier
specific commitment
date of entry
financial service
refer to
first meeting
entry into force
commercial presence
separate custom
same time
sector concerned
term of reference
basic telecommunication
environmental purpose
custom administration
adopt set
work programme
multilateral system
dispute settlement
environmental measure
efficient operation
annex schedule
multilateral discipline
coverage offer
enter into
submit schedule
Cit: 1 of 3
Type: Noun - Noun(head)
Target language: Spanish
Sort alphabetically
Filter ...
Validate ...
Filter Collocations
Collocation Filter
Key1:
Key2:
Collocation type:
Adjective - Noun
Noun - Adjective
Noun - Noun(head)
Subject - Verb
Verb - Object
Verb - Prep
Verb - Prep - Argument
Noun - Adjective
Noun - Prep - Noun
Noun(head) - Noun
Adjective - Prep - Noun
Compound
Order by:
score
frequency
Threshold
score: 0.0
frequency: 1
Display:
new collocations
validated only
collocations in lexic
maximum
collocations
Validate Collocations
Collocations for Validation
dispute settlement
Edit Details
Collocation: dispute settlement
Key1: dispute
Key2: settlement
Type: Noun - Noun(head)
Translation: Spanish solución de diferencias
French règlement des différends
Sample context: Ministers recognize, with respect to "" dispute "" settlement pursuant to the Agreement on Implementation of Article VI of GATT 1994 or Part V of
Context translations:
Spanish Los Ministros reconocen, con respecto a la solución de diferencias de conformidad con el Acuerdo relativo a la Aplicación del Artículo VI del
French règlement des différends résultant de l'application de mesures antidumping et de mesures compensatoires
Validate All
Validate Selected
Clear List
Cancel
Allocated Memory: 17'650'800 Bytes

Collocate Discovery using Web Search. Frequency Counts

- Web corpus: availability, coverage, search tools
- comparing hits number

collocate candidate	base	hits*
widely	available	**880,000**
highly	available	**245,000**
largely	available	**3,690**

frequency suggests the collocate

*Google search engine, 4 May 2004

Syntactic Approach

- simple frequency counts - noisy
  - same context by chance (e.g., headings, not the same sentence)
  - unwanted category
  - no inflection
- aim: perform syntactic analysis of snippets
  - extract only co-occurrences:
    - syntactically-bound
    - the desired collocate category (co-occurrence type, e.g. Adjective - Noun)
  - afford morpho-syntactic variation

Method

- perform the search with the base word only as query
- build corpus of Web instances (search result snippets)
- syntactic analysis of sentences containing the base word
- extract co-occurrences of given type(s)
- apply statistical collocation test
- show the ordered list of co-occurrences and display context (sentence + link)

Details

- Google search engine
  - highest number of indexed pages
    - (3 billions -> 4.28 billions)
  - API access to search service
- advanced search parameters
  - retrieve only pages in a given language (homography)
- Observations:
  - French and English only (parser's languages)
  - limited access to Google results (key - 1'000 queries/day, only first 1'000 snippets)
  - time expensive:
    - mainly server search time and downloading
    - pre-processing (sentence boundaries)
    - parsing

Results. Evaluation Methods

- number of different bigrams and processing time for different results strata (average for 20 base words)

Snippets	100	200	300	500	750	1000
Bigrams	35	69.6	103.8	158.6	221.2	263.1
Search time	3.65	9.18	16.82	37.01	114.8	143.17
Parsing time	12.9	32.2	34.96	59.3	72.76	92.58

- interesting results even for few snippets (200)
- evaluation:
  - against BBI dictionary
  - students solving cloze exercises without/by using the tool

Example. Discussions

- collocates for "civilization"
- comparison with the BBI dictionary entry

	Verb-Object	Adjective-Noun	Other types
BBI only	spread, stamp out	advanced	
Common	introduce, create, destroy	ancient, modern	cradle of ~
Our tool only	develop	early, flourishing, human, new, noble	~ rise, ~ emerge, ~ extend, ~ grow, ~ fall, expansion of ~, fall of ~, collapse of ~, founder of ~, era of ~, development of ~

- improvements:
  - inflection for the base word
  - page source (directory category, page ranking)
  - speed up the results retrieval

Future Work

- corpus-driven investigation of collocations syntactic and semantic idiosyncrasy
- discovery of syntactic collocation types
  - extracting generic relations (specification, complementation, …)
  - compiling a list of interesting syntactic patterns
- evaluation of extraction
  - recall measure

    #collocations identified/#collocations in corpus
  - annotated resources (annotation tool)

References

- **Benson, M., Benson, E., and Ilson, R.** 1986. The BBI Dictionary of English Word Combinations. Amsterdam: John Benjamins.
- **Benson, M.** 1990. Collocations and general-purpose dictionaries. International Journal of Lexicography, 3(1):23--35.
- **Choueka, Y, Klein, S. T. and Neuwitz, E.** 1983. ``Automatic Retrieval of Frequent Idiomatic and Collocational Expressions in a Large Corpus''. Journal of the Association for Literary and Linguistic Computing, 4:1.34-38.
- **Church, K. and Hanks, P.** 1990. Word association norms, mutual information, and lexicography. Computational Linguistics, 16(1):22–29.
- **Dunning, T.** 1993. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1):61--74.
- **Harris, Z. S.** 1988. Language and Information. New York: Columbia University Press.
- **Laenzlinger, C. and Wehrli, E.**. 1991. Fips, un analyseur interactif pour le français. TA informations, 32(2):35--49.
- **Manning, C. and Schütze, H.** 1999. Foundations of Statistical Natural Language Processing. Cambridge, Mass.: MIT Press.
- **Mel'cuk, I. A. et al.** 1984, 1988, 1992, 1999. Dictionnaire explicatif et combinatoire du français contemporain: Recherches lexico-sémantiques I, II, III, IV. Montréal: Presses de l'Université de Montréal.
- **Sinclair, J.** 1987. Collins-Cobuild English language dictionary. Ed. by J. Sinclair. London: Collins. (CCELD).
- **Sinclair, J.** 1991. Corpus, Concordance, Collocation. Oxford: Oxford University Press.
- **Smadja, F.**. 1993. Retrieving collocations form text: Xtract. Computational Linguistics, 19(1):143--177.