NLP

Introduction to NLP

Collocations

Collocations (phrases)

- Dictionary definitions
  - Meaning of words in isolation
- "Know a word by the company that it keeps"
  - Firth 1935
- Examples
  - dead end
  - strong tea
  - Benazir Bhutto
  - Fabry disease

Collocations

- Properties
  - Common use
  - No general syntactic or semantic rules
  - Important for non-native speakers
- Collocation acquisition
  - Important for NLP

Types Of Multiword Sequences

- Idioms
- Free-word combinations
- Collocations

Examples

Idioms	Collocations	Free-word combinations
To kick the bucket	To trade actively	To take the bus
Dead end	Table of contents	The end of the road
To catch up	Orthogonal projection	To buy a house

Properties

- Arbitrariness: substitutions are usually not allowed:
  - Make an effort vs. *make an exertion
  - Running commentary vs. *running discussion
  - Commit treason vs. *commit treachery
- Language- and dialect-specific
  - Régler la circulation = direct traffic
  - Russian, German, Serbo-Croatian: direct translation of regulate is used
  - AE: set the table, make a decision
  - BE: lay the table, take a decision
  - “semer le désarroi” – “to sow disarray” – “to wreak havoc”
- Common in technical language
- Recurrent in context

Uses

- Disambiguation (e.g, “bank”/“loan”,“river”)
- Translation
- Generation

Types of Collocations

- Grammatical
  - come to, put on; afraid that, fond of, by accident, witness to
- Semantic
  - only certain synonyms
- Flexible
  - find/discover/notice by chance

Base-Collocator Pairs

- Base – bears most of the meaning of the collocation. Writers think of the base first. Foreign language speakers search by base. For decoding purposes, it is more appropriate to store the collocation under the collocator.

Base	Collocator	Example
Noun	verb	Set the table
Noun	adjective	Warm greetings
Verb	adverb	Struggle desperately
Adjective	adverb	Sound asleep
Verb	preposition	Put on

Extracting Collocations

- Most-common bigrams?
- Drop function words?
- Look at POS sequences?

Extracting Collocations

- Mutual information

$$I(x;y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

- Larger means stronger
- What if I(x;y) = 0?
- What if I(x;y) < 0?

Yule's Coefficient

A – frequency of pairs involving both W and X

B – frequency of pairs involving W only

C – frequency of pairs involving X only

D – frequency of pairs involving neither

$$Y = \frac{AD - BC}{AD + BC}$$

$$-1 \leq Y \leq 1$$

Example

	W	w			
X	A=800	C=180		A	800
x	B=160	D=80		B	160
				C	180
				D	80
				AD-BC	35200
				AD+BC	92800
					0.38

Example From The Hansard Corpus (Brown, Lai, And Mercer) – "Prime"

French word	Mutual information
sein	5.63
bureau	5.63
trudeau	5.34
premier	5.25
résidence	5.12
intention	4.57
no	4.53
session	4.34

Flexible And Rigid Collocations

- Example (from Smadja): “free” and “trade”

Total	p-5	p-4	p-3	p-2	p-1	p+1	p+2	p+3	p+4	p+5
8031	7	6	13	5	7918	0	12	20	26	24

Xtract (Smadja)

- The Dow Jones Industrial Average
- The NYSE's composite index of all its listed common stocks fell *NUMBER* to *NUMBER*

Translating Collocations

- Brush up a lesson, repasser une leçon
- Bring about/осуществлять
- Hansards:
  - late spring
  - fin du printemps
  - Atlantic Canada Opportunities Agency
  - Agence de promotion économique du Canada atlantique

Links

- Sample phrasal collocations
  - http://en.wiktionary.org/wiki/Appendix:Collocations_of_do,_have,_make,_and_take
- List of English language idioms
  - http://en.wikipedia.org/wiki/List_of_English-language_idioms
- Idiomsite
  - http://www.idiomsite.com/

NLP