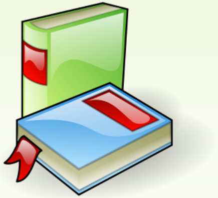
Getting our hands dirty with the mwetoolkit

Carlos Ramisch & Silvio Cordeiro

Outline

Lexical resources

- Essential to any NLP application
- Contain information about the ***lexical units***
- Structured data, more than a list of words
- Dictionaries, terminologies, thesauri, ontologies



Lexical units

- Language basic elements
- Building blocks
- Conveys a meaning (not single morphemic units like affixes and function words)
- Inflection does not make a new lexical unit (lexeme)
- Examples: *mouse*, *washing machine*, *pull off*

How does one build a lexicon? I

The standard approach

- years of work
- dozens of highly trained professionals
- thousands of dollars
- for humans or for machines (or for both)?
- high quality result

How does one build a lexicon? II

The "lazy" approach

- automatically learn lexical information from texts
- language independent
- cheap and dirty
- requires large amounts of text and high computational power
- contains noise and silence

Our goal

To discover a tool that automatically finds interesting MWEs in corpora, which can in turn help lexicon construction

Outline

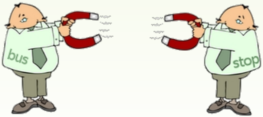
1. Warming up
2. Multiword expressions
3. The mwetoolkit

What are MWEs?

- *loan shark*
- *French kiss*
- *open mind*
- *vacuum cleaner*
- *voice mail*
- *high heel shoe*
- *make sense*
- *good morning*
- *take a shower*
- *upside down*
- ...

- *esprit ouvert*
- *à poil*
- *coup de main*
- *machine à laver*
- *talon aiguille*
- *avoir du sens*
- *bonne journée*
- *se rendre compte*
- *tant pis*
- *au revoir*
- ...

- *quebrar um galho*
- *lavar roupa suja*
- *cara de pau*
- *amigo da onça*
- *aspirador de pó*
- *fazer sentido*
- *tomar banho*
- *dar-se conta*
- *nem te conto*
- *depois de amanhã*
- ...



MWE: definitions

What is a word? What is a MWE? [Church, 2011]

- Habitual and customary places of a word [Firth, 1957]
- A unit whose exact meaning cannot be derived directly from the meaning of its parts [Choueka, 1988]
- Arbitrary and recurrent word combinations [?]
- A MWE has to be listed in a lexicon [Evert, 2004]
- Idiosyncratic interpretations that cross word boundaries (or spaces) [Sag et al., 2002]
- A combination of lexemes that must be treated as a unit at some level of linguistic processing. [Calzolari et al., 2002]

MWE: definitions

What is a word? What is a MWE? [Church, 2011]

- Habitual and customary places of a word [Firth, 1957]
- A unit whose exact meaning cannot be derived directly from the meaning of its parts [Choueka, 1988]
- Arbitrary and recurrent word combinations [?]
- A MWE has to be listed in a lexicon [Evert, 2004]
- Idiosyncratic interpretations that cross word boundaries (or spaces) [Sag et al., 2002]
- A combination of lexemes that must be treated as a unit at some level of linguistic processing. [Calzolari et al., 2002]

My favourite definition

Multiword expressions are. . .

. . . lexical items that:

- can be decomposed into multiple lexemes, and;
- display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity

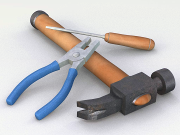
[Baldwin and Kim, 2010]

Automatic MWE lexicon construction

- Idea: capture regularities in word combinations
- Combinations: contiguous or not contiguous
- Number of words (> 2)
- Use of POS and syntax patterns for ***candidate extraction***
- Use of association measures and learning for ***candidate filtering***
- Support and speed up manual lexicographic work

Tools for MWE acquisition

- LocalMaxs – `hlt.di.fct.unl.pt/luis/multiwords/`
- Text::NSP – `search.cpan.org/dist/Text-NSP`
- UCS – `www.collocations.de/software.html`
- jMWE – `projects.csail.mit.edu/jmwe`
- Varro – `sourceforge.net/projects/varro/`
- Terminology extraction tools
- Many more



Limitations

- Focus on part of acquisition pipeline
- Depend on given language, formalism or tool
- Choice a priori of level of analysis
- Lack of integrated and systematic framework

Outline

1. Warming up
2. Multiword expressions
3. The mwetoolkit

The mwetoolkit

`http://mwetoolkit.sf.net`

Theory and Applications of Natural Language Processing Monographs

Carlos Ramisch

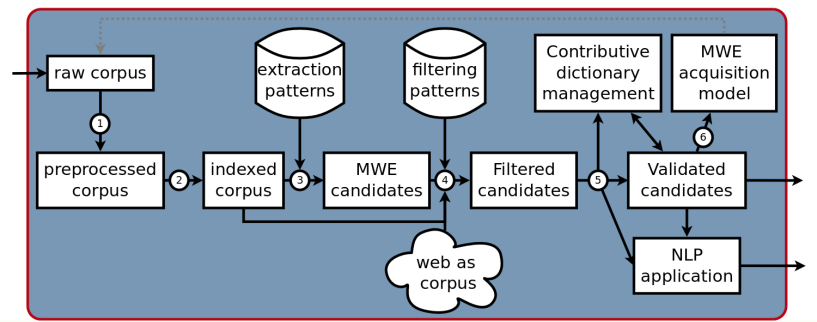
Multiword Expressions Acquisition

A Generic and Open Framework

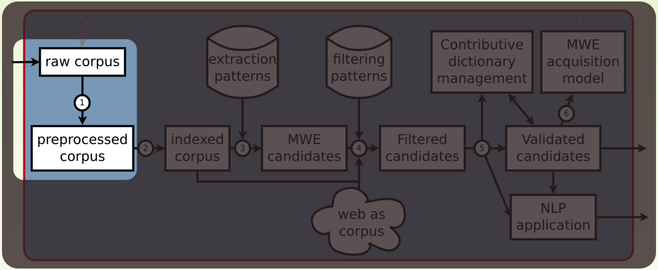
Springer

- Developed since 2009 [Ramisch et al., 2010b, Ramisch et al., 2010a, Araujo et al., 2011, Ramisch, 2015]

Acquisition pipeline



Preprocessing



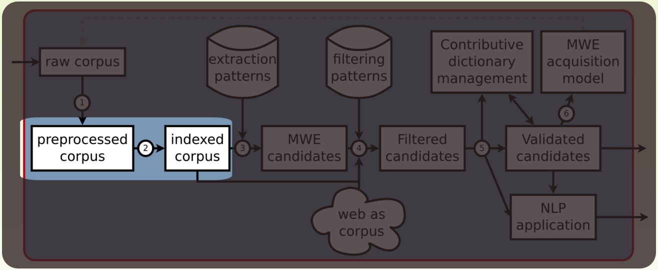
Preprocessing (external)

External tools + import

1. Tokenisation
2. Lemmatisation
3. POS tagging
4. Dependency parsing

Supports several file formats for corpora: CONLL, PlainText, TreeTagger, etc

2. Indexing


raw corpus

preprocessed corpus

indexed corpus

extraction patterns

filtering patterns

MWE candidates

Filtered candidates

Validated candidates

Contributive dictionary management

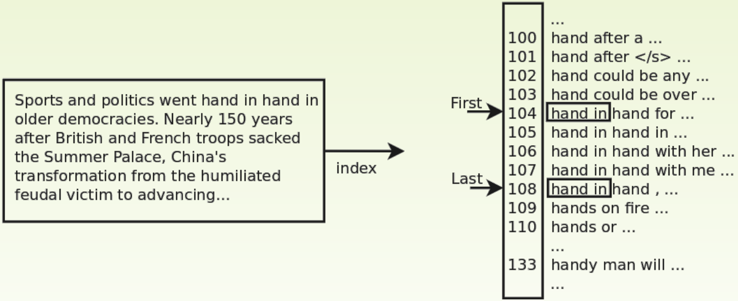
MWE acquisition model

NLP application

web as corpus

Indexing

- Suffix array



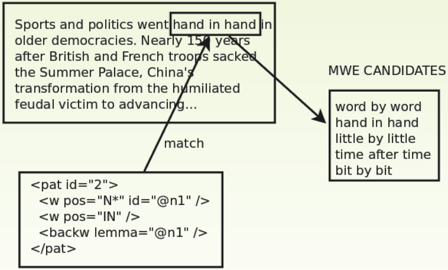
Sports and politics went hand in hand in older democracies. Nearly 150 years after British and French troops sacked the Summer Palace, China's transformation from the humiliated feudal victim to advancing...

	...
100	hand after a ...
101	hand after </s> ...
102	hand could be any ...
103	hand could be over ...
104	hand in hand for ...
105	hand in hand in ...
106	hand in hand with her ...
107	hand in hand with me ...
108	hand in hand , ...
109	hands on fire ...
110	hands or ...
	...
133	handy man will ...
	...

Candidate extraction

- *Inputs:* indexed corpus, extraction patterns
- *Outputs:* MWE candidates

- RegExp pattern
- Multilevel
- Back reference
- Wildcard
- NEW: negation, match length



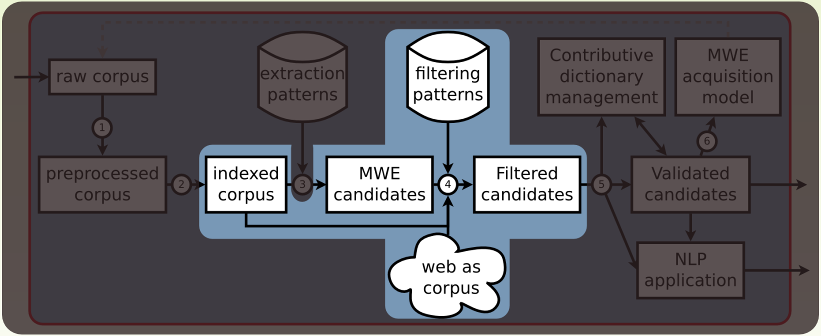
Sports and politics went hand in hand in older democracies. Nearly 150 years after British and French troops sacked the Summer Palace, China's transformation from the humiliated feudal victim to advancing...

```
<pat id="2">
  <w pos="N*" id="@n1" />
  <w pos="IN" />
  <backw lemma="@n1" />
</pat>
```

MWE CANDIDATES

word by word
hand in hand
little by little
time after time
bit by bit

Candidate filtering



Association measures I

- Compare expected count $E(w_1^n)$ and observed count $c(w_1^n)$

$$E(w_1^n) = \frac{c(w_1)c(w_2)\dots c(w_n)}{N^{n-1}}$$

- Some popular association measures

$$\begin{aligned} \text{t-score} &= \frac{c(w_1^n) - E(w_1^n)}{\sqrt{c(w_1^n)}} \\ \text{pmi} &= \log_2 \frac{c(w_1^n)}{E(w_1^n)} \\ \text{dice} &= \frac{n \times c(w_1^n)}{\sum_{i=1}^{n} c(w_i)} \end{aligned}$$

- Use of thresholds to remove noise

Association measures II

Contingency tables

	w_2	$\neg w_2$	
w_1	$c(w_1 w_2)$	$c(w_1 \neg w_2)$ $= c(w_1) - c(w_1 w_2)$	$c(w_1)$
$\neg w_1$	$c(\neg w_1 w_2)$ $= c(w_2) - c(w_1 w_2)$	$c(\neg w_1 \neg w_2)$ $= N - c(w_1) - c(w_2) + c(w_1 w_2)$	$c(\neg w_1)$ $= N - c(w_1)$
	$c(w_2)$	$c(\neg w_2)$ $= N - c(w_2)$	N

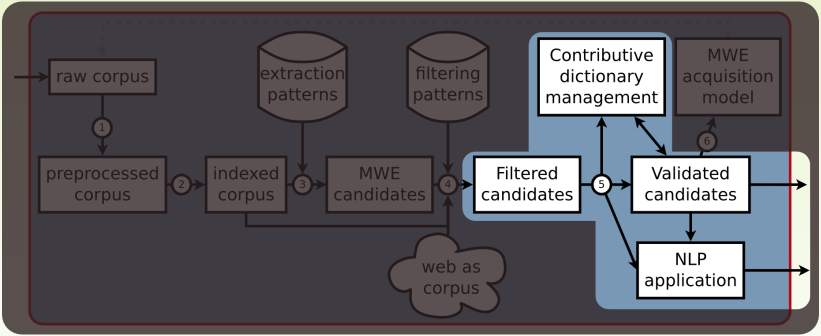
$$\mathrm{LL} = \sum_{w_i w_j} \log \left[\frac{c(w_i w_j)}{E(w_i w_j)} \right]^{c(w_i w_j)}$$

Stefan Evert's website `http://www.collocations.de`

Corpus annotation

1. Use information about source sentence generated by candidates.py (expressive regexp)
2. Project a lexicon on a corpus (independent resources)

Validation, evaluation, application



Evaluation of MWE acquisition

1. What are the acquisition goals (that is, the target applications) of the resulting MWEs?
2. What is the nature of the evaluation measures that we intend to use?
3. What is the cost of the resources (dictionaries, reference lists, human experts) required for the desired evaluation?
4. How ambiguous are the target MWE types?

Acquisition context

Generalisation of evaluation results depends on parameters of acquisition context

- Characteristics of target MWEs
  - Type
  - Language
  - Domain
- Characteristics of corpora
  - Size
  - Nature
  - Level of analysis
- Existing resources

Time to get our hands dirty...

1. Download the mwetoolkit tutorial files from the website
2. Open `commands.txt`, read and try the commands
3. For more detailed documentation, consult the mwetoolkit website

Acknowledgements

Silvio Cordeiro, Aline Villavicencio, Magali Sanches Duran, Evita Linardaki, Vitor de Araujo, Sandra Castellanos, CAMELEON & AIM-WEST projects

Merci beaucoup !
Muito obrigado!
Thank you!

Getting our hands dirty with the mwetoolkit

mwetoolkit team (`mwetoolkit@gmail.com`)

References I

Araujo, V. D., Ramisch, C., and Villavicencio, A. (2011).
Fast and flexible MWE candidate generation with the mwetoolkit.
In Kordoni, V., Ramisch, C., and Villavicencio, A., editors, *Proc. of the ALC Workshop on MWEs: from Parsing and Generation to the Real World (MWE 2011)*, pages 134–136, Portland, OR, USA. ACL.

Baldwin, T. and Kim, S. N. (2010).
Multiword expressions.
In Indurkhya, N. and Damerau, F. J., editors, *Handbook of Natural Language Processing*, pages 267–292. CRC Press, Taylor and Francis Group, Boca Raton, FL, USA, 2 edition.

Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lenci, A., Macleod, C., and Zampolli, A. (2002).
Towards best practice for multiword expressions in computational lexicons.
In *Proc. of the Third LREC (LREC 2002)*, pages 1934–1940, Las Palmas, Canary Islands, Spain. ELRA.

References II

Choueka, Y. (1988).
Looking for needles in a haystack or locating interesting collocational expressions in large textual databases.
In Fluhr, C. and Walker, D. E., editors, *Proceedings of the 2nd International Conference on Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications - RIA 1988)*, pages 609–624, Cambridge, MA, USA. CID.

Church, K. (2011).
How many multiword expressions do people know?
In Kordoni, V., Ramisch, C., and Villavicencio, A., editors, *Proc. of the ALC Workshop on MWEs: from Parsing and Generation to the Real World (MWE 2011)*, pages 137–144, Portland, OR, USA. ACL.

Evert, S. (2004).
The Statistics of Word Cooccurrences: Word Pairs and Collocations.
PhD thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart, Stuttgart, Germany.
353 p.

References III

Firth, J. R. (1957).
Papers in Linguistics 1934-1951.
Oxford UP, Oxford, UK.
233 p.

Ramisch, C. (2015).
Multiword Expressions Acquisition: A Generic and Open Framework, volume XIV of *Theory and Applications of Natural Language Processing*.
Springer.

Ramisch, C., Villavicencio, A., and Boitet, C. (2010a).
Multiword expressions in the wild? the mwetoolkit comes in handy.
In Liu, Y. and Liu, T., editors, *Proc. of the 23rd COLING (COLING 2010) — Demonstrations*, pages 57–60, Beijing, China. The Coling 2010 Organizing Committee.

Ramisch, C., Villavicencio, A., and Boitet, C. (2010b).
mwetoolkit: a framework for multiword expression identification.
In *Proc. of the Seventh LREC (LREC 2010)*, pages 662–669, Valetta, Malta. ELRA.

References IV

Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002).
Multiword expressions: A pain in the neck for NLP.
In *Proc. of the 3rd CICLing (CICLing-2002)*, volume 2276/2010 of *LNCS*, pages 1–15, Mexico City, Mexico. Springer.