MWE vs. NLP

MWEs from a Natural Language Processing perspective

PARSEME/ENeL workshop on MWE e-lexicons

Héctor Martínez Alonso
University of Paris-Diderot & INRIA (France)
hector.martinez-alonso@inria.fr

Overview

1. Common ground
2. MWE for NLP
   - Machine translation
   - Relation extraction
3. NLP for MWE, word association
   - Some applications
   - Pointwise mutual Information
4. Wrap-up

MWE Definition 2.1 from Ramisch (2015)

MWEs are lexical items that:

1. Are decomposable into **multiple lexemes**,
2. Present **idiomatic behaviour** at **some level** of linguistic analysis and, as a consequence,
3. Must be **treated as a unit** at some level of computational processing.

	lexical	syntactic	semantic	pragmatic	statistical
bye bye	+	+		+	+
ad hoc	+	+			+
give up		+	+		+
rely on		+			+
rocket science			+		+
washing machine			+		+
give a try			+		+
and so on		+	+		+
every now and then		+	+		+
drastically drop					+
yellow dress					
give a present					
several options					

1) Tokenization

Don't you know I'm John Mayer's taken-for-dead son, ma'am?

1) Tokenization and wordness status

To day (until XVI century)
To-day (until early XX century)
Today (well, *today*)

2) Idiomaticity: Morphosyntactic

By and large, they were criminals *at large*.

2) Variation in morphosyntactic fixedness

Ulica Obi-Wana Kenobiego in Grabowiec, Poland

MWE for NLP

1. Statistical Machine Translation
2. Relation Extraction

1) Statistical Machine Translation

It's raining cats and dogs

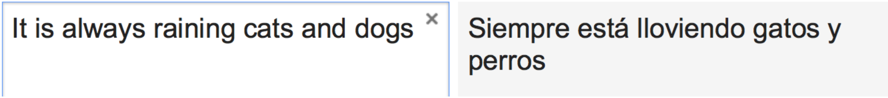
Lueve a cántaros

1) Statistical Machine Translation

It is always raining cats and dogs ×

Siempre está lloviendo gatos y perros

1) Statistical Machine Translation



(Counterargument: Maybe the idiom is already fixed at *It's*.)

2) Relation extration

We were trying to extract e.g. profession-product/activity pairs. Using patterns like *Person Created Entity*, with

1. *Person*, list of human terms, e.g. *plumber, child, Galileo.*
2. *Created*, list of creation verbs, e.g. *invent, make.*
3. *Entity*, the product or activity we want to identify.

E.g. *Galileo invented the telescope.*

2) Relation extraction: *Person Created Entity*

1. True Positive: Cobblers **made** shoes
2. True Negative: Mankind **brought** conflict
3. False positive: Teenagers **made** *out with* their classmates
4. False negative: Diplomats **brought** *about* negotiations

2) Relation extraction: *Person Created Entity*

1. True Positive: Cobblers **made** shoes
2. True Negative: Mankind **brought** conflict
3. False positive: Teenagers **made** *out with* their classmates
4. False negative: Diplomats **brought** *about* negotiations

Ignoring MWEs limited our predictive power.

NLP for MWE lexicography

1. Estimate compositionality
2. Help find glosses and examples
3. Identify syonymy
4. **Detect MWEs**

A two-word idiom

red herring (noun):

1. a dried smoked herring, turned red by the smoke.
2. a clue or information which is misleading or distracting.

bluff, ruse, feint, deception, subterfuge, hoax, trick...

Association between words: Pointwise Mutual Information

$$PMI(x;y) = \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$$

PMI, with words w_1 and w_2

$$PMI(w_1; w_2) = \log\left(\frac{p(w_1, w_2)}{p(w_1)\,p(w_2)}\right)$$

PMI, contribution of terms

$$PMI(w_1; w_2) = \log\left(\frac{p(w_1, w_2)}{p(w_1)p(w_2)}\right)$$

PMI, $w_1 = red$ and $w_1 = herring$

$$PMI(red; herring) = \log\left(\frac{p(red\ herring)}{p(red)p(herring)}\right)$$

What is the contribution of the numerator and the two terms of denominator and to the score?

Association between words: Mutual Information

$$PMI(x;y) = \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$$

1. Related but not equal to conditional prob. $P(x|y) = \frac{P(x,y)}{P(y)}$
2. PMI is not a prob and can be < 0 and > 1
3. $PMI(x;y) \neq PMI(y;x)$

Association between words: Mutual Information

Compare associations of *red car*, *red herring*, and *fresh herring*

w	p(w)	$w_1\ w_2$	$p(w_1\ w_2)$
red	0.00012	red car	0.00000004
fresh	0.00006	red herring	0.00000018
car	0.00007	fresh herring	0.000000015
herring	0.0000025	...	...

Association between words: Mutual Information

w	p(w)	$w_1\ w_2$	$p(w_1\ w_2)$
red	0.00012	red car	0.00000004
fresh	0.00006	red herring	0.00000018
car	0.00007	fresh herring	0.000000015
herring	0.0000025	...	...

$$MI(x;y) = p(x,y)\log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$$

MI(red herring) = 6.4

MI(red car) = 1.6

MI(fresh herring) = 4.3

A single metric does not explain it all... but it explains a lot!

★ ▽ ▽	puerto	rico	10.03
	hong	kong	9.73
	los	angeles	9.56
★ △ ▽	carbon	dioxide	9.10
	prize	laureate	8.86
	san	francisco	8.83
	nobel	prize	8.69
★ △ △	ice	hockey	8.66
	star	trek	8.64
	car	driver	8.41
■ △ △			...
■ △ △	and	of	-2.80
	a	and	-2.92
	of	and	-3.71

Wrapping up

1. NLP benefits from MWE knowledge
2. Lexicography

Questions and remarks

Thank you!