CSE528

Natural Language Processing

Venue:ADB-405 Topic: Multi-word Expressions

Prof. Tulasi Prasad Sariki,

SCSE, VIT Chennai Campus

www.learnersdesk.weebly.com

Contents

- What are Multi Word Expressions (MWE) ?
- Why care about MWEs ?
- MWE Characteristics & Classification
- MWE Extraction Methods
- MWE Extraction Evaluation

What are Multi Word Expressions (MWE) ?

- A language word - lexical unit in the language that stands for a concept.
  e.g. train, water, ability.
- However, that may not be true.
  e.g. Prime Minister.
- Due to institutionalized usage, we tend to think of
  Prime Minister as a single concept.
- Here the concept crosses word boundaries.

Defining a Multi Word Expression

A sequence, continuous or discontinuous, of words or other elements, which is or appears to be prefabricated: that is stored and retrieved whole from memory at the time from use, rather than being subject to generation or analysis by language grammar.

Defining a Multi Word Expression

In languages such as English, the conventional interpretation of the requirement of decomposability into lexemes is that MWEs must in themselves be made up of multiple whitespace-delimited words.

For example, *marketing manager* is potentially a MWE as it is made up of two lexemes (*marketing* and *manager*), while fused words such as *lighthouse* are conventionally not classified as MWEs.

Defining a Multi Word Expression

- Simply put, a multiword expression (MWE):
  - crosses word boundaries
  - is lexically, syntactically, semantically, pragmatically and/or statistically idiosyncratic.
- E.g. traffic signal, Real Madrid, green card, fall asleep, leave a mark, ate up, figured out, kick the bucket, spill the beans, ad hoc.

Defining a Multi Word Expression

Statistical idiosyncrasies (frequent use)

- Usage of the multiword has been conventionalized, though it is still semantically decomposable
- E.g. traffic signal, good morning

Lexical idiosyncrasies (one or more components of an MWE are not part of the conventional English lexicon)

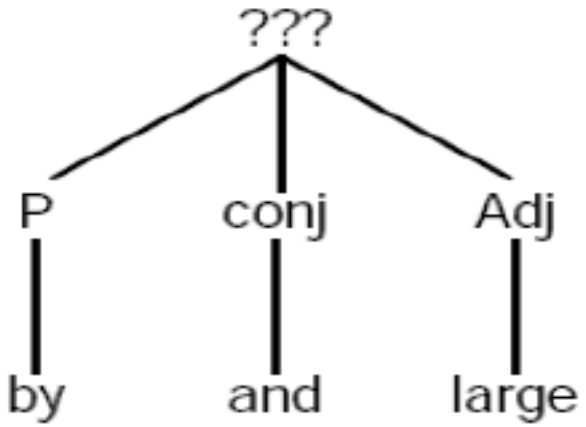
- Lexical items generally not seen in the language, probably borrowed from other languages
- E.g. ad hoc, ad hominem

Defining a Multi Word Expression

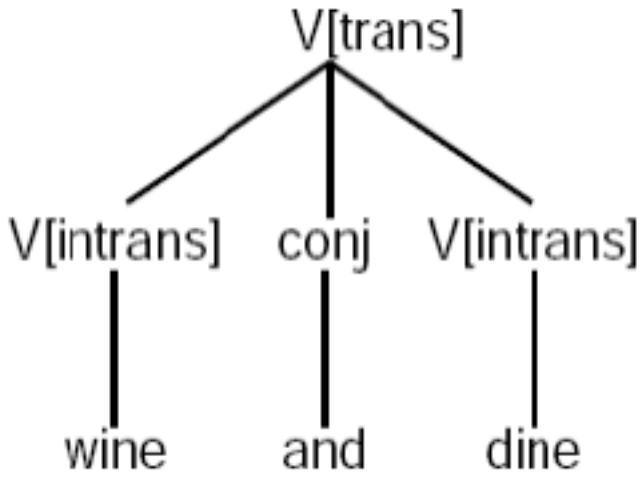
- Syntactic idiosyncrasy (occurs when the syntax of the MWE is not derived directly from that of its components).

Conventional grammar rules don't hold, these multiwords exhibit peculiar syntactic behaviour.

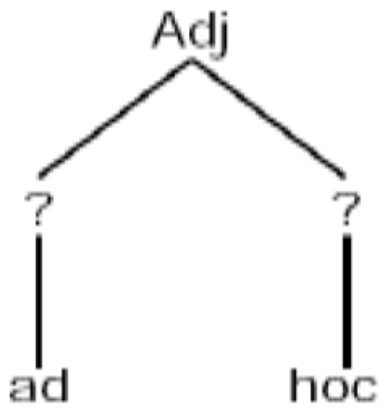
by and large



wine and dine



ad hoc



Defining a Multi Word Expression

- Semantic Idiosyncrasy (not being explicitly derivable from its parts)

The meaning of the multi word is not completely composable from those of its constituents.

This arises from figurative or metaphorical usage (literal usage)

The degree of compositionality varies

E.g. blow hot and cold – keep changing opinions

spill the beans – reveal secret

run for office – contest for an official post

Defining a Multi Word Expression

Semantic Idiosyncrasy

For example, *middle of the road* usually signifies “non-extremism, especially in political views,” which we could not readily predict from either *middle* or *road.*

Pragmatic idiomaticity is the condition of a MWE being associated with a fixed set of situations or a particular context

Defining a Multi Word Expression

Crosslingual variation

- There is remarkable variation in MWEs across languages

Single-word paraphrasability

- Single-word paraphrasability is the observation that significant numbers of MWEs can be paraphrased with a single word

Why care about MWEs?

- A large fraction of words in English are MWEs (41% in Wordnet). Other languages too exhibit this behaviour.
- Conventional grammars and parsers fail.
- Semantic interpretation not possible through compositional methods
- Pains for machine translation – word by word translation will not work
- New terminology in various domains likely to be multiword. Implications for information extraction.
- In IR, multiword queries mean multiword indexing

MWE Characteristics

- **Non-Compositionality**

  Non-decomposable – e.g. blow hot and cold

  Partially decomposable – e.g. spill the beans

- **Syntactic Flexibility**

  Can undergo inflections, insertions, passivizations

  e.g. promise(d/s) him the moon

  The more non-compositional the phrase, the less syntactically flexible it is

MWE Characteristics

- **Substitutability**

MWEs resist substitution of their constituents by similar words

E.g. ‘many thanks’ cannot be expressed as ‘several thanks’ or ‘many gratitudes’

- **Institutionalization**

Results in statistical significance of collocations

- **Paraphrasability**

Sometimes it is possible to replace the MWE by a single word

E.g. leave out replaced by omit

Classifying Multi Word Expressions

Based on syntactic forms and compositionality

Institutionalized Noun collocations

E.g. traffic signal, George Bush, green card

Phrasal Verbs (Verb-Particle constructions)

E.g. call up, eat up

Classifying Multi Word Expressions

Light verb constructions (V-N collocations)

E.g. fall asleep, give a demo

Verb Phrase Idioms

E.g. sweep under the rug

Extracting Multi Word Expressions

Basic Tasks

1. Extract Collocations

   Statistical evidence of institutionalization

   Use of hypothesis testing

   Maintain reasonably high recall

The Pointwise Mutual Information between two words is a measure of the strength of their collocation.

Pearson's chi-square test, log-lokely hood test

Extracting Multi Word Expressions

2. Establish linguistic validity of collocation

Not all collocations make linguistic sense

Use filters to remove invalid collocations

Use of POS tags and Use of Parsers

3. Measure semantic decompositionality of the MWE

Semantic idiosyncrasy an important characteristic of MWEness

Latent Semantic Indexing (Cosien Similarity)

END