Multiword Expressions

Dr. G. Bharadwaja Kumar

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.

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

- NLP tasks and applications may generate ungrammatical or unnatural output if they do not handle MWEs correctly

- Word sense disambiguation (WSD)

MWEs tend to be less polysemous than the individual words in it:

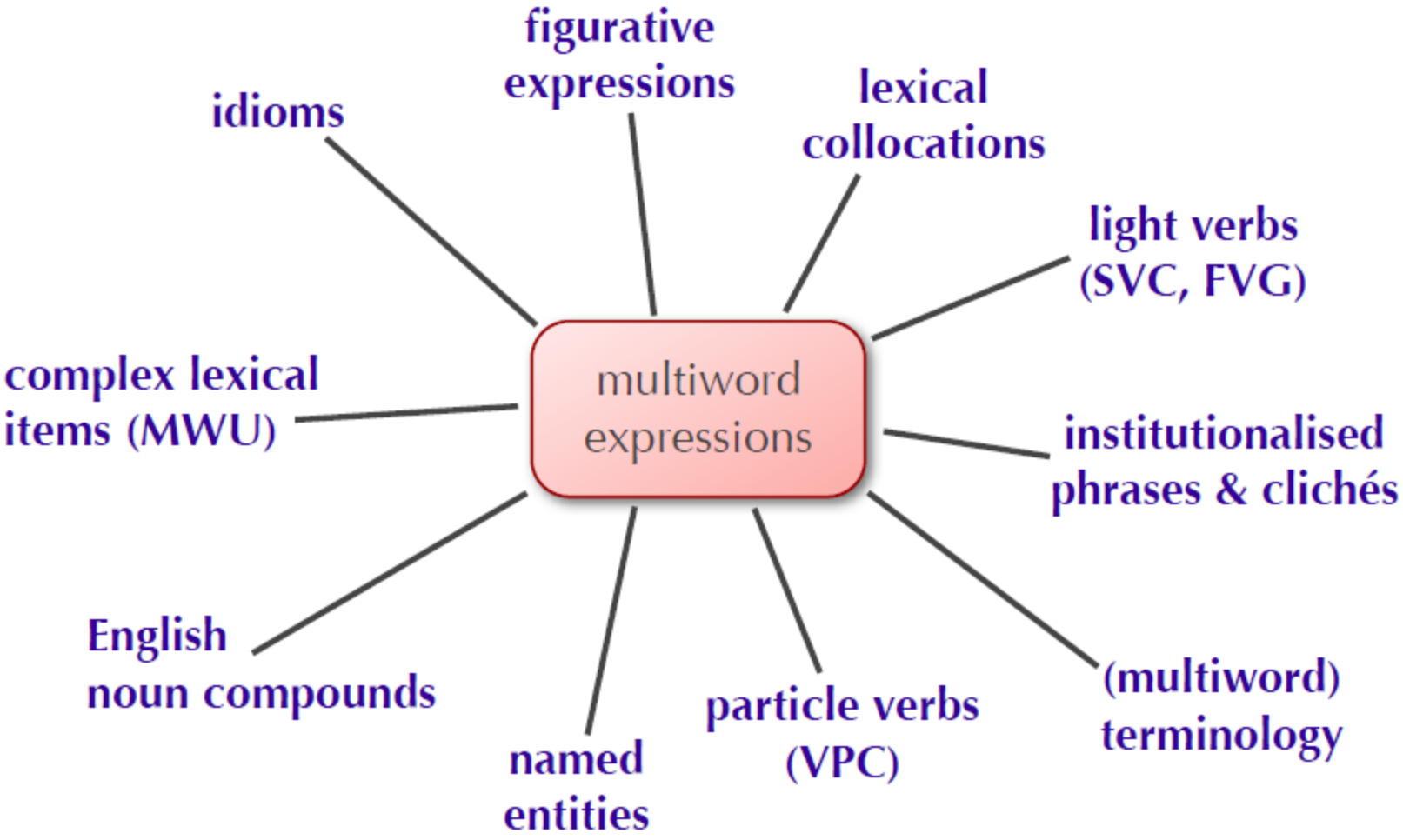
  - world = 9 senses in Wordnet
  - record = 14 senses
  - world record = 1 sense
- the importance of MWEs for WSD and for other semantic tasks like the annotation of semantic role labels
  - take a walk -> a walk is not complement of take

Information retrieval (IR)

- indexing MWEs to improve query accuracy:

- for rock star avoid retrieving irrelevant pages with:
  - geological descriptions of rocks or
  - astronomy websites about stars

Subtypes of multiword expressions



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
  - traffic signal, good morning
  - Black and White, Traffic Signal

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

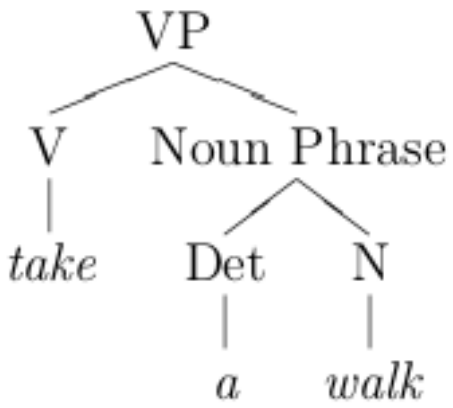
- Lexical items generally not seen in the language, probably borrowed from other languages
  - E.g. ad hoc, ad hominem

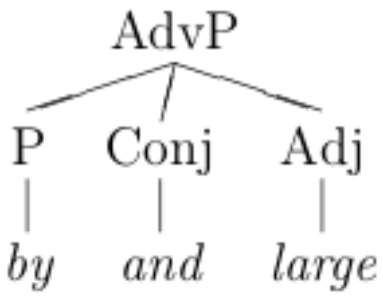
Defining a Multi Word Expression

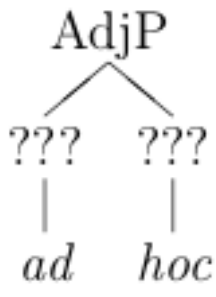
- Syntactic idiosyncrasy (occurs when the syntax of the MWE is not derived directly from that of its components).

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

- By and large, wine and dine







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.

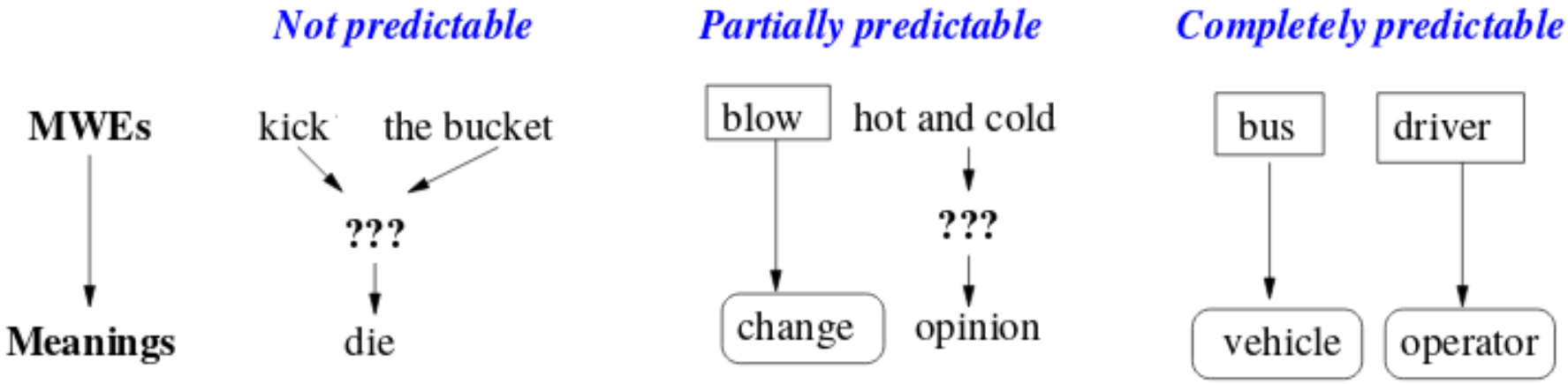
This arises from figurative or metaphorical usage (literal usage)

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


Not predictable
Partially predictable
Completely predictable
MWEs
kick
the bucket
???
Meanings
die
blow
hot and cold
???
change
opinion
bus
driver
vehicle
operator

Defining a Multi Word Expression

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

- Fall back

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

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**

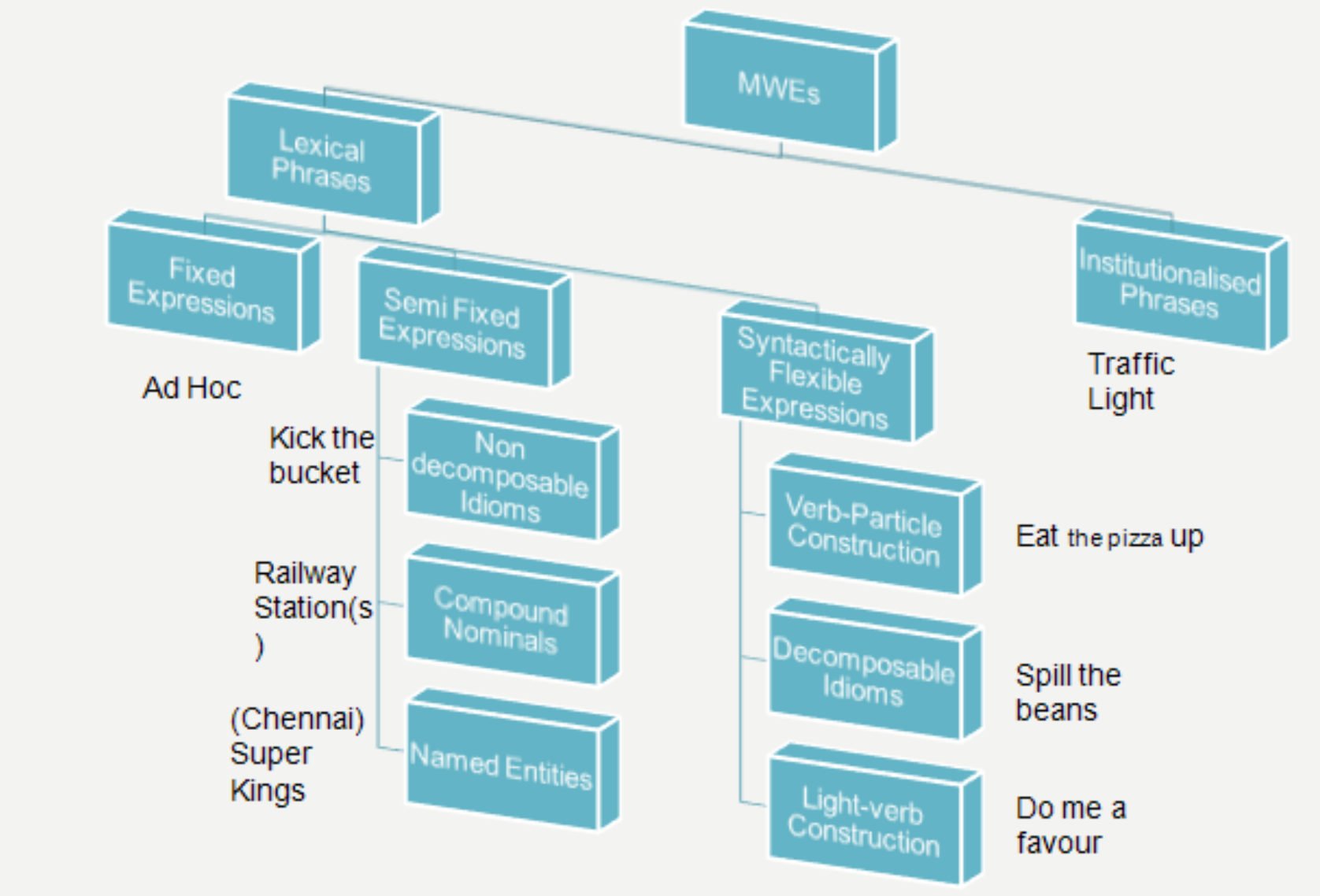
Sometimes it is possible to replace the MWE by a single word

E.g. leave out replaced by omit

Classifying Multi Word Expressions

Light verb constructions (V-N collocations)
E.g. fall asleep, give a demo
Verb Phrase Idioms
E.g. sweep under the rug

Classifying Multi Word Expressions

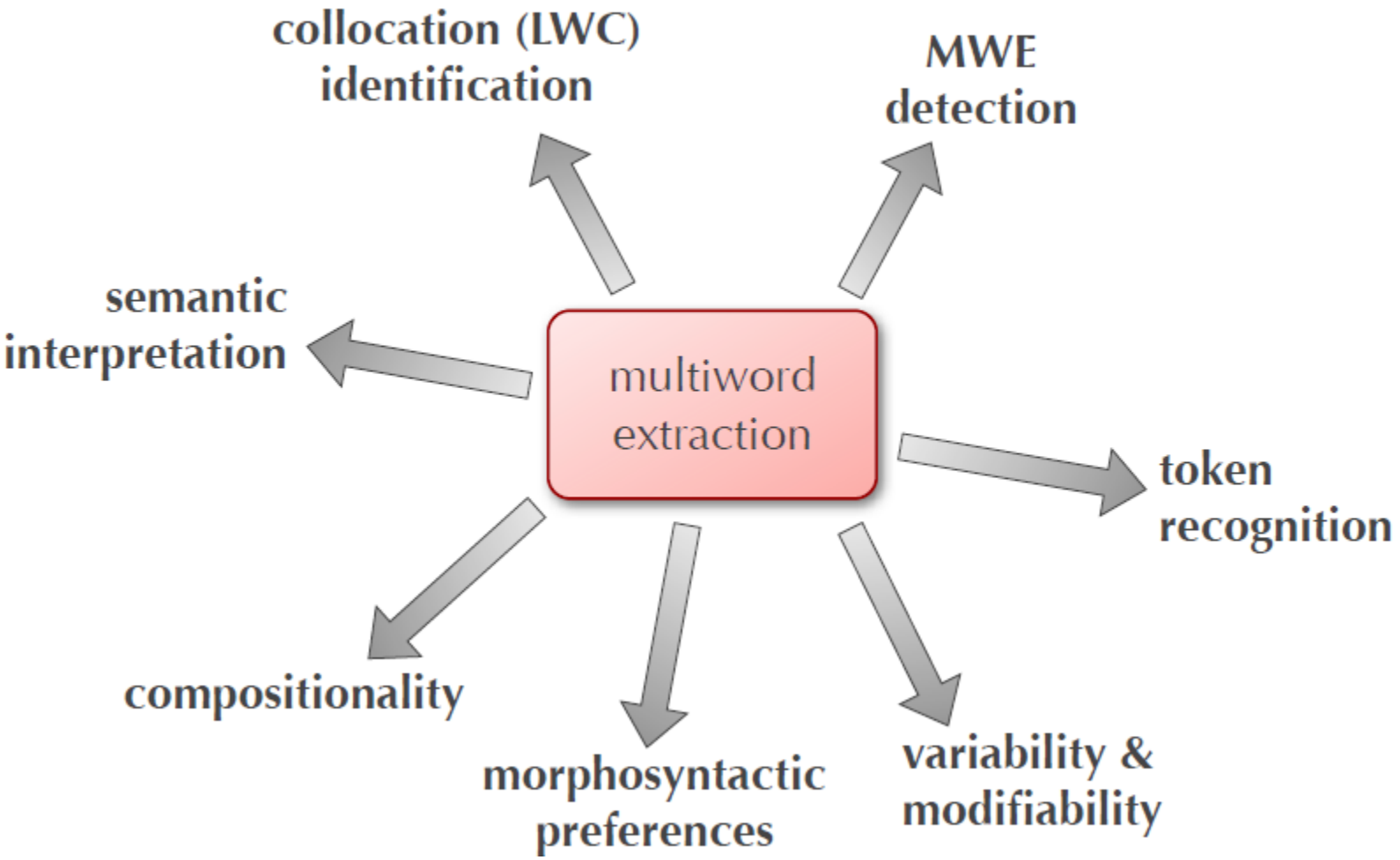


Testing an Expression for MWEhood

Classification of MWEs in Terms of Their Idiomaticity

	Lexical	Syntactic	Semantic	Pragmatic	Statistical
all aboard	–	–	–	+	+
bus driver	–	–	+	–	+
by and large	–	+	+	–	+
kick the bucket	–	–	+	–	+
look up	–	–	+	–	+
shock and awe	–	–	–	+	+
social butterfly	–	–	+	–	+
take a walk	–	–	+	–	?
to and fro	?	+	–	–	+
traffic light	–	–	+	–	+
eat chocolate	–	–	–	–	–

Multiword extraction tasks



Most Popular

- Shallow Statistical approaches because
  - for many languages scarcity of resources and tools
  - need for language independent approaches
  - for many MWE types lack of resources with linguistic patterns
  - need for type independent approaches
- development of knowledge-free general-purpose approaches
  - frequency
  - association measures

Extracting collocations

Basic Tasks

1. Extract Collocations

   Statistical evidence of institutionalization

   Use of hypothesis testing

   Maintain reasonably high recall

t-test

For example, in the corpus, new occurs 15,828 times, companies 4,675 times, and "new companies" occurs 8 times among the 14,307,668 bigrams

$$H_0 : P(newcompanies) = P(new)P(companies)$$

$$= \frac{15828}{14307668} * \frac{4675}{14307668}$$

$$\approx 3.675 * 10^{-7}$$

The observed frequency of occurence of new companies is 8 in the corpus.

$$\bar{x} = \frac{8}{14307668} \approx 5.591 * 10^{-7}$$

Now applying the t-test:

$$t = \frac{\bar{x} - \mu}{\sqrt{s^2/n}}$$

$$\approx \frac{5.591*10^{-7} - 3.675*10^{-7}}{\sqrt{\frac{5.591*10^{-7}}{14307668}}}$$

$$\approx .999932$$

This t value of 0.999932 is not larger than 2.576, the critical value for $\alpha = 0.005$. So we cannot reject the null hypothesis that new and companies occur independently and do not form a collocation.

Association Based measures

Pointwise mutual information

	without corpus level significance
word-based	PMI: $log \frac{f(x,y)}{f(x)*f(y)/W}$
document-based	PMId: $log \frac{d(x,y)}{d(x)*d(y)/D}$

C-Value

$$\text{c-value}(a) = \log_2 |a| (f_a - \frac{1}{|T_a|} \sum_{b \in T_a} f_b)$$

f_a is the frequency of a MWE candidate a , |a| is the number of words of a and T_a is the set of all MWE candidates that contains a

.

There are lot of association methods available in literature.

Two best read

Lyse, Gunn Inger, and Gisle Andersen. "Collocations and statistical analysis of n-grams: Exploring Newspaper Language: Using the Web to Create and Investigate a Large Corpus of Modern Norwegian" Studies in Corpus Linguistics, John Benjamins Publishing, Amsterdam (2012): 79-109.

Hoang, Hung Huu, Su Nam Kim, and Min-Yen Kan. "A re-examination of lexical association measures." Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications. Association for Computational Linguistics, 2009.

Uploaded on the website

Named Entity Recognition

NER Definition

Named entity recognition (NER) (also known as **entity identification (EI)** and **entity extraction**) is the task that locate and classify atomic elements in text into predefined categories such as the names of persons ations, express netary values,

Figure 1: An example of NER application on an example text

Named Entity Recognition

Main Categories of NERs are

- Person names
- Organizations (companies, government organisations, committees, etc)
- Locations (cities, countries, rivers, etc)
- Temporal (Date and time expressions)
- Number Expressions (money, number, ordinal and percent)

Example from MUC-7

<ENAMEX TYPE=„LOCATION“>Italy</ENAMEX>‘s business world was rocked by
the announcement <TIMEX TYPE=„DATE“>last Thursday</TIMEX> that Mr.
<ENAMEX TYPE=„PERSON“>Verdi</ENAMEX> would leave his job as vice-president
of <ENAMEX TYPE=„ORGANIZATION“>Music Masters of Milan, Inc</ENAMEX>
to become operations director of
<ENAMEX TYPE=„ORGANIZATION“>Arthur Andersen</ENAMEX>.

Named Entity Recognition

- Key part of Information Extraction system
- Information linking & Relation extraction
- Question Answering
- Machine Translation
- Summary generation
- Indexing
- increase accuracy of Internet search results (location Clinton/South Carolina vs. President Clinton)

Two Primary Approaches

- Rule based NER systems rely on handcrafted local grammatical rules written by linguists which is labor intensive and requires highly skilled labor.

- Grammar rules make use of gazetteers and lexical triggers in the context in which NEs appear.

- ML based systems on the other hand utilize learning algorithms that require large tagged data sets for training and testing.
  - Hidden Markov Model
  - Maximum Entropy
  - Conditional Random Field
  - Support Vector Machine

Problems in NE Task Definition

Category definitions are intuitively quite clear, but there are many grey areas.

- ✓Person vs. Artefact: "The **ham sandwich** wants his bill." vs "Bring me a **ham sandwich**."
- ✓Organisation vs. Location : "**England** won the World Cup" vs. "The World Cup took place in **England**".
- ✓Company vs. Artefact: "shares in **MTV**" vs. "watching **MTV**"
- ✓Location vs. Organisation: "she met him at **Heathrow**" vs. "the **Heathrow** authorities"

Entity Name Types

- **Persons** are entities limited to humans. A person may be a single individual or a group. Individual refer to names of each individual person. Group refers to set of individual
- **Location** entities are limited to geographical entities such as geographical areas like names of countries, cities, continents and landmasses, bodies of water, and geological formations.
- **Organization** entities are limited to corporations, agencies, and other groups of people defined by an established organizational structure

Named Entity Recognition Demo Results

The Named Entity Recognizer has identified the following named entities.

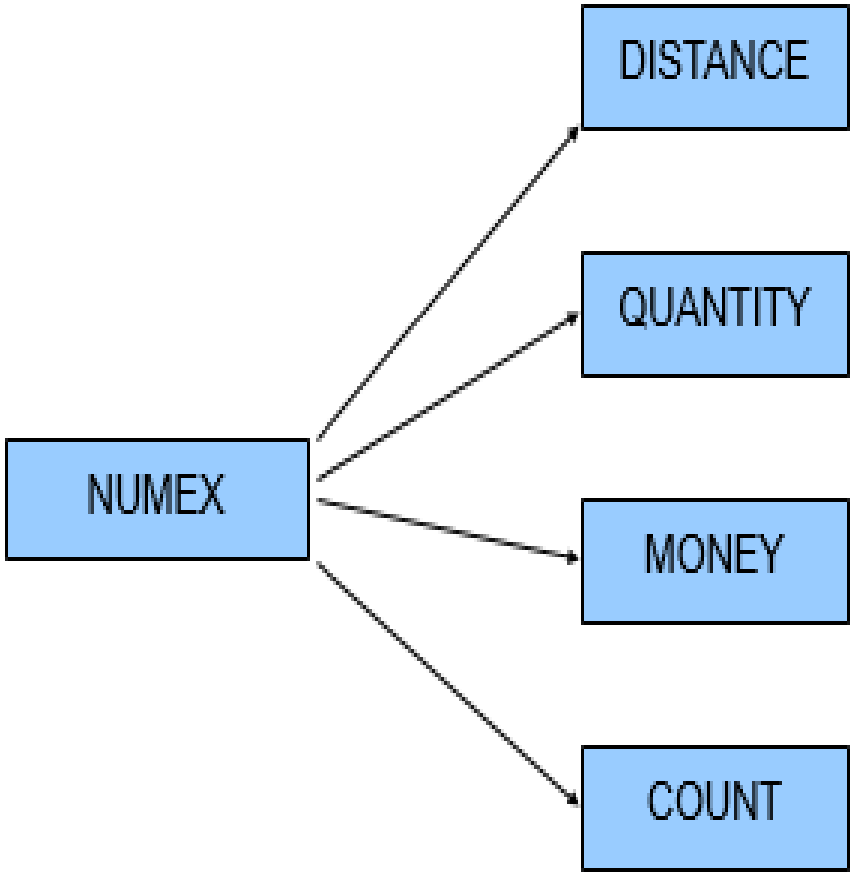
[LOC Houston] , Monday, July 21 -- Men have landed and walked on the moon. Two [MISC Americans] , astronauts of [MISC Apollo 11] , steered their fragile four-legged lunar module safely and smoothly to the historic landing yesterday at 4:17:40 P.M., Eastern daylight time. [PER Neil A. Armstrong] , the 38-year-old civilian commander, radioed to earth and the mission control room here: "[LOC Houston] , [ORG Tranquility Base] here; the Eagle has landed."

The first men to reach the moon -- [PER Mr. Armstrong] and his co-pilot, Col. [PER Edwin E. Aldrin] , Jr. of the [ORG Air Force] -- brought their ship to rest on a level, rock-strewn plain near the southwestern shore of the arid [ORG Sea of Tranquility] . About six and a half hours later, [PER Mr. Armstrong] opened the landing craft's hatch, stepped slowly down the ladder and declared as he planted the first human footprint on the lunar crust: "That's one small step for man, one giant leap for mankind."

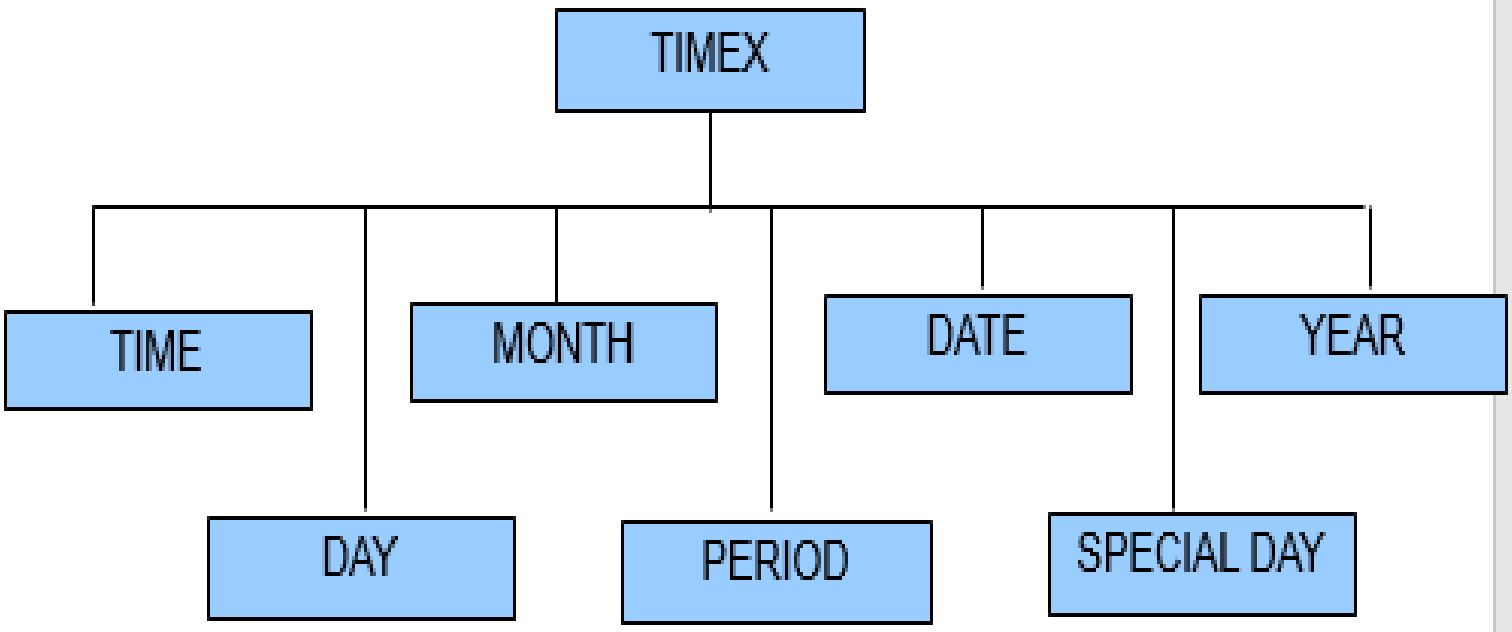
Key:

- PER - Person
- ORG - Organization
- LOC - Location
- MISC - Miscellaneous

Numerical Expressions



Time Expressions



Baseline: list lookup approach

System that recognises only entities stored in its lists (gazetteers).

Advantages - Simple, fast, language independent, easy to retarget (just create lists)

Disadvantages – impossible to enumerate all names, collection and maintenance of lists, cannot deal with name variants, cannot resolve ambiguity

Creating Gazetteer Lists

Online phone directories and yellow pages for person and organisation names (e.g. [Paskaleva02])

Locations lists

US GEOnet Names Server (GNS) data – 3.9 million locations with 5.37 million names (e.g., [Manov03])

UN site: http://unstats.un.org/unsd/citydata

Global Discovery database from *Europa technologies Ltd*, UK (e.g., [Ignat03])

Automatic collection from annotated training data

- Potential set of NE is too numerous to include in dictionaries/Gazetteers
- Names changing constantly
- Names appear in many variant forms
- Subsequent occurrences of names might be abbreviated

❑ list search/matching does not perform well

❑ context based pattern matching needed

Shallow Parsing Approach

Internal evidence – names often have internal structure. These components can be either stored or guessed.

location:

CapWord + {City, Forest, Center}

e.g. Sherwood Forest

Cap Word + {Street, Boulevard, Avenue, Crescent, Road}

e.g. *Portobello Street*

Shallow Parsing Approach

External evidence - names are often used in very predictive local contexts

Location:

"to the" COMPASS "of" CapWord

e.g. *to the south of* ***Loitokitok***

"based in" CapWord

e.g. *based in* ***Loitokitok***

CapWord "is a" (ADJ)? GeoWord

e.g. ***Loitokitok*** *is a friendly city*

Difficulties in Shallow Parsing Approach

Ambiguously capitalised words (first word in sentence)
[All American Bank] vs. All [State Police]

Semantic ambiguity
"John F. Kennedy" = airport (location)
"Philip Morris" = organisation

Structural ambiguity
[Cable and Wireless] vs. [Microsoft] and [Dell]
[Center for Computational Linguistics] vs. message from [City Hospital] for [John Smith].

Machine Learning NER

NED: Identify named entities using BIO tags

- B beginning of an entity
- I continues the entity
- O word outside the entity

NEC: Classify into a predefined set of categories

- Person names
- Organizations (companies, governmental organizations, etc.)
- Locations (cities, countries, etc.)

k Nearest Neighbor for classification

Learning is just storing the representations of the training examples.

Testing instance x_p:

compute similarity between x_p and all training examples

take vote among x_p k nearest neighbours

assign x_p with the category of the most similar example in T

Distance measures

Nearest neighbor method uses similarity (or distance) metric.

Given two objects x and y both with n values

$$x = (x_1, x_2, ..., x_n) \quad y = (y_1, y_2, ..., y_n)$$

calculate the Euclidean distance as

$$d(x, y) = \sqrt[2]{\sum_{i=1}^{n} |x_i - y_i|^2}$$

An Example

	isPersonName	isCapitalized	isLiving	teachesCS544
Jerry Hobbs	1	1	1	1
USC	0	1	0	0
eduard hovy	1	0	1	1
Kevin Knight	1	1	1	1

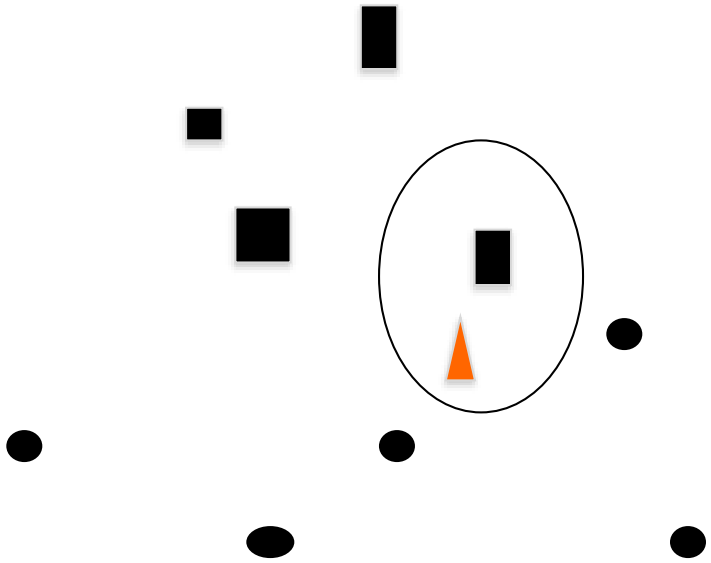
Euclidean distance:

$$d(JerryHobbs, USC) = \sqrt[2]{(1^2 + 0 + 1^2 + 1^2)} = 1.73$$

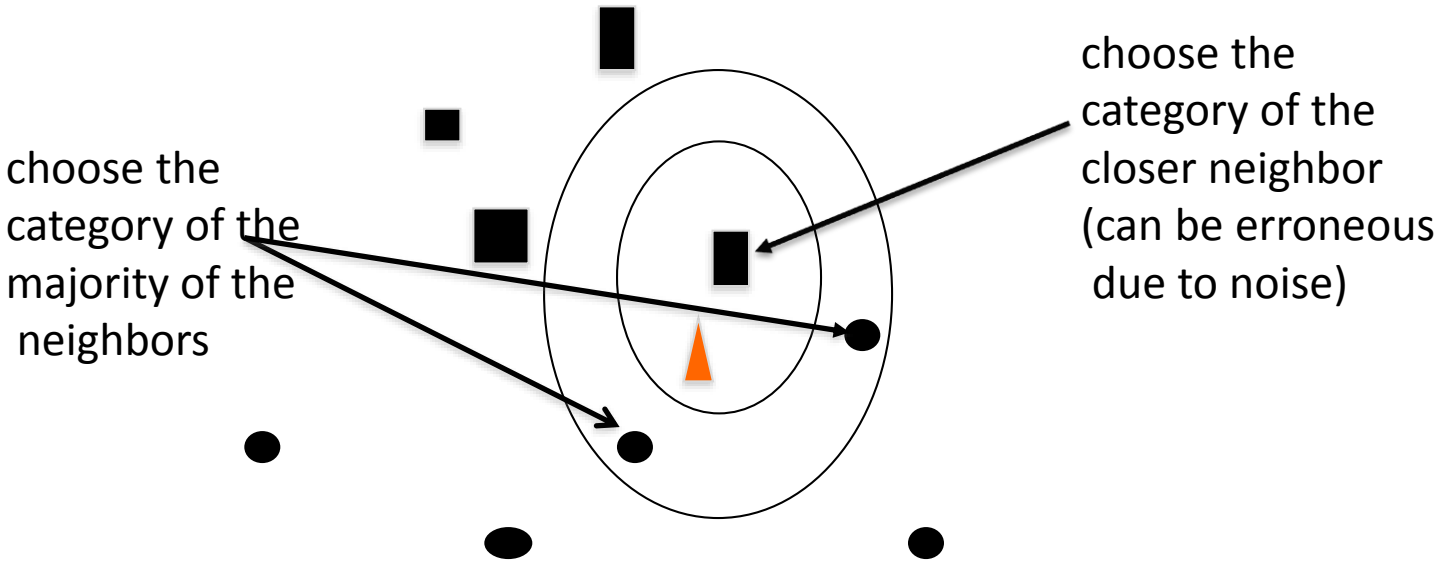
$$d(JerryHobbs, eduardhovy) = \sqrt[2]{(0 + 1^2 + 0 + 0)} = 1$$

$$d(JerryHobbs, KevinKnight) = \sqrt[2]{(0 + 0 + 0 + 0)} = 0$$

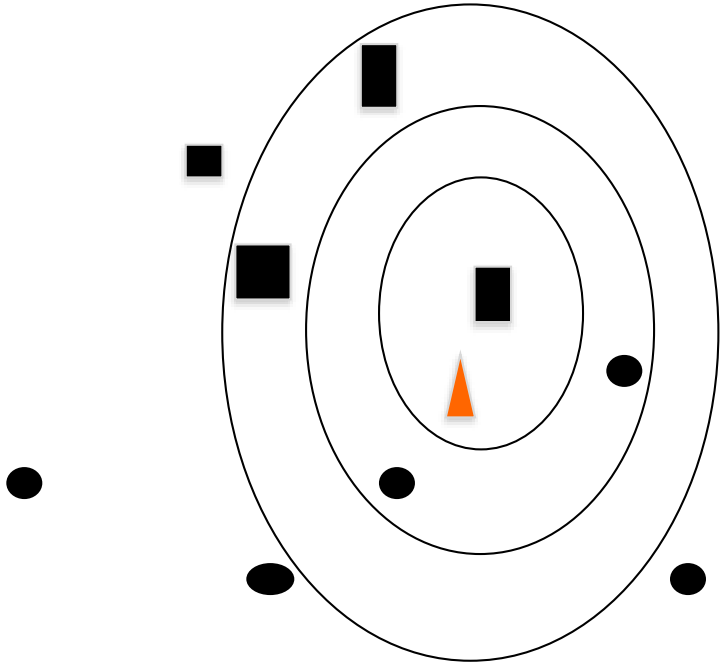
1-Nearest Neighbor



3-Nearest Neighbor



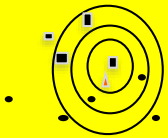
5-Nearest Neighbor



the value of *k* is typically odd to avoid ties

k Nearest Neighbours

Pros

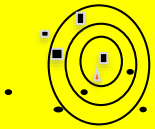


+ robust

+ simple

+ training is very fast (storing examples)

Cons



- depends on similarity measure & k-NNs

- easily fooled by irrelevant attributes

- testing is computationally expensive

Training Corpus

Steven	B-PER
Paul	I-PER
Jobs	I-PER
,	O
co-founder	O
of	O
Apple	B-ORG
Inc	I-ORG
,	O
was	O
born	O
in	O
California	B-LOC
.	O

Hidden-Markov Models (HMM) same way as used for POS tagging