NICF -TEXT ANALYTICS

MODULE 8: LINGUISTIC RESOURCES TO IMPROVE CONCEPTUALIZATION

Dr. Fan Zhenzhen

Institute of Systems Science

National University of Singapore

Email: zhenzhen@nus.edu.sg

Objectives

At the end of this module, you can

- Identify common text analytics artifacts or resources
- Develop such artifacts/resources based on domain knowledge

Outline

- Linguistic/knowledge resources and their roles in text analytics
- Dictionaries
  - General dictionaries
  - Synonym dictionaries
  - WordNet
  - Sentiment/Opinion Lexicon
- Defining patterns using regular expressions

Linguistic Resources

- In machine readable form
  - **Dictionaries** - valid terms, POS information, list of stop words, or words to be filtered
  - **Terminologies** – special domain words and phrases
  - **Patterns/rules** – for information extraction

Dictionaries

- Text analytics systems may be equipped with dictionaries in different languages for various purposes.
  - Terminology dictionaries for special domains or tasks
    - e.g. Biomedical domain
    - Customer Relation Management
    - IT
    - Market Intelligence
    - Opinions Mining, etc.

Valid Term Dictionary

- A list of valid terms
  - Only terms in the dictionary appear in the document term vector or matrix.
  - It helps to restrict the dimension of the matrix a priori and to focus on specific terms for distinct text mining contexts.
- It may include useful information such as POS

Filter Dictionary

- A list of invalid terms
  - stopword list (not complete):

a
about
above
after
again
against
all
am
an
and
any
are
aren't
as
at
be
because
been
before
being
below
between
both
been
before
being
below
between
both
but
by
…

From *http://www.ranks.nl/resources/stopwords.html*

Synonym Dictionaries

- Typically for known synonyms, user-defined synonyms

dislike, *detest*

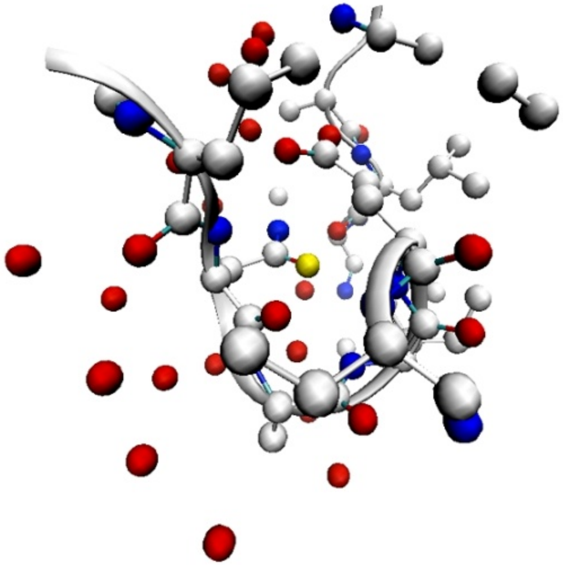
- Can be used as a hard way to deal with inflections if no stemmer is used

like, *likes, liked*

Knowledge Resources

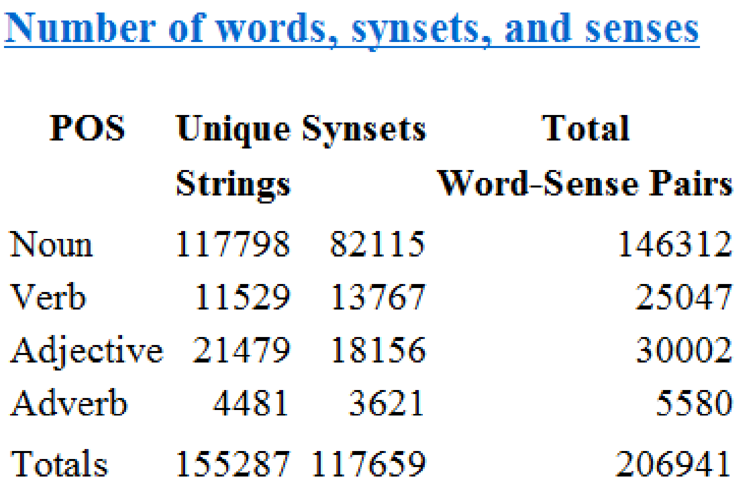
- Taxonomy and ontology – a hierarchical conceptual model to map terms to concepts
- Prerequisite for advance text mining, together with terminology lexicon

An ontology identifies and distinguishes concepts and their relationships; it describes content and relationships.
A taxonomy formalizes the hierarchical relationships among concepts and specifies the term to be used to refer to each; it prescribes structure and terminology



WordNet

- A large lexical database of English
- Created and maintained by the Cognitive Science Laboratory of Princeton University
- *Nouns*, *verbs*, *adjectives* and *adverbs* are grouped into sets of cognitive synonyms (*synsets*), each expressing a distinct concept



Number of words, synsets, and senses

POS	Unique Strings	Synsets	Total Word-Sense Pairs
Noun	117798	82115	146312
Verb	11529	13767	25047
Adjective	21479	18156	30002
Adverb	4481	3621	5580
Totals	155287	117659	206941

Statistics from WordNet website
http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html

WordNet

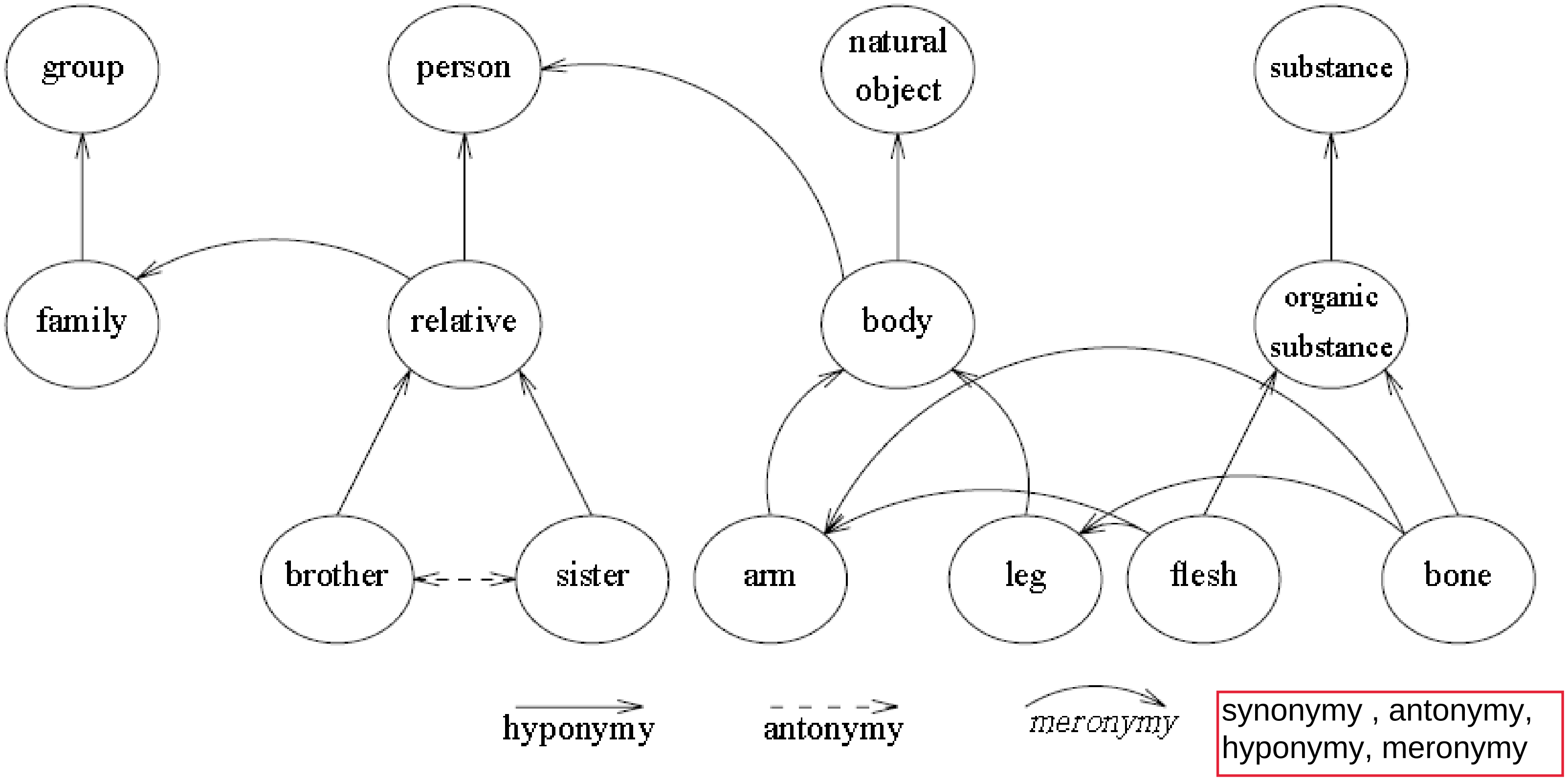
- Synsets are linked by conceptual-semantic and lexical relations
  - Lexical relations
    - Synonymy – e.g. *shut* and *close, happy* and *joyful*
    - Antonymy – e.g. *wet* and *dry*, *young* and *old*, *happy* and *sad*
    - Morphological relations
  - Semantic relations
    - Hyponymy (or ISA relation, super-subordinate relation) – e.g. *apple* and *fruit*, *bed* and *furniture*, *communicate* and *talk* and *whisper*
    - Meronymy (part-whole relation) – e.g. *leg* and *chair*
  - And more…

Semantics is a branch of linguistics that looks at the meanings of words and language, including the symbolic use of language. It also refers to the multiple meanings of words as well.
Examples of Semantics: A toy block could be called a block, a cube, a toy.

WordNet

Lexical representations, or rather more technically, lexical concepts, represent the semantic pole of linguistic units, and are the mentally- instantiated abstractions which language users derive from conceptions and the specific semantic contribution perceived to be associated with particular forms.

Figure 2. Network representation of three semantic relations among an illustrative variety of lexical concepts



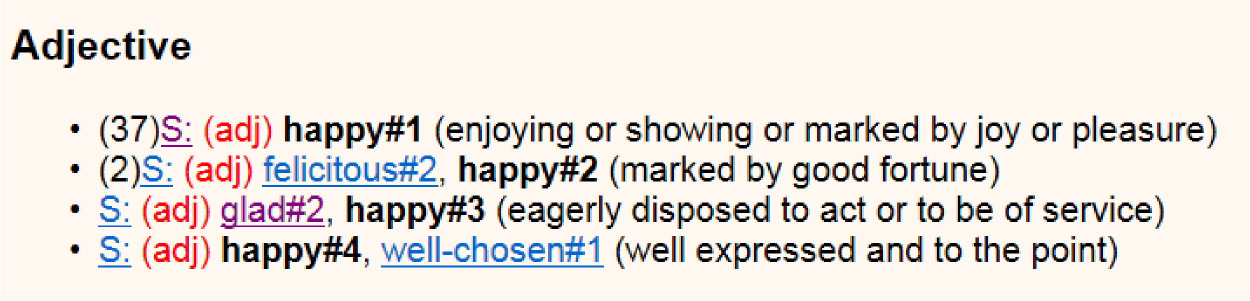
synonymy , antonymy, hyponymy, meronymy

同义词，反义词，下位词，代名词

From *Nouns in WordNet: A Lexical Inheritance System*

WordNet

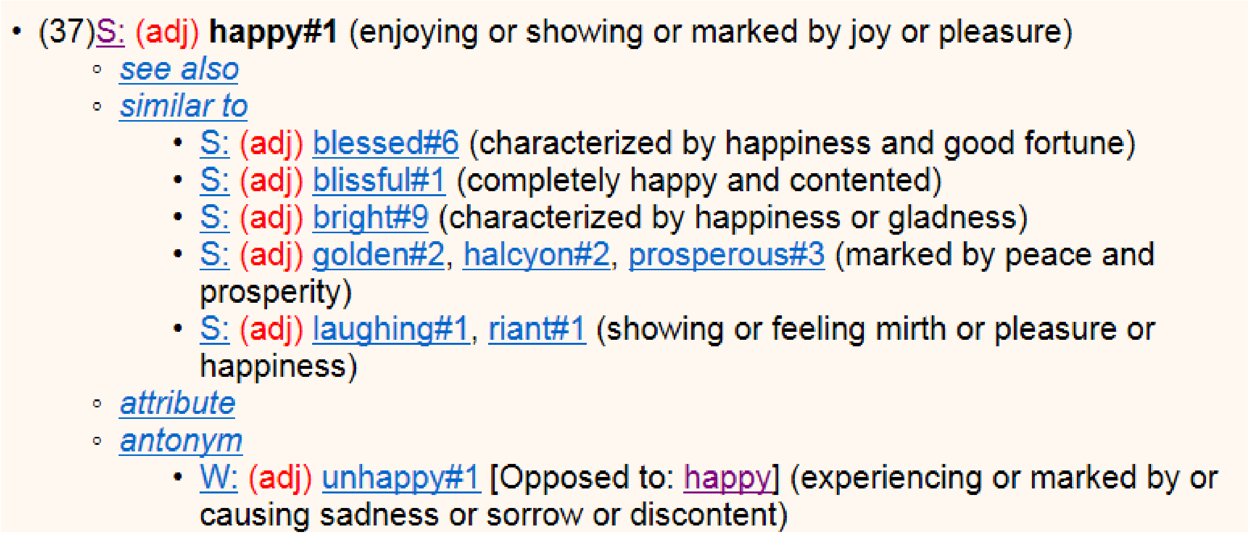
- Example information in Wordnet for “happy”:



Adjective

- (37)S: (adj) **happy#1** (enjoying or showing or marked by joy or pleasure)
- (2)S: (adj) felicitous#2, **happy#2** (marked by good fortune)
- S: (adj) glad#2, **happy#3** (eagerly disposed to act or to be of service)
- S: (adj) **happy#4**, well-chosen#1 (well expressed and to the point)

- Expanded view:



- (37)S: (adj) **happy#1** (enjoying or showing or marked by joy or pleasure)
    - *see also*
    - *similar to*
        - S: (adj) blessed#6 (characterized by happiness and good fortune)
        - S: (adj) blissful#1 (completely happy and contented)
        - S: (adj) bright#9 (characterized by happiness or gladness)
        - S: (adj) golden#2, halcyon#2, prosperous#3 (marked by peace and prosperity)
        - S: (adj) laughing#1, riant#1 (showing or feeling mirth or pleasure or happiness)
    - *attribute*
    - *antonym*
        - W: (adj) unhappy#1 [Opposed to: happy] (experiencing or marked by or causing sadness or sorrow or discontent)

WordNet

- Free and open source
- Proved useful for a wide range of Natural Language Processing applications
  - Word sense disambiguation
  - Word semantic distance measuring
  - Mono- and cross-lingual Information retrieval,
  - Question-answering systems
  - Machine translation
  - Document structuring and categorisation

Sentiment/Opinion Lexicon

- *Sentiment words, opinion words, polar words,* or *opinion-bearing words*.
- Lexicons or dictionaries of words or phrases that convey *positive* or *negative* sentiments, for example:

beautiful, wonderful, amazing…
bad, poor, awful…

Defining Patterns using Regular Expressions

Defining patterns/rules

- With regular expression, we can extract strings containing certain characters, or not containing certain characters, or strings with pre-specified patterns of letters or numbers.
- Such patterns can be defined in a very compact way
  - E.g. regular expression for email addresses

    [A-Z0-9._-]+@([A-Z0-9.-]+\.)+[A-Z]{2,4}
  - Strings matching this expression can then be extracted
    - E.g. *zhenzhen@nus.edu.sg*

Regular expressions are very useful in extracting concepts expressed in a certain way, e.g. *currency*, *dates*, *e-mail addresses, phone numbers*, etc.

Common Operators

- Special characters (operators) are used to define character patterns

Operator	Purpose
. (period)	Match any single character E.g. .in matches both ***Win**dows*, and ***Lin**ux*
^	Match the empty string that occurs at the beginning of a line or string E.g. ^tre will not match *s**tre**tch*
$	Match the empty string that occurs at the end of a line
\d	Match any single digit
\D	Match any single non-digit character
\w	Match any single alphanumeric character

Common Operators

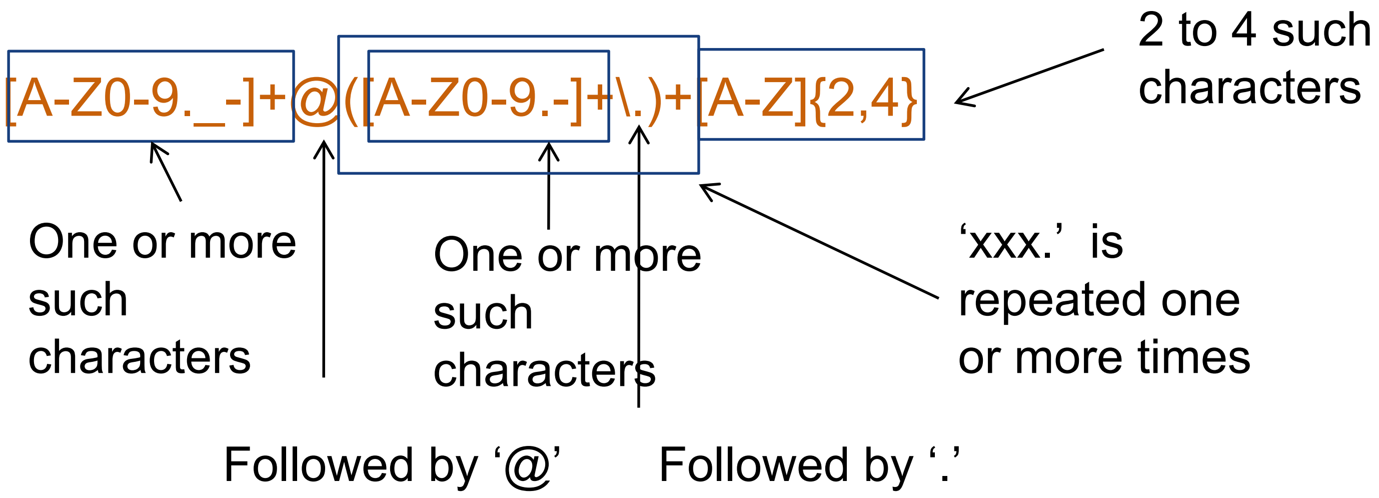
Operator	Purpose
?	Match the preceding character 0 or 1 time E.g. colou?r matches ***color*** *(0)* and ***colour*** *(1)*
*	Zero or more of the preceding character E.g. tre* matches ***tree*** *(2)*, ***tre****ad (1)*, and ***tr****ough (0)*
+	Match the preceding character 1 or more times E.g. tre+ matches ***tree***, and ***tre****ad*
[…]	Match anything inside the square brackets for one character position once E.g. [0-9] matches any character in the range 0-9 [abc] matches ***a***, ***b***, or ***c***
[^…]	Match any character excluding those in the square brackets E.g. [^A-M]in matches ***Win****dows*, but not ***Lin****ux*

Common Operators

Operator	Purpose
{n}	Match the preceding character, or character range, n times E.g. [0-9]{3}-[0-9]{4} matches local phone number like *123-4567*
{n,m}	Match the preceding character at least n times but not more than m times E.g. [A-Z]{2,4} matches *com*, *sg*, but not *abcde*
()	Group parts of search expression together
|	Separate two alternative values E.g. gr(a|e)y matches both *gray* and *grey*
\b	Match empty string, frequently used to indicate a word boundary E.g. \bhis\b matches *his* only, not *this* or *history*

Regular Expression

- Take a look at our email pattern regex again:



[A-Z0-9._-]+@([A-Z0-9.-]+\.)+[A-Z]{2,4}

2 to 4 such characters

One or more such characters

One or more such characters

'xxx.' is repeated one or more times

Followed by '@'

Followed by '.'

Reference and Resources

- GA Miller. WordNet: A Lexical Database for English, *Communications of the ACM*, 1995
- GA Miller. Nouns in WordNet: A Lexical Inheritance System, *International Journal of lexicography*, Oxford University Press, 1990
- Morato, Marzal, Llorens and Moreiro. WordNet Applications, in Proceedings of Global WordNet Conference, pp. 270-278, 2004.
- B. Liu. *Sentiment Analysis and Opinion Mining*, Morgan & Claypool, 2012.
- Regular Expression Tutorial: http://www.zytrax.com/tech/web/regex.htm