

Master of Technology in Knowledge Engineering

Text Mining

Linguistic and Knowledge Resources

Fan Zhenzhen
Institute of Systems Science
National University of Singapore
email: zhenzhen@nus.edu.sg

© 2015 NUS. The contents contained in this document may not be reproduced in any form or by any means, without the written permission of ISS, NUS, other than for the purpose for which it has been supplied.

Agenda

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

Objectives

- To be introduced to different types of resources
- To understand the roles of linguistic and knowledge resources
- To learn how to define such resources

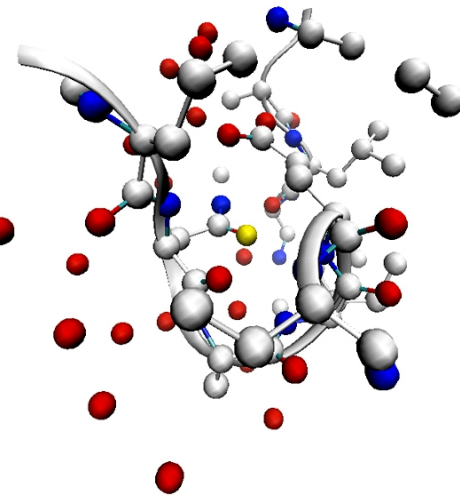
Linguistic Resources

- Linguistic resources: sets of language data and descriptions in machine readable form
- Used for building text mining systems
 - Corpora - to provide examples for statistical methods to work
- Or for improving text mining systems, needed by various processing steps
 - **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

Knowledge Resources

- Taxonomy and ontology – a hierarchical conceptual model to map terms to concepts
- Prerequisite for advance TM, together with terminology lexicon
 - E.g. to derive complex information such as temporal, causal, conditional and other types of semantic relations between biomedical entities instead of simple associations
- More details in module

“Advanced Topics in Text Mining”



Dictionaries

Dictionaries

- Text analytics systems may be equipped with dictionaries in different languages for various purposes.
 - General domain dictionaries for more accurate tokenization, stemming, and POS tagging.
 - 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 in the language in concern
- Or as dictionary for terms to be used in the term vector (e.g. R Text Mining package)
 - 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

- Also known as Stopword List / exclusion dictionary
- To support the stopwords removal step in preprocessing
- A list of very common words
 - usually functional words like *preposition*, *conjunction*, etc.
 - or words that are unimportant for the mining task
- Example stopwords list (not complete):

<i>a</i>	<i>an</i>	<i>because</i>	<i>before</i>
<i>about</i>	<i>and</i>	<i>been</i>	<i>being</i>
<i>above</i>	<i>any</i>	<i>before</i>	<i>below</i>
<i>after</i>	<i>are</i>	<i>being</i>	<i>between</i>
<i>again</i>	<i>aren't</i>	<i>below</i>	<i>both</i>
<i>against</i>	<i>as</i>	<i>between</i>	<i>but</i>
<i>all</i>	<i>at</i>	<i>both</i>	<i>by</i>
<i>am</i>	<i>be</i>	<i>Been</i>	<i>...</i>

From <http://www.ranks.nl/resources/stopwords.html>

Synonym Dictionaries

- Also known as substitution dictionary
- to group similar words under one term
- Typically for known synonyms, user-defined synonyms

dislike, *detest*

- Also a direct way to deal with common misspellings with the correct spelling

dislike, *dilike*

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

like, *likes*, *liked*

Synonym Dictionaries

- Typically synonym words are listed in a file for string match
- Some tools/applications allow some flexibility in stating whether the synonyms should be matched
 - Strictly as it appears in the definition, disallowing inflected forms
 - With any word starting with the term
 - With any word ending with the term



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

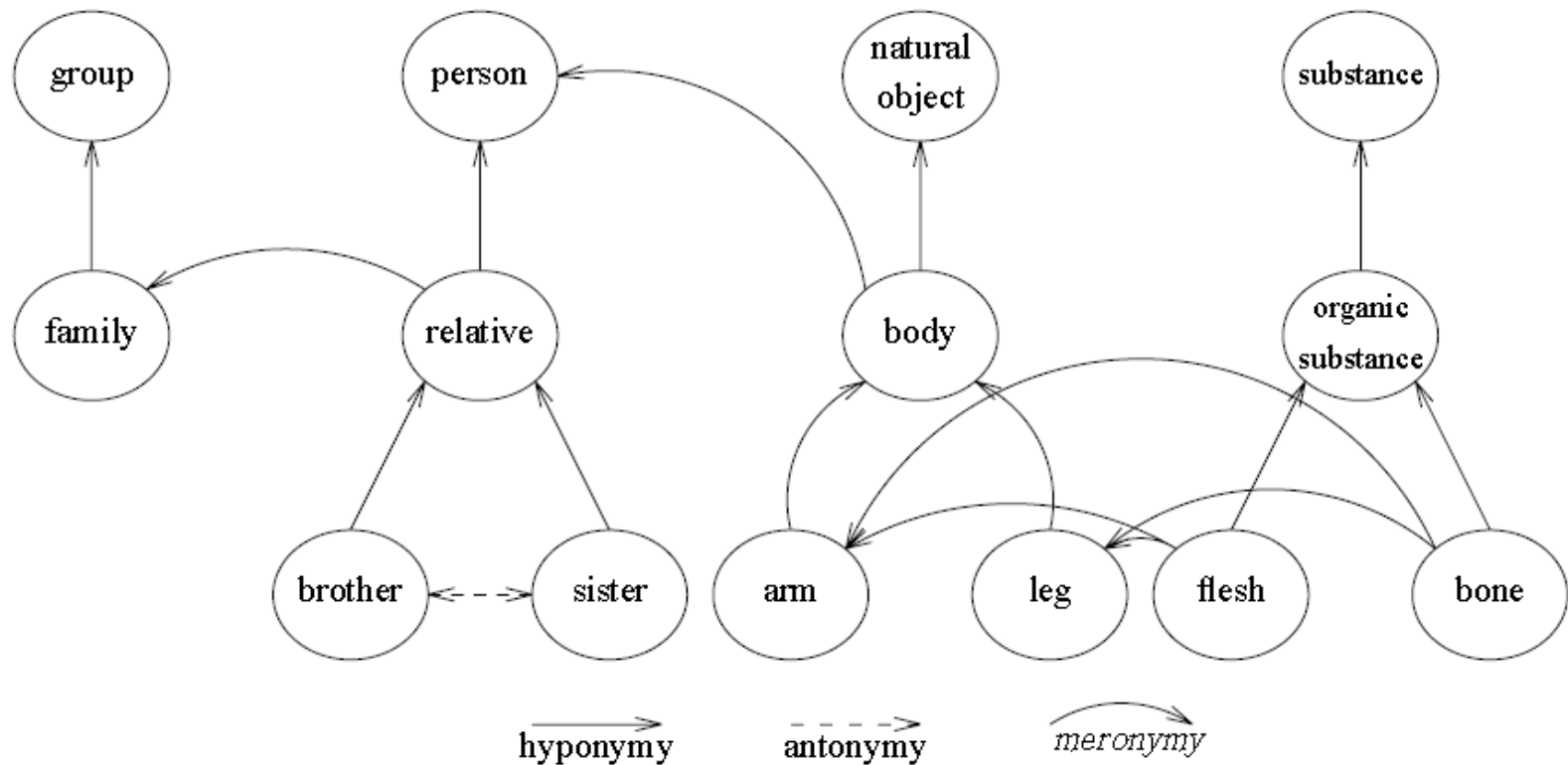
Statistics from WordNet website
<http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html>

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...

WordNet

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



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

- Essential resources required for Opinion Mining to detect sentences containing subjective opinions.
- also known as *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...

- Such sentiment/opinion lexicon can be manually compiled, which can be labor intensive and time consuming.

Sentiment/Opinion Lexicon

- Or they can be generated automatically from dictionaries
 - Start with a small set of manually collected seed sentiment words
 - Search in WordNet or other online dictionaries for their synonyms and antonyms
 - Add the newly found words to the seed list
 - Begin the next iteration until no more new words can be found
- Another approach is to generate from corpus
 - Start with some seed words and identify more sentiment words and their orientation using linguistic rule or conventions on connectives (*and, or*)
 - More complex

Challenges in Using Opinion Lexicon

- An opinion word's opinion orientation can be sensitive to its context.
 - E.g. *long* – **positive** or **negative**?
 - “The battery life is very *long*” 
 - “The queue at the counter is very *long*” 
- Sarcasm, in which the speakers say the opposite of what they mean
 - E.g. “What a *great* phone! It stopped working in two days.”



Sentiment/Opinion Lexicon

- Some Sentiment Lexicons are publically available
 - General Inquirer lexicon:
http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm
 - Sentiment lexicon (Liu Bing):
<http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon>
 - OpinionFinder subjectivity lexicon:
<http://www.cs.pitt.edu/mpqa/lexicons.html>
 - SentiWordNet – assign each synset of WordNet sentiment scores
<http://sentiwordnet.isti.cnr.it/>

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 Windows , and Linux
^	Match the empty string that occurs at the beginning of a line or string E.g. ^tre will not match stretch
\$	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

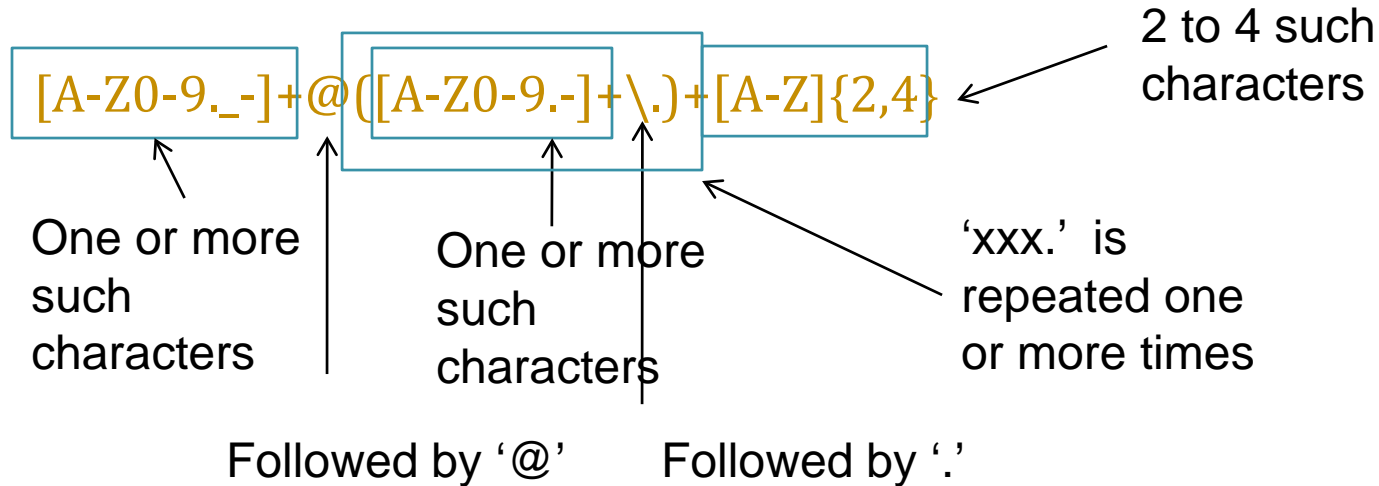
Operator	Purpose
?	Match the preceding character 0 or 1 time E.g. colou?r matches color (0) and colour (1)
*	Zero or more of the preceding character E.g. tre* matches tree (2), tread (1), and trough (0)
+	Match the preceding character 1 or more times E.g. tre+ matches tree , and tread
[...]	Match anything 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 Windows , but not Linux

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 <i>com</i> , <i>sg</i> , but not <i>abcde</i>
()	Group parts of search expression together
	Separate two alternative values E.g. gr(a e)y matches both <i>gray</i> and <i>grey</i>
\b	Match empty string, frequently used to indicate a word boundary E.g. \bhis\b matches <i>his</i> only, not <i>this</i> or <i>history</i>

Regular Expression

- Take a look at our email pattern regex again:



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>