CLPL4a: Semantics and Lexicon

Tafseer Ahmed

Semantic Relations (between words)

Synonyms: e.g. big/large

Antonyms: e.g. hot/cold

Hyponyms: e.g. chair/furniture

Meronyms: e.g. wheel/car

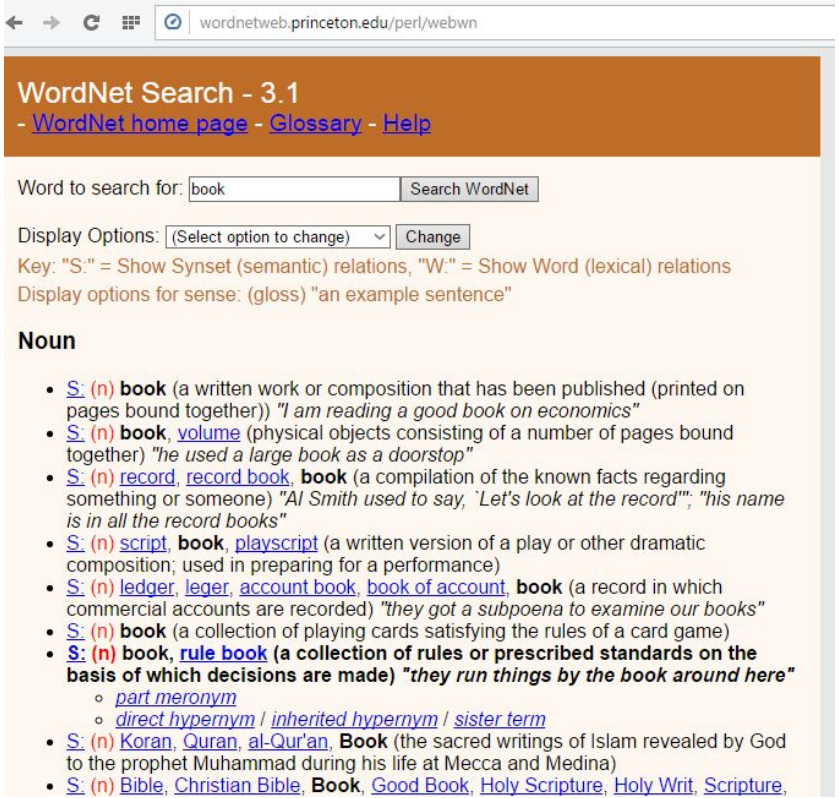
Semantic Fields: e.g. restaurants (waiter, menu, plate, food, chef, ...)

How to get Semantic Relations

- Manual Knowledge Engineering
  - WordNet
  - SentiLex (Sentiment Lexicon)
  - UCREL Semantic Lexicon
  - .....
- Machine Learning using Corpus

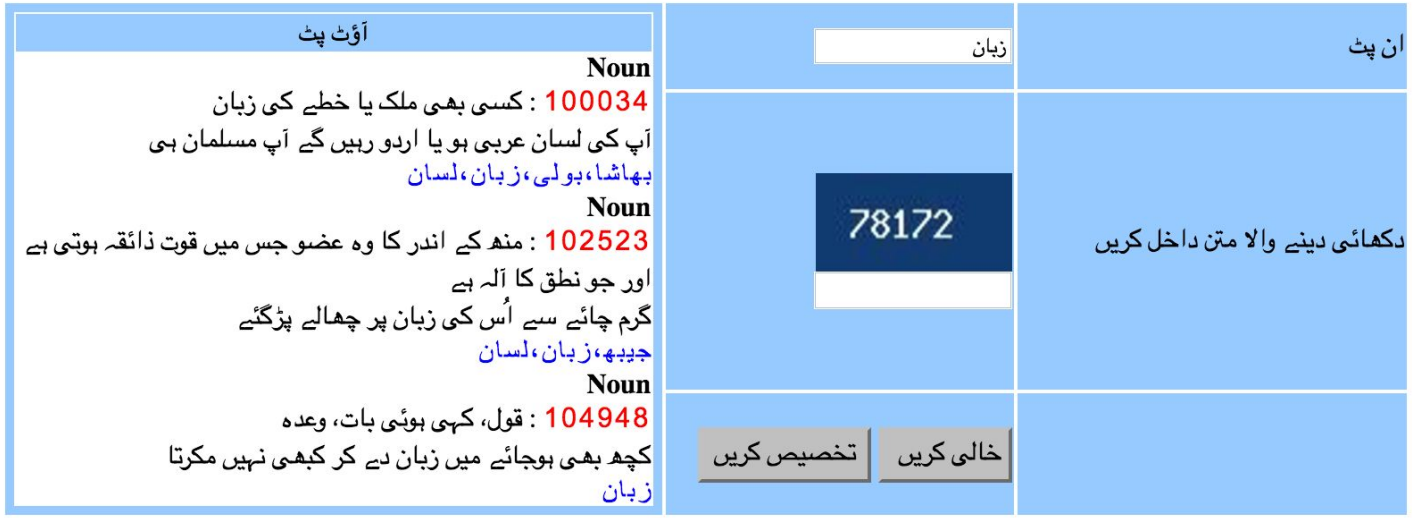
Manual Knowledge Engineering

Wordnet


wordnetweb.princeton.edu/perl/webwn
WordNet Search - 3.1
- WordNet home page - Glossary - Help
Word to search for: book
Search WordNet
Display Options: (Select option to change)
Change
Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"
Noun
S: (n) book (a written work or composition that has been published (printed on pages bound together)) "I am reading a good book on economics"
S: (n) book, volume (physical objects consisting of a number of pages bound together) "he used a large book as a doorstop"
S: (n) record, record book, book (a compilation of the known facts regarding something or someone) "Al Smith used to say, `Let's look at the record'"; "his name is in all the record books"
S: (n) script, book, playscript (a written version of a play or other dramatic composition; used in preparing for a performance)
S: (n) ledger, leger, account book, book of account, book (a record in which commercial accounts are recorded) "they got a subpoena to examine our books"
S: (n) book (a collection of playing cards satisfying the rules of a card game)
S: (n) book, rule book (a collection of rules or prescribed standards on the basis of which decisions are made) "they run things by the book around here"
part meronym
direct hypernym / inherited hypernym / sister term
S: (n) Koran, Quran, al-Qur'an, Book (the sacred writings of Islam revealed by God to the prophet Muhammad during his life at Mecca and Medina)
S: (n) Bible, Christian Bible, Book, Good Book, Holy Scripture, Holy Writ, Scripture,

Urdu Wordnet

- http://www.cle.org.pk/clestore/urduwordnet.htm



Sentiment Based Lexicon

Bing Liu Opinio Lexicon (for Customer Reviews)

http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar

6786 words (2006 positive, 4783 negative)

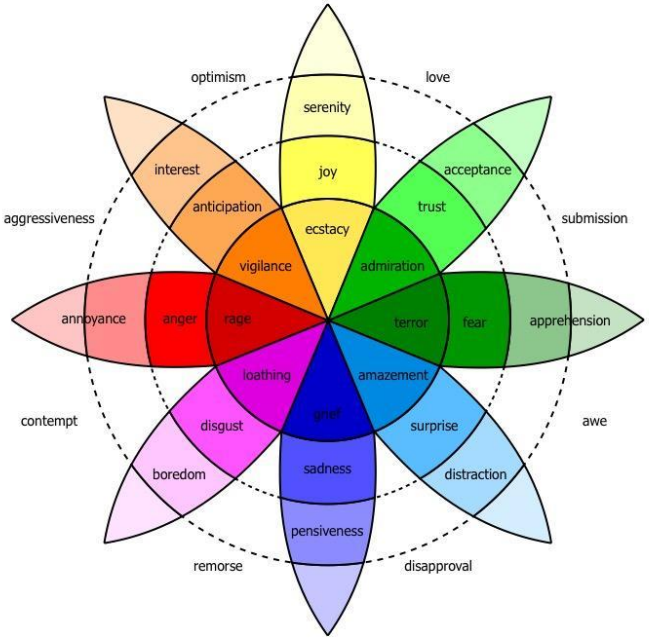
SentiWordNet

- http://sentiwordnet.isti.cnr.it/
- All WordNet synsets automatically annotated for degrees of positivity, negativity, and neutrality/objectiveness

Emotion based Lexicon

Plutchick's wheel of emotion

- 8 basic emotions
- in four opposing pairs:
  - joy–sadness
  - anger–fear
  - trust–disgust
  - anticipation–surprise



- **Ratings for 14,000 words for emotional dimensions:**
  - **valence** (the pleasantness of the stimulus)
  - **arousal** (the intensity of emotion provoked by the stimulus)
  - **dominance** (the degree of control exerted by the stimulus)
  -

Valence		Arousal		Dominance	
vacation	*8.53*	*rampage*	*7.56*	*self*	*7.74*
happy	*8.47*	*tornado*	*7.45*	*incredible*	*7.74*
whistle	*5.7*	*zucchini*	*4.18*	*skillet*	*5.33*
consciou	*5.53*	*dressy*	*4.15*	*concur*	*5.29*
torture	*1.4*	*dull*	*1.67*	*earthquak*	*2.14*

Emotion Based Lexicon

NRC Lexicon

- translated into over 100 languages
- sentiments: negative, positive
- emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, trust

https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

To discuss in next session

- Semantic Roles
  - PropBank
- FrameNet
- Verb Net

Machine Learning

Discovering Semantics using Corpus

“You shall know a word by the company it keeps.”

(J. R. Firth)

Guessing Game 1: What can be say about Ong-choi?

- Ong-choi is delicious sautéed with garlic.
- Ong-choi is superb over rice
- Ong-choi leaves with salty sauces

What can be say about Ong-choi?

- Ong-choi is delicious sautéed with garlic.
- Ong-choi is superb over rice
- Ong-choi leaves with salty sauces

Ong-choi is a leafy green like spinach.

Guessing Game 2: Which words behaves similar?

Term-Document Matrix

	Novel 1	Novel 2	Novel 3	Novel 4
Word 1	1	0	7	13
Word 2	14	80	62	89
Word 3	36	58	1	4
Word 4	20	15	2	3

Guessing Game 2: Which words behaves similar?

Term-Document Matrix

	Novel 1	Novel 2	Novel 3	Novel 4
Word 1	1	0	7	13
Word 2	14	80	62	89
Word 3	36	58	1	4
Word 4	20	15	2	3

Guessing Game 2: Which words behaves similar?

Term-Document Matrix

	Novel 1	Novel 2	Novel 3	Novel 4
Word 1	1	0	7	13
Word 2	14	80	62	89
Word 3	36	58	1	4
Word 4	20	15	2	3

Guessing Game 2: Which words behaves similar?

	Novel 1	Novel 2	Novel 3	Novel 4
Word 1	1	0	7	13
Word 2	14	80	62	89
Word 3	36	58	1	4
Word 4	20	15	2	3

Conclusions drawn by Algorithm/Computer:

- Word 3 & Word 4 behaves similarly.
- Word 1 behaves differently than Word 3 & Word 4.
- Novel 1 and 2 are similar and Novel 3 & 4 are similar.

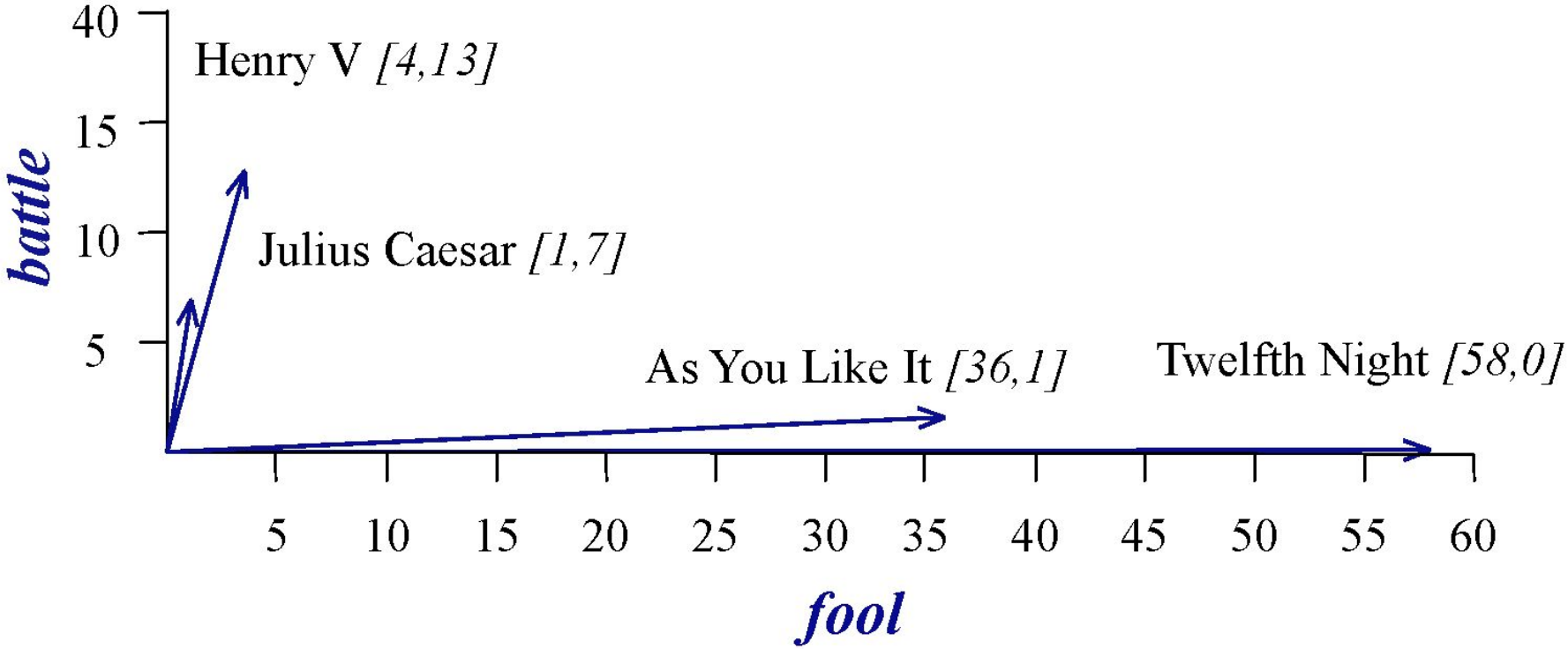
Guessing Game 2: Which words behaves similar?

	As You Like It	Twelfth Night	Julius Ceaser	Henry V
Battle	1	0	7	13
Good	14	80	62	89
Fool	36	58	1	4
Wit	20	15	2	3

Conclusions:

- The words “Fool” & “Wit” behaves similarly.
- The word “Battle” behaves differently than “Fool” & “Wit”.
- “As You Like It” and “Twelfth Night” are similar
  and “Julius Ceaser” & “Henry V” are similar.

Vector Semantics



Term-Term Matrix

sugar, a sliced lemon, a tablespoonful of **apricot** preserve or jam, a pinch each of,
their enjoyment. Cautiously she sampled her first **pineapple** and another fruit whose taste she likened
well suited to programming on the digital **computer**. In finding the optimal R-stage policy from
for the purpose of gathering data and **information** necessary for the study authorized in the

	aardvark	computer	data	pinch	result	sugar	...
apricot	0	0	0	1	0	1	
pineapple	0	0	0	1	0	1	
digital	0	2	1	0	1	0	
information	0	1	6	0	4	0	
...	...						

Source: SLPW

What we learnt?

- On the basis of context, Words can be represented as vectors.
- Word Similarity or relatedness can be calculated by means of these vectors.

- *Now, we will not discuss more about how to create word vectors or how to calculate word similarity/relatedness.*
- *We will only list the names of tools and some of their interesting outputs.*

(Major) Tools for Semantic Relations

Using Neural Networks

- Word2Vec (by Google)
- FastText (by Facebook)
- Elmo

Using Probabilities

- Topic Model (LDA)
- Guided LDA

Word2Vec Output (for Urdu Corpus)

```
In [22]: model.most_similar( 'سورج', topn = 15)

Out[22]: [('0.8165232539176941 ,'چاند),
          ('0.8143321275711106 ,'سُورج),
          ('0.7290103435516357 ,'آفتاب),
          ('0.7133749723434448 ,'ماہتاب),
          ('0.7077307701110084 ,'تارے),
          ('0.7029957771301270 ,'ستاروں),
          ('0.6967627406120300 ,'ستارے),
          ('0.6915234327316284 ,'کرنوں),
          ('0.6877450942993164 ,'مریخ),
          ('0.6747698783874512 ,'غروب),
          ('0.6725152730941772 ,'بادلوں),
          ('0.6688274741172790 ,'تاروں),
          ('0.6525486707687378 ,'گہن),
          ('0.6493672132492065 ,'شفق),
          ('0.6463797092437744 ,'شعاعوں)]
```

Word2Vec Output (for Urdu Corpus)

Word2Vec Output (for Urdu Corpus)

- What would be the answer of:

King – Man + Woman

Word2Vec Output (for Urdu Corpus)

- What would be the answer of:

King – Man + Woman

```
In [24]: word_add = ['بادشاہ' ,'عورت']
         word_sub = ['مرد']

         model.most_similar(positive=word_add, negative=word_sub)

Out[24]: [('0.6132516264915466 ,'ملکہ),
```

Topic Model

- Probability based
- Latent Dirichlet Allocation (LDA)
- Discovers sets of related words called topics

Topic Model for Urdu newspapers

روپے, ٹیکس, پاکستان, ایف	سیاسی, پارٹی, عوام, انتخابات	کام, فلم, بات, کہا
پاکستان,بھارت,بھارتی,کشمیر	کراچی, پولیس, سندھ, افراد	حکومت, شریف, وزیر, نواز
ملک, دنیا, ترقی, نظام	جنرل, مشرف, صدر, پرویز	پاکستان, دہشت, طالبان, حکومت

Joint Work with King Abdul Aziz University, Jeddah and NED UET, Karachi

Guided LDA

Seed words are provided along with the Corpus

- https://github.com/vi3k6i5/GuidedLDA

Guided LDA

- Seeds for Topic 0:
  game, team, win, player, season, second, victory
- Seeds for Topic 1:
  percent, company, market, price, sell, business, stock, share
- Seeds for Topic 2:
  music, write, art, book, world, film
- Seeds for Topic 3: .....

Example from https://github.com/vi3k6i5/GuidedLDA

Guided LDA

- Discovered Topic 0:

  game, play, team, win, season, player, second, point, start, victory
- Discovered Topic 1:

  company, percent, market, price, business, sell, executive, pay, plan, sale
- Discovered Topic 2:

  play, life, man, music, place, write, turn, woman, old, book
- Discovered Topic 3: ....

Applications in Corpus Linguistics

- **Finding Semantic Fields**
  - **Topic Models**
  - **Clustering Word2Vec vectors**
- **Extending Word Lists**
  - **Finding similar words by word2vec**
  - **Guided LDA**

Limitations

- The methods find related words
- The methods does not (directly) distinguish different relations i.e. synonyms, antonyms, hyponyms etc.
- Hence, post-editing of word lists by human expert is required.
- The methods should be used for shortlisting (not finalizing) the word lists.

Rule Based Acquisition

Early intuition from **Hearst (1992)**

- "Agar is a substance prepared from a mixture of **red algae, such as Gelidium,** for laboratory or industrial use"

- What does *Gelidium* mean?
- How do you know?`

Rule Based Acquisition - hyponyms

Hearst pattern	Example occurrences
X and other Y	...temples, treasuries, and other important civic buildings.
X or other Y	Bruises, wounds, broken bones or other injuries...
Y such as X	The bow lute, such as the Bambara ndang...
Such Y as X	...such authors as Herrick, Goldsmith, and Shakespeare.
Y including X	...common-law countries, including Canada and England...
Y , especially X	European countries, especially France, England, and Spain...