Taal- en spraaktechnologie

Sophia Katrenko

Utrecht University, the Netherlands

June 1, 2012

Outline

1. Lexical acquisition: resources
2. Distributional similarity
3. WordNet similarity

Focus

This part of the course focuses on

- meaning representation
- lexical semantics
- distributional similarity
- intro to machine learning
- word sense disambiguation
- information extraction

Today

- Chapter 19 (Lexical semantics)
- Chapter 20 (Computational lexical semantics: from section 6)
- Have a look at Homework 2

Lexical acquisition

Thematic roles (1)

Examples

Pat opened the door.

$\exists e, x, y \; Opening(e) \wedge Opener(e, Pat) \wedge OpenedThing(e, y) \wedge Door(y)$

I broke the window.

$\exists e, x, y$
$Breaking(e) \wedge Breaker(e, Speaker) \wedge BrokenThing(e, y) \wedge Window(y)$

Breaker and *Opener* are **deep roles** and subjects are **agents**.

Thematic roles (2)

More thematic roles:

Role	Example
AGENT	**I** broke the window.
EXPERIENCER	**John** has a headache.
FORCE	**The wind** blows leaves.
THEME	I broke **the window**.
RESULT	We made **a table**.
CONTENT	He asked " **You wrote this poem yourself?**".
INSTRUMENT	A dentist uses many **tools.**
BENEFICIARY	We wrote this poem for **Andrew.**
SOURCE	I came from **Amsterdam**.
GOAL	I went to **Utrecht**.

Thematic roles (3)

Why thematic roles?

- to generalize over predicate arguments
- can be useful for applications, such as machine translation

Examples

$John_{AGENT}$ broke the $window_{THEME}$.

$John_{AGENT}$ broke the $window_{THEME}$ with a $rock_{INSTRUMENT}$.

The $rock_{INSTRUMENT}$ broke the $window_{THEME}$.

The $window_{THEME}$ broke.

Thematic roles (4)

Thematic grid (θ-grid, case frame)

The set of thematic role arguments taken by a verb.

Thematic grid: example

AGENT: Subject, *THEME*: Object

AGENT:Subject, *THEME*: Object, *INSTRUMENT* : $\text{PP}_{\textit{with}}$

INSTRUMENT:Subject, *THEME*: Object

THEME:Subject

Thematic roles (5)

- It is difficult to fix the inventory for thematic roles (e.g., there are *intermediary* instruments that can appear as subjects and *enabling* instruments that can't).
- An alternative to thematic roles: *generalized semantic roles* defined by a set of heuristic features.
- Some models define semantic roles specifically for a verb in question.

PropBank (1)

PropBank - sentences annotated with semantic roles:

- Semantic roles are defined with respect to a particular verb sense.
- Roles are given numbers as in $Arg0$ (often Proto-Agent), $Arg1$ (often Proto-Patient).
- Some models define semantic roles specifically for a verb in question.

PropBank (2)

[From Palmer et al.]

Frameset **kick.01** "drive or impel with the foot"
  Arg0: Kicker
  Arg1: Thing kicked
  Arg2: Instrument (defaults to foot)
Ex1: [$_{\text{ArgM-DIS}}$ But] [$_{\text{Arg0}}$ two big New York banks$_i$] seem [$_{\text{Arg0}}$ *trace*$_i$] to have *kicked* [$_{\text{Arg1}}$ those chances] [$_{\text{ArgM-DIR}}$ away], [$_{\text{ArgM-TMP}}$ for the moment], [$_{\text{Arg2}}$ with the embarrassing failure of Citicorp and Chase Manhattan Corp. to deliver $7.2 billion in bank financing for a leveraged buy-out of United Airlines parent UAL Corp]. (wsj_1619)
Ex2: [$_{\text{Arg0}}$ John$_i$] tried [$_{\text{Arg0}}$ *trace*$_i$] to *kick* [$_{\text{Arg1}}$ the football], but Mary pulled it away at the last moment.

FrameNet (1)

FrameNet (Baker et al.) - sentences annotated with semantic roles:

- Focusing on corpus evidence for semantic and syntactic generalizations.
- Valences of words are represented, semantic roles are specific to frames.
- Types of roles: core roles (e.g., Item or Attribute) and non-core roles (Duration, Speed).
- Several domains covered (e.g., healthcare, time, communication, etc.).
- Different from dictionaries because it presents multiple annotated examples of each sense of a word (i.e. each lexical unit). The set of examples (approximately 20 per LU) illustrates all of the combinatorial possibilities of the lexical unit.

FrameNet (2)

More on FrameNet:
https://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf

... [$_{\text{Cook}}$ the boys] ... GRILL [$_{\text{Food}}$ their catches] [$_{\text{Heating\_instrument}}$ on an open fire].

[$_{\text{Avenger}}$ I] 'll GET EVEN [$_{\text{Offender}}$ with you] [$_{\text{Injury}}$ for this]!

Current trends

- Research on bilingual FrameNets (e.g., English-Chinese, *Bengfeng and Fung*, 2004), also for applications, e.g. machine translation (*Boas*, 2011).
- Mapping across different resources on semantic roles, e.g. between PropBank and VerbNet, *Loper et al.*, 2007).
- Numerous challenges on labeling semantic roles automatically, in different flavours, e.g. spatial role labeling this year: `http://www.cs.york.ac.uk/semeval-2012/task3/`.

Similarity and Relatedness Measures

Words

Mark Twain's Speeches (1910)

An average English word is four letters and a half. By hard, honest labor I've dug all the large words out of my vocabulary and shaved it down till the average is three and a half... I never write "metropolis" for seven cents, because I can get the same money for "city". I never write "policeman", because I can get the same price for "cop"... I never write "valetudinarian" at all, for not even hunger and wretchedness can humble me to the point where I will do a word like that for seven cents; I wouldn't do it for fifteen.

Distributional hypothesis

Distributional similarity (Firth, 1957; Harris, 1968)

"You shall know a word by the company it keeps"

(words found in the similar contexts tend to be semantically similar).

Mohammed and Hirst, 2005

Distributionally similar words tend to be semantically similar, where two words w_1 and w_2 are said to be distributionally similar if they have many common co-occurring words and these co-occurring words are ech related to w_1 and w_2 by the same syntactic relation.

Motivation

Semantic similarity is useful for various applications:

- **information retrieval**, **question answering**: to retrieve documents whose words have similar meanings to the query words.
- **natural language generation**, **machine translation**: to know whether two words are similar to know if we can substitute one for the other in particular contexts.
- **language modeling**: can be used to cluster words for class-based models.

Similarity measures

Similarity between two lexical items can be measured in many ways, e.g.

- using distributional information (corpora counts)
- using WordNet structure

Questions

Several questions to be addressed when measuring distributional similarity:

1. How the co-occurrence terms are defined (e.g., on the level of a sentence, an *n*-gram, using dependency triples from syntactic analysis)?
2. How the terms are weighted (what is the value of features: binary, frequency, mutual information)?
3. What vector distance metric to use.

Representation

Example 1 from JM book:

	arts	boil	data	function	large	sugar	summarized	water
apricot	0	1	0	0	1	1	0	1
pineapple	0	1	0	0	1	1	0	1
digital	0	0	1	1	1	0	1	0
information	0	0	1	1	1	0	1	0

Figure 19.9 Co-occurrence vectors for four words, computed from the Brown corpus, showing only 8 of the (binary) dimensions (hand-picked for pedagogical purposes to show discrimination). Note that *large* occurs in all the contexts and *arts* occurs in none; a real vector would be extremely sparse.

Representation

Example 2 from JM book:

	subj-of, absorb	subj-of, adapt	subj-of, behave	...	pobj-of, inside	pobj-of, into	...	nmod-of, abnormality	nmod-of, anemia	nmod-of, architecture	...	obj-of, attack	obj-of, call	obj-of, come from	obj-of, decorate	...	nmod, bacteria	nmod, body	nmod, bone marrow
cell	1	1	1		16	30		3	8	1		6	11	3	2		3	2	2

Figure 19.10 Co-occurrence vector for the word *cell*, from Lin (1998a), showing grammatical function (dependency) features. Values for each attribute are frequency counts from a 64-million word corpus, parsed by an early version of MINIPAR.

Association measures (1)

Let w be a target word, f be each element of its co- occurrence vector that consists of a relation r and a related word w'; $f = (r, w')$. Then, the maximum likelihood estimate (MLE) is as follows:

$$P(f|w) = \frac{count(f,w)}{count(w)} \tag{1}$$

and

$$P(f,w) = \frac{count(f,w)}{\sum_{w'} count(f,w')} \tag{2}$$

Association measures (2)

Association measures based on

- probability itself:

$$assocprob(w, f) = P(f|w) \tag{3}$$

- pointwise mutual information

$$assoc_{PMI}(w, f) = log_2 \frac{P(w, f)}{P(w)P(f)} \tag{4}$$

Similarity measures

A note on measure vs. metric

A metric on a set X is a function d, such that $d : X \times X \rightarrow \mathrm{R}$ and which has the following properties:

- $d(x, y) \geq 0$
- $d(x, y) = 0$ iff $x = y$
- $d(x, y) = d(y, x)$
- $d(x, z) \leq d(x, y) + d(y, z)$

Similarity measures

For two binary vectors **w** and **v**, the most common measures are as follows:

measure	**definition**
matching coefficient	$\|X \cap Y\|$
Dice coefficient	$\frac{2\|X \cap Y\|}{\|X\|+\|Y\|}$
Jaccard coefficient	$\frac{\|X \cap Y\|}{\|X \cup Y\|}$
Overlap coefficient	$\frac{\|X \cap Y\|}{min(\|X\|,\|Y\|)}$
cosine	$\frac{\|X \cap Y\|}{\sqrt{\|X\| \times \|Y\|}}$

Similarity measures

If we move to frequency counts:

word	*context*$_1$	*context*$_2$	...	*context*$_n$
w	w_1	w_2	...	w_n
v	v_1	v_2	...	v_n

$$d_{Dice} = \frac{2|X \cap Y|}{|X| + |Y|} \tag{5}$$

$$d_{Dice} = \frac{2\sum_{i=1}^{n} min(w_i, v_i)}{\sum_{i=1}^{n} w_i + \sum_{i=1}^{n} v_i} \tag{6}$$

Similarity measures

If we move to frequency counts:

word	$context_1$	$context_2$	...	$context_n$
w	w_1	w_2	...	w_n
v	v_1	v_2	...	v_n

Jaccard coefficient

$$d_{Jaccard} = \frac{|X \cap Y|}{|X \cup Y|} \tag{7}$$

$$d_{Jaccard} = \frac{\sum_{i=1}^{n} min(w_i, v_i)}{\sum_{i=1}^{n} max(w_i, v_i)} \tag{8}$$

Similarity measures

If we move to frequency counts:

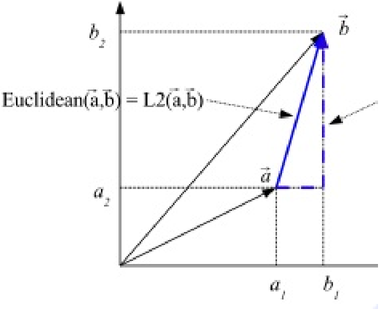
word	$context_1$	$context_2$	...	$context_n$
w	w_1	w_2	...	w_n
v	v_1	v_2	...	v_n

$$d_{Manhattan} = \sum_{i=1}^{n} |w_i - v_i| \tag{9}$$

$$d_{Euclidean} = \sqrt{\sum_{i=1}^{n} (w_i - v_i)^2} \tag{10}$$

Representation

Euclidean and Manhattan measures from JM book:



Similarity measures

If we move to frequency counts:

word	$context_1$	$context_2$	$\ldots$	$context_n$
w	w_1	w_2	$\ldots$	w_n
v	v_1	v_2	$\ldots$	v_n

$$d_{cosine} = \frac{\sum_{i=1}^{n} w_i v_i}{\sqrt{\sum_{i=1}^{n} w_i^2}\sqrt{\sum_{i=1}^{n} v_i^2}} \tag{11}$$

WordNet-based measures

How to use WordNet to measure relatedness/similarity? The following notions are used:

- Path between two synsets c_1 and c_2, $pathlen(c1, c2)$ (the number of edges in the shortest path in the thesaurus graph between the sense nodes c_1 and c_2)
- The lowest common subsumer $lcs(c_1, c_2)$ (the lowest node in the hierarchy that subsumes (is a hypernym of) both c_1 and c_2)

WordNet-based measures

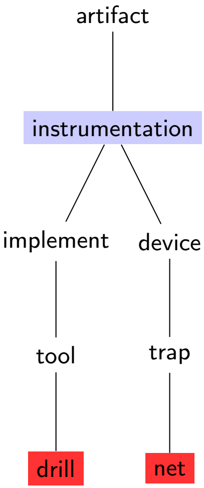


Figure: Part of the WordNet hierarchy

WordNet-based measures

The following notions are used:

- The probability that a randomly selected word in a corpus is an instance of concept c, $P(c)$ (Resnik, 1995)

$$P(c) = \frac{\sum_{w \in words(c)} count(w)}{N} \quad (12)$$

$words(c)$ = the set of words subsumed by concept c,

N = the total number of words in the corpus that are also present in the thesaurus.

- Information content

$$IC(c) = -\log P(c) \quad (13)$$

WordNet-based measures

Definitions

- Leacock and Chodorow, 1998 (lch)

$$sim_{path}(c_1, c_2) = -\log pathlen(c_1, c_2) \tag{14}$$

- Resnik measure (Resnik, 1995) (res)

$$sim_{resnik}(c_1, c_2) = -\log P(lcs(c_1, c_2)) \tag{15}$$

WordNet-based measures

Definitions

- Wu and Palmer, 1998 (wup)

$$sim_{wup}(c_1, c_2) = \frac{2 * dep(lcs(c_1, c_2))}{len(c_1, lcs(c_1, c_2)) + len(c_2, lcs(c_1, c_2)) + 2 * dep(lcs(c_1, c_2))}$$

WordNet-based measures

Lin (1998) has compared two object A and B given their

- commonality: the more information A and B have in common, the more similar they are $(IC(common(A, B)))$.
- difference: the more differences between the information in A and B, the less similar they are $(IC(description(A, B)) - IC(common(A, B)))$.

$$sim_{Lin}(A, B) = \frac{\log P(common(A, B))}{\log P(description(A, B))} \tag{16}$$

WordNet-based measures

How to apply it to WordNet?

$$sim_{Lin}(c_1, c_2) = \frac{2 \log P(lcs(c_1, c_2))}{\log P(c_1) + \log P(c_2)} \tag{17}$$

Jiang-Conrath distance (Jiang and Conrath, 1997)

$$dist_{JC}(c_1, c_2) = 2 \log P(lcs(c_1, c_2)) - (\log P(c_1) + \log P(c_2)) \tag{18}$$

Measures

So, what measure is the best?

- there is no best measure apriori (similarly as there is no machine learning method that *always* performs the best - so-called No-free lunch theorem).
- different applications may require different measures to be used.

Measures

So, what measure is the best?

- there is no best measure apriori (similarly as there is no machine learning method that *always* performs the best - so-called No-free lunch theorem).
- different applications may require different measures to be used.

Measures

So, what measure is the best?

- there is no best measure apriori (similarly as there is no machine learning method that *always* performs the best - so-called No-free lunch theorem).
- different applications may require different measures to be used.

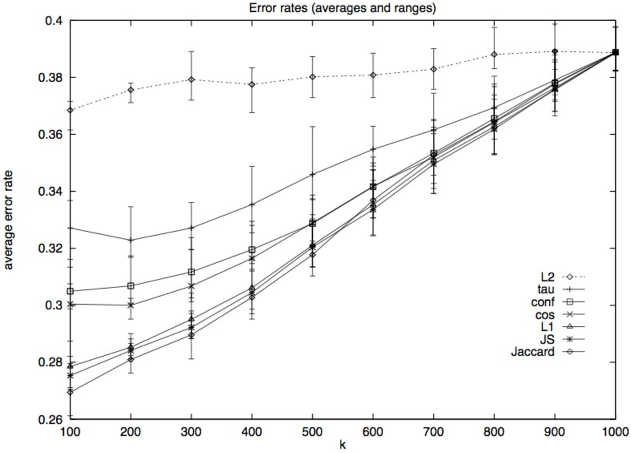
Measures

L. Lee. *Measures of Distributional Similarity*. In Proceedings of the 37th ACL, 1999.
`http://acl.ldc.upenn.edu/P/P99/P99-1004.pdf`

- Data: verb-object co-occurrence pairs in the 1988 Associated Press newswire (1000 most frequent nouns).
- various distributional measures (cosine, Euclidean, others).
- Goal: improving probability estimation for unseen co-occurrences: "replaced each noun- verb pair (n, v_1) with a noun-verb-verb triple (n, v_1, v_2) such that $P(v_2) \approx P(v_1)$. The task for the language model under evaluation was to reconstruct which of (n, v_1) and (n, v_2) was the original cooccurrence."

Measures

L. Lee. *Measures of Distributional Similarity*. In Proceedings of the 37th ACL, 1999.



WordNet measures (1)

S. Katrenko et al.. *Using Local Alignments for Relation Recognition*. In JAIR, 2010.
http://www.aaai.org/Papers/JAIR/Vol38/JAIR-3801.pdf

- **Data**: Annotated relation instances in text (for 7 relation types, e.g. part-whole as in *There are many* **trees** *in this* **forest**).
- **Method**: Using alignment of syntactic structures while elements of these structure that correspond to words are aligned using either distributional or WordNet similarity.
- **Goal**: Predict if a certain relation takes place (binary predictions per relation type).

So is there any difference in performance based on the WordNet measure being used?

Relatedness measure	Accuracy	Precision	Recall	F-score
wup	72.91	71.20	72.56	71.62
lch	72.96	72.31	70.93	71.02
lin	65.27	62.01	67.07	63.65
res	62.94	62.51	59.66	60.46
jcn	55.55	52.25	69.28	57.07
random	56.57	53.10	52.94	52.83

WordNet measures (3)

So is there any difference in performance based on the WordNet measure being used?

Relation type	Ranking
Cause - Effect	$wup \sim lch > lin > res \sim random > jcn$
Instrument - Agency	$wup \sim lch > lin > res > jcn \sim random$
Product - Producer	$wup \sim lch > lin \sim jcn \sim res > random$
Origin - Entity	$wup \sim lch > lin > res \sim jcn > random$
Theme - Tool	$lch > lin \sim wup > res > jcn > random$
Part - Whole	$wup \sim lin \sim lch > res > jcn \sim random$
Content - Container	$wup > lch > lin \sim res > jcn \sim random$

: Ranking of the relatedness measures with respect to their accuracy on the training sets ($\sim$ stands for measure equivalence, $a > b$ indicates that the measure a significantly outperforms b).

WordNet measures (4)

Conclusions

- *wup*, *lch*, and *lin* almost always yield the best results, no matter what relation is considered.
- *wup* and *lch* explore the WordNet taxonomy using a length of the paths between two concepts, or their depth in the WordNet hierarchy and, consequently, belong to the path-based measures.
- *res*, *lin* and *jcn* are information content based measures, and here relatedness between two concepts is defined through the amount of information they share.
- path-based measures outperform information content measures on this task but it may not be true for other applications.

Your homework #2

Free association word pairs (First, Hapax and Random categories), e. g.

hate love: FIRST
else something: HAPAX
digital revolt: RANDOM

`http://wordspace.collocations.de/doku.php/data:esslli2008:correlation_with_free_association_norms`

`http://www.phil.uu.nl/tst/2012/Werk/huiswerk2.pdf`

To summarize (1)

Today, we have looked at

- other resources for lexical semantics (e.g., PropBank)
- distributional and WordNet similarity measures

To summarize (1)

Today, we have looked at

- other resources for lexical semantics (e.g., PropBank)
- distributional and WordNet similarity measures

To summarize (2)

- read at home (if you haven't done it yet) chapter 19 and 20 (from section 6) from Jurafsky.
- next class: June 13 on machine learning concepts and methods.