Distributional Semantics and Automatic Semantic Information Extraction: An Introduction

Theory: Weaver, Firth, Z. Harris

- Warren Weaver („Translation“, 1949/1955)
  - “But if one lengthens the slit in the opaque mask, until one can see not only the central word in question, but also say N words on either side, then if N is large enough one can unambiguously decide the meaning of the central word.”[1]
- John Rupert Firth and Zellig Harris
  - the most precise way of determining a word’s meaning is by investigating the meanings of the words that occur along with that word.[2]
- Distributional Semantics
  - “linguistic items with similar distributions have similar meanings”[3]

Word-Sense Disambiguation/Induction

- This theoretical basis is used to automatically determine the sense of a word
- E.g., the English word “dog”
  - Noun, verb, adjective?
  - Noun: the animal? Something ugly?
  - Verb: to follow closely? To be lazy?
  - Adjective: “The dog days of summer”
- Machine translation

Problems with WSD

- Have you used Google Translate?
- Easier to tell apart homographs (different words spelled the same) than various senses of the same word
  - 90-96% on homographs
  - 59.1% to 69.0% for various senses
  - baseline accuracy, choosing the most frequent sense = 51.4% and 57%”[4]
- Significant difference but still poor results for senses

Tracking similarity/differences

- Words differ where distribution differs
- Example: difference between “big” and “large”?
  - Big occurs frequently with “sister”, large does not
  - We see that “big” has a sense that “large” does not
- Words with fewer differences are closer
- This is essentially how topic modeling works

What is the “context”

- We are talking about meaning, so meaning units
- i.e., words that rely on each other to create meaning
- Document?
- Paragraph?
- Sentence?
- Sinclair
  - “The text is the sentence that is in front of us when an act of reading is in progress. Each sentence then is a new beginning to the text.”[5]

Problem with ancient texts

- Where are the sentence boundaries?
  - Codex Sinaiticus
- Closely related words with *tend* to occur close to each other
- Research suggests between 2 and 5 words left and right

Which features

- For English, types should work well
- For Greek, et al.
  - The Greek verb
- Lemmas (dictionary forms) might be better
  - But it always depends on the size of your corpus
  - If you dilute your information too much, you will get good results only for the most common words

What can you do with this information?

- Topic Modeling
- Machine Translation
- Semantic Drift
  - i.e., calculate which words change meanings,
  - by how much,
  - and in which direction.

Count Co-Occurrence

L4	L3	L2	L1	Target	R1	R2	R3	R4
ἐν	ἀρχή	ποιέω	ὁ	θεός	ὁ	οὐρανός	καί	ὁ
ὁ	ἄβυσσος	καί	πνεῦμα	θεός	ἐπιφέρω	ἐπάνω	ὁ	ὕδωρ

Counts:
ὁ - 5
καί - 2
ἐν - 1
ἀρχή - 1
ποιέω - 1
ἄβυσσος - 1
πνεῦμα - 1
ἐπιφέρω - 1
ἐπάνω - 1
ὕδωρ - 1

The Co-occurrence Matrix

- “the dog bit the man” and “the bat hit the ball”

	the	dog	bit	man	bat	hit	ball
the	4	2	2	2	2	2	2
dog	2	0	1	1	0	0	0
bit	2	1	0	1	0	0	0
man	2	1	1	0	0	0	0
bat	2	0	0	0	0	1	1
hit	2	0	0	0	1	0	1
ball	2	0	0	0	1	1	0

Your Homework!

- Construct a 4L-4R co-occurrence matrix for every document in the “input” folder under the “Week 6” homework folder in “Course_Materials”

Works Cited

1. Warren Weaver. "Translation." 1955. http://www.mt-archive.info/Weaver-1949.pdf. 30 October 2013. 8.
2. Zellig S. Harris, “How Words Carry Meaning.” Language and Information: The Bampton Lectures, Columbia University, 1986. Lecture. http://www.ircs.upenn.edu/zellig/3_2.mp3. See also John Rupert Firth, "A synopsis of linguistic theory 1930-1955.“ in *Selected Papers of J.R. Firth, 1952-1959*. Ed. F.R. Palmer. Harlow: Longmans, 1968. P. 179.
3. http://en.wikipedia.org/wiki/Distributional_semantics
4. http://en.wikipedia.org/wiki/Word-sense_disambiguation
5. John Sinclair, “Trust the Text”, in *Trust the Text: Language, Corpus and Discourse.* Ed. John Sinclair and Ronald Carter. London: Routledge, 2004. 9-23. P. 14.