Semantic NLP
Clinical NLP Systems

BMI701 Introduction of Biomedical Informatics
Lab Session 6

Wei-Hung Weng

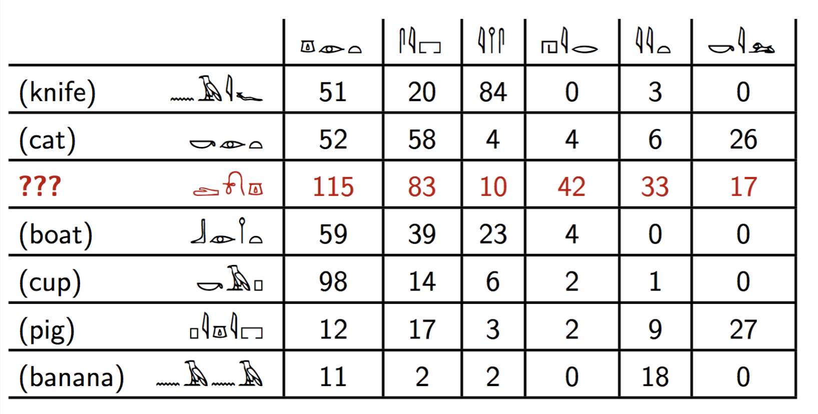
October 26, 2016

HMS DBMI — MGH LCS

HARVARD MEDICAL SCHOOL

MASSACHUSETTS GENERAL HOSPITAL

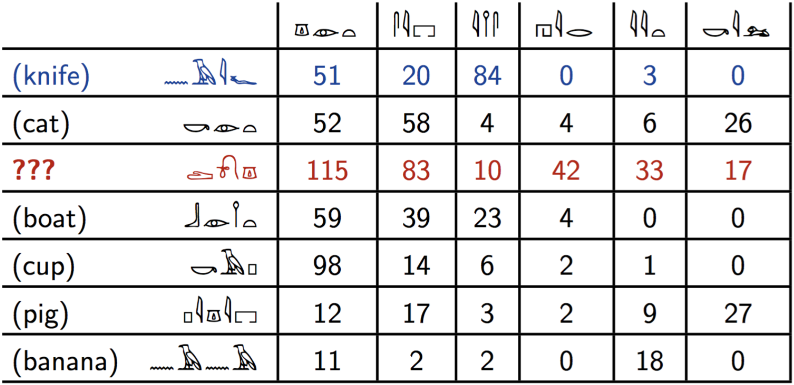
Deciphering Hieroglyphs



(knife)	51	20	84	0	3	0
(cat)	52	58	4	4	6	26
???	115	83	10	42	33	17
(boat)	59	39	23	4	0	0
(cup)	98	14	6	2	1	0
(pig)	12	17	3	2	9	27
(banana)	11	2	2	0	18	0

Evert 2010

Deciphering Hieroglyphs

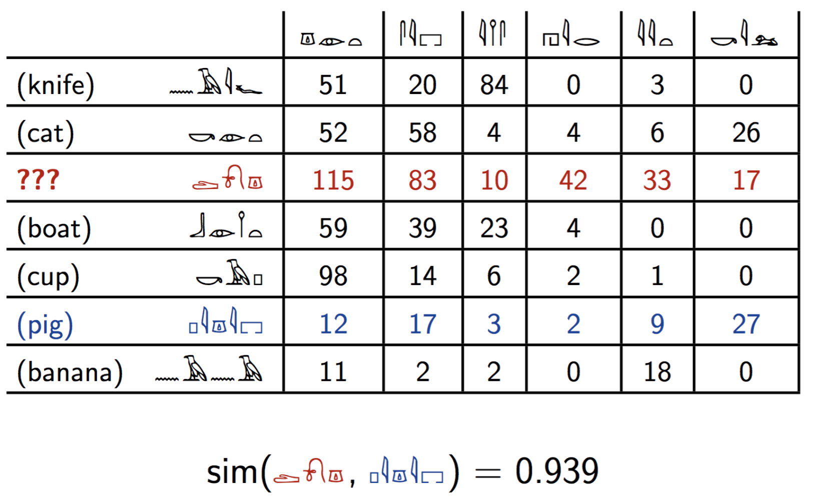


(knife)	51	20	84	0	3	0
(cat)	52	58	4	4	6	26
???	115	83	10	42	33	17
(boat)	59	39	23	4	0	0
(cup)	98	14	6	2	1	0
(pig)	12	17	3	2	9	27
(banana)	11	2	2	0	18	0

sim(,) = 0.770

Evert 2010

Deciphering Hieroglyphs

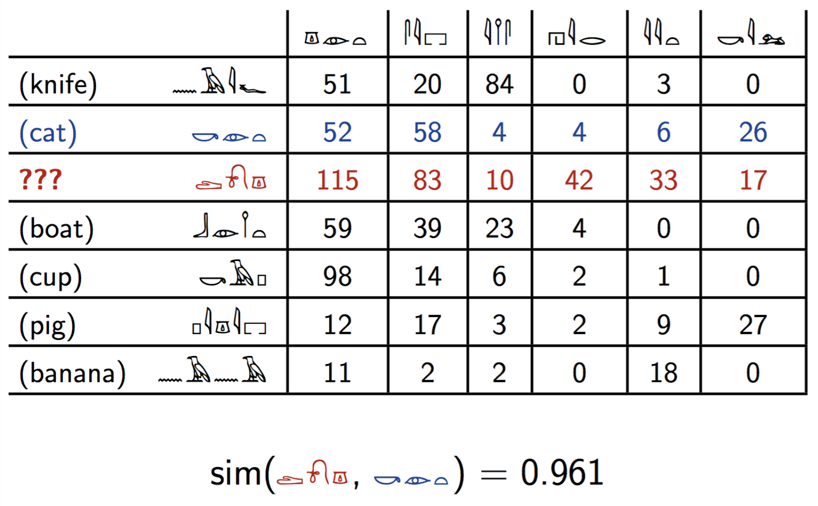


(knife)	51	20	84	0	3	0
(cat)	52	58	4	4	6	26
???	115	83	10	42	33	17
(boat)	59	39	23	4	0	0
(cup)	98	14	6	2	1	0
(pig)	12	17	3	2	9	27
(banana)	11	2	2	0	18	0

sim(,) = 0.939

Evert 2010

Deciphering Hieroglyphs

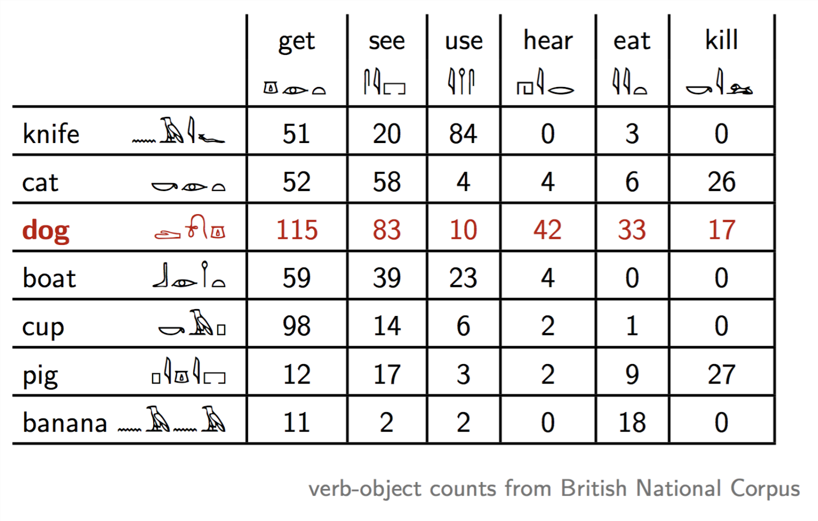


		𓏌𓁹𓏏	𓇌𓊪	𓇋𓋹𓇋	𓉐𓇋𓂋	𓇌𓏏	𓎡𓇋𓆑
(knife)	𓈖𓄿𓇋𓆑	51	20	84	0	3	0
(cat)	𓎡𓁹𓏏	52	58	4	4	6	26
???	𓂧𓎛𓇋𓏌	115	83	10	42	33	17
(boat)	𓆓𓁹𓋹𓏏	59	39	23	4	0	0
(cup)	𓎡𓄿𓊪	98	14	6	2	1	0
(pig)	𓊪𓇋𓏌𓇋𓊪	12	17	3	2	9	27
(banana)	𓈖𓄿𓈖𓄿	11	2	2	0	18	0

$$\mathrm{sim}(\text{𓂧𓎛𓇋𓏌}, \text{𓎡𓁹𓏏}) = 0.961$$

Evert 2010

Deciphering Hieroglyphs



	get	see	use	hear	eat	kill
knife	51	20	84	0	3	0
cat	52	58	4	4	6	26
dog	115	83	10	42	33	17
boat	59	39	23	4	0	0
cup	98	14	6	2	1	0
pig	12	17	3	2	9	27
banana	11	2	2	0	18	0

verb-object counts from British National Corpus

Evert 2010

Tf-idf Weighting

- Importance of the term in the corpus
- For term i in document j

$$w_{i,j} = tf_{i,j} \times \log(\frac{N}{df_i})$$

- $tf_{i,j}$: frequency of i in j
- df_i: number of documents have i
- N: number of all documents

Tf-idf Weighting Example

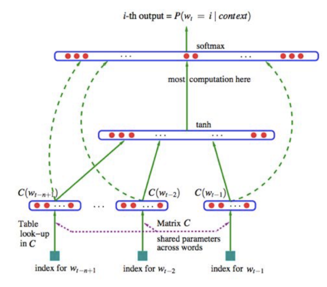
- A: "Dog is so cute."
- B: "I like dog."
- $tfidf_{('dog',A)} = \frac{1}{4} \times \log(\frac{2}{2}) = 0$
- $tfidf_{('dog',B)} = \frac{1}{3} \times \log(\frac{2}{2}) = 0$
- $tfidf_{('cute',A)} = \frac{1}{4} \times \log(\frac{2}{1}) = \frac{\log 2}{4}$
- $tfidf_{('cute',B)} = \frac{0}{3} \times \log(\frac{2}{0}) = 0$

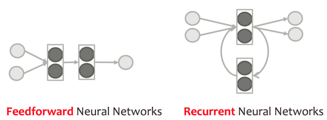
Semantic Approaches

- Matrix decomposition
  - LSI (Deerwester 1990), NMF (Lee 1999), NTF (Cruys 2010)
  - Using SVD: $U \Sigma V$
  - Fast, unless using NTF
- Language model
  - PLSI (Hofmann 1999), LDA (Blei 2003)
  - Topic modeling, using probability
  - Heavy computation

Semantic Approaches

- Neural network model
  - NNLM (Bengio 2003), RNN (Mikolov 2010), skip-gram / CBOW (Mikolov 2013)
  - Heavy computation, hard to implementation
  - Interpretation...?

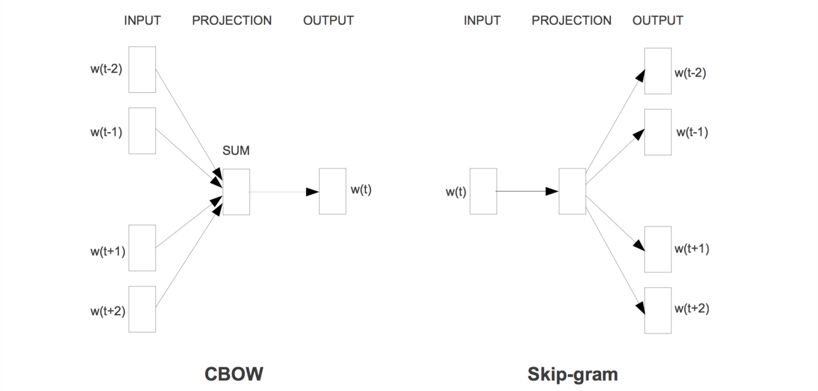




Bengio 2003

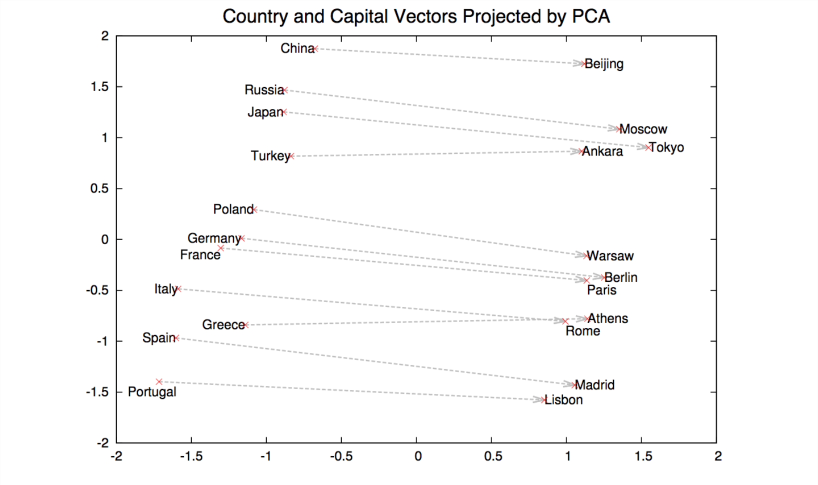
Chang 2015

Semantic Approaches



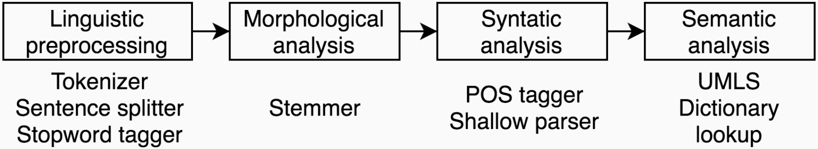
Mikolov 2013

Semantic Approaches



Mikolov 2013

NLP Workflow



MetaMap / cTAKES workflow

MetaMap

- Developed by NLM (Aronson 1994)
- Web application of MetaMap
- Java API
- Locally execution
- Download

cTAKES

- Developed by Mayo NLP (Savova 2010)
- Modularized
- CLI
- Download

Demonstration

- Topic modeling (`topicmodels`)
- MetaMap / cTAKES
  - Need to download in advance
  - Use CLI or `system()` in R
  - Further processing

Some Advanced NLP Courses

- NLTK book (very useful!)
- Coursera NLP (Jurafusky)
- Coursera NLP (Radev)
- Coursera NLP provided by Michael Collins is also good, but it's gone now
- Coursera NLP (Collins)
- CS287: Natural Language Processing
- 6.864: Advanced Natural Language Processing

Take Home Message

- More text mining techniques
  - Topic modeling, vector space model
- MetaMap / cTAKES
- Contact
  - Github repository
  - ckbjimmy@gmail.com
  - Linkedin: Wei-Hung Weng