Lexical Acquisition

Lecture #9

Introduction to Natural Language Processing

CMPSCI 585, Fall 2004

University of Massachusetts Amherst

Andrew McCallum

Words and their meaning

Three lectures:

- Collocations
  - multiple words together, different meaning than than the sum of its parts
- Word disambiguation
  - one word, multiple meanings
- *This time:* Lexical Acquisition
  - verb subcategorization
  - attachment ambiguity
  - selectional preference
  - semantic similarity
    - multiple words, “same” meaning

Today’s Main Points

- What is Lexical Acquisition and why is it useful.
- Verb subcategorization.
- Attachment ambiguity
- Selectional preference
- Clustering words into semantically similar classes.

Lexical Acquisition

- Acquiring the properties of words
- Practical: filling holes in dictionaries
  - Lots of useful information isn’t in dictionaries anyway e.g. “associated with” versus “associated to”
- Claim: most knowledge of language is encoded in words and their properties.
- Acquiring collocations and word sense disambiguation are examples of lexical acquisition, but there are many other types.

Why Lexical Acquisition

- Language evolves. i.e., new words and new uses of old words are constantly invented.
- Traditional Dictionaries were written for the needs of human users. Lexicons are dictionaries formatted for computers.
- In addition to the format, lexicons can be useful if they contain quantitative information. Lexical acquisition can provide such information.

Verb Phrase and Subcategorization

- Verb phrase consists of
  - Verb
  - a number of constituents
- Examples
  - VP → V disappear
  - VP → V NP prefer a morning flight
  - VP → V NP PP leave Boston in the morning
  - VP → V PP leave on Thursday
  - VP → V S said you had a $200 fare
    - Sentential complement

Different verbs, different constituents

- A verb phrase can have many possible kinds of constituents, but
- Not every verb is compatible with every verb phrase
- Examples
  - "want" VP → V NP "I want a flight"
  - "want" VP → V VPto "I want to fly to…"
  - "find" VP → V NP "I found a flight"
  - "find" VP → V VPto * "I found to fly to…"
- *Transitive*, take a direct object
  - "find" "I found a flight"
- *Intransitive*, do not take a direct object
  - "disappear" * "I disappeared a flight"
- Transitive and Intransitive are simple examples of verb *subcategorization*.

Verb Subcategorization

- Verbs express their semantic arguments with different syntactic means.
- "frame" = slots for arguments of the verb
- "category" = verbs that take the same semantic args
  e.g. verbs with semantic arguments theme and recipient
- "subcategory" = verbs that use the same syntactic means to express these semantic arguments.

- Additional examples:

  subcategory #1: prepositional phrase
  "He donated a large sum of money to the church."

  subcategory #2: double-object
  "He gave the church a large sum of money."

Examples of subcategorization frames

- **Intransitive verb**
  - NP[subject]
  - *"The woman walked."*
- **Transitive verb**
  - NP[subject] NP[object]
  - *"John loves Mary."*
- **Ditransitive verb**
  - NP[subject], NP[direct object], NP[indirect object]
  - *"Mary gave Peter flowers."*
- **Intransitive with PP**
  - NP[subject], PP
  - *"I rent in Northampton."*
- **Sentential complement**
  - NP[subject], clause
  - *"I know (that) she likes you."*
- **Transitive with sentential complement**
  - NP[subject], NP[object], clause
  - *"She told me that Gary is coming."*

One verb, multiple subcategorizations

- One verb can take different subcategorization frames

- Example: "find"

  - VP → V NP …find a flight
  - VP → V NP NP …find me a flight

Subcategorization needed for parsing

- *She told the man where Peter grew up.*
- *She found the place where Peter grew up.*

- *She told [the man] [where Peter grew up].*
- *She found [the place [where Peter grew up]].*

  Helps us get attachment right.

- Unfortunately most dictionaries don't contain subcategorization frames, and those that do are horribly incomplete.

Learning subcategorization frames
[Brent 1993]

- Does some particular verb take direct object frame VP → V NP?
- **Cues for frames**
  e.g. assume that pattern
  "verb (pronoun | capitalized word) punctuation"
  identifies *direct object frame* with error rate e=0.1
- **Count occurrences**
  n = number of occurrences of verb in question
  m = number of occurrences of cue with verb
- Hypothesis testing, H0 = verb does not take frame

$$P(H0|\text{cue count} \geq m) = \sum_{r=m}^{n} \binom{n}{r} e^{r} (1-e)^{n-r}$$

Learning subcategorization frames
[Brent 1993] [Manning 1993]

- Brent's system does well at precision, but not well at recall.
- (Manning, 93)'s system addresses this problem by using a tagger and running the cue detection on the output of the tagger.
  - e.g. say "find/V DET NP" indicates direct object frame
- Manning's method can learn a large number of subcategorization frames, even those that have only low-reliability cues.

Learned subcategorization frames
[Manning 1993]

Verb	Correct	Incorrect	Oxford AL Dictionary
bridge	1	1	1
burden	2		2
depict	2		3
emanate	1		1
leak	1		5
occupy	1		3
remark	1	1	4
retire	2	1	5

Error in remark: attributed intransitive frame, probably due to
"And here we are 10 years later with the same problems," Mr. Smith remarked.

Attachment Ambiguity

- Where to attach a phrase in the parse tree?
- *"I saw the man with the telescope."*
  - What does "with a telescope" modify?
  - Is the problem AI complete? Yes, but…
  - Proposed simple structural factors
    - Right association [Kimball 1973]
      'low' or 'near' attachment = 'early closure' of NP
    - Minimal attachment [Frazier 1978]
      (depends on grammar) = 'high' or 'distant' attachment = 'late closure' (of NP)

Attachment Ambiguity

- Such simple structural factors dominated in early psycholinguistics, and are still widely invoked.
- In the V NP PP context, right attachment gets right 55-76% of the cases…
- But this means that it gets wrong 33-45% of the cases!

Attachment Ambiguity

- "The children ate the cake <u>with a spoon</u>."
- "The children ate the cake <u>with frosting</u>."

- "Joe included the package <u>for Susan</u>."
- "Joe carried the package <u>for Susan</u>."

- *Ford, Bresnan and Kaplan (1982):*
  "It is quite evident, then, that the closure effects in these sentences are induced in some way by the choice of the lexical items."

Simple model

- (Log) likelihood ratio
  - A common and good way of comparing between two exclusive alternatives
  - Same idea as a naïve Bayes classifier

$$\log \frac{P(\text{preposition}|\text{verb})}{P(\text{preposition}|\text{noun})}$$

  - if >0, attach to verb, if <0 attach to noun
  - For example,
    P(with a spoon | ate) > P(with a spoon | cake)

Attachment, Problematic Example

- *"Chrysler confirmed that it would **end** its troubled **venture** with Maserati."*

w	C(w)	C(w, with)
end	5156	607
venture	1442	155

- Get wrong answer:
  P(with|end) = (607/5156) = 0.118
  P(with|venture) = (155/1442) = 0.107

- Should also express preference for attaching 'low'.

Attachment Method
[Hindle & Rooth 1993]

- Event space: all V NP PP* sequences
  but PP must modify V or first N
- Don't directly decide whether PP modifies V or N
- Rather look at binary random variables
  - VA_p: Is there a PP headed by p which attaches to v
  - NA_p: Is there a PP headed by p which attaches to n
- Both can be 1:
  "He put the book on World War II on the table."

Attachment Method
[Hindle & Rooth 1993]

- Independence assumptions
  $P(VA_p, NA_p \mid v, n) = P(VA_p \mid v,n)\, P(NA_p \mid v,n)$
  $= P(VA_p \mid v)\, P(NA_p \mid n)$
- Decision space: first PP after NP. [NB!]
- $P(Attch(p)=n|v,n) = P(VA_p=0 \vee VA_p=1|v)\, P(NA_p=1|n)$
  $= 1.0\, P(NA_p=1|n)$
  $= P(NA_p=1|n)$

- It doesn't matter what VA_p is! If both are true, the first PP after the NP must modify the noun (in phrase structure trees, lines don't cross).

Attachment Method
[Hindle & Rooth 1993]

- But conversely, in order for the first PP headed by the preposition p to attach to the verb, both VA_p=1 and NA_p=0 must hold.
- $P(Attach(p)=v|v,n) = P(VA_p=1, NA_p=0|v,n)$
  $= P(VA_p=1|v)\, P(NA_p=0|n)$

- We assess which is more likely by a (log) likelihood ratio:

$$\lambda(v,n,p) = \log_2 \frac{P(\mathrm{Attach}(p) = v|v,n)}{P(\mathrm{Attach}(p) = n|v,n)}$$
$$= \log_2 \frac{P(VA_p = 1|v)P(NA_p = 0|v)}{P(NA_p = 1|n)}$$

- If large positive, decide verb attachment; if large negative, decide noun attachment.

Attachment Method
[Hindle & Rooth 1993]

- How do we learn probabilities?
  From (smoothed) MLEs:

  $P(VA_p=1|v) = C(v,p) / C(v)$
  $P(NA_p=1|n) = C(n,p) / C(n)$

- How do we get estimates from unlabeled corpus?
  Use partial parser, and look for unambiguous cases:
  - "The road to London is long and winding."
  - "She sent him to the nursery to gather up his toys."

Attachment Method
[Hindle & Rooth 1993]

- Hindle and Rooth heuristically determine C(v,p), C(n,p) and C(n,0) from unlabeled data:

1. Build an initial model by counting all unambiguous cases.
2. Apply initial model to all ambiguous cases and assign them to the appropriate count if l exceeds a threshold (2/-2).
3. Divide the remaining ambiguous cases evenly between the counts (increase C(v,p) and C(n,p) by 0.5 for each).

Attachment Method Example
[Hindle & Rooth 1993]

- “Moscow sent more than 100,000 soldiers into Afghanistan…”

Other attachment issues

- There are attachment questions other than prepositional phrases
  - adverbial, participial, noun compounds
  - Examples
    door bell manufacturer
    [door bell] manufacturer
    Unix system administrator
    Unix [system administrator]
  - Data sparseness is a bigger problem with many of these
- In general, indeterminacy is quite common
  - “We have not **signed** a settlement **agreement** with them.”
  - Either reading seems equally plausible.

Lexical acquisition, semantic similarity

- Previous models give same estimate to all unseen events.
- Unrealistic - could hope to refine that based on semantic classes of words
- Examples
  - “Susan had never eaten a fresh durian before.”
  - Although never seen “eating pineapple” should be more likely than “eating holograms” because pineapple is similar to apples, and we have seen “eating apples”.

An application: selectional preferences

- Most verbs prefer arguments of a particular type. Such regularities are called *selectional preferences* or *selectional restrictions*.
- “Bill drove a…” Mustang, car, truck, jeep
- Selectional preference strength: how strongly does a verb constrain direct objects
- “see” versus “unknotted”

Measuring selectional preference strength

- Assume we are given a clustering of (direct object) nouns. Resnick (1993) uses WordNet.

$$S(v) = D(P(C|v)||P(C) = \sum_c P(c|v) \log \frac{P(c|v)}{P(c)}$$

- Selectional association between a verb and a class

$$A(v,c) = \frac{P(c|v) \log \frac{P(c|v)}{P(c)}}{S(v)}$$

  Proportion that its summand contributes to preference strength.
- For nouns in multiple classes, disambiguate as most likely sense: $A(v,n) = \max_{c \in \text{classes}(n)} A(v,c)$

Selection preference strength (made up data)

Noun class c	P(c)	P(c|eat)	P(c|see)	P(c|find)
people	0.25	0.01	0.25	0.33
furniture	0.25	0.01	0.25	0.33
food	0.25	0.97	0.25	0.33
action	0.25	0.01	0.25	0.01
SPS S(v)		**1.76**	**0.00**	**0.35**

A(eat, food) = 1.08
A(find, action) = -0.13

Selectional Preference Strength example
(Resnick, Brown corpus)

Verb v	Noun n	$A(v,n)$	Class	Noun n	$A(v,n)$	Class
answer	*request*	4.49	speech act	*tragedy*	3.88	communication
find	*label*	1.10	abstraction	*fever*	0.22	psych. feature
hear	*story*	1.89	communication	*issue*	1.89	communication
remember	*reply*	1.31	statement	*smoke*	0.20	article of commerce
repeat	*comment*	1.23	communication	*journal*	1.23	communication
read	*article*	6.80	writing	*fashion*	−0.20	activity
see	*friend*	5.79	entity	*method*	−0.01	method
write	*letter*	7.26	writing	*market*	0.00	commerce

But how might we measure word similarity for word classes?

- Vector spaces

A document-by-word matrix A.

	cosmonaut	astronaut	moon	car	truck
d_1	1	0	1	1	0
d_2	0	1	1	0	0
d_3	1	0	0	0	0
d_4	0	0	0	1	1
d_5	0	0	0	1	0
d_6	0	0	0	0	1

But how might we measure word similarity for word classes?

- Vector spaces

word-by-word matrix B

	cosmonaut	astronaut	moon	car	truck
cosmonaut	2	0	1	1	0
astronaut	0	1	1	0	0
moon	1	1	2	1	0
car	1	0	1	3	1
truck	0	0	0	1	2

A modifier-by-head matrix C

	cosmonaut	astronaut	moon	car	truck
Soviet	1	0	0	1	1
American	0	1	0	1	1
spacewalking	1	1	0	0	0
red	0	0	0	1	1
full	0	0	1	0	0
old	0	0	0	1	1

Similarity measures for binary vectors

Similarity measure	Definition
matching coefficient	$\lvert X \cap Y \rvert$
Dice coefficient	$\frac{2\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert}$
Jaccard coefficient	$\frac{\lvert X \cap Y \rvert}{\lvert X \cup Y \rvert}$
Overlap coefficient	$\frac{\lvert X \cap Y \rvert}{\min(\lvert X \rvert, \lvert Y \rvert)}$
cosine	$\frac{\lvert X \cap Y \rvert}{\sqrt{\lvert X \rvert \times \lvert Y \rvert}}$

Cosine measure

$$\cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

maps vectors onto unit circle by dividing through by lengths:

$$|\vec{x}| = \sqrt{\sum_{i=1}^{n} x_i^2}$$

Example of cosine measure on word-by-word matrix on NYT

Focus word	Nearest neighbors							
garlic	*sauce*	.732	*pepper*	.728	*salt*	.726	*cup*	.726
fallen	*fell*	.932	*decline*	.931	*rise*	.930	*drop*	.929
engineered	*genetically*	.758	*drugs*	.688	*research*	.687	*drug*	.685
Alfred	*named*	.814	*Robert*	.809	*William*	.808	*W*	.808
simple	*something*	.964	*things*	.963	*You*	.963	*always*	.962

Probabilistic measures

(Dis-)similarity measure	Definition
KL divergence	$D(p\|q) = \sum_i p_i \log \frac{p_i}{q_i}$
Skew	$D(q\|\alpha r + (1-\alpha)q)$
Jensen-Shannon (was IRad)	$\frac{1}{2}D(p\|\frac{p+q}{2}) + D(q\|\frac{p+q}{2})$
L_1 norm (Manhattan)	$\sum_i \lvert p_i - q_i \rvert$

Neighbors of word "company" [Lee]

Skew ($\alpha = 0.99$)	J.-S.	Euclidean
airline	business	city
business	airline	airline
bank	firm	industry
agency	bank	program
firm	state	organization
department	agency	bank
manufacturer	group	system
network	govt.	today
industry	city	series
govt.	industry	portion

Examples of Verb Subcategorization

Frame	Functions	Verb	Example
NP NP	subject, object	greet	She greeted me.
NP S	subject, clause	hope	She hopes he will attend.
NP INF	subject, infinitive	hope	She hopes to attend.
NP NP S	subject, object, clause	tell	She told me he will attend.
NP NP INF	subject, object, infinitive	tell	She told him to attend.