Lexical Acquisition

FSNLP, chapter 8

Christopher Manning and
Hinrich Schütze

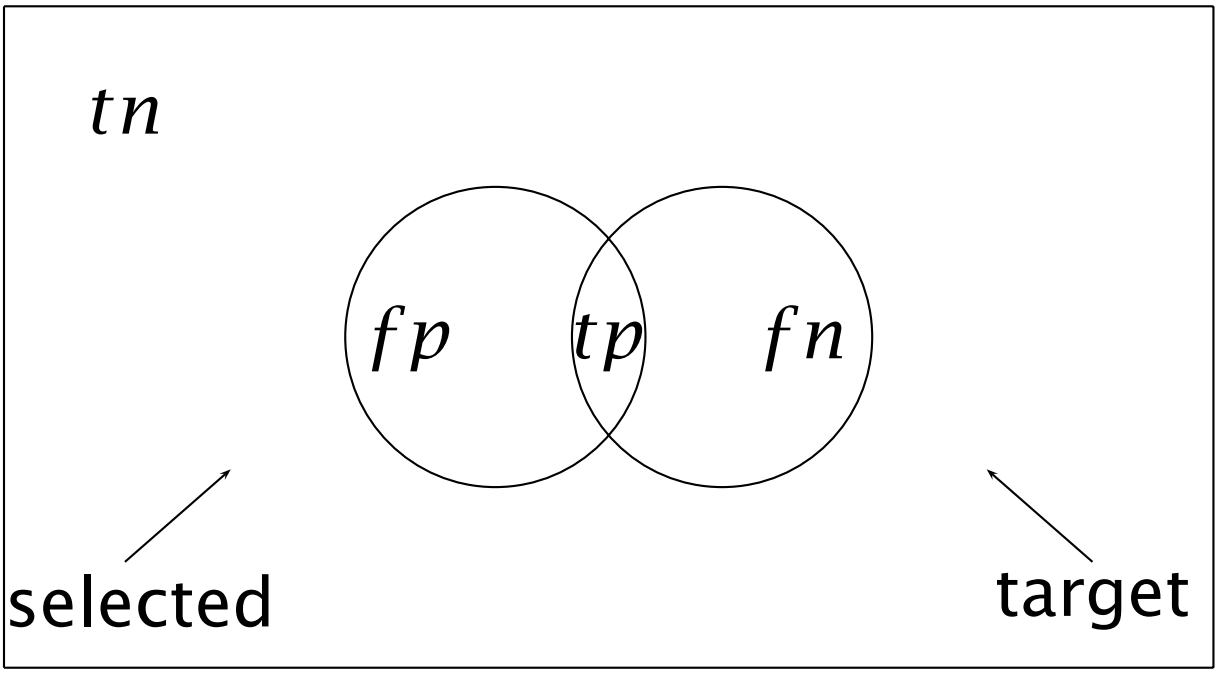
Lexical acquisition

- Language acquisition: acquiring the properties of words
- Practical: filling holes in dictionaries
  - Language is productive
  - Lots of stuff isn't in dictionaries anyway
- Claim: most knowledge of language is encoded in words.

The 2×2 contingency matrix

	Actual	
System	target	¬ target
selected	tp	fp
¬selected	fn	tn

A diagram motivating the measures of precision and recall.



Precision is defined as a measure of the proportion of selected items that the system got right:

$$\text{precision} = \frac{tp}{tp + fp}$$

Recall is defined as the proportion of the target items that the system selected:

$$\text{recall} = \frac{tp}{tp + fn}$$

Combined measures

Does one just add them? Bad, because the measures aren't independent.

What's a sensible model?

(see http://www.dcs.gla.ac.uk/Keith/Preface.html)
Rijsbergen (1979:174) defines and justifies the usually used alternative.

Assumptions:

- Interested in document proportions not absolute numbers
- Decreasing marginal effectiveness of recall and precision, e.g.:

$$(R+1, P-1) > (R, P)$$

but

$$(R+1, P) > (R+2, P-1)$$

Makes curves convex towards origin.

The F measure (where $F = 1 - E$):

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

where P is precision, R is recall and α weights precision and recall. (Or in terms of β, where $\alpha = 1/(\beta^2 + 1)$.)

A value of $\alpha = 0.5$ is often chosen.

$$F = 2PR/(R + P)$$

Subcategorization frames

Here are some subcategorization frames that are common in English.

- **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 Paddington*.
- **Transitive with PP.** NP[subject], NP[object], PP. *She put the book on the table*.

- **Sentential complement.** NP[subject], clause. *I know (that) she likes you*.
- **Transitive with sentential complement.** NP[subj], NP[obj], clause. *She told me that Gary is coming on Tuesday*.

(1) a. She told the man where Peter grew up.

b. She found the place where Peter grew up.

(2) a. She told [the man] [where Peter grew up].

b. She found [the place [where Peter grew up]].

(Info in learner's dictionaries.)

Brent (1993):

- Cues for frames.

  e.g., pronoun or capitalized followed by punctuation

- Hypothesis testing

$$
\begin{aligned}
p_E &= P(v^i(f^j) = 0 | C(v^i, c^j) \geq m) \\
&= \sum_{r=m}^{n} \binom{n}{r} {\epsilon_j}^r (1 - \epsilon_j)^{n-r}
\end{aligned}
$$

verb v^i occurs n times; there are $m \leq n$ occurrences with a cue for frame f^j $C(v^i, c^j)$ is the number of times that v^i occurs with cue c^j, and ϵ_j is the error rate for cue f^j,

Manning (1993)

Uses tagger. More errorful, but much more abundant cues.

- *He relies **on** relatives.*
- *She compared the results **with** earlier findings.*

Learned subcategorization frames

Verb	Correct	Incorrect	OALD
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
shed	1		2
troop	0		3

Two of the errors are prepositional phrases (PPs): *to bridge between* and *to retire in*.

One could argue that *retire* subcategorizes for the PP *in Malibu* in a sentence like *John retires in Malibu* since the verb and the PP-complement enter into a closer relationship than mere adverbial modification.

The third error in the table is the incorrect assignment of the intransitive frame to *remark*. This is probably due to sentences like (3) which look like *remark* is used without any arguments (except the subject).

(3) "And here we are 10 years later with the same problems," Mr. Smith remarked.

Attachment ambiguities

- *I saw the man with a 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 ambiguities (2)

- Such simple structural factors dominated in early psycholinguistics, and are still widely invoked.
- In the V NP PP context, right attachment gets right 55–67% of cases.
- But that means it gets wrong 33–45% of cases

Attachment ambiguities (3)

- The children ate the cake with a spoon.
- The children ate the cake with frosting.
- Moscow sent more than 100,000 soldiers into Afghanistan . . .
- Sydney Water breached an agreement with NSW Health . . .

Words are good predictors (even absent understanding).

Importance of lexical factors

Ford, Bresnan and Kaplan (1982) [as part of the promotion of 'lexicalist' linguistic theories]

- Order of grammatical rule processing (by human) determines closure effects
- Ordering is jointly determined by strengths of alternative lexical forms, strengths of alternative syntactic rewrite rules, and the sequence of hypotheses in the parsing process

Importance of lexical factors (2)

Ford, Bresnan and Kaplan (1982):

- *Joe included the package for Susan.*
- *Joe carried the package for Susan.*

"It is quite evident, then, that the closure effects in these sentences are induced in some way by the choice of the lexical items." (Psycholinguistic studies show this is true *independent* of discourse context.)

Simple model

(Log) Likelihood Ratio [a common and good way of comparing between two exclusive alternatives]

$$\lambda(v, n, p) = \log \frac{P(p|v)}{P(p|n)}$$

Problem: ignores preference for attaching “low”.

Problematic example (*NYT*)

- Chrysler confirmed that it would end its troubled venture with Maserati.
w	$C(w)$	$C(w, \textit{with})$
end	5156	607
venture	1442	155
- Get wrong answer:

$$P(p|v) = \frac{607}{5156} \approx 0.118$$

$$> \quad P(p|n) = \frac{155}{1442} \approx 0.107$$

Hindle and Rooth (1993 [1991])

- 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 RVs:
  - 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

Independence assumptions:

$$\begin{aligned} P(\mathsf{VA}_p, \mathsf{NA}_p | v, n) &= P(\mathsf{VA}_p | v, n) P(\mathsf{NA}_p | v, n) \\ &= P(\mathsf{VA}_p | v) P(\mathsf{NA}_p | n) \end{aligned}$$

Decision space: first PP after NP. [NB!]

$$\begin{aligned} P(\mathsf{Attach}(p) = n | v, n) &= P(\mathsf{VA}_p = 0 \vee \mathsf{VA}_p = 1 | v) \\ &\quad \times P(\mathsf{NA}_p = 1 | n) \\ &= 1.0 \times P(\mathsf{NA}_p = 1 | n) \\ &= P(\mathsf{NA}_p = 1 | n) \end{aligned}$$

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).

But conversely, in order for the first PP headed by the preposition p to attach to the verb, both $\mathsf{VA}_p = 1$ and $\mathsf{NA}_p = 0$ must hold:

$$\begin{aligned} P(\mathsf{Attach}(p) = v | v, n) &= P(\mathsf{VA}_p = 1, \mathsf{NA}_p = 0 | v, n) \\ &= P(\mathsf{VA}_p = 1 | v) P(\mathsf{NA}_p = 0 | n) \end{aligned}$$

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

$$\begin{aligned} \lambda(v, n, p) &= \log_2 \frac{P(\mathsf{Attach}(p) = v | v, n)}{P(\mathsf{Attach}(p) = n | v, n)} \\ &= \log_2 \frac{P(\mathsf{VA}_p = 1 | v) P(\mathsf{NA}_p = 0 | v)}{P(\mathsf{NA}_p = 1 | n)} \end{aligned}$$

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

Building the model

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

$$P(\text{VA}_p = 1|v) = \frac{C(v,p)}{C(v)}$$

$$P(\text{NA}_p = 1|n) = \frac{C(n,p)}{C(n)}$$

How do we get estimates from an unlabelled corpus? Use partial parser, and look for unambiguous cases:

- The road *to London* is long and winding.
- She sent him *into the nursery* to gather up his toys.

Hindle and Rooth heuristically determining $C(v,p)$, $C(n,p)$, and $C(n,\varnothing)$ 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 λ exceeds a threshold $(2/-2)$.
3. Divide the remaining ambiguous cases evenly between the counts (increase both $C(v,p)$ and $C(n,p)$ by 0.5 for each).

Example

Moscow sent more than 100,000 soldiers into Afghanistan . . .

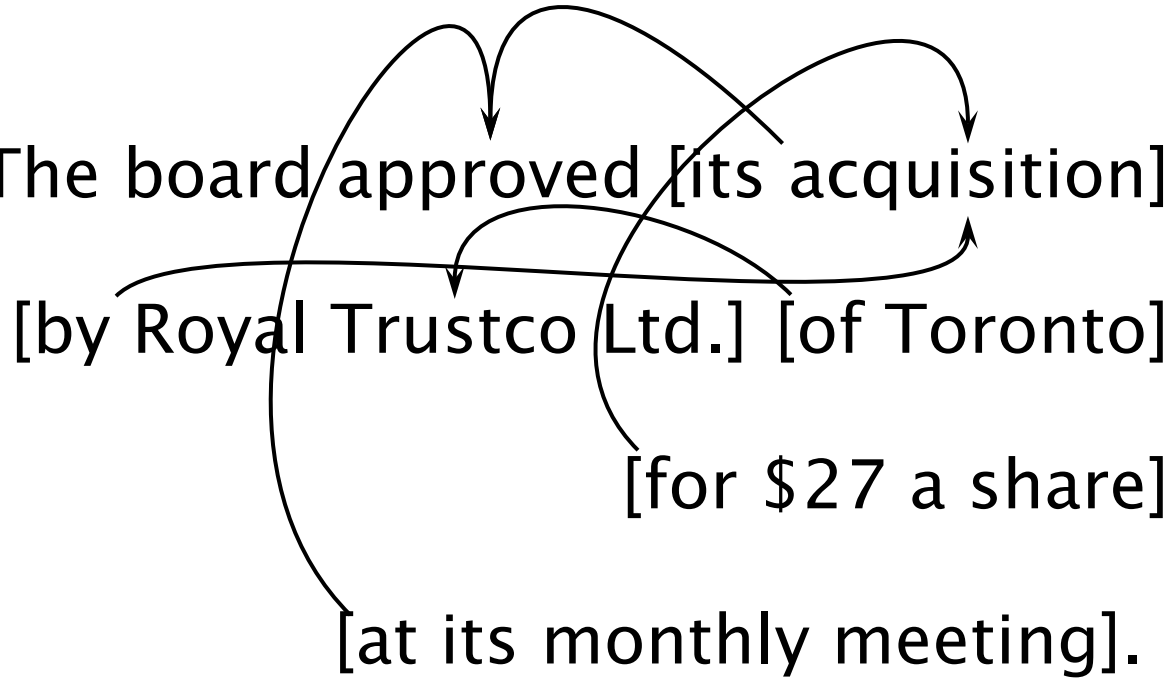
$$
\begin{aligned}
P(\mathrm{VA}_{into} = 1 | send) &= \frac{C(send, into)}{C(send)} \\
&= \frac{86}{1742.5} \approx 0.049 \\
P(\mathrm{NA}_{into} = 1 | soldiers) &= \frac{C(soldiers, into)}{C(soldiers)} \\
&= \frac{1}{1478} \approx 0.0007 \\
P(\mathrm{NA}_{into} = 0 | soldiers) &= 1 - P(\mathrm{NA}_{into} = 1 | soldiers) \\
&\approx 0.9993 \\
\lambda(send, soldiers, into) &\approx \log_2 \frac{0.049 \times 0.9993}{0.0007} \approx 6.13
\end{aligned}
$$

Attachment to verb is about 70 times more likely.

Overall accuracy is about 80% (forced choice); 91.7% correct at 55.2% recall ($\lambda = 3.0$).

Final remarks

- Ignores other conditioning factors (noun head in PP, superlative adjective)
- Just doing the simplest V NP PP case
- Gibson and Pearlmutter (1994) argue that overuse of this simple case has greatly biased psycholinguistic studies



Final remarks (2)

- There are other attachment cases: coordination, adverbial and participial phrases, noun compounds. Data sparseness is a bigger problem with many of these (more open class heads needed).
- In general, indeterminacy is quite common:

  We have not signed a settlement agreement with them.

  Either reading seems equally plausible.

Lexical acquisition

- Previous models give same estimate to all unseen events
- Unrealistic – could hope to refine that based on semantic classes of words
- E.g, although never seen, *eating pineapple* should be more likely than *eating holograms* because *pineapple* is similar to *apples*, and we have seen *eating apples*
- It's the same data. Why are classes useful?

An application: selectional preferences

- Verbs take arguments of certain types (usually! – remember metaphor)
- *Bill drove a . . .*
- *Mustang, car, truck, jeep, . . .*
- Resnik (1993) uses KL divergence for verb objects distributions
- Selectional preference strength: how strongly does a verb constrain direct objects
- *see* vs. *unknotted*
- Model via using head words only – a usually correct assumption
- Use a class-based model of nouns – for generalization. Resnik uses WordNet.

Selectional preference strength (how strongly does verb select?)

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

Selectional association between verb and 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:

$$A(v,n) = \max_{c \in \text{classes}(n)} A(v,c)$$

SPS example (made-up data)

Noun class c	$P(c)$	$P(c\|\textit{eat})$	$P(c\|\textit{see})$	$P(c\|\textit{find})$
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

$$A(\textit{eat}, \text{food}) = 1.08$$

$$A(\textit{find}, \text{action}) = -0.13$$

SPS example (Resnik, 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

A 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	$\frac{\lvert X \cap Y \rvert}{\sqrt{\lvert X \rvert \times \lvert Y \rvert}}$

Real-valued vector spaces

Vector dot product (how much do they have in common):

$$\vec{x} \cdot \vec{y} = \sum_{i=1}^{n} x_i y_i$$

0 if orthogonal – like matching coefficient, not normalized.

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:

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

Euclidean distance gives same ordering for normalized vectors:

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

Example: cosine as semantic similarity 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)	$\sum_i \|p_i - q_i\|$

Neighbors of *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

Evaluation

- Qualitative
- Task-based
  - Language models (Dagan, Pereira, and Lee)
  - Resnik
  - . . .