Applications (2 of 2):
Recognition, Transduction, Discrimination, Segmentation, Alignment, etc.

Kenneth Church

Kenneth.Church@jhu.edu

Solitaire → Multiplayer Games: Auctions (Ads)

http://www.scienceoftheweb.org/15-396/lectures/lecture09.pdf

A Single Auction → A Stream of Continuous Auctions

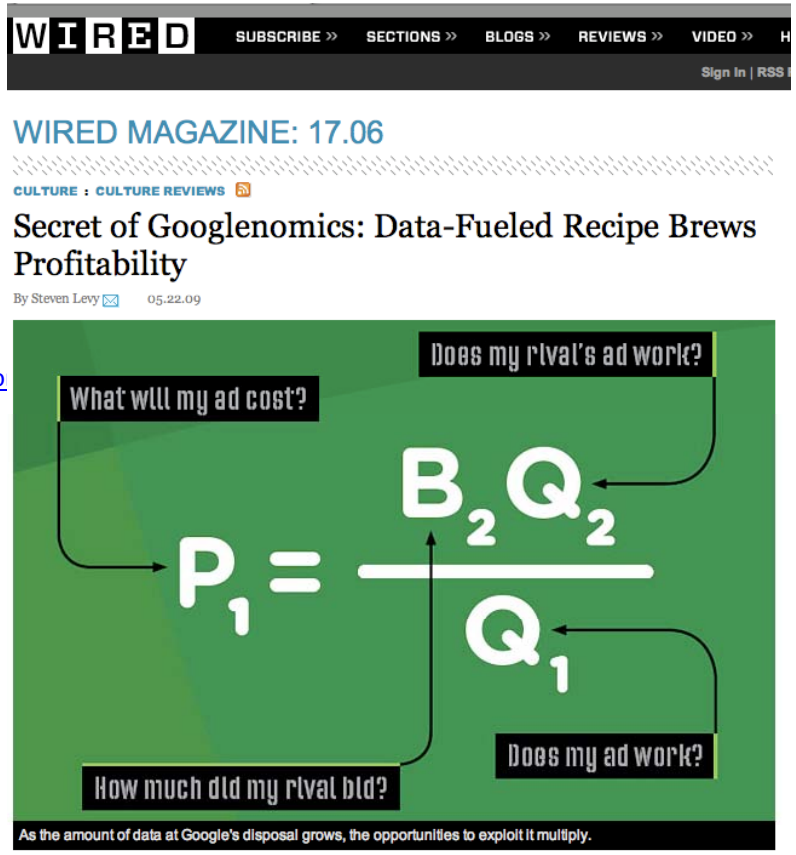
- Standard Example of Second Price Auction
  - Single Auction for a Single Apple
- Theoretical Result
  - Second Price Auction → Truth Telling
  - http://en.wikipedia.org/wiki/Vickrey_auction
  - Optimal Strategy:
    - Bid what the apple is worth to you
    - Don't worry about what it is worth to others
  - First Price Auction → ~~Truth Telling~~
- Does theory generalize to a continuous stream?

Pricing: Cost Per Click (CPC)

- B_i = your bid
- B_{i+1} = next bid
- CTR_i = your click through rate
- CTR_{i+1} = next click through rate
- CPC_i = your price
  - (if we show your ad and user clicks)
- Improvement: CTR → Q (Prior)
- Single Auction:
  - $CPC_i = B_{i+1}$
- Continuous Stream:
  - $CPC_i = B_{i+1} CTR_{i+1} / CTR_i$

- Equilibrium
  - Advertisers
    - Awareness
    - Sales
    - New Customers
    - ROI
  - Users
    - Minimize pain
    - Obtain Value
  - Market Maker
    - Maximize Revenue
- Truth Telling?

Multi-Player Games → Many Technical Opportunities



- Economics
  - http://www.wired.com/culture/culturereviews/magazine/17-06/nep_googleno
- Machine Learning
  - Learning to Rank
  - Estimate CTR (Q/Priors)
  - Sparse Data:
    - What is the CTR for a new ad?
  - Errors can be expensive
    - If CTR is too low for new ad → Penalize Growth
    - If too high → Reward Bad Guys to do Bad Things
- Truth Telling for Continuous Auctions?
  - Probably not, especially if participants can estimate Q better than market maker
- Machine Learning: Solitaire → Multi-Player Games
  - Can I estimate Q better than you can? Man-eating tiger

Applications

- Recognition: Shannon's Noisy Channel Model
  - Speech, Optical Character Recognition (OCR), Spelling
- Transduction
  - Part of Speech (POS) Tagging
  - Machine Translation (MT)
- Parsing: ???
- Ranking
  - Information Retrieval (IR)
  - Lexicography
- Discrimination:
  - Sentiment, Text Classification, Author Identification, Word Sense Disambiguation (WSD)
- Segmentation
  - Asian Morphology (Word Breaking), Text Tiling
- Alignment: Bilingual Corpora, Dotplots
- Compression
- Language Modeling: good for everything

Speech → Language

Shannon's: Noisy Channel Model

- *I* → *Noisy Channel* → *O*
- $I' \approx ARGMAX_I\ Pr(I|O) = ARGMAX_I\ Pr(I)\ Pr(O|I)$

Language Model

Channel Model

Trigram Language Model

Word	Rank	More likely alternatives
We	9	The This One Two A Three Please In
need	7	are will the would also do
to	1	
resolve	85	have know do…
all	9	The This One Two A Three Please In
of	2	The This One Two A Three Please In
the	1	
important	657	document question first…
issues	14	thing point to

Application Independent

Channel Model

Application	Input	Output
Speech Recognition	writer	rider
OCR (Optical Character Recognition)	all	a1l
Spelling Correction	government	goverment

Speech → Language
Using (Abusing) Shannon's Noisy Channel Model: Part of Speech Tagging and Machine Translation

- Speech
  - *Words → Noisy Channel → Acoustics*
- OCR
  - *Words → Noisy Channel → Optics*
- Spelling Correction
  - *Words → Noisy Channel → Typos*
- Part of Speech Tagging (POS):
  - *POS → Noisy Channel → Words*
- Machine Translation: "Made in America"
  - *English → Noisy Channel → French*

Didn't have the guts to use this slide at Eurospeech (Geneva)

$W_i \rightarrow Noisy\ Channel \rightarrow W_o$

Channel Model Depends on Application

Application	Input	Output
Speech Recognition	writer	rider
OCR	all	a11 (*A-one-L*)
	of	o{
	form	farm
Spelling Correction	government	goverment
	occurred	occured
	commercial	commerical
	similar	similiar

sub[X, Y] = Sub of X (incorrect) for Y (correct)

X	**Y (correct)**					
	a	b	c	d	e	f
a	0	0	7	2	342	1
b	1	0	9	9	3	3
c	7	6	0	16	1	9
d	2	10	13	0	12	1
e	388	0	4	11	0	3
f	0	15	1	4	2	0

Spelling Correction

echo absorbant adusted ambitios afte | spell | correct

absorbant	absorbent	
adusted	adjusted	100%
	dusted	0%
afte	after	100%
	fate	0%
	aft	0%
	ate	0%
	ante	0%
ambitios	ambitious	77%
	ambitions	23%
	ambition	0%

$$\underset{c}{ARGMAX}\ Pr(c)\ Pr(t|c)$$

$P(c)$ is a unigram model (no context for now)

$$Pr(t|c) \approx \begin{cases} del[c_{p-1}, c_p]/chars[c_{p-1}, c_p] & \text{if deletion} \\ add[c_{p-1}, t_p]/chars[c_{p-1}] & \text{if insertion} \\ sub[t_p, c_p]/chars[c_p] & \text{if substitution} \\ rev[c_p, c_{p+1}]/chars[c_p, c_{p+1}] & \text{if reversal} \end{cases}$$

Typo	Correction	Transformation			
acress	actress	@	t	2	deletion
acress	cress	a	#	0	insertion
acress	caress	ac	ca	0	reversal
acress	access	r	c	2	substitution
acress	across	e	o	3	substitution
acress	acres	s	#	4	insertion
acress	acres	s	#	5	insertion

c	%	Raw	freq(c)	Pr(t|c)		
actress	37%	.16	1343	55.	/	470,000
cress	0%	.00	0	46.	/	32,000,000
caress	0%	.00	4	.95	/	580,000
access	0%	.00	2280	.98	/	4,700,000
across	18%	.077	8436	93.	/	10,000,000
acres	21%	.092	2879	417.	/	13,000,000
acres	23%	.098	2879	205.	/	6,000,000

#	Freq	Typo	Corrections
0	3937	admininistration	
1	6993	absorbant	absorbent
2	1562	adusted	adjusted dusted
3	639	ambitios	ambitious, ambitions, ambition
4	367	compatability	compatibility, compactability, comparability, computability
5	221	afte	after, fate, aft, ate, ante
6	157	dialy	daily, diary, dials, dial, dimly, dilly
7	94	poice	police, price, voice, poise, pice, ponce, poire
8	82	piots	pilots, pivots, riots, plots, pits, pots, pints, pious
9	77	spash	splash, smash, slash, spasm, stash, swash, sash, pash, spas
10+	613		
Total	14,742		

2000 typos per month in AP

#	March	April	May	June	July	Aug	Sept	Total
0	720	604	542	606	492	465	508	3937
1	1120	997	1037	1007	958	944	930	6993
2	269	224	209	223	199	224	214	1562
3	109	92	89	101	79	87	82	639
4	58	57	62	45	43	59	43	367
5	54	41	20	26	28	24	28	221
6	35	22	19	19	22	17	23	157
7	20	11	13	7	11	15	17	94
8	19	14	14	5	7	7	16	82
9	15	11	6	11	10	8	16	77
10+	154	97	79	75	53	77	78	613
Total	2573	2170	2090	2125	1902	1927	1955	14,742

- lots of typos to train on
- 2000 / month (6% of lowercase word types)
  - Types vs. Tokens
  - Vocabulary Size (V) vs. Corpus Size (N)

Some typos are frequent

AP Freq (44 M words)	WSJ Freq (22 M words)	Typo	Correction
106	15	goverment	government
71	21	occured	occurred
61	6	responsiblity	responsibility
47	2	negotations	negotiations
45	8	benefitted	benefited
45	13	commerical	commercial
41	0	assocations	associations
39	26	televison	television
38	1	millenium	millennium
38	9	possiblity	possibility
34	3	accomodate	accommodate
32	16	similiar	similar

‘‘goverment’’ is more frequent than many words

AP Freq	Word	AP Freq	Word
99	extinct	93	standby
99	pellets	92	attends
98	remorse	92	condors
97	lighted	91	coaches
97	marital	88	averted

Evaluation

absurb, absorb, absurd

... financial community. ‘‘It is **absurb** and probably obscene for any person so engaged to ...

	Judge 1	**Judge 2**	**Judge 3**
choice 0 (spell error)	99	124	93
choice 1	188	176	167
choice 2	175	159	151
other	28	26	30
?	74	79	123
total	564	564	564

The Judges found the task harder than anticipated.

Performance

Method	**Discrimination**	**%**
correct	286/329	87 ± 1.9
Judge 1	271/273	99 ± 0.5
Judge 2	271/275	99 ± 0.7
Judge 3	271/281	96 ± 1.1
channel-only	263/329	80 ± 2.2
prior-only	247/329	75 ± 2.4
chance	172/329	52 ± 2.8

The Task is Hard without Context

Typo	Choice 1	Choice 2
actuall	actual	actually
constuming	consuming	costuming
conviced	convicted	convinced
confusin	confusing	confusion
workern	worker	workers

Easier with Context

- actuall, actual, actually
  - … in determining whether the defendant actually will die.
- constuming, consuming, costuming
- conviced, convicted, convinced
- confusin, confusing, confusion
- workern, worker, workers

actuall, actual, actually

... in determining whether the defendant **actuall** will die. In the 1985 decision, the ...

Easier with Context

constuming, consuming, costuming

... on Friday night, a show as lavish in **constuming** and lighting as those the late Liberace used to ...

conviced, convicted, convinced

... of the area. ‘‘When we’re **conviced** and the Peruvians are convinced (the base camp) ...

confusin, confusing, confusion

... The political situation grew more **confusin** today, with an official media report indicating ...

workern, worker, workers

... for the attacks. The **workern**, who was unloading a car at a job site in a ...

Context Model

- Bigram model of context
- Dynamic programming isn't necessary

$$Pr(l,r,t|c)\ Pr(c) \approx Pr(l|c)\ Pr(r|c)\ Pr(t|c)\ Pr(c)$$

- All four factors should be independent (if properly estimated)

E/E: A Poor Estimate of Context

$$Pr(l|c) = \frac{Pr(lc)}{Pr(c)}$$

$$\approx \frac{(freq(lc) + 0.5)/d_1}{(freq(c) + 0.5)/d_2}$$

$$\propto \frac{freq(lc) + 0.5}{freq(c) + 0.5}$$

A poor estimate of context is worse than none

	chance	E/E
wrong	164.5	169
uninformative	0	4
right	164.5	156

Five Methods of Estimating Context

$$Pr(l|c) = \frac{Pr(lc)}{P(c)} \approx \frac{freq(lc) + 0.5}{freq(c) + 0.5} \quad \text{E/E}$$

$$Pr(l|c) = \frac{Pr(lc)}{P(c)} \approx \frac{freq(lc)}{freq(c) + 0.5} \quad \text{M/E}$$

$$Pr(l|c) \approx \frac{freq(lc) + 0.5}{freq(c) + V/2} \quad \text{E}$$

$$Pr(l|c) \approx \frac{freq(lc) + 0.5\sqrt{freq(c)}^{-}}{freq(c) + 0.5V\sqrt{freq(c)}^{-}} \quad \text{MM}$$

$$Pr(l|c) \approx \frac{(r+1)\,\frac{N_{r+1}}{N_r}}{freq(c) + 0.5} \quad \text{G/E}$$

Better Estimates of Context Exist

	E	MM	G/E
wrong	62	59	45
uninformative	0	0	4
right	267	270	280

Context is Useless Unless Carefully Measured

		disastrous		useless		useful
	no context	+M/E context	+E/E context	+E context	+MM context	+G/E context
wrong	43	11	61	39	40	34
useless	0	136	0	0	0	0
right	286	182	268	290	289	295
%	86.9%	55.3%	81.5%	88.1%	87.8%	89.7%
± σ	1.9%	2.7%	2.1%	1.8%	1.8%	1.7%

Each Factor Helps

	Model	%
1	channel	80
1	prior	76
1	left	78
1	right	77
2	channel + prior	87
2	channel + left	87
2	channel + right	88
2	prior + left	83
2	prior + right	80
2	left + right	86
3	channel + prior + left	90
3	channel + prior + right	88
3	channel + left + right	90
3	prior + left + right	86
4	channel + prior + left + right	90
	Judge 1	99
	Judge 2	99
	Judge 3	96

Future Improvements

- Add More Factors
  - Trigrams
  - Thesaurus Relations
  - Morphology
  - Syntactic Agreement
  - Parts of Speech
- Improve Combination Rules
  - Shrink (Meaty Methodology)

Shrinks (Robustness Statistics)

- Standard Example: Baseball batting averages
- For one player: $batting\ \ average = \dfrac{hits}{at\ \ bats}$
- MLE (maximum likelihood estimate): optimal for one average, but not for many.
- Problem: imagine a rookie goes to the plate for first time and gets a hit. Is he the best player there ever was?
- A standard fix: shrink the individual player's average, x, toward the team's average, $\bar{x}$:
  - $\hat{x} = (1-\alpha)\,x + \alpha \cdot \bar{x}$
  - Shrinking, α, depends on lack of belief, σ^2.
    - More shrinking for rookies (small counts),
    - less shrinking for seasoned players (large counts)
- Lots of other shrinking formulas such as: $\hat{x} = \alpha \cdot x^{\beta}$, where $0 < \beta < 1$ (β increases with belief/robustness).
- Trade-off Random Error (variance) for Bias (mean)

Conclusion (Spelling Correction)

- There has been a lot of interest in smoothing
  - Good-Turing estimation
  - Knesser-Ney
- Is it worth the trouble?
- Ans: Yes (at least for recognition applications)

Transition: First Speech, then Language

- Many of the very same methods are being applied to problems in natural language processing by many of the very same researchers.
- Noisy Channel Model: $I \rightarrow Noisy\ Channel \rightarrow O$
- Recognition: Speech, (OCR), Spelling Correction
- Training is better than Guessing
  - Language Modeling: ngrams
    - Nobody likes them, but hard to beat.
  - Channel Modeling: confusion matrices
- Smoothing (meaty methodology): important, but poor estimates of context can be worse than none.
- More apps
  - Transduction: part of speech tagging, MT
  - Ranking: Information Retrieval, Lexicography
  - Discrimination: Word Sense Disambiguation

Recasting Part-of-Speech Tagging as a Noisy Channel Problem

- The empirical approach has been adopted by almost all contemporary part-of-speech programs: Bahl and Mercer (1976), Leech *et al*. (1983), Jelinek (1985), Deroualt and Merialdo (1986), Garside *et al* (1987), Church (1988), DeRose (1988), Hindle (1989), Kupiec (1989, 1992), Ayuso *et al*. (1990), deMarcken (1990), Karlsson (1990), Boggess *et al*. (1991), Merialdo (1991), Voutilainen *et al*. (1992).
- Part of Speech Tagging Task
  - Input (seq of words): *The chair will table the motion*
  - Output (seq of tags): *art noun modal verb art noun*
  - [A/AT former/AP top/NN aide/NN] to/IN [Attorney/NP/NP General/NP/NP Edwin/NP/NP Meese/NP/NP] interceded/VBD to/TO extend/VB [an/AT aircraft/NN company/NN 's/$ government/NN contract/NN] ,/, then/RB went/VBD into/IN [business/NN] with/IN [a/AT lobbyist/NN] [who/WPS] worked/VBD for/IN [the/AT defense/NN contractor/NN] ,/, according/IN to/IN [a/AT published/VBN report/NN] ./.

- Performance:
  - Accuracy: ~95% correct by word on unrestricted text
  - Modest time & space:
    - linear time, constant space, reasonable constants
  - Massive citations, but few convincing applications
- Imagine that a sequence of parts of speech, P, is presented at the input to the channel and for some crazy reason, it appears at the output of the channel in a corrupted form as a sequence of words, W.
  - Our job is to determine P given W.
- $P \rightarrow Noisy\ Channel \rightarrow W$
- $\hat{P} = \underset{P}{ARGMAX}\ Pr(P)\ Pr(W|P)$
- Parameters of this model (dictionary + grammar):
  1. Lexical probabilities, $Pr(W_i|P_i)$, and
  2. Contextual probabilities, $Pr(P_i|P_{i-2}P_{i-1})$

Is 95% good enough?

- On the one hand, it is better than we have been doing before n-gram part of speech taggers came into fashion,
- but on the other hand, it still means that a large fraction of sentences will contain at least one fatal error.
- If subsequent processing (e.g., parsing, semantic analysis) require perfect part of speech analysis, then 95% performance is clearly not nearly good enough, and probably 99% isn't either.
- Perhaps we need to modify these subsequent steps so they can tolerate an error rate of 1-5%. Alternatively, we may need to aim for somewhat higher levels of tagging performance than we can currently achieve.

How Hard is the Problem?

- 95% might sound good,
- but really dumb methods do almost as well.
- If we simply ignore the context, and just select the most likely part of speech given the word, we will achieve nearly 90% correct.
- (Some methods manage to fall below this baseline by focusing on the grammar rather than the lexicon.)
- 95% may not sound so good when we realize that the lexicon gives you the first 90%, and context contributes only about half of the remaining 10%.

Intuition

- Many people who have not worked in computational linguistics have a strong intuition that lexical ambiguity is usually not much of a problem.

- It is commonly believed that most words have just one part of speech, and that the few exceptions such as ‘‘table’’ are easily disambiguated by context in most cases.

- This intuition is largely supported by the numbers just cited.

- That is, most cases can be resolved without context (e.g., 90%), and that simple n-gram models of context are sufficient for more than half of the remainder.

Why Traditional Methods Failed

- Traditional grammar-based methods ignore lexical prefs,
- Which are important
  - Lexical prefs (without grammar/ngrams): ~90% correct
  - Grammar/ngrams (without lexical prefs): much worse
- Trivial Example: *I see a bird*.
- Easy for stat methods because desired tags have huge lexical probs:

Lexical Probabilities
(based on Brown Corpus)

Pr(PPSS | "I")	5837/5838
Pr(VB | "see")	771/772
Pr(AT | "a")	23013/23019
Pr(NN | "bird")	26/26

- However, if we ignore freqs (as most parsers do), then...

Lexical Possibilities (based on Websters)

Word	**Parts of Speech** (common)	(rare)
I	pronoun	noun (letter of the alphabet)
see	verb	noun (e.g., *the Holy See*)
a	article	noun (letter of the alphabet)
bird	noun	verb (used by bird watchers)

- Dictionaries focus on (unlikely) possibilities

The Non-deterministic Non-Solution

- Traditional parsers try all possibilities and hope the bad ones are ungrammatical.
- (punt and return all possibilities)
- One might hope the bad tags in the trival example could be ruled out by the parser as syntactically ill-formed.

- But *no*.

- If the parser is going to accept noun phrases of the form:
  - [NP [N city] [N school] [N committee] [N meeting]]
- then it can't rule out (among others)
  - [NP [N I] [N see] [N a] [N bird]]
- The ''bad'' part of speech assignments aren't impossible;
  - they are just (extremely) improbable.

The Proposed Method

- Conceptually, enumerate all assignments
- Score each path (product of lexical and contextual probabilities)
- Select best
- Suppose *I*, *see* and *a* are each two ways ambiguous. Then there are 8 paths:

	I	see	a	bird
1.	PPSS	VB	AT	NN
2.	PPSS	VB	IN	NN
3.	PPSS	UH	AT	NN
4.	PPSS	UH	IN	NN
5.	NP	VB	AT	NN
6.	NP	VB	IN	NN
7.	NP	UH	AT	NN
8.	NP	UH	IN	NN

	.	.	I	see	a	bird	.	.	
A1	.	.	PPSS	VB	AT	NN	.	.	
context	0.99	0.20	0.07	0.07	0.23	0.25	1.00	1.00	e-4
lex	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
A2	.	.	PPSS	VB	IN	NN	.	.	
context	0.99	0.20	0.08	0.03	0.13	0.25	1.00	1.00	e-9
lex	1.00	1.00	1.00	1.00	e-4	1.00	1.00	1.00	
A3	.	.	PPSS	UH	AT	NN	.	.	
context	0.99	1.00	0.00	0.00	0.23	0.25	1.00	1.00	0
lex	1.00	1.00	1.00	e-3	1.00	1.00	1.00	1.00	
A4	.	.	PPSS	UH	IN	NN	.	.	
context	0.99	1.00	0.00	0.00	0.13	0.25	1.00	1.00	0
lex	1.00	1.00	1.00	e-3	e-4	1.00	1.00	1.00	
A5	.	.	NP	VB	AT	NN	.	.	
context	0.97	0.03	0.01	0.07	0.23	0.25	1.00	1.00	e-10
lex	1.00	1.00	e-4	1.00	1.00	1.00	1.00	1.00	
A6	.	.	NP	VB	IN	NN	.	.	
context	0.97	0.03	0.01	0.03	0.13	0.25	1.00	1.00	e-15
lex	1.00	1.00	e-4	1.00	e-4	1.00	1.00	1.00	
A7	.	.	NP	UH	AT	NN	.	.	
context	0.97	0.00	0.00	0.00	0.23	0.25	1.00	1.00	0
lex	1.00	1.00	e-4	e-3	1.00	1.00	1.00	1.00	
A8	.	.	NP	UH	IN	NN	.	.	
context	0.97	0.00	0.00	0.00	0.13	0.25	1.00	1.00	0
lex	1.00	1.00	e-4	e-3	e-4	1.00	1.00	1.00	

Dynamic Programming/Viterbi Search (Meaty Methodology)

- Conceptually, there could be k^n part of speech sequences, where n is the length of the input sentence, and k is the (worst case) lexical ambiguity.
- Fortunately, there is a linear time dynamic programming solution.
- If two paths are the same within the ngram window of 3 words, then keep the just better one.
- This way, there will be at most nk^3 paths to consider.
- k is small

Smoothing Issues (Meaty Methodology)

- Must do something with Zeros
- Zipf's Law: there will always be a large tail of low frequency words
- 40,000 words (80%) in the Brown Corpus have freq < 5
- If ''yawn'' appears once as a noun and once as a verb, what is the probability that it could be an adjective?

Lesson from speech recognition research

- Statistical methods are often helpful when:
  - data rates are high,
  - there is plenty of training material, and
  - nothing else seems to work very well
    - because we don't know what we're doing.
- Probability vs Possibility
- Computational linguistics doesn't like to use word frequencies, but any psycholinguist knows that they they swamp out syntactic factors
- Breadth vs Depth

Problems

- Flying Planes and friends
  [Time/NN] flies/VBZ like/CS [an/AT arrow/NN] ./.
  [Fruit/NN] flies/VBZ like/CS [a/AT banana/NN] ./.

  [Flying/VBG planes/NNS] can/MD be/BE dangerous/JJ ./.
  [They/PPSS] are/BER flying/VBG [planes/NNS] ./.

- Inadequate window size
  [The/AT horse/NN] has/HVZ slipped/VBN ./.
  [The/AT horse/NN] has/HVZ raced/VBN past/IN [the/AT barn/NN] and/CC slipped/VBD ./.

- Unknown words
  Do/DO [you/PPSS] know/VB [what/WDT] [a/AT xxx/NN] is/BEZ ?/.
  [I/PPSS] know/VB [care/NN] if/CS [you/PPSS] xxx/VB !/.
  [I/PPSS] need/MD xxx/VB ./.

- Lack of word association norms, semantics, pragramatics
  [I/PPSS] like/VB to/TO work/VB ./.
  [I/PPSS] went/VBD to/TO work/VB ./.
  [I/PPSS] went/VBD to/IN [school/NN] ./.

Conclusions: First Speech, Then Language

- Noisy Channel Model: $I \rightarrow Noisy\ Channel \rightarrow O$
- Recognition: Speech, (OCR), Spelling Correction
- Transduction: part of speech tagging, MT
  - $P \rightarrow Noisy\ Channel \rightarrow W$
  - Imagine that a sequence of parts of speech, P, is presented at the input to the channel and for some crazy reason, it appears at the output of the channel in a corrupted form as a sequence of words, W.
  - Task: given ''corrupted'' output (words)
    - Recover ''clean'' input (parts of speech).
  - Machine Translation (MT): even crazier
  - $E \rightarrow Noisy\ Channel \rightarrow F$
  - Task: given ''corrupted'' output (French)
    - Recover ''clean'' input (English).
  - Controversial for MT, but not for lexicography
- More apps: ranking, discrimination

Historical Note (Lots of Citations)

- Early example: stats → performance
- Controversy: Stats better than traditional methods?
  - Many alternatives soon caught up.
- Practical applications: hope, but...
- Field needed a success (AI Winter)
- Great term project! (Meaty methodology)

☞ But not a lot of exciting recent literature...

- Hard to improve performance
- Upper bound: machines as good as people
  - Mindless Metrics (standard eval)
    - Two people disagree → difference of opinion
    - Machine disagrees → machine is wrong
- Recommendation: progress is limited by eval
  - Fix eval: distinguish man from machine.

→ More exciting literature

Transition: Bounds

- Is 95% good enough? (Engineering considerations)
  - How good are people? (Turing Test)
- How hard is the problem?
- Upper and lower bounds
- Lower bound: performance of a dumb method
- Upper bound: human performance
  - Shannon's method of estimating the entropy of English
  - Ask human subjects to guess the next letter.
- Apply these arguments to another application
  - Word Sense Disambiguation
- Fix eval: distinguish man from machine.

Estimating Upper and Lower Bounds on the Performance of Word-Sense Disambiguation Programs

William Gale
Kenneth Ward Church
David Yarowsky

AT&T Bell Laboratories
600 Mountain Ave.
Murray Hill, NJ 07974
kwc@research.att.com

- Two new word-sense disambiguation systems:
  1. Trained on bilingual text (the Canadian Hansards), and
  2. Trained on monolingual text (Roget's & Grolier's).
- Need a credible evaluation methodology

Bounds Estimates

- Lower bound: 75% (averaged over ambiguous types)
  - Straw-man: ignore context
    - assume *crane* is always *animal*, never *machine*
    - assume *sentence* is always *syntax*, never *punishment*
- Upper bound: 96.8%
  - Limited by ability to obtain reliable judgments from human informants.
  - Depends on task.
  - Jorgensen used a difficult classification task, and found only 68% agreement among judges.
  - 68% is unusable → upper bound < lower bound
  - We have developed a much easier discrimination task that produces more usable results: 96.8%

Word-Sense Disambiguation: Lots of Background

- Lots of History: Kaplan (1950), Yngve (1955), Bar-Hillel (1960), Masterson (1967)

- Lots of Recent Work: Black (1988), Brown *et al.* (1991), Choueka and Lusignan (1985), Clear (1989), Dagan *et al.* (1991), Gale *et al.* (to appear), Hearst (1991), Lesk (1986), Smadja and McKeown (1990), Walker (1987), Veronis and Ide (1990), Yarowsky (1992), Zernik (1990, 1991).

- Lots of Applications: text-to-speech (TTS), machine translation (MT), information retrieval (IR), etc.

- Lots of Potential: might soon have sense-taggers that work as well as current part-of-speech taggers.

Knowledge Acquisition Bottleneck

- Previous studies have been stymied by a lack of data.
- As a result, AI-approaches have tended to focus on ‘‘toy’’ domains, because they couldn’t get enough data (knowledge) to cover a real domain.
  - ‘‘The expert for THROW is currently six pages long... but it should be 10 times that size.’’
    – Small and Reiger (1982)
  - ‘‘The number of facts we human beings know is, in a certain very pregnant sense, infinite.’’
    – Bar-Hillel (1960)
- Similarly, statistical approaches, e.g., Kelly and Stone (1975), have had to depend on relatively small amounts of hand-labeled text for testing and training, because such testing and training material is fairly hard to come by.

Parallel-Text: An Alternative Source of Testing and Training Materials

- Following Brown *et al*. and Dagan *et al*., we have achieved considerable progress recently by taking advantage of a new source of testing and training materials.

- Rather than depending on small amounts of hand-labeled text, we have been making use of relatively large amounts of parallel text (e.g., Hansards).

- The translation can often be used in lieu of hand-labeling:

  1. *sentence* → *peine* (''judicial'' sense)
  2. *sentence* → *phrase* (''syntactic'' sense)

- In this way, we have been able to acquire a considerable amount of testing and training material for developing and testing our disambiguation algorithms.

Outline of Algorithm (Bilingual Method)

- Sentence Alignment
- Word Correspondence
- Train Context Models: $Pr(token \mid sense)$
- Test on New Data

$$score(d) = \prod_{token \text{ in } d} \frac{Pr(token \mid rel)}{Pr(token \mid irrel)}$$ IR

$$score(d) = \prod_{token \text{ in } d} \frac{Pr(token \mid author_1)}{Pr(token \mid author_2)}$$ Author

$$score(c) = \prod_{token \text{ in } c} \frac{Pr(token \mid sense_1)}{Pr(token \mid sense_2)}$$ Sense

Sentence Alignment

English	French
According to our survey, 1988 sales of mineral water and soft drinks were much higher than in 1987, reflecting the growing popularity of these products. Cola drink manufacturers in particular achieved above-average growth rates.	Quant aux eaux minérales et aux limonades, elles rencontrent toujours plus d'adeptes. En effet, notre sondage fait ressortir des ventes nettement supérieures à celles de 1987, pour les boissons à base de cola notamment.
The higher turnover was largely due to an increase in the sales volume.	La progression des chiffres d'affaires résulte en grande partie de l'accroissement du volume des ventes.
Employment and investment levels also climbed.	L'emploi et les investissements ont également augmenté.
Following a two-year transitional period, the new Foodstuffs Ordinance for Mineral Water came into effect on April 1, 1988. Specifically, it contains more stringent requirements regarding quality consistency and purity guarantees.	La nouvelle ordonnance fédérale sur les denrées alimentaires concernant entre autres les eaux minérales, entrée en vigueur le 1er avril 1988 après une période transitoire de deux ans, exige surtout une plus grande constance dans la qualité et une garantie de la pureté.

Aligning Words

- *English*: We took the initiative in assessing and amending current legislation and policies to ensure that they reflect a broad interpretation of the charter.

- *French*: Nous avons pris l'initiative d'évaluer et de modifier des lois et des politiques en vigueur afin qu'elles correspondent à une interprétation généreuse de la charte.

We took the initiative in assessing and amending

pris *initiative* *evaluer* *modifier*

Word	Sense	Contextual Clues
sentence	peine	inmate, parole, serving, a, released, prison, mandatory, judge, after, years, who, death, his, murder
sentence	phrase	I, read, second, amended, ‘‘, ’’, protects, version, just, letter, quote, word, ..., last, amendment, insults, assures, quotation, first
drug	medicaments	prices, prescription, patent, increase, generic, companies, upon, consumers, higher, price, consumer, multinational, pharmaceutical, costs
drug	drogues	abuse, paraphernalia, illicit, use, trafficking, problem, food, sale, alcohol, shops, crime, cocaine, epidemic, national, narcotic, strategy, head, control, marijuana, welfare, illegal, traffickers, controlled, fight, dogs

Context

- Most researchers have focused on small contexts ± 5 words
- Because people don't need any more
- But, we use ± 50 words
- because we find that the larger contexts are useful,
- and the machine needs all the help it can get.

Results (Bilingual Method)

Word	Sense 1	Sense 2	%
sentence	judicial	syntactic	98%
duty	tax	obligation	91%
drug	medical	illicit	91%
language	medium	style	91%
land	property	country	87%
position	location	job	84%
average			90%

Problems with Training on Parallel Text

1. Monolingual ≠ Bilingual
   - *interest* → *interêt*
   - Complex for monolingual purposes, but not for bilingual purposes.
   - *wear* (English → Japanese)
   - Complex for bilingual purposes, but not for monolingual purposes.
2. Availability
   - Very few sources: Canadian Hansards, ISSCO,...
   - Hansards are not ‘‘balanced’’

Monolingual Version

- Replace *sense* with *Roget Category*
  - Fewer Parameters: 1042 << V
  - And therefore, easier to obtain robust estimates.
  - Also, can be trained on untagged material
- Testing:

$$\prod_{w \text{ in context}} Pr(w \mid Roget\ Category_i)$$

- Training (on untagged material):
  - Make a set of all words that are listed in Roget's under $Roget\ Category_i$
  - Use all the context of all instances in Grolier's of all of these words (appropriately weighted) as evidence for $Roget\ Category_i$
- See Yarowsky's 1992 Coling paper for details.

Tagging Unrestricted Text

	Input	Output
Treadmills attached to	*cranes* were used to lift heavy obje	TOOLS
and for supplying power for	*cranes* , hoists , and lifts .SB The	TOOLS
Above this height , a tower	*crane* is often used .SB This compri	TOOLS
elaborate courtship rituals	*cranes* build a nest of vegetation o	ANIMAL
are more closely related to	*cranes* and rails .SB They range in	ANIMAL
low trees .PP At least five	*crane* species are in danger of exti	ANIMAL

Tagging Dictionary Definitions

Dict	Sense	Input	Output
COBUILD	1.1	a machine with a long movable arm that...	TOOLS
COBUILD	1.2	large bird with a long neck and long...	ANIMAL
CED1	1	any large long-necked long-legged wading...	ANIMAL
CED1	2	any similar bird , such as a heron .	ANIMAL
CED1	3	a device for lifting and moving heavy ob...	TOOLS
CED1	4	a large trolley carrying a boom, on the...	TOOLS

Results (Monolingual Method)

Word		Previous
bow	91%	< 67% (Clear, 1989)
bass	99%	100% (Hearst, 1991)
galley	99%	50-70% (Lesk, 1986)
mole	99%	N/A (Hirst, 1987)
sentence	98%	90% (Gale *et al.*)
slug	97%	N/A (Hirst, 1987)
star	96%	N/A (Hirst, 1987)
duty	96%	96% (Gale *et al.*)
issue	94%	< 70% (Zernik, 1990)
taste	93%	< 65% (Clear, 1989)
cone	77%	50-70% (Lesk, 1986)
interest	72%	72% (Black, 1988); 70% (Zernik, 1990)
Average	92%	N/A

Problems with this kind of Evaluation

- Words were sampled over literature, not over vocabulary.
- Therefore, experiment is more appropriate for predicting performance over systems, not over new inputs.
- Moreover, one feels uncomfortable comparing results across experiments since there many potentially important differences including:
  - test and training materials,
  - judges,
  - genre,
  - and many more.

General Bounds

- What is the State-of-the-Art?
  - No clear consensus
  - Zernik suggests that *interest* is relatively easy; we believe that it is relatively hard.
- What level of performance would be adequate?
  - Bar-Hillel (1960) left the field when he couldn't see how to beat 75%, which didn't seem to be good enough.
  - Are we there yet?

Lower Bound

- Straw-man: ignore context
  - assume *crane* is always *animal*, never *machine*
  - assume *sentence* is always *syntax*, never *punishment*
- Hopefully, any reasonable system will beat this baseline...

Word	Baseline	
issue	96%	94%
duty	87%	96%
galley	83%	99%
star	83%	96%
taste	74%	93%
bass	70%	99%
slug	62%	97%
sentence	62%	98%
interest	60%	72%
mole	59%	99%
cone	51%	77%
bow	48%	91%
Average	70%	92%

- These words are harder than average.
- They are representative of word-sense literature, not of vocabulary.

More Representative Sample

Baseline Performance

	Tokens	Types
All 97 Words	93%	92%
30 Ambig Words	81%	75%

Word	S	F	B	Word	S	F	B	Word	S	F	B	Word	S	F	B
acid	1	937	100%	gold	1	391	100%	pottery	1	175	100%	deposit	2	570	88%
annexation	1	7	100%	interface	1	6	100%	projector	1	22	100%	hour	4	181	87%
benzene	1	50	100%	interruption	1	6	100%	regiment	1	13	100%	path	2	84	86%
berry	1	37	100%	intrigue	1	3	100%	relaxation	1	3	100%	view	3	359	86%
capacity	1	168	100%	journey	1	19	100%	reunification	1	12	100%	pyramid	3	119	82%
cereal	1	64	100%	knife	1	52	100%	shore	1	73	100%	antenna	2	171	81%
clock	1	99	100%	label	1	12	100%	sodium	1	319	100%	trough	3	26	77%
coke	1	54	100%	landscape	1	381	100%	specialty	1	39	100%	tyranny	2	12	75%
colon	1	35	100%	laurel	1	26	100%	stretch	1	6	100%	figure	6	594	73%
commander	1	206	100%	lb	1	276	100%	summer	1	328	100%	institution	4	559	71%
consort	1	12	100%	liberty	1	113	100%	testing	1	71	100%	crown	4	87	64%
contract	1	216	100%	lily	1	30	100%	tungsten	1	35	100%	drum	2	124	63%
cruise	1	21	100%	locomotion	1	12	100%	universe	1	360	100%	pipe	4	189	60%
cultivation	1	88	100%	lynx	1	8	100%	variant	1	14	100%	processing	2	125	59%
delegate	1	21	100%	marine	1	316	100%	vigor	1	3	100%	coverage	2	19	58%
designation	1	3	100%	memorial	1	14	100%	wire	1	140	100%	execution	2	7	57%
dialogue	1	67	100%	menstruation	1	14	100%	worship	1	86	100%	min	2	28	57%
disaster	1	31	100%	miracle	1	13	100%	virus	2	410	98%	interior	4	236	56%
equation	1	327	100%	monasticism	1	21	100%	device	3	507	97%	campaign	2	306	51%
esophagus	1	18	100%	mountain	1	1129	100%	direction	2	347	96%	output	2	188	51%
fact	1	200	100%	nitrate	1	46	100%	reader	2	75	96%	gin	3	42	50%
fear	1	37	100%	orthodoxy	1	4	100%	core	3	188	94%	drive	3	72	49%
fertility	1	51	100%	pest	1	44	100%	hull	2	48	94%				
flesh	1	14	100%	planning	1	86	100%	right	5	1014	94%				
fox	1	58	100%	possibility	1	27	100%	proposition	2	38	89%				

Upper Bound

- Limited by ability to obtain reliable judgments from human informants.
- Depends on task.
- Jorgensen used a difficult classification task, and found only 68% agreement among judges.
- 68% is unusable → upper bound < lower bound
- 68% is also below Bar-Hillel's min of 75%
- We have developed a much easier discrimination task that produces more usable results: 96.8%
- Would rather not change task like this,
  - but seems necessary to do so.

A Discrimination Experiment

Experiment originally designed to test
One-Sense-Per-Discourse Hypothesis

> antenna
> 1. jointed organ found in pairs on the heads of insects and crustaceans, used for feeling, etc. → the illus at insect.
> 2. radio or TV aerial.

lack eyes, legs, wings, **antennae**, and distinct mouthparts and
The Brachycera have short **antennae** and include the more evolved

silk moths passes over the **antennae**. Only males that detect
relatively simple form of **antenna** is the dipole, or doublet

96.8% Agreement

Judge	n	%
1	82	100.0%
2	72	87.8%
3	81	98.7%
4	82	100.0%
5	80	97.6%
Average		96.8%
Average (without Judge 2)		99.1%

Conclusions

- Two new word-sense disambiguation systems:
  1. Trained on bilingual text (the Canadian Hansards), and
  2. Trained on monolingual text (Roget's & Grolier's).
- Needed a credible evaluation paradigm
- Lower Bound (75%): performance of baseline system
- Upper Bound (96.8%): agreement among judges
- Similar bounds arguments have been used in part-of-speech tagging (90-95, incl ambig)
- Bounds arguments should be more robust to minor variations in test materials, phase of the moon, etc.

Applications: Foil for Discussing Techniques (Meaty Methodology)

- Recognition:
  - Speech, Optical Character Recognition (OCR), Spelling Correction
- Transduction:
  - Part of Speech Tagging, Machine Translation (MT)
- Parsing: ???
- Ranking:
  - Lexicography, Information Retrieval (IR)
- Discrimination:
  - Text Classification, Author Identification, Word Sense Disambiguation
- Segmentation: Asian Morphology, Text Tiling
- Alignment: Bilingual Corpora, Dotplots
- Compression
- Language Modeling: good for everything