Applications of Text Analysis

Kenneth Ward Church
AT&T Bell Laboratories
Murray Hill, NJ. 07974
kwc@research.att.com

- Speech Recognition
- Speech Synthesis
- OCR

- Machine Translation (MT)
  - Interpretation: speech $\rightarrow$ speech
  - Translation: writing $\rightarrow$ writing

- Spelling Correction & grammar checking
- Part-of-Speech Tagging

- Word-Sense Disambiguation
- Information Retrieval
- Author Identification

- Text Compression

Shannon's Noisy Channel Model

- Information Theory
- Developed at Bell Labs in the 1940s
- Recognition Applications (Speech, OCR, Spelling)

$$I \rightarrow Noisy\ Channel \rightarrow O$$

$$\hat{I} = \underset{I}{ARGMAX}\ Pr(I \mid O) = \underset{I}{ARGMAX}\ Pr(I)\ Pr(O \mid I)$$

- Language Model: $Pr(I)$
- Channel Model: $Pr(O|I)$

Language Model: $Pr(I)$ **(Trigram Approximation)**

- Collect a sample of text
- Count freq of trigrams & bigrams
- $Pr(I) \approx \prod_{i=i}^{N} Pr(I_i | I_{i-2} I_{i-1})$
- $Pr(I_i | I_{i-2}\ I_{i-1}) \approx \dfrac{freq(I_{i-2}\ I_{i-1}\ I_i)}{freq(I_{i-2}\ I_{i-1})}$

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 these problems ...
of	2	the
the	1	
important	657	document question first ...
issues	14	thing point to ...

Channel Model: $Pr(O|I)$

- $Pr(0|I)$ is high if O is ‘‘similar’’ to I

- where def of ‘‘similar’’ depends on application.

Examples of Channel Confusions
in Different Applications

Application	Input	Output
Speech Recognition	writer	rider
	here	hear
Optical Character Recognition	all	a1l (*A*-*one*-*L*)
	of	o{
	form	farm
Spelling Correction	government	goverment
	occurred	occured
	commercial	commerical

Spelling Correction

- Analogous to speech recognition,
- but easier to study

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%

Some typos are frequent

AP Freq (per Mwords)	**WSJ Freq** (per Mwords)	**Typo**	**Correction**
2.4	0.7	goverment	government
1.6	1.0	occured	occurred
1.4	0.3	responsiblity	responsibility
1.1	0.1	negotations	negotiations
1.0	0.4	benefitted	benefited
1.0	0.6	commerical	commercial
0.9	0.0	assocations	associations
0.9	1.2	televison	television
0.9	0.0	millenium	millennium
0.9	0.4	possiblity	possibility
0.8	0.1	accomodate	accommodate
0.7	0.7	similiar	similar
13.6	5.5		

The typo *goverment* is more common than many correctly spelled words: *extinct, pellets, remorse, lighted, marital, standby, attends, condors, coaches, averted*.

The Spelling Correction Model

- c = correctly spelled word
  t = observed typo

- $c \rightarrow$ Noisy Channel $\rightarrow t$

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

The Language Model for Spelling Correction

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

- Collect sample of text (e.g., a year of AP)

- $Pr(c) \approx \dfrac{freq(c)}{N}$

The Channel Model for Spelling Correction

- Identify sources of noise
    1. Insertions: cress $\rightarrow$ *a*cress
    2. Deletions: ac*t*ress $\rightarrow$ acress
    3. Substitutions: ac*c*ess $\rightarrow$ ac*r*ess
    4. Reversals: *ca*ress $\rightarrow$ *ac*ress

- Collect a large sample of typos

- 2000 typos / month in AP news

- Estimate confusion matrices

- Estimate $Pr(t|c)$

A Confusion Matrix

sub[t, c] = Substitution of t for c

t	**c (correct)**								
	a	b	c	d	e	f	g	h	i
a	0	0	7	1	342	0	0	2	118
b	0	0	9	9	2	2	3	1	0
c	6	5	0	16	0	9	5	0	0
d	1	10	13	0	12	0	5	5	0
e	388	0	3	11	0	2	2	0	89
f	0	15	0	3	1	0	5	2	0
g	4	1	11	11	9	2	0	0	0
h	1	8	0	3	0	0	0	0	0
i	103	0	0	0	146	0	1	0	0

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

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	Pr(c)	Pr(t|c)
actress	37%	.16/N	1343/N	55 / 470,000
cress	0%	.00/N	0/N	46 /32,000,000
caress	0%	.00/N	4/N	1 / 580,000
access	0%	.00/N	2280/N	1 / 4,700,000
across	18%	.077/N	8436/N	93 /10,000,000
acres	21%	.092/N	2879/N	417 /13,000,000
acres	23%	.098/N	2879/N	205 / 6,000,000

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 **actuall** will die. In the 1985 decision, the ...

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

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

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

- Poor estimates of context
  are worse than none.

	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

Other Noisy Channel Model Applications

- Speech Recognition
- Optical Character Recognition (OCR)
- Part-of-Speech Tagging
- Machine Translation

Part-of-Speech Tagging

I see a bird
pro verb art noun

pos $\rightarrow$ Noisy Channel $\rightarrow$ words

$$\hat{pos} \approx \underset{pos}{ARGMAX}\ Pr(pos)\ Pr(words|pos)$$

Under certain independence assumptions, $\hat{pos} \approx$

$$\underset{pos}{ARGMAX} \prod_i Pr(pos_i \mid pos_{i-1}\ pos_{i-2})\ Pr(words_i \mid pos_i)$$

where $Pr(pos_i|word_i)$ is a dictionary,

and $Pr(pos_i|pos_{i-1}\ pos_{i-2})$ is a grammar

Lexical Probabilities of Some Common Words

Word	Pos	P(Pos | Word)	Pos	P(Pos | Word)
costs	VBZ	0.08	NNS	0.92
human	NN	0.12	JJ	0.88
humans	NNS	1.00		
amount	VB	0.18	NN	0.82
amounts	VBZ	0.46	NNS	0.54
meeting	VBG	0.20	NN	0.80
attack	VB	0.25	NN	0.75
mark	VB	0.30	NN	0.70
support	VB	0.31	NN	0.69
change	VB	0.32	NN	0.68
sacrifice	VB	0.37	NN	0.63
use	VB	0.38	NN	0.62
gathering	VBG	0.45	NN	0.55
strike	VB	0.46	NN	0.54
sink	VB	0.48	NN	0.52
travel	VB	0.49	NN	0.51
landing	VBG	0.52	NN	0.48
count	VB	0.58	NN	0.42
finish	VB	0.58	NN	0.42

Typical Contextual Probabilities

log P	(Pos |	__ Pos,	Pos)	log P	(Pos |	__ Pos,	Pos)
–0.81	IN	AT	NNS	–4.31	VBN	AT	NN
–0.83	IN	AT	NN	–4.47	VBN	AT	NNS
–2.28	.	AT	NNS	–4.49	NN	AT	NN
–2.51	.	AT	NN	–4.67	NN	AT	NNS
–2.54	VB	AT	NNS	–5.27	NNS	AT	NN
–2.72	VB	AT	NN	–5.60	JJ	AT	NN
–2.97	,	AT	NNS	–5.64	NNS	AT	NNS
–2.98	,	AT	NN	–5.95	NP	AT	NN
–3.12	VBD	AT	NN	–5.96	NP	AT	NNS
–3.28	VBD	AT	NNS	–6.26	JJ	AT	NNS
–3.41	VBG	AT	NNS	–8.21	PPSS	AT	NNS
–3.62	VBG	AT	NN	–8.90	AT	AT	NNS
–3.88	RB	AT	NN	–9.42	PPSS	AT	NN
–4.17	RB	AT	NNS				

The Proposed Method

- Conceptually, enumerate all assignments
- Score each path (prod of lexical and contextual probs)
- Select best

I	see	a	bird
PPSS	VB	AT	NN
PPSS	VB	IN	NN
PPSS	UH	AT	NN
PPSS	UH	IN	NN
NP	VB	AT	NN
NP	VB	IN	NN
NP	UH	AT	NN
NP	UH	IN	NN

- This alg takes exponential time because there may be k^n paths for an input of n words, where each word has up to k parts-of-speech.
- Fortunately, there is a linear time dynamic programming solution.

	.	.	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	

Traditional Part-of-Speech Methods

- Enumerate each part-of-speech sequence,
  and rule it out as syntactically ill-formed,
  if it can't be parsed.

- *I/noun see/verb a/prep bird/verb

- Ignore lexical probabilities

- Pass-the-buck onto semantics and pragmatics,
  if syntactic constraints aren't adequate.

Motivation for Traditional Methods

- Although most naive people think that most words are unambiguous, so-called ‘‘experts’’ know better.
- Local ngram context is not always sufficient.
- The literature is full of examples where no amount of context will help:

Time	flies	like	an	arrow
noun	*verb*	*sconj*	*art*	*noun*
verb	*noun*	*sconj*	*art*	*noun*

Time	flies	like	an	arrow
Fruit	flies	like	a	banana

Flying	planes	can	be	dangerous
verb	*noun*	*verb*	*verb*	*adj*
adj	*noun*	*verb*	*verb*	*adj*

Why Traditional Methods Failed

Consider the trivial sentence: *I see a bird.*

Easy, if we use lexical probs (Brown Corpus)	
Prob(*I* is a pronoun)	5131/5132
Prob(*see* is a verb)	771/772
Prob(*a* is an article)	22938/22944
Prob(*bird* is a noun)	25/25

Hard, if we ignore lexical probs (Websters)

I	pronoun	noun (letter of alphabet)
see	verb	noun (*the Holy See*)
a	article	noun (letter of alphabet)
bird	noun	verb

- It has been hoped that a parser could eliminate many of these tags.

- But if the parser is going to accept:

  [NP [N city] [N school] [N committee] [N meeting]]

- then it can’t rule out

  [NP [N I] [N see] [N a] [N bird]]

- Attempts to resort to semantics & pragmatics have not worked well (outside of a ‘‘toy’’ domain).

- The noisy channel approach works better because lexical probabilities are important,

- and consequently, it has become the tagging method of choice.

Noisy Channel Approach to Machine Translation (MT)

- Originally proposed by Weaver (1949), and recently revived at IBM, but remains controversial.

- Possibly too outrageous for MT, but may be useful for other purposes.

- The IBM French-to-English MT system uses the following model:

- E $\rightarrow$ Noisy Channel $\rightarrow$ F

- $\hat{E} = \underset{E}{ARGMAX}\ Pr(E)\ Pr(F|E)$

- Use a trigram model for $Pr(E)$, and various stats computed over previously translated texts for $Pr(F|E)$.

A Probabilistic Bilingual Dictionary

E	F	Pr(F|E)
the	le	0.610
the	la	0.178
the	l’	0.083
the	les	0.023
the	ce	0.013
the	il	0.012

- Self-organizing:
    - Hansards $\rightarrow$ Black Box Stats $\rightarrow$ MT
    - No linguists and no French speakers

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.

Word Correspondences

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’ ´evaluer et de modifier des lois et des politiques en vigueur afin qu’ elles correspondent`a une interpŕetation ǵeńereuse de la charte .

We	took	the	initiative	in	assessing	and	amending
	pris		*initiative*		*evaluer*		*modifier*

Top 10 Choices

	drug		sentence
1023	medicaments	217	peine
297	pharmaceutiques	105	phrase
285	drogues	92	sentence
262	medicament	27	*NULL*
145	drogue	14	condamnation
118	*NULL*	10	condamne
96	pharmaceutique	9	peines
84	assurance-medicaments	9	condamnees
61	toxicomanie	7	emprisonnement
43	stupefiants	6	detenu

Terminology

- Although Noisy Channel approach may be too outrageous for MT, it may be useful for other purposes.
- Human translators need help with terminology
    1. *magnetic fields* → *magnetic meadows*
    2. *cascade* (of electrical components) → *waterfall*
    3. *store* (computer sense) → *store* (Grocery sense)
- Terminology makes use of shared knowledge between reader and writer
- Translators tend to be less familiar with the topic than readers and writers.
- Solution: use stats computed over parallel texts

  existing translations contain more solutions to more translation problems than any other existing resource
  -- Pierre Isabelle

Discrimination Applications

- Word-Sense Disambiguation: distinguish ‘‘river’’ bank from ‘‘money’’ bank
- Author Identification: distinguish the Federalist Papers written by Madison from those written by Hamilton
- Information Retrieval (IR): distinguish documents that are relevant to a query from those that are not.

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

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

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

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

Word-Sense Disambiguation

- Required: large amounts of testing and training materials, with labels indicating the correct sense.
- Solution: use previously translated texts
- The translation can often be used in lieu of hand-labeling:
  1. *sentence* $\rightarrow$ *peine* (‘‘judicial’’ sense)
  2. *sentence* $\rightarrow$ *phrase* (‘‘syntactic’’ sense)
- Outline of Algorithm
  1. Sentence Alignment
  2. Word Correspondence
  3. Train Context Models: $Pr(word \mid sense)$
  4. Test on New Data

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

Text Compression

Entropy: $H = - \sum_{p} p \log p$

- size of search space for a recognizer
- measure of disk space
- measure of information
- bandwidth of a channel (telephone)
- efficiency of a code

Cross Entropy of Various Language Models

Model	Bits / Character
ASCII	8
Huffman code each char	5
Lempel-Ziv (Unix™ *compress*)	4.43
Unigram (Huffman code each word)	2.1
Trigram	1.76
Human Performance	1.25 (Shannon, 1951)

Conclusion

- Shannon's Noisy Channel Model
    1. Language Model
    2. Channel Model

- Applications
    1. Recognition: Speech, OCR, Spelling
    2. Part-of-Speech Tagging
    3. MT, Terminology and Bilingual Lexicography
    4. Discrimination: Word-sense, Author, IR
    5. Text Compression

- Text Analysis: pragmatic application of well-understood weak methods (e.g., ngrams, Bayesian discrimination methods) and very large data-sets to address problems of practical concern, as well as classic academic puzzles.