COMS6998 Topics in Human Language Technology (HLT)

Kenneth Church

Kenneth.Ward.Church@gmail.com

http://www.columbia.edu/~kc3109/

Too many papers are boring

- Survey papers
  - (and most conference papers)
  - tend to be boring
- It is ok to be wrong,
  - but please don't be boring
- Main assignment:
  - Write a survey paper
  - Due at the end of the term
- The survey paper should discuss 1+ seminal papers,
  - and the impact on subsequent literature

Weekly Assignments

(To encourage everyone to keep up with the reading)

- Imagine you are writing a survey paper on the paper(s) assigned for that week and their impact on the subsequent literature
- For the imaginary survey paper, I want to see:
  - an abstract of no more than 200 words
    - Outlining the argument in the imaginary paper
  - Citation counts for the assigned paper(s) (from Google Scholar)
  - A partial bibliography of 10-30 references of papers that would be appropriate to discuss in the imaginary survey
  - 1-3 tweets (< 140 chars) pitching the imaginary survey
    - The imaginary tweets should identify an audience that might not read the imaginary survey (if it existed) without more motivation to do so
  - A review of the imaginary survey paper (using a standard review form), from the perspective of an imaginary reviewer

How to Find Citations with Google Scholar

Google Scholar Citations

A statistical approach to machine translation

Search within citing articles

The mathematics of statistical machine translation: Parameter estim
PF Brown, VJD Pietra, SAD Pietra… - Computational linguistics, 1993 - dl.acm.org
Abstract We describe a series of five statistical models of the translation process and giv
algorithms for estimating the parameters of these models given a set of pairs of sentence
that are translations of one another. We define a concept of word-by-word alignment
Cited by 4601 Related articles All 49 versions Cite Save

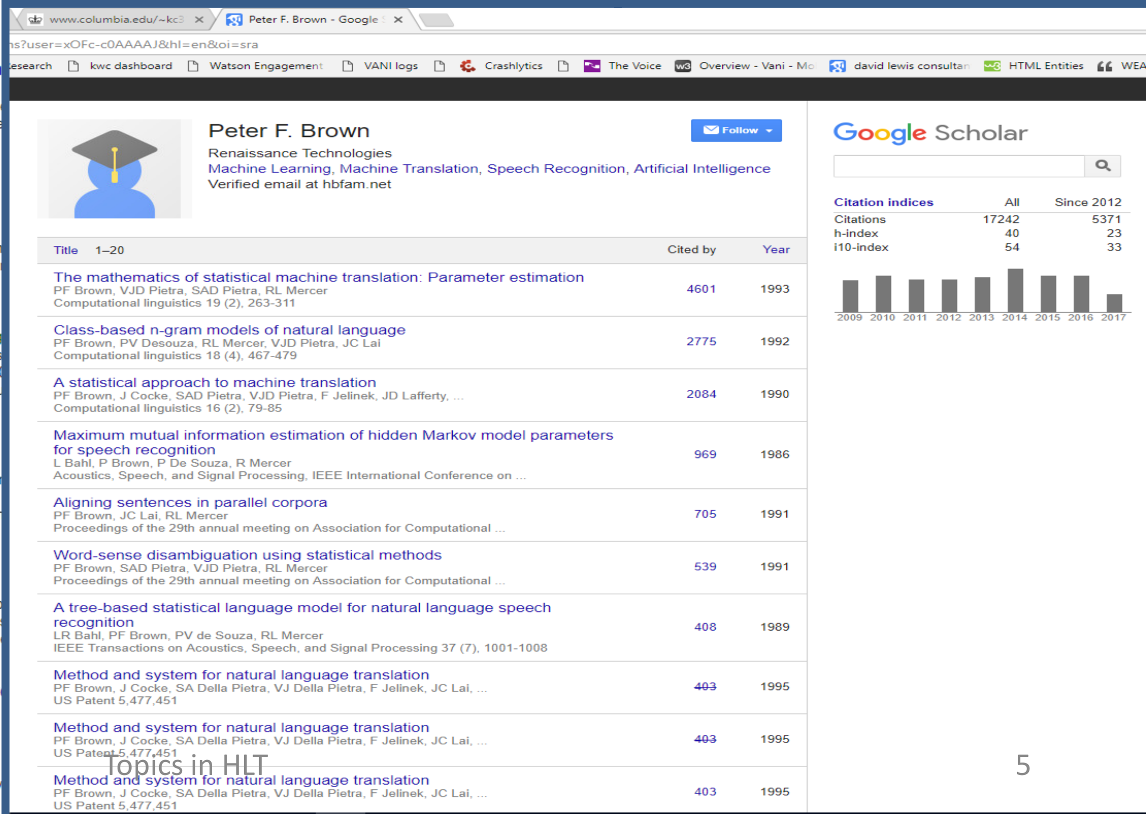
A maximum entropy approach to natural language processing
AL Berger, VJD Pietra, SAD Pietra - Computational linguistics, 1996 - dl.acm.org
Abstract The concept of maximum entropy can be traced back along multiple threads to
Biblical times. Only recently, however, have computers become powerful enough to perm
the widescale application of this concept to real world problems in statistical estimation a
Cited by 3584 Related articles All 68 versions Cite Save More

An empirical study of smoothing techniques for language modeling
SF Chen, J Goodman - Proceedings of the 34th annual meeting on …, 1996 - dl.acm.org
Abstract We present an extensive empirical comparison of several smoothing techniques
the domain of language modeling, including those described by Jelinek and Mercer (198
Katz (1987), and Church and Gale (1991). We investigate for the first time how factors su
Cited by 2946 Related articles All 54 versions Cite Save

Verbs semantics and lexical selection
Z Wu, M Palmer - Proceedings of the 32nd annual meeting on …, 1994 - dl.acm.org
Abstract This paper will focus on the semantic representation of verbs in computer syster
and its impact on lexical selection problems in machine translation (MT). Two groups of
English and Chinese verbs are examined to show that lexical selection must be based or
Cited by 2937 Related articles All 21 versions Cite Save More

Class-based n-gram models of natural language
PF Brown, PV Desouza, RL Mercer, VJD Pietra… - Computational …, 1992 - dl.acm.org
Abstract We address the problem of predicting a word from previous words in a sample o
text. In particular, we discuss n-gram models based on classes of words. We also discus
several statistical algorithms for assigning words to classes based on the frequency of th
Cited by 2775 Related articles All 44 versions Cite Save

Transformation-based error-driven learning and natural language pr
case study in part-of-speech tagging
E Brill - Computational linguistics, 1995 - dl.acm.org
Abstract Recently, there has been a rebirth of empiricism in the field of natural language
processing. Manual encoding of linguistic information is being challenged by automated
corpus-based learning as a method of providing a natural language processing system w
Cited by 2139 Related articles All 50 versions Cite Save



www.columbia.edu/~kc3 | Peter F. Brown - Google

ns?user=xOFc-c0AAAAJ&hl=en&oi=sra

esearch | kwc dashboard | Watson Engagement | VANI logs | Crashlytics | The Voice | Overview - Vani - Mo | david lewis consultan | HTML Entities | WEA

Peter F. Brown
Renaissance Technologies
Machine Learning, Machine Translation, Speech Recognition, Artificial Intelligence
Verified email at hbfam.net

Follow

Title 1–20	Cited by	Year
The mathematics of statistical machine translation: Parameter estimation PF Brown, VJD Pietra, SAD Pietra, RL Mercer Computational linguistics 19 (2), 263-311	4601	1993
Class-based n-gram models of natural language PF Brown, PV Desouza, RL Mercer, VJD Pietra, JC Lai Computational linguistics 18 (4), 467-479	2775	1992
A statistical approach to machine translation PF Brown, J Cocke, SAD Pietra, VJD Pietra, F Jelinek, JD Lafferty, ... Computational linguistics 16 (2), 79-85	2084	1990
Maximum mutual information estimation of hidden Markov model parameters for speech recognition L Bahl, P Brown, P De Souza, R Mercer Acoustics, Speech, and Signal Processing, IEEE International Conference on ...	969	1986
Aligning sentences in parallel corpora PF Brown, JC Lai, RL Mercer Proceedings of the 29th annual meeting on Association for Computational ...	705	1991
Word-sense disambiguation using statistical methods PF Brown, SAD Pietra, VJD Pietra, RL Mercer Proceedings of the 29th annual meeting on Association for Computational ...	539	1991
A tree-based statistical language model for natural language speech recognition LR Bahl, PF Brown, PV de Souza, RL Mercer IEEE Transactions on Acoustics, Speech, and Signal Processing 37 (7), 1001-1008	408	1989
Method and system for natural language translation PF Brown, J Cocke, SA Della Pietra, VJ Della Pietra, F Jelinek, JC Lai, ... US Patent 5,477,451	~~403~~	1995
Method and system for natural language translation PF Brown, J Cocke, SA Della Pietra, VJ Della Pietra, F Jelinek, JC Lai, ... US Patent 5,477,451	~~403~~	1995
Method and system for natural language translation PF Brown, J Cocke, SA Della Pietra, VJ Della Pietra, F Jelinek, JC Lai, ... US Patent 5,477,451	403	1995

Google Scholar

Citation indices	All	Since 2012
Citations	17242	5371
h-index	40	23
i10-index	54	33

Peter F. Brown

Follow

Renaissance Technologies
Machine Learning, Machine Translation, Speech Recognition, Artificial Intelligence
Verified email at hbfam.net

Title 1–20	Cited by	Year
Shelf life expiration date management P Brown US Patent 9,547,851		2017
Is tip-apex distance related to radiation use? P Brown, M El-Sobky, V Peter International Journal of Surgery 36, S114		2016
Shelf life expiration date management P Brown US Patent 9,208,520	1	2015
Seafarers get aboard new course structure P Brown		2015
System and method for executing synchronized trades in multiple exchanges RL Mercer, PF Brown US Patent App. 14/451,356		2014
Simulated training for Antarctic seafarers P Brown, M Lutzhoft		2014
The Ice is right: Antarctic ice modelling in maritime training simulators PE Brown, M Lutzhoft 15th Annual general assembly International Association of Maritime ...		2014
Quantifying the hidden benefits of high-performance building B Birkenfeld, P Brown, N Kresse, J Sullivan, P Thiam International Society of Sustainability Professionals	5	2011
Multilayered coated infra-red surface plasmon resonance fibre sensors for aqueous chemical sensing T Allsop, R Neal, C Mou, P Brown, S Rehman, K Kalli, DJ Webb, D Mapps, ... Optical Fiber Technology 15 (5), 477-482	9	2009
Multilayered coated infra-red surface plasmon resonance fibre sensors for chemical sensing T Allsop, R Neal, C Mou, P Brown, S Rehman, K Kalli, DJ Webb, D Mapps, ... The European Conference on Lasers and Electro-Optics, CH5 4	2	2009

Google Scholar

Citation indices	All	Since 2012
Citations	17242	5371
h-index	40	23
i10-index	54	33

Google isn’t Perfect

h-index ≥ 40 →

http://web.cs.ucla.edu/~palsberg/h-number.html

Encourage Student Presentations

- I'd like to encourage students to practice presentation skills in a supportive environment,
  - so it isn't a good idea to grade presentations
- In addition, it may not be practical for everyone to give a presentation,
  - especially if there are too many students
  - and too little time.

Point: Emphasize Diverse Perspectives

- Think about how other people would think about these issues
- Too many people are too focused on their own immediate needs and there own perspectives,
  – and not enough about
    - how things will stand up to the test of time,
    - from lots of different perspectives

Why Survey Papers?

- My first presentation was on my own research
- Many rock bands start out doing covers
  - before they write their own music
- So too, I believe students should start out presenting covers of seminal papers
  - before presenting their own research

First Assignment (Due WED midnight)

http://www.columbia.edu/~kc3109/

- Vote here on papers/videos/topics
  - you would like to cover in the course
  - as well as papers you would like to present
  - NOTE: there are ***TWO*** tabs
- The first assignment involves skimming as much of this work as possible
  - so you can cast reasonably informed votes.

Topic 1: Ketchup

https://www.superlectures.com/interspeech2016/

- Jurafsky uses history of ketchup (& ice cream elsewhere)
  - to shed light on currently popular methods in speech and language
- He traces etymology of "ketchup" from an Asian fish sauce
  - Advances in (sailing) technology made it possible to replace anchovies with less expensive tomatoes and sugar from the west
  - The ice cream story combines fruit syrups (Sharbat) from Persia
    - with gun powder from China and advances in refrigeration technology

The **ketchup model** of innovation:

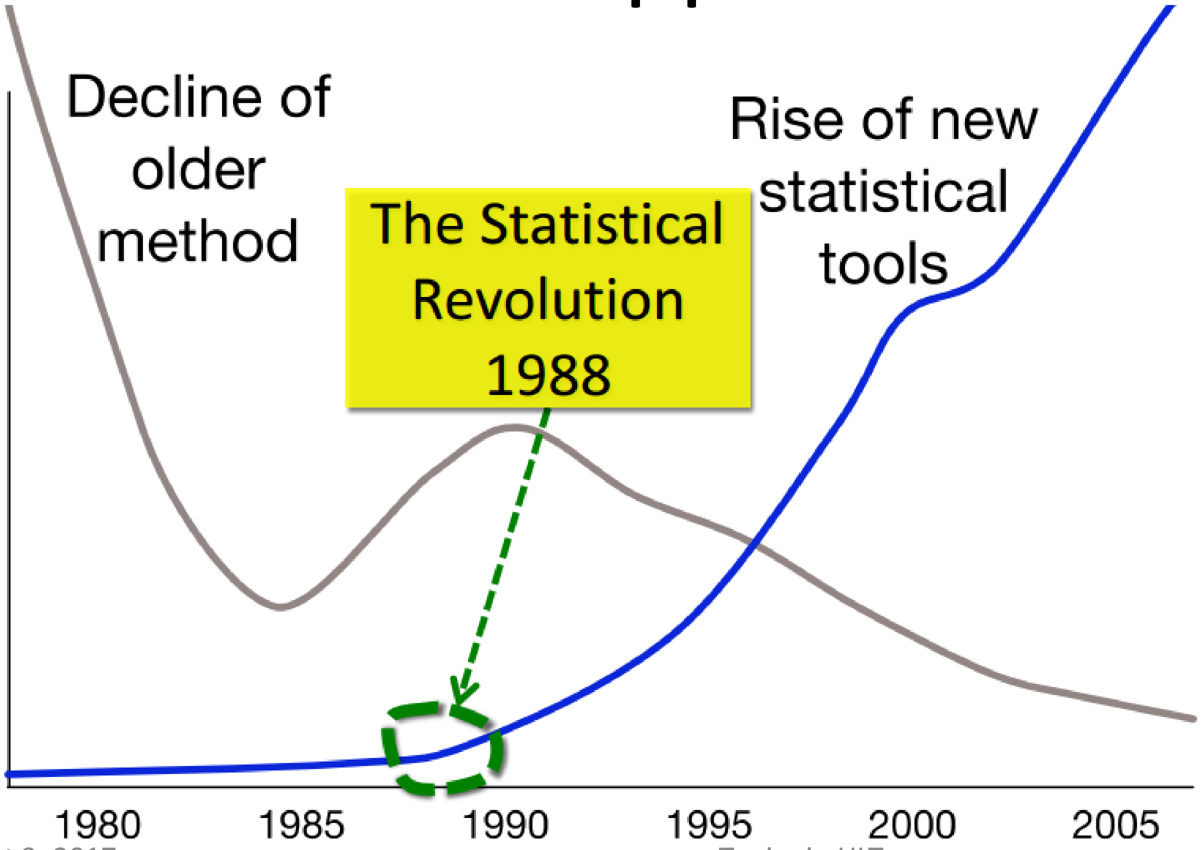
- We borrow technology from the neighbors
- Interdisciplinarity plays a key role

Big Tent

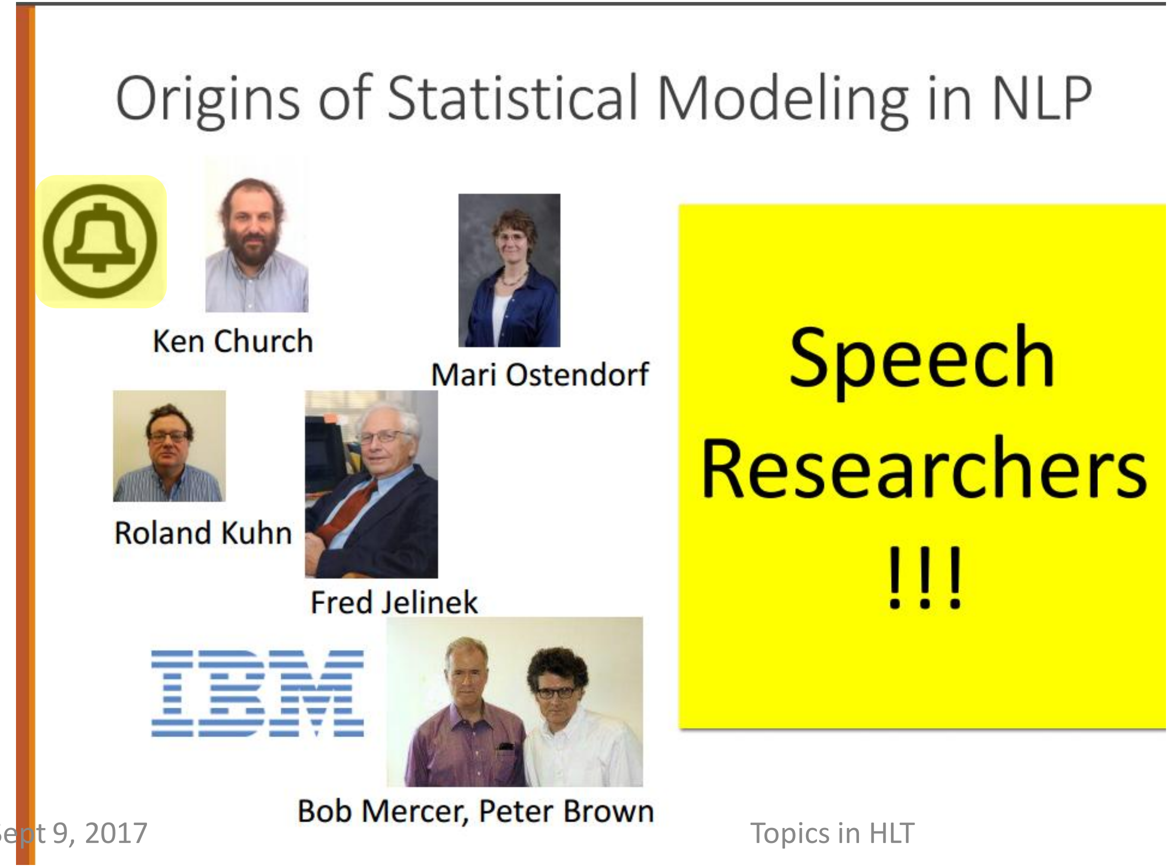
The Speech Invasion

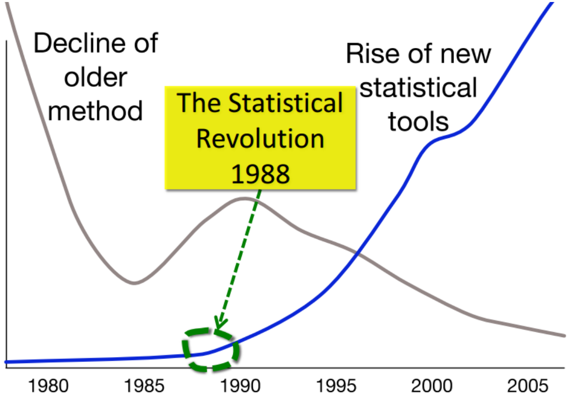
- At speech meetings (Interspeech-2016, as opposed to NAACL-2009),
  - Jurafsky credits speech researchers for transferring currently popular techniques from speech to language
  - Some of these people were probably also involved in transferring similar methods from information theory into speech (and perhaps hedge funds)

What happened in 1988?



What happened in1988?





Some of my Best Friends are Linguists

(LREC 2004)

Frederick Jelinek
Johns Hopkins University

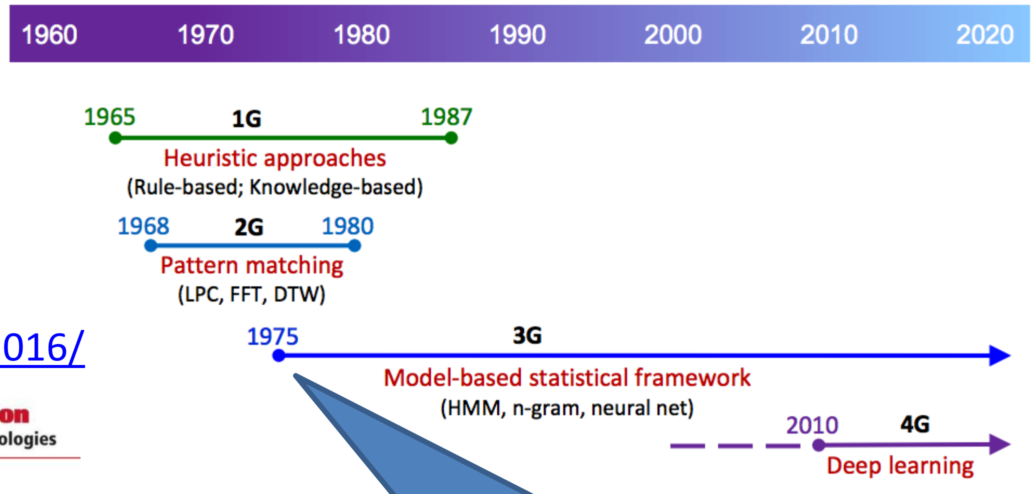
THANKS TO: E. Brill, L. Burzio, W. Byrne, C. Cieri, J. Eisner, R. Frank, L. Guthrie, S. Khudanpur, G. Leech, M. Liberman, M. Marcus, M. Palmer, P. Smolensky, and D. Yarowsky

May 28, 2004 Johns Hopkins

h2016/

ytheon
Technologies

Generations of Automatic Speech Recognition (ASR)



(Adapted from Furui)

A 50-year personal Retrospective on Speech and Language Processing

John Makhoul
makhoul@bbn.com

Interspeech, San Francisco, CA
9 September 2016

What happened in 1975?

The same thing that happened to language in 1988 (and to hedge funds in 1990s)?

Some of my Best Friends are Linguists

(LREC 2004)

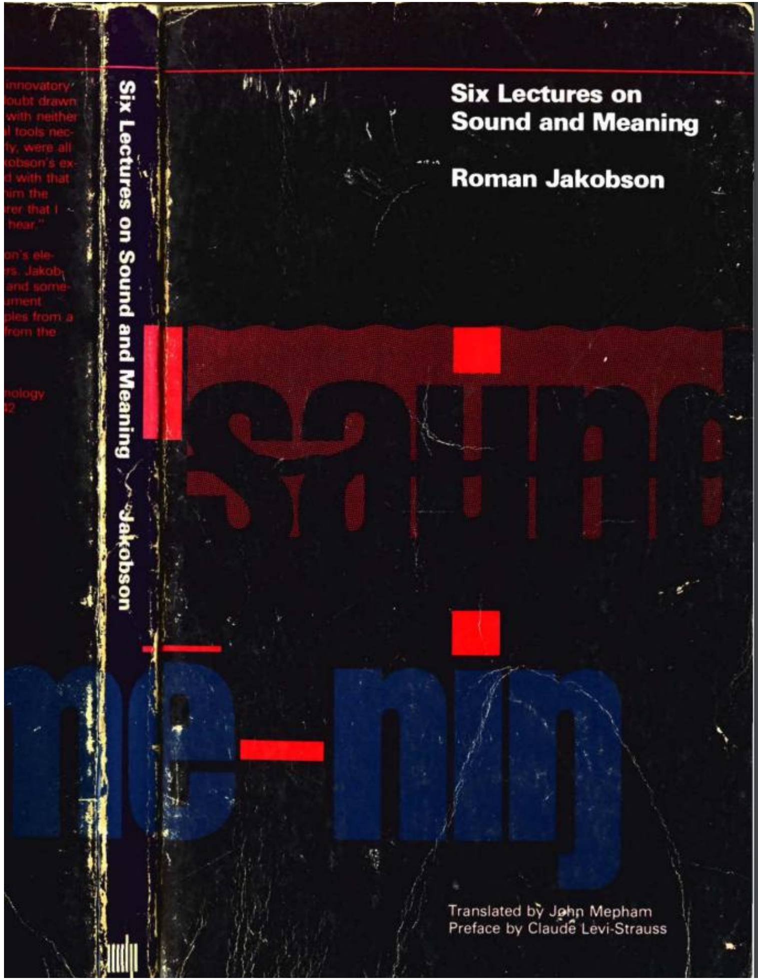
Frederick Jelinek
Johns Hopkins University

THANKS TO: E. Brill, L. Burzio, W. Byrne, C. Cieri, J. Eisner, R. Frank, L. Guthrie, S. Khudanpur, G. Leech, M. Liberman, M. Marcus, M. Palmer, P. Smolensky, and D. Yarowsky

May 28, 2004 Johns Hopkins

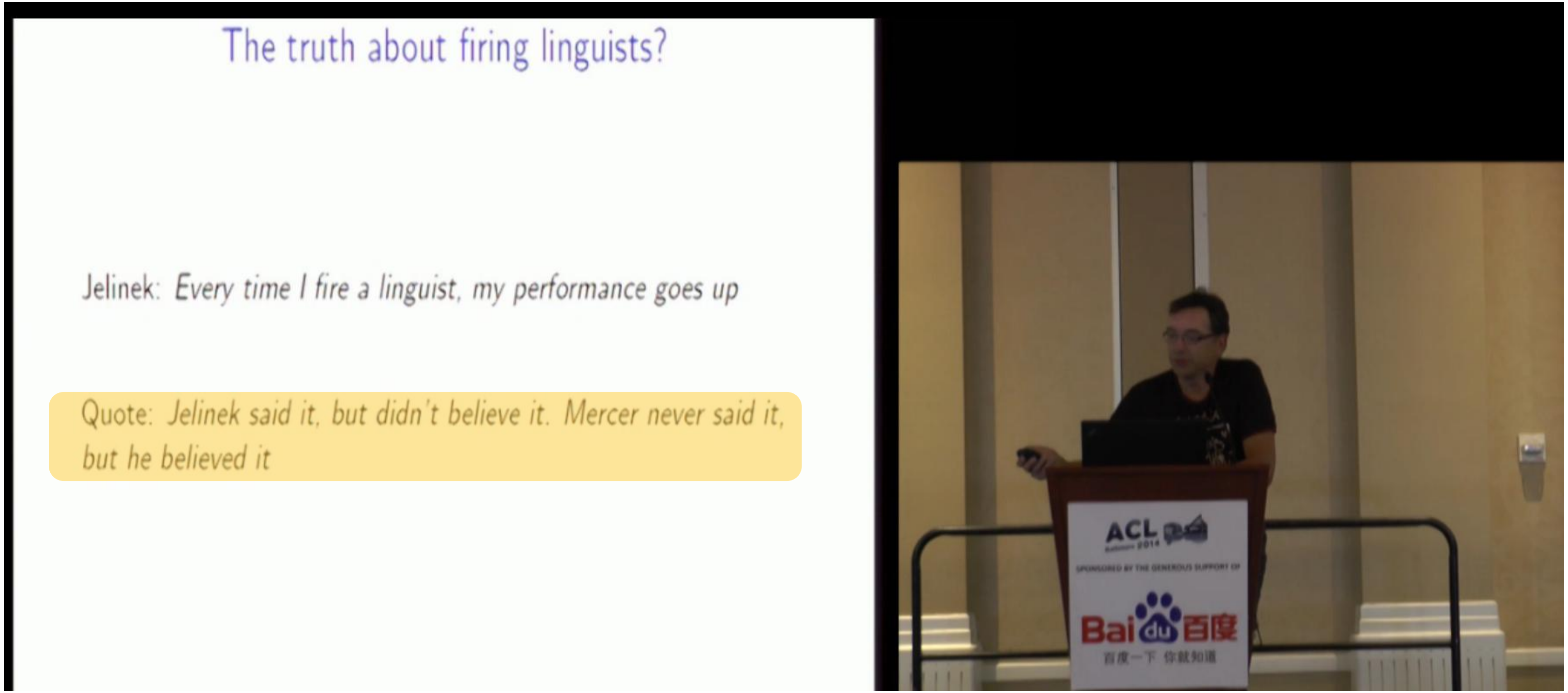
Frederick Jelinek

Born	Bedřich Jelínek November 18, 1932 Kladno, now Czech Republic
Died	September 14, 2010 (aged 77) Baltimore, United States
Citizenship	American
Fields	Information theory, natural language processing
Institutions	Cornell University, IBM Research, Johns Hopkins University
Alma mater	Massachusetts Institute of Technology
Doctoral advisor	Robert Fano
Notable students	Neil Sloane
Known for	Advancement of natural language processing techniques
Influences	Roman Jakobson
Notable awards	• James L. Flanagan Award (2005) • ACL Lifetime Achievement Award (2009)
Spouse	Milena Jelinek



Robert Mercer
ACL Lifetime Achievement

http://techtalks.tv/talks/closing-session/60532/



The Case for Empiricism (With and Without Statistics)

Kenneth Church
1101 Kitchawan Road
Yorktown Heights, NY 10589
USA
Kenneth.Ward.Church@gmail.com

Abstract

These days we tend to use terms like *empirical* and *statistical* as if they are interchangeable, but it wasn't always this way, and probably for good reason. In *A Pendulum Swung Too Far* (Church, 2011), I argued that graduate programs should make room for both Empiricism and Rationalism. We don't know which trends will dominate the field tomorrow, but it is a good bet that it won't be what's hot today. We should prepare the next generation of students for all possible futures, or at least all probable futures. This paper argues for a diverse interpretation of Empiricism, one that makes room for everything from Humanities to Engineering (and then some).

points that I made in my introduction to Chuck's LTA talk at ACL-2012.

I had the rather unusual opportunity to see his talk (a few times) before writing my introduction because Chuck video-taped his talk in advance.[1] I knew that he was unable to make the trip, but I had not appreciated just how serious the situation was. I found out well after the fact that the LTA meant a lot to him, so much so that he postponed an operation that he probably shouldn't have postponed (over his doctor's objection), so that he would be able to answer live questions via Skype after the showing of his video tape.

I started my introduction by crediting Lily Wong Fillmore, who understood just how much Chuck wanted to be with us in Korea, but also, just how impossible that was. Let me take this opportunity to thank her once again for her contributions to the video (technical lighting, editing, encouragement and so much more).

For many of us in my generation, C4C, Chuck's "The Case for Case" (Fillmore, 1968) was the introduction to a world beyond Rationalism and Chomsky. This was especially the case for me, since I was studying at MIT, where we learned many things (but not Empiricism).

After watching Chuck's video remarks, I was struck by just how nice he was. He had nice things to say about everyone from Noam Chomsky to Roger Schank. But I was also struck by just how difficult it was for Chuck to explain how important C4C was (or even what it said and why it mattered). To make sure that the international audience wasn't misled by his up-

On firing linguists…

Introduction to the Special Issue on Computational Linguistics Using Large Corpora

Kenneth W. Church*
AT&T Bell Laboratories

Robert L. Mercer†
IBM T.J. Watson Research Center

- Finally, they removed the dictionary lookup HMM,
  - taking for the pronunciation of each word its spelling.
  - Thus, a word like *t-h-r-o-u-g-h* was assumed to have a pronunciation like *tuh huh ruh oh uu guh huh*.
- After training, the system learned that
  - with words like *l-a-t-e* the front end often missed the *e*.
  - Similarly, it learned that *g*'s and *h*'s were often silent.
  - This crippled system was still able to recognize
    - 43% of 100 test sentences correctly as compared with
    - 35% for the original Raleigh system.

On firing linguists… (2 of 2)

- These results firmly established the importance of a ***coherent, probabilistic*** approach to speech recognition and the importance of data for estimating the parameters of a probabilistic model.
  - One by one, pieces of the system that had been assiduously assembled by speech experts yielded to probabilistic modeling.
  - Even the elaborate set of hand-tuned rules for segmenting the frequency bank outputs into phoneme-sized segments would be replaced with training (Bakis 1976; Bahl et al. 1978).
- By the summer of 1977, performance had reached **95%** correct by sentence and 99.4% correct by word,
  - a considerable improvement over the same system with hand-tuned segmentation rules
  - (**73%** by sentence and 95% by word).
- Progress in speech recognition at Yorktown and almost everywhere else as well has continued along the lines drawn in these early experiments.
  - As computers increased in power, ever greater tracts of the ***heuristic wasteland*** opened up for colonization by probabilistic models.
  - As greater quantities of recorded data became available,
    - these areas were tamed by automatic training techniques.

Sound & Meaning >> Spelling

Topic 2: The Speech Invasion

- [The Mathematics of Statistical Machine Translation: Parameter Estimation]
  - The foundation of most MT work today
  - Peter Brown’s daughter asked him if he knew that machines can translate…
- [A stochastic parts program and noun phrase parser for unrestricted text]
  - As a grad student, I was told that everything we could do in NLP had already been done (so I should work on things we couldn’t do like pragmatics)
  - So, it wasn’t easy to present a paper on something that had been “solved”
- [An Introduction to Hidden Markov Models]
  - The classic reference is often the most accessible (not first, last or best)
- [Shannon's Theory of Communication]
  - Amazingly accessible
- [Shannon's Estimate of the Entropy of English]
  - Even more accessible

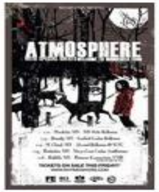
Topic 3: Case for Case

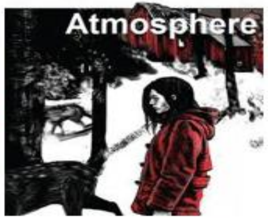
- Speech vs. Language and the Case for Case
  - It is interesting to contrast Fillmore's views on spelling here
  - with Mercer's in their ACL Lifetime Achievement here
- Fillmore is a linguist who believes that sound and meaning are better sources of evidence than spelling,
  - whereas Mercer believes every time I fire a linguist, my performance goes up (as discussed in the introduction to video, Jelinek said it, but Mercer believes it)

LTA-2012: Charles J Fillmore

- Highlights
  - Case for Case
    - 6k citations in Google Scholar
  - Framenet
    - 2 papers with 1k citations each
- "Minnesota Nice"
  - Nice things to say about everyone: Chomsky/Schank
  - Self-deprecating humor
    - (but don't you believe it)

Minnesota Nice



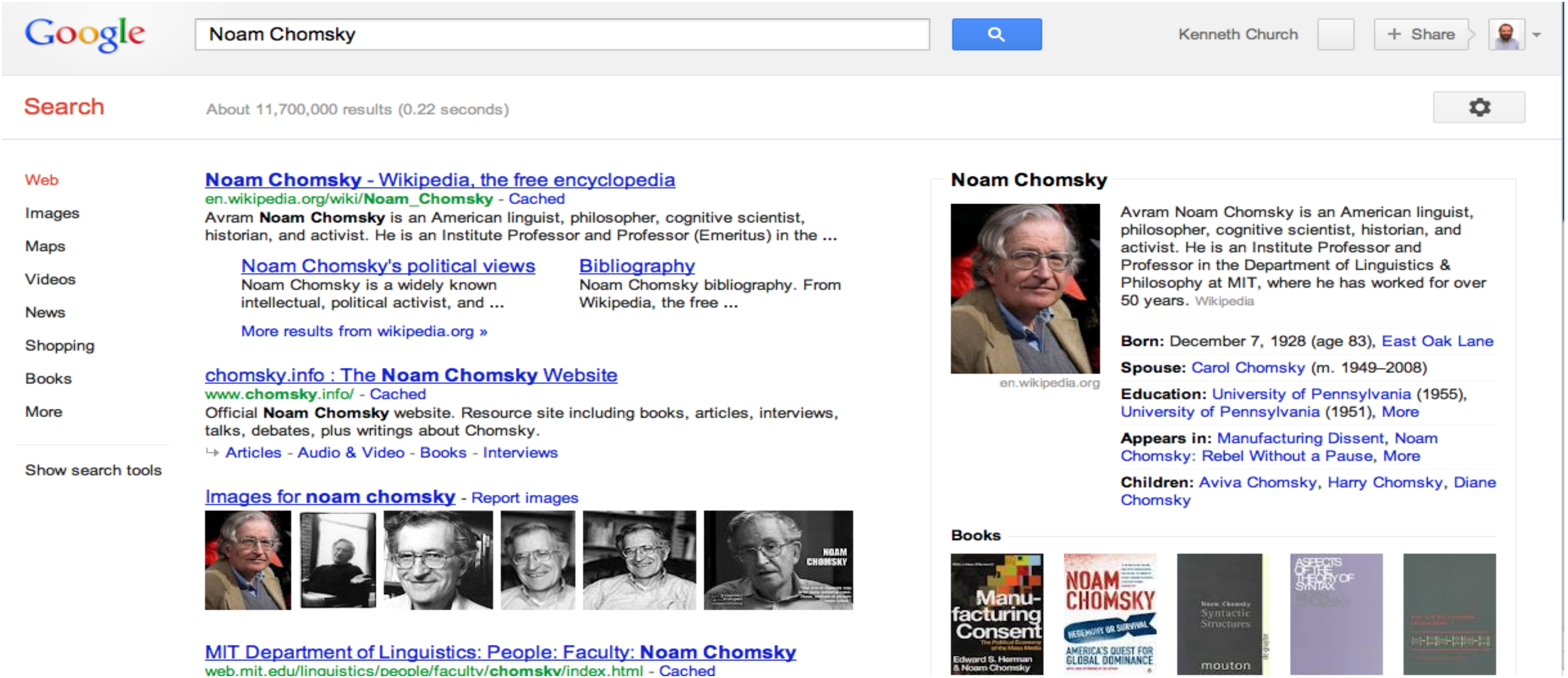


Case for Case (C4C): Practical Apps

- Information Extraction (MUC)
- Semantic Role Labeling

- Key Question: Who did what to whom?
  – Not: What is the NP and the VP of S?

Commercial Information Extraction



Do Read “Case for Case”

- Great arg but also
  - Demonstrates strong command of
    - Classic literature as well as
    - Linguistic facts
- Our field:
  - Too “silo”-ed
  - Too few citations to
    - Classic literature, other fields and other types of facts
- We could use more “Minnesota Nice”

Historical Motivation: A Case for Case

From Morphology → MUC

- Context Free Grammar is attractive for
  - Langs with more word order and less morphology (English)
- But Case Grammar is attractive for
  - Langs with more morphology and less word order
  - Examples: Latin, Greek & Japanese
- Latin (over-simplified):
  - Subject: Nominative case
  - Object: Accusative case
  - Indirect Object: Dative case
  - Other args: Ablative case

Japanese I/Vocabulary/Case Markers

< Japanese I | Vocabulary

These are to be placed after a wo

- wa（は） - topic marker
- ga（が） - subject
- (w)o（を） - direct object
- mo（も） - "also" (substitutes
- no（の） - possessive (revers
- na（な） - marks an adjective
- de（で） - "by means of", "in"
- ni（に） - indirect object, "in"/
- to（と） - "and", object of "sa
- ya（や） - "and" for a list
- (h)e（へ） - destination "to"
- ka（か） - question mark (poli

Using these semantic features valency patterns of the basic predicates necessary in the task domain are defined. As an example, the predicate '**okuru**' ('send' in English) is given the following valency patterns:

N[con/-tra]'***wo***' + V,
N[con/-tra]'***wa***' + N[loc]'***ni***' + V,
N[con/-tra]'***wa***' + N[hum]'***ni***' + V,
N[con/-tra]'***wa***' + N[tim/pro]'**made*ni***' + V,
N[tim/pro]'**made*ni***' + N[con/-tra]'***wo***' + V,
N[hum]'***ni***' + N[con/-tra]'***wo***' + V,
N[hum]'***ga***' + N[con/-tra]'***wo***' + V,
N[hum]'***ga***' + N[con/-tra]'***wo***' + N[hum/-pro]'***ni***' + V,
N[hum/-pro]'***wa***' + N[con/-tra]'***wo***' + N[hum]'***ni***' + V,

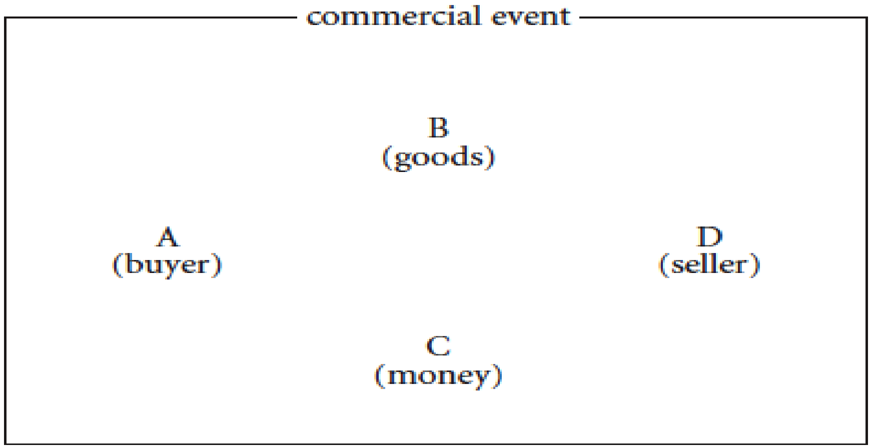
Nouns: instrument	***Jitensha de** ikimashō.* 自転車で行きましょう。	Let's go **by bicycle.**
Nouns: location	***Koko de** yasumitai.* ここで休みたい。	I want to rest **here.**

Case Grammar → Frames / Lexicography

Valency → Scripts (Roger Schank) / Lexicography (Sue Atkins)

- Valency: Predicates have args (optional & required)
  - Example: “give” requires 3 args:
    - Agent (A), Object (O), and Beneficiary (B)
    - Jones (A) gave money (O) to the school (B)
  - Latin Morphology:
    - Nominative, Accusative & Dative

Harry bought the puppy from Mr. Smith for $60.
Harry bought the puppy with the $60 that his mot
Mr. Smith sold the puppy.
Mr. Smith sold the puppy to Harry.
Mr. Smith sold the puppy for $60.



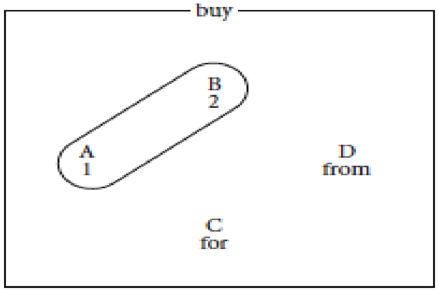


Fig. 3

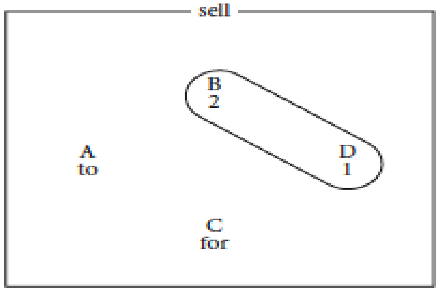


Fig. 4

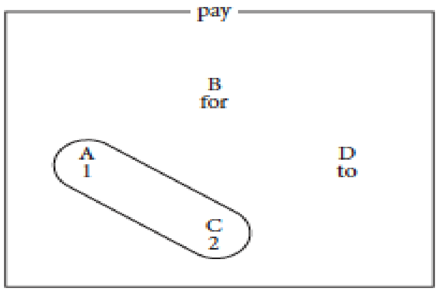


Fig. 5

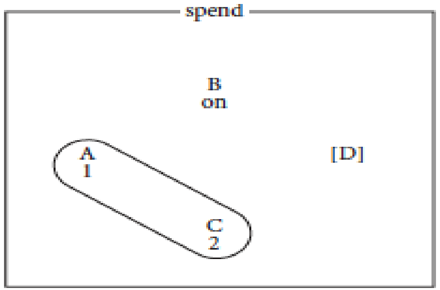


Fig. 6

to him.)

Case Grammar → Frames / Lexicography

Valency → Scripts (Roger Schank) / Lexicography (Sue Atkins)

- Valency: Predicates have args (optional & required)
  - Example: “give” requires 3 arguments:
    - Agent (A), Object (O), and Beneficiary (B)
    - Jones (A) gave money (O) to the school (B)
  - Latin Morphology: Nominative, Accusative & Dative
- Frames
  - Commercial Transaction Frame: Buy/Sell/Pay/Spend
  - Save <good thing> from <bad situation>
  - Risk <valued object> for <situation>|<purpose>|<beneficiary>|<motivation>
- Collocations & Typical predicate argument relations:
  - Save whales from extinction (not vice versa)
  - Ready to risk everything for what he believes
- Representation Challenges: What matters for practical apps/NLU?
  - Stats on POS? Word order? Frames (typical predicate-args/collocations)?

Challenge for Next Generation:

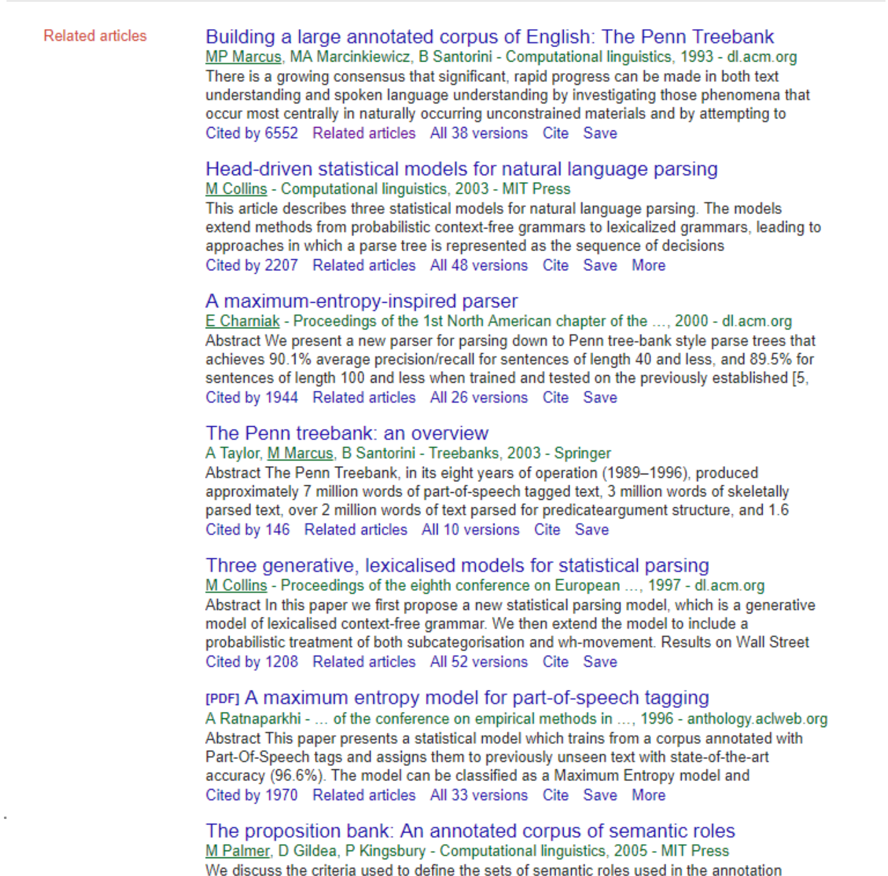
General Linguistics → Computational Linguistics

- Challenge:
  - Do corpus-based lexicography methods scale up?
  - Are they too manually intensive?
  - If so, could we use machine learning methods to speed up manual methods?
  - Just as statistical parsers learn phrase structure rules (S → NP VP)
    - Can we learn valency?
    - Collocations?
    - Typical predicate argument relations?

- Corpus-size requirements: When can we expect to learn frames?
  - Many content words have freq in parts per million
  - 1970s Corpora: 1 M words (Brown)
    - Large enough to make a list of content words
  - 1990s: 100 M words (British National Corpora)
    - Large enough to see associations of common predicates with function words
    - “save” + “from”
  - Coming soon: (1M)^2 words (Google?)
    - Large enough to see associations of pairs of content words (collocations)
      - “give” + $$
      - “save” + “whale” and “save” + “extinction”
      - “risk” valued object for purpose

Topic 4: Parsing & Part of Speech Tagging

- Parsing and part of speech tagging used to be one of the most important topics in computational linguistics
- The Penn Treebank can be downloaded from here.
- The Penn Treebank has a huge number of citations,
  - and many of the papers that cite this paper are also highly cited
  - Google scholar makes it easy to find these papers with links like this.


Related articles

Building a large annotated corpus of English: The Penn Treebank
MP Marcus, MA Marcinkiewicz, B Santorini - Computational linguistics, 1993 - dl.acm.org
There is a growing consensus that significant, rapid progress can be made in both text understanding and spoken language understanding by investigating those phenomena that occur most centrally in naturally occurring unconstrained materials and by attempting to
Cited by 6552 Related articles All 38 versions Cite Save

Head-driven statistical models for natural language parsing
M Collins - Computational linguistics, 2003 - MIT Press
This article describes three statistical models for natural language parsing. The models extend methods from probabilistic context-free grammars to lexicalized grammars, leading to approaches in which a parse tree is represented as the sequence of decisions
Cited by 2207 Related articles All 48 versions Cite Save More

A maximum-entropy-inspired parser
E Charniak - Proceedings of the 1st North American chapter of the ..., 2000 - dl.acm.org
Abstract We present a new parser for parsing down to Penn tree-bank style parse trees that achieves 90.1% average precision/recall for sentences of length 40 and less, and 89.5% for sentences of length 100 and less when trained and tested on the previously established [5,
Cited by 1944 Related articles All 26 versions Cite Save

The Penn treebank: an overview
A Taylor, M Marcus, B Santorini - Treebanks, 2003 - Springer
Abstract The Penn Treebank, in its eight years of operation (1989–1996), produced approximately 7 million words of part-of-speech tagged text, 3 million words of skeletally parsed text, over 2 million words of text parsed for predicateargument structure, and 1.6
Cited by 146 Related articles All 10 versions Cite Save

Three generative, lexicalised models for statistical parsing
M Collins - Proceedings of the eighth conference on European ..., 1997 - dl.acm.org
Abstract In this paper we first propose a new statistical parsing model, which is a generative model of lexicalised context-free grammar. We then extend the model to include a probabilistic treatment of both subcategorisation and wh-movement. Results on Wall Street
Cited by 1208 Related articles All 52 versions Cite Save

[PDF] A maximum entropy model for part-of-speech tagging
A Ratnaparkhi - ... of the conference on empirical methods in ..., 1996 - anthology.aclweb.org
Abstract This paper presents a statistical model which trains from a corpus annotated with Part-Of-Speech tags and assigns them to previously unseen text with state-of-the-art accuracy (96.6%). The model can be classified as a Maximum Entropy model and
Cited by 1970 Related articles All 33 versions Cite Save More

The proposition bank: An annotated corpus of semantic roles
M Palmer, D Gildea, P Kingsbury - Computational linguistics, 2005 - MIT Press
We discuss the criteria used to define the sets of semantic roles used in the annotation

Topic 5: Whatever you measure, you get

- When BLEU was first introduced,
  - it was impressive that the community could agree on a single evaluation metric,
  - and that that metric might be correlated with human judgments.
- But when Och proposed optimizing the metric (here),
  - there was widespread concern that the optimization might find a way to game the metric,
  - resulting in a solution that would score high on the objective metric (BLEU),
  - but less well on subjective evaluations with humans.
- Apparently, the metric stood up remarkably well to optimizing, though a number of alternatives have been proposed subsequently.

Topic 6: Machine Learning has a long history

- Much of the early work is based on simple methods such as linear separators,
  - which can be used to distinguish relevant documents from irrelevant documents in
    - Information Retrieval,
    - positive sentiment from negative sentiment,
    - Hamilton's essays from Madison's, and so on.
- Massive citations
  - Methods such as support vector machines and logistic regression tend to be cited even more than applications
  - Tools such as sklearn and libshorttext are also heavily cited,
    - as well as datasets such as this.

Topic 7: ALPAC & Whither Speech Recognition

- There is a long tradition of questioning Machine Learning and Artificial Intelligence
  - Pierce chaired the ALPAC committee (which is credited for defunding research on Machine Translation)
  - He is also credited with defunding speech research with this letter to JASA: Whither Speech Recognition
- It is convenient to dismiss Pierce's criticisms (because they are so inconvenient),
  - but Pierce is a force that should be taken seriously.
- At Bell Labs,
  - He made contributions to speech including a coding standard, still used in WAV files today.
  - He was also involved in more significant projects such as the transistor and satellite telecommunications, and served as Vice President of Research.
- Many of the arguments in the ALPAC report are still relevant today.
  - The full report is well worth reading, and can be found here.
- It is ok for the next generation to reject positions held by previous generations,
  - but it isn't right to reject a position without reading it first.

Topic 8: WMD

- A recent best seller, Weapons of Math Destruction (WMD),
  - continues the tradition of questioning Machine Learning
  - by pointing out that machines (opaque black boxes) are making lots of important decisions like:
    - Who gets into a good school
    - Who gets a loan
    - Who goes to jail
- If machines are merely optimizing an objective function like $$
  - They will do so for better or for worse.
- How can society enforce policies about fairness
  - If even those of us who built the machines
  - Have little understanding of what the optimizations are doing and why?

Topic 9: Minsky & Chomsky

- Implications for currently popular methods in machine learning
  - Minsky and Chomsky were both at Harvard in the 1950s,
    - and they both started their careers by questioning the received wisdom at the time.
  - Perceptrons is a reaction to machine learning
    - and Syntactic Structures is a reaction to ngram language models.
  - While Minsky and Chomsky disagree on many/most issues,
    - their work had a chilling impact on various research directions for several decades.
  - Many of the methods that they objected to have since regained popularity,
    - and many of their objections are being ignored and forgotten
    - (perhaps for good reasons, and perhaps not)
- On a more positive note, in addition to questioning the received wisdom at the time,
  - Syntactic Structures also inspired some great work in other fields (see how Knuth spent his honeymoon)
  - Work that appeals across multiple fields has a better chance of becoming highly cited
- Legacy & Popular Press
  - Chomsky has published an amazing number of books
    - A number of videos are available on Youtube and Netflix.
  - Minsky is less organized/disciplined (and more eclectic).
    - This TED talk ends with Minsky objecting to neural nets, as he ran out of time.

Topic 10: . How does science (and engineering) progress?

- Kuhn suggests the process has made significant progress over time
  - (unlike a pendulum swinging back and forth),
  - and that the progression is dramatic and far from incremental.
- According to Wikipedia,
  - Kuhn argues that the evolution of scientific theory does not emerge from the straightforward accumulation of facts,
    - but rather from a set of changing intellectual circumstances and possibilities.
- Kuhn dated the genesis of his book to 1947,
  - when he was a graduate student at Harvard University
  - and had been asked to teach a science class for humanities undergraduates
  - with a focus on historical case studies
- Kuhn later commented that until then,
  - *I'd never read an old document in science*
- Aristotle's Physics was astonishingly unlike
  - Isaac Newton's work in its concepts of matter and motion
- Kuhn wrote
  - *... as I was reading him,*
    - *Aristotle appeared not only ignorant of mechanics,*
    - *but a dreadfully bad physical scientist as well.*
  - *About motion, in particular, his writings seemed to me full of*
    - *egregious errors, both of logic and of observation.*
  - This was in an apparent contradiction with the fact that Aristotle was a brilliant mind
- While perusing Aristotle's Physics,
  - Kuhn formed the view that in order to properly appreciate Aristotle's reasoning,
    - one must be aware of the scientific conventions of the time.
  - Kuhn concluded that Aristotle's concepts were not “bad Newton,” just different.

Topic 11: Word2vec & Embeddings

- Embeddings have become popular recently with word2vec.
  - A popular NIPS paper has suggested that word2vec is a factored version of
    - my work with Hanks on pointwise mutual information (PMI).
  - Scatter plots comparing word2vec with PMI show a modest correlation, but far from perfect.
    - That is, word pairs with large PMI scores also tend to receive large word2vec scores (and vice versa),
      - but there are plenty of exceptions.

- Factoring is typically performed using an iterative optimization such as Stochastic gradient descent.
  - It is easy to download code (and precomputed vectors) for both Word2vec and GloVe.
  - It is popular these days to factor PMI with stochastic gradient descent,
    - though it isn't obvious why that should be better than
      - singular value decomposition and
      - Latent Semantic Analysis

Opportunities for future work

- Are embeddings better than PMI?
  - Maybe we don't want to fit PMI "too well"
  - Word2vec & GloVe has a bunch of hyper parameters,
    - and they tend to matter in practice (TACL)
- PMI ≈ M M^T
  - Imaginary solutions:
    - If PMI isn't positive definite,
    - then M might not exist
  - Workarounds:
    - imaginary numbers
    - left & right eigenvectors (U ≠ V)
      - PMI ≈ U D V^T
      - but SVD of PMI doesn't seem to work well in practice (don't fit PMI "too well")
- Degrees of freedom:
  - It is common for embeddings to use K=300 dimensions,
    - but most words don't appear K times in the corpus
  - It is hard to justify K parameters for a word that doesn't appear K times
- Alternatives to Embeddings
  - Sketches
    - Li & Church "A Sketch Algorithm for Estimating Two-Way and Multi-Way Associations" *Computational Linguistics* (2007)

Topic 12: The Resources Debate

- Examples of resources:
  - British National Corpus (BNC)
  - WordNet
- Embeddings can be viewed as a reaction to lexicography (e.g., building lexical resources by hand)
  - In my work with Hanks, we hoped to automate some of the drudgery,
  - but we didn't expect to replace the lexicographer with bots, or with turkers.
- It was a surprise when Wikipedia could compete with traditional encylopedias (see Nature article),
  - and soon thereafter, traditional enclopedias quickly disappeared.
- Rise of Resources
  - Conferences on linguistic resources such as LREC have done well over the years.
  - WordNet and FrameNet are highly cited.
  - NLTK offers convenient tools for estimating word similarity
    - (in ways that are quite different from pointwise mutual information and word2vec)
- Will interest in resources decline
  - with the rise of neural nets and unsupervised methods?

Topic 13: Smoothing

- When the counts, c, are large, p ≈ c/N,
  - But What is the probability of something we haven't seen (much)?
- Statistical transforms and smoothing probably remain important today in contexts such as embeddings.
  - Both Word2vec and GloVe talk about raising something to the ¾ power
  - GloVe introduces another transform to downweight small counts
- There is a long history on smoothing (going back to WW II)
  - Good Turing smoothing
  - Good-Turing Frequency Estimation Without Tears
  - A Bit of Progress in Language Modeling

Topic 14: PageRank & Algorithms of Graphs of Social Media

- Starting with Page Rank and the founders of Google,
  - it has become popular to model the web and social media as graphs.
  - Page rank can be viewed as eigen values of a graph (ref)
- Kleinberg has published a number of highly cited papers including:
  - Authoritative Sources in a Hyperlinked Environment
  - Maximizing the Spread of Influence through a Social Network
  - The Link Prediction Problem for Social Networks
  - The Small-World Phenomenon: An Algorithmic Perspective

First Assignment (Due WED midnight)

http://www.columbia.edu/~kc3109/

- Vote here on papers/videos/topics
  – you would like to cover in the course
  – as well as papers you would like to present
  – NOTE: there are ***TWO*** tabs
- The first assignment involves skimming as much of this work as possible
  – so you can cast reasonably informed votes.