

# Studying the History of Ideas Using Topic Models

**David Hall**  
Symbolic Systems  
Stanford University  
Stanford, CA 94305, USA  
dlwh@stanford.edu

**Daniel Jurafsky**  
Linguistics  
Stanford University  
Stanford, CA 94305, USA  
jurafsky@stanford.edu

**Christopher D. Manning**  
Computer Science  
Stanford University  
Stanford, CA 94305, USA  
manning@stanford.edu

## Abstract

How can the development of ideas in a scientific field be studied over time? We apply unsupervised topic modeling to the ACL Anthology to analyze historical trends in the field of Computational Linguistics from 1978 to 2006. We induce topic clusters using Latent Dirichlet Allocation, and examine the strength of each topic over time. Our methods find trends in the field including the rise of probabilistic methods starting in 1988, a steady increase in applications, and a sharp decline of research in semantics and understanding between 1978 and 2001, possibly rising again after 2001. We also introduce a model of the diversity of ideas, *topic entropy*, using it to show that COLING is a more diverse conference than ACL, but that both conferences as well as EMNLP are becoming broader over time. Finally, we apply Jensen-Shannon divergence of topic distributions to show that all three conferences are converging in the topics they cover.

## 1 Introduction

How can we identify and study the exploration of ideas in a scientific field over time, noting periods of gradual development, major ruptures, and the waxing and waning of both topic areas and connections with applied topics and nearby fields? One important method is to make use of citation graphs (Garfield, 1955). This enables the use of graph-based algorithms like PageRank for determining researcher or paper centrality, and examining whether their influence grows or diminishes over time.

However, because we are particularly interested in the change of *ideas* in a field over time, we have chosen a different method, following Kuhn (1962). In Kuhn’s model of scientific change, science proceeds by shifting from one paradigm to another. Because researchers’ ideas and vocabulary are constrained by their paradigm, successive incommensurate paradigms will naturally have different vocabulary and framing.

Kuhn’s model is intended to apply only to very large shifts in scientific thought rather than at the micro level of trends in research foci. Nonetheless, we propose to apply Kuhn’s insight that vocabulary and vocabulary shift is a crucial indicator of ideas and shifts in ideas. Our operationalization of this insight is based on the unsupervised topic model Latent Dirichlet Allocation (LDA; Blei et al. (2003)).

For many fields, doing this kind of historical study would be very difficult. Computational linguistics has an advantage, however: the ACL Anthology, a public repository of all papers in the *Computational Linguistics* journal and the conferences and workshops associated with the ACL, COLING, EMNLP, and so on. The ACL Anthology (Bird, 2008), and comprises over 14,000 documents from conferences and the journal, beginning as early as 1965 through 2008, indexed by conference and year. This resource has already been the basis of citation analysis work, for example, in the ACL Anthology Network of Joseph and Radev (2007). We apply LDA to the text of the papers in the ACL Anthology to induce topics, and use the trends in these topics over time and over conference venues to address questions about the development of the field.

Venue	# Papers	Years	Frequency
Journal	1291	1974–Present	Quarterly
ACL	2037	1979–Present	Yearly
EACL	596	1983–Present	~2 Years
NAACL	293	2000–Present	~Yearly
Applied NLP	346	1983–2000	~3 Years
COLING	2092	1965–Present	2 Years
HLT	957	1986–Present	~2 Years
Workshops	2756	1990–Present	Yearly
TINLAP	128	1975–1987	Rarely
MUC	160	1991–1998	~2 Years
IJCNLP	143	2005	—
Other	120	—	—

Table 1: Data in the ACL Anthology

Despite the relative youth of our field, computational linguistics has witnessed a number of research trends and shifts in focus. While some trends are obvious (such as the rise in machine learning methods), others may be more subtle. Has the field gotten more theoretical over the years or has there been an increase in applications? What topics have declined over the years, and which ones have remained roughly constant? How have fields like Dialogue or Machine Translation changed over the years? Are there differences among the conferences, for example between COLING and ACL, in their interests and breadth of focus? As our field matures, it is important to go beyond anecdotal description to give grounded answers to these questions. Such answers could also help give formal metrics to model the differences between the many conferences and venues in our field, which could influence how we think about reviewing, about choosing conference topics, and about long range planning in our field.

## 2 Methodology

### 2.1 Data

The analyses in this paper are based on a text-only version of the Anthology that comprises some 12,500 papers. The distribution of the Anthology data is shown in Table 1.

### 2.2 Topic Modeling

Our experiments employ Latent Dirichlet Allocation (LDA; Blei et al. (2003)), a generative latent variable model that treats documents as bags of words generated by one or more topics. Each document is char-

acterized by a multinomial distribution over topics, and each topic is in turn characterized by a multinomial distribution over words. We perform parameter estimation using collapsed Gibbs sampling (Griffiths and Steyvers, 2004).

Possible extensions to this model would be to integrate topic modelling with citations (e.g., Dietz et al. (2007), Mann et al. (2006), and Jo et al. (2007)). Another option is the use of more fine-grained or hierarchical model (e.g., Blei et al. (2004), and Li and McCallum (2006)).

All our studies measure change in various aspects of the ACL Anthology over time. LDA, however, does not explicitly model temporal relationships. One way to model temporal relationships is to employ an extension to LDA. The Dynamic Topic Model (Blei and Lafferty, 2006), for example, represents each year’s documents as generated from a normal distribution centroid over topics, with the following year’s centroid generated from the preceding year’s. The Topics over Time Model (Wang and McCallum, 2006) assumes that each document chooses its own time stamp based on a topic-specific beta distribution.

Both of these models, however, impose constraints on the time periods. The Dynamic Topic Model penalizes large changes from year to year while the beta distributions in Topics over Time are relatively inflexible. We chose instead to perform *post hoc* calculations based on the observed probability of each topic given the current year. We define  $\hat{p}(z|y)$  as the empirical probability that an arbitrary paper  $d$  written in year  $y$  was about topic  $z$ :

$$\begin{aligned}
\hat{p}(z|y) &= \sum_{d:t_d=y} \hat{p}(z|d)\hat{p}(d|y) \\
&= \frac{1}{C} \sum_{d:t_d=y} \hat{p}(z|d) \\
&= \frac{1}{C} \sum_{d:t_d=y} \sum_{z'_i \in d} I(z'_i = z)
\end{aligned} \tag{1}$$

where  $I$  is the indicator function,  $t_d$  is the date document  $d$  was written,  $\hat{p}(d|y)$  is set to a constant  $1/C$ .

## 3 Summary of Topics

We first ran LDA with 100 topics, and took 36 that we found to be relevant. We then hand-selected seed

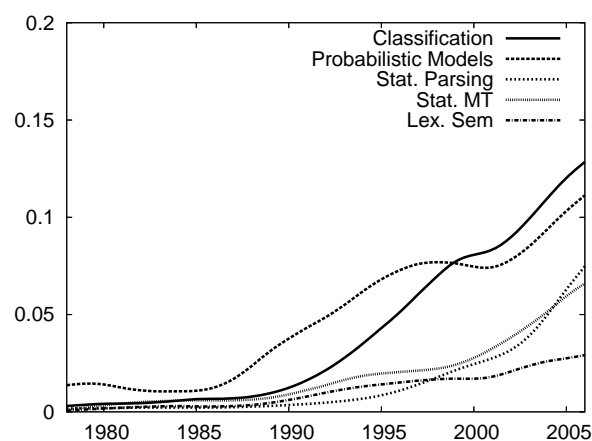


Figure 1: Topics in the ACL Anthology that show a strong recent increase in strength.

words for 10 more topics to improve coverage of the field. These 46 topics were then used as priors to a new 100-topic run. The top ten most frequent words for 43 of the topics along with hand-assigned labels are listed in Table 2. Topics deriving from manual seeds are marked with an asterisk.

## 4 Historical Trends in Computational Linguistics

Given the space of possible topics defined in the previous section, we now examine the history of these in the entire ACL Anthology from 1978 until 2006. To visualize some trends, we show the probability mass associated with various topics over time, plotted as (a smoothed version of)  $\hat{p}(z|y)$ .

### 4.1 Topics Becoming More Prominent

Figure 1 shows topics that have become more prominent more recently.

Of these new topics, the rise in *probabilistic models* and *classification/tagging* is unsurprising. In order to distinguish these two topics, we show 20 of the strongly weighted words:

**Probabilistic Models:** *model word probability set data number algorithm language corpus method figure probabilities table test statistical distribution function al values performance*

**Classification/Tagging:** *features data corpus set feature table word tag al test accuracy pos classification performance tags tagging text task information class*

Some of the papers with the highest weights for the *probabilistic models* class include:

- N04-1039 Goodman, Joshua. Exponential Priors For Maximum Entropy Models (HLT-NAACL, 2004)
- W97-0309 Saul, Lawrence, Pereira, Fernando C. N. Aggregate And Mixed-Order Markov Models For Statistical Language Processing (EMNLP, 1997)
- P96-1041 Chen, Stanley F., Goodman, Joshua. An Empirical Study Of Smoothing Techniques For Language Modeling (ACL, 1996)
- H89-2013 Church, Kenneth Ward, Gale, William A. Enhanced Good-Turing And CatCal: Two New Methods For Estimating Probabilities Of English Bigrams (Workshop On Speech And Natural Language, 1989)
- P02-1023 Gao, Jianfeng, Zhang, Min Improving Language Model Size Reduction Using Better Pruning Criteria (ACL, 2002)
- P94-1038 Dagan, Ido, Pereira, Fernando C. N. Similarity-Based Estimation Of Word Cooccurrence Probabilities (ACL, 1994)

Some of the papers with the highest weights for the *classification/tagging* class include:

- W00-0713 Van Den Bosch, Antal Using Induced Rules As Complex Features In Memory-Based Language Learning (CoNLL, 2000)
- W01-0709 Estabrooks, Andrew, Japkowicz, Nathalie A Mixture-Of-Experts Framework For Text Classification (Workshop On Computational Natural Language Learning CoNLL, 2001)
- A00-2035 Mikheev, Andrei. Tagging Sentence Boundaries (ANLP-NAACL, 2000)
- H92-1022 Brill, Eric. A Simple Rule-Based Part Of Speech Tagger (Workshop On Speech And Natural Language, 1992)

As Figure 1 shows, *probabilistic models* seem to have arrived significantly before *classifiers*. The probabilistic model topic increases around 1988, which seems to have been an important year for probabilistic models, including high-impact papers like A88-1019 and C88-1016 below. The ten papers from 1988 with the highest weights for the *probabilistic model* and *classifier* topics were the following:

- C88-1071 Kuhn, Roland. Speech Recognition and the Frequency of Recently Used Words (COLING)
- J88-1003 DeRose, Steven. Grammatical Category Disambiguation by Statistical Optimization. (CL Journal)
- C88-2133 Su, Keh-Yi, and Chang, Jing-Shin. Semantic and Syntactic Aspects of Score Function. (COLING)
- A88-1019 Church, Kenneth Ward. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. (ANLP)
- C88-2134 Sukhotin, B.V. Optimization Algorithms of Deciphering as the Elements of a Linguistic Theory. (COLING)
- P88-1013 Haigh, Robin, Sampson, Geoffrey, and Atwell, Eric. Project APRIL: a progress report. (ACL)
- A88-1005 Boggess, Lois. Two Simple Prediction Algorithms to Facilitate Text Production. (ANLP)
- C88-1016 Peter F. Brown, et al. A Statistical Approach to Machine Translation. (COLING)
- A88-1028 Oshika, Beatrice, et al.. Computational Techniques for Improved Name Search. (ANLP)
- C88-1020 Campbell, W.N. Speech-rate Variation and the Prediction of Duration. (COLING)

What do these early papers tell us about how

<b>Anaphora Resolution</b>	resolution anaphora pronoun discourse antecedent pronouns coreference reference definite algorithm
<b>Automata</b>	string state set finite context rule algorithm strings language symbol
<b>Biomedical</b>	medical protein gene biomedical wkh abstracts medline patient clinical biological
<b>Call Routing</b>	call caller routing calls destination vietnamese routed router destinations gorin
<b>Categorial Grammar</b>	proof formula graph logic calculus axioms axiom theorem proofs lambek
<b>Centering*</b>	centering cb discourse cf utterance center utterances theory coherence entities local
<b>Classical MT</b>	japanese method case sentence analysis english dictionary figure japan word
<b>Classification/Tagging</b>	features data corpus set feature table word tag al test
<b>Comp. Phonology</b>	vowel phonological syllable phoneme stress phonetic phonology pronunciation vowels phonemes
<b>Comp. Semantics*</b>	semantic logical semantics john sentence interpretation scope logic form set
<b>Dialogue Systems</b>	user dialogue system speech information task spoken human utterance language
<b>Discourse Relations</b>	discourse text structure relations rhetorical relation units coherence texts rst
<b>Discourse Segment.</b>	segment segmentation segments chain chains boundaries boundary seg cohesion lexical
<b>Events/Temporal</b>	event temporal time events tense state aspect reference relations relation
<b>French Function</b>	de le des les en une est du par pour
<b>Generation</b>	generation text system language information knowledge natural figure domain input
<b>Genre Detection</b>	genre stylistic style genres fiction humor register biber authorship registers
<b>Info. Extraction</b>	system text information muc extraction template names patterns pattern domain
<b>Information Retrieval</b>	document documents query retrieval question information answer term text web
<b>Lexical Semantics</b>	semantic relations domain noun corpus relation nouns lexical ontology patterns
<b>MUC Terrorism</b>	slot incident tgt target id hum phys type fills perp
<b>Metaphor</b>	metaphor literal metonymy metaphors metaphorical essay metonymic essays qualia analogy
<b>Morphology</b>	word morphological lexicon form dictionary analysis morphology lexical stem arabic
<b>Named Entities*</b>	entity named entities ne names ner recognition ace nes mentions mention
<b>Paraphrase/RTE</b>	paraphrases paraphrase entailment paraphrasing textual para rte pascal entailed dagan
<b>Parsing</b>	parsing grammar parser parse rule sentence input left grammars np
<b>Plan-Based Dialogue</b>	plan discourse speaker action model goal act utterance user information
<b>Probabilistic Models</b>	model word probability set data number algorithm language corpus method
<b>Prosody</b>	prosodic speech pitch boundary prosody phrase boundaries accent repairs intonation
<b>Semantic Roles*</b>	semantic verb frame argument verbs role roles predicate arguments
<b>Yale School Semantics</b>	knowledge system semantic language concept representation information network concepts base
<b>Sentiment</b>	subjective opinion sentiment negative polarity positive wiebe reviews sentence opinions
<b>Speech Recognition</b>	speech recognition word system language data speaker error test spoken
<b>Spell Correction</b>	errors error correction spelling ocr correct corrections checker basque corrected detection
<b>Statistical MT</b>	english word alignment language source target sentence machine bilingual mt
<b>Statistical Parsing</b>	dependency parsing treebank parser tree parse head model al np
<b>Summarization</b>	sentence text evaluation document topic summary summarization human summaries score
<b>Syntactic Structure</b>	verb noun syntactic sentence phrase np subject structure case clause
<b>TAG Grammars*</b>	tree node trees nodes derivation tag root figure adjoining grammar
<b>Unification</b>	feature structure grammar lexical constraints unification constraint type structures rule
<b>WSD*</b>	word senses wordnet disambiguation lexical semantic context similarity dictionary
<b>Word Segmentation</b>	chinese word character segmentation corpus dictionary korean language table system
<b>WordNet*</b>	synset wordnet synsets hypernym ili wordnets hypernyms eurowordnet hyponym ewn wn

Table 2: Top 10 words for 43 of the topics. Starred topics are hand-seeded.

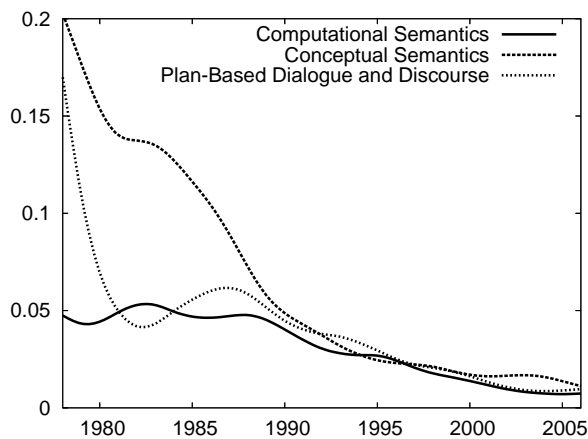


Figure 2: Topics in the ACL Anthology that show a strong decline from 1978 to 2006.

probabilistic models and classifiers entered the field? First, not surprisingly, we note that the vast majority (9 of 10) of the papers appeared in conference proceedings rather than the journal, confirming that in general new ideas appear in conferences. Second, of the 9 conference papers, most of them appeared in the COLING conference (5) or the ANLP workshop (3) compared to only 1 in the ACL conference. This suggests that COLING may have been more receptive than ACL to new ideas at the time, a point we return to in Section 6. Finally, we examined the background of the authors of these papers. Six of the 10 papers either focus on speech (C88-1010, A88-1028, C88-1071) or were written by authors who had previously published on speech recognition topics, including the influential IBM (Brown *et al.*) and AT&T (Church) labs (C88-1016, A88-1005, A88-1019). Speech recognition is historically an electrical engineering field which made quite early use of probabilistic and statistical methodologies. This suggests that researchers working on spoken language processing were an important conduit for the borrowing of statistical methodologies into computational linguistics.

## 4.2 Topics That Have Declined

Figure 2 shows several topics that were more prominent at the beginning of the ACL but which have shown the most precipitous decline. Papers strongly associated with the *plan-based dialogue* topic include:

- J99-1001 Carberry, Sandra, Lambert, Lynn. A Process Model For Recognizing Communicative Acts And Modeling Negotiation Subdialogues (CL, 1999)
- J95-4001 McRoy, Susan W., Hirst, Graeme. The Repair Of Speech Act Misunderstandings By Abductive Inference (CL, 1995)
- P93-1039 Chu, Jennifer. Responding To User Queries In A Collaborative Environment (ACL, 1993)
- P86-1032 Pollack, Martha E. A Model Of Plan Inference That Distinguishes Between The Beliefs Of Actors And Observers (ACL, 1986)
- T78-1017 Perrault, Raymond C., Allen, James F. Speech Acts As A Basis For Understanding Dialogue Coherence (Theoretical Issues In Natural Language Processing, 1978)
- P84-1063 Litman, Diane J., Allen, James F. A Plan Recognition Model For Clarification Subdialogues (COLING-ACL, 1984)

Papers strongly associated with the *computational semantics* topic include:

- J90-4002 Haas, Andrew R. Sentential Semantics For Propositional Attitudes (CL, 1990)
- P83-1009 Hobbs, Jerry R. An Improper Treatment Of Quantification In Ordinary English (ACL, 1983)
- J87-1005 Hobbs, Jerry R., Shieber, Stuart M. An Algorithm For Generating Quantifier Scopings (CL, 1987)
- C90-1003 Johnson, Mark, Kay, Martin. Semantic Abstraction And Anaphora (COLING, 1990)
- P89-1004 Alshawi, Hiyun, Van Eijck, Jan. Logical Forms In The Core Language Engine (ACL, 1989)

Papers strongly associated with the *conceptual semantics/story understanding* topic include:

- C80-1022 Ogawa, Hitoshi, Nishi, Junichiro, Tanaka, Kokichi. The Knowledge Representation For A Story Understanding And Simulation System (COLING, 1980)
- A83-1012 Pazzani, Michael J., Engelman, Carl. Knowledge Based Question Answering (ANLP, 1983)
- P82-1029 McCoy, Kathleen F. Augmenting A Database Knowledge Representation For Natural Language Generation (ACL, 1982)
- H86-1010 Ksiezzyk, Tomasz, Grishman, Ralph. An Equipment Model And Its Role In The Interpretation Of Nominal Compounds (Workshop On Strategic Computing - Natural Language, 1986)
- P80-1030 Wilensky, Robert, Arens, Yigal. PHRAN - A Knowledge-Based Natural Language Understander (ACL, 1980)
- A83-1013 Boguraev, Branimir K., Sparck Jones, Karen. How To Drive A Database Front End Using General Semantic Information (ANLP, 1983)
- P79-1003 Small, Steven L. Word Expert Parsing (ACL, 1979)

The declines in both computational semantics and *conceptual semantics/story understanding* suggests that it is possible that the entire field of natural language understanding and computational semantics broadly construed has fallen out of favor. To see if this was in fact the case we created a metatopic called *semantics* in which we combined various semantics topics (not including pragmatic topics like anaphora resolution or discourse coherence) including: *lexical semantics*, *conceptual semantics/story*

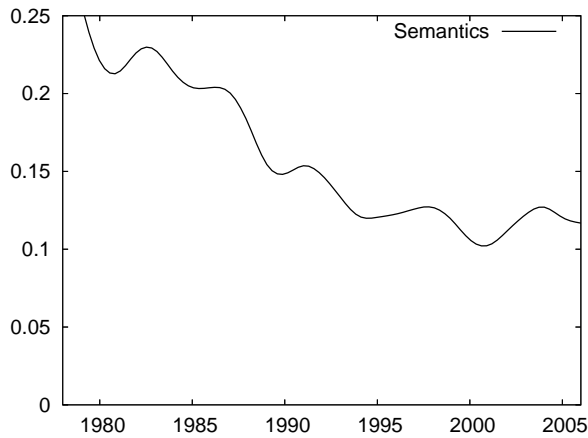


Figure 3: Semantics over time

*understanding, computational semantics, WordNet, word sense disambiguation, semantic role labeling, RTE and paraphrase, MUC information extraction, and events/temporal.* We then plotted  $\hat{p}(z \in S|y)$ , the sum of the proportions per year for these topics, as shown in Figure 3. The steep decrease in semantics is readily apparent. The last few years has shown a levelling off of the decline, and possibly a revival of this topic; this possibility will need to be confirmed as we add data from 2007 and 2008.

We next chose two fields, Dialogue and Machine Translation, in which it seemed to us that the topics discovered by LDA suggested a shift in paradigms in these fields. Figure 4 shows the shift in translation, while Figure 5 shows the change in dialogue.

The shift toward statistical machine translation is well known, at least anecdotally. The shift in dialogue seems to be a move toward more applied, speech-oriented, or commercial dialogue systems and away from more theoretical models.

Finally, Figure 6 shows the history of several topics that peaked at intermediate points throughout the history of the field. We can see the peak of *unification* around 1990, of *syntactic structure* around 1985 of *automata* in 1985 and again in 1997, and of *word sense disambiguation* around 1998.

## 5 Is Computational Linguistics Becoming More Applied?

We don't know whether our field is becoming more applied, or whether perhaps there is a trend towards new but unapplied theories. We therefore

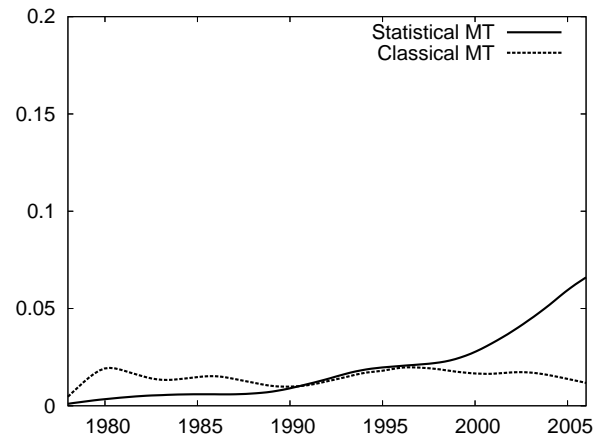


Figure 4: Translation over time

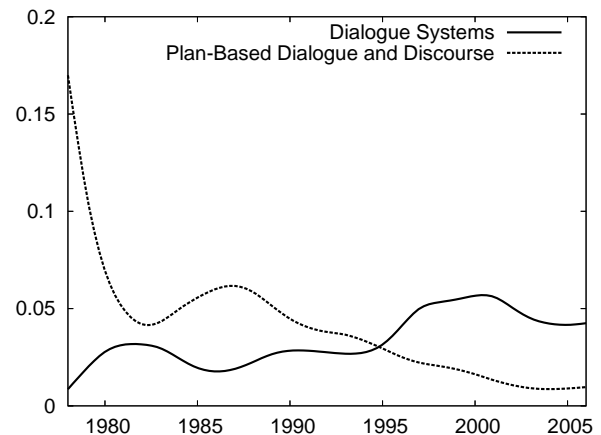


Figure 5: Dialogue over time

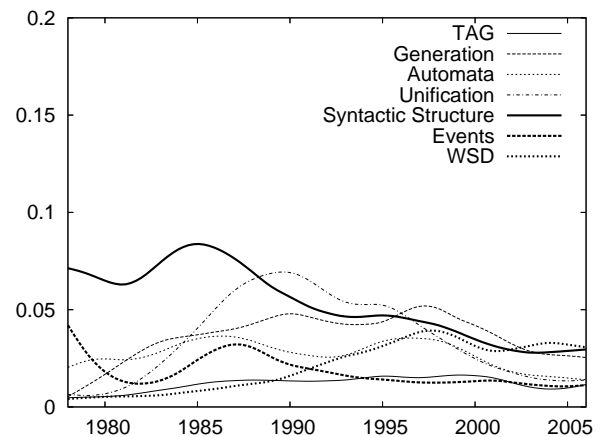


Figure 6: Peaked topics

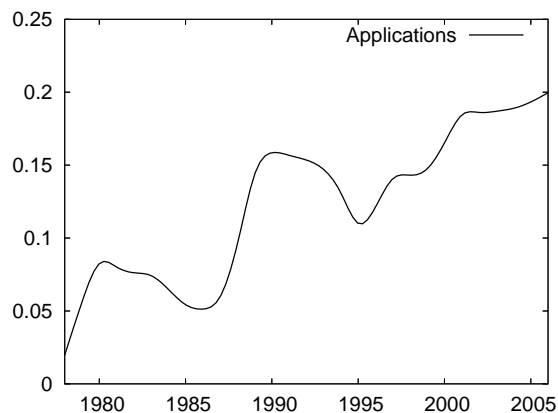


Figure 7: Applications over time

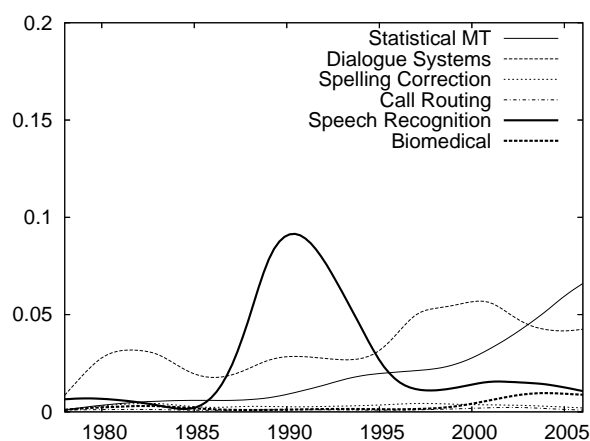


Figure 8: Six applied topics over time

looked at trends over time for the following applications: *Machine Translation*, *Spelling Correction*, *Dialogue Systems*, *Information Retrieval*, *Call Routing*, *Speech Recognition*, and *Biomedical* applications.

Figure 7 shows a clear trend toward an increase in applications over time. The figure also shows an interesting bump near 1990. Why was there such a sharp temporary increase in applications at that time? Figure 8 shows details for each application, making it clear that the bump is caused by a temporary spike in the *Speech Recognition* topic.

In order to understand why we see this temporary spike, Figure 9 shows the unsmoothed values of the *Speech Recognition* topic prominence over time.

Figure 9 clearly shows a huge spike for the years 1989–1994. These years correspond exactly to the DARPA Speech and Natural Language Workshop,

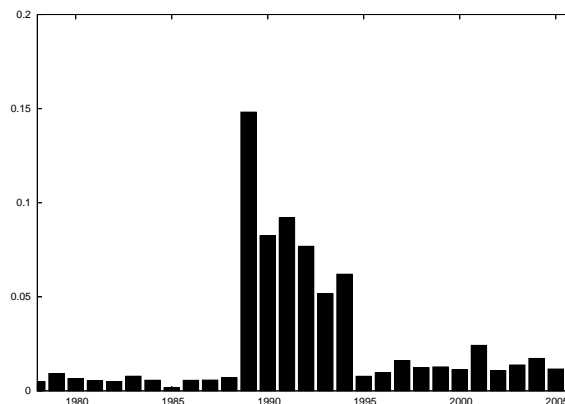


Figure 9: Speech recognition over time

held at different locations from 1989–1994. That workshop contained a significant amount of speech until its last year (1994), and then it was revived in 2001 as the Human Language Technology workshop with a much smaller emphasis on speech processing. It is clear from Figure 9 that there is still some speech research appearing in the Anthology after 1995, certainly more than the period before 1989, but it's equally clear that speech recognition is not an application that the ACL community has been successful at attracting.

## 6 Differences and Similarities Among COLING, ACL, and EMNLP

The computational linguistics community has two distinct conferences, COLING and ACL, with different histories, organizing bodies, and philosophies. Traditionally, COLING was larger, with parallel sessions and presumably a wide variety of topics, while ACL had single sessions and a more narrow scope. In recent years, however, ACL has moved to parallel sessions, and the conferences are of similar size. Has the distinction in breadth of topics also been blurred? What are the differences and similarities in topics and trends between these two conferences?

More recently, the EMNLP conference grew out of the Workshop on Very Large Corpora, sponsored by the Special Interest Group on Linguistic Data and corpus-based approaches to NLP (SIGDAT). EMNLP started as a much smaller and narrower

conference but more recently, while still smaller than both COLING and ACL, it has grown large enough to be considered with them. How does the breadth of its topics compare with the others?

Our hypothesis, based on our intuitions as conference attendees, is that ACL is still more narrow in scope than COLING, but has broadened considerably. Similarly, our hypothesis is that EMNLP has begun to broaden considerably as well, although not to the extent of the other two.

In addition, we're interested in whether the topics of these conferences are converging or not. Are the probabilistic and machine learning trends that are dominant in ACL becoming dominant in COLING as well? Is EMNLP adopting some of the topics that are popular at COLING?

To investigate both of these questions, we need a model of the topic distribution for each conference. We define the empirical distribution of a topic  $z$  at a conference  $c$ , denoted by  $\hat{p}(z|c)$  by:

$$\begin{aligned}\hat{p}(z|c) &= \sum_{d:c_d=c} \hat{p}(z|d)\hat{p}(d|c) \\ &= \frac{1}{C} \sum_{d:c_d=c} \hat{p}(z|d) \\ &= \frac{1}{C} \sum_{d:c_d=c} \sum_{z'_i \in d} I(z'_i = z)\end{aligned}\quad (2)$$

We also condition on the year for each conference, giving us  $\hat{p}(z|y, c)$ .

We propose to measure the breadth of a conference by using what we call *topic entropy*: the conditional entropy of this conference topic distribution. Entropy measures the average amount of information expressed by each assignment to a random variable. If a conference has higher topic entropy, then it more evenly divides its probability mass across the generated topics. If it has lower, it has a far more narrow focus on just a couple of topics. We therefore measured *topic entropy*:

$$H(z|c, y) = - \sum_{i=1}^K \hat{p}(z_i|c, y) \log \hat{p}(z_i|c, y) \quad (3)$$

Figure 10 shows the conditional topic entropy of each conference over time. We removed from the ACL and COLING lines the years when ACL

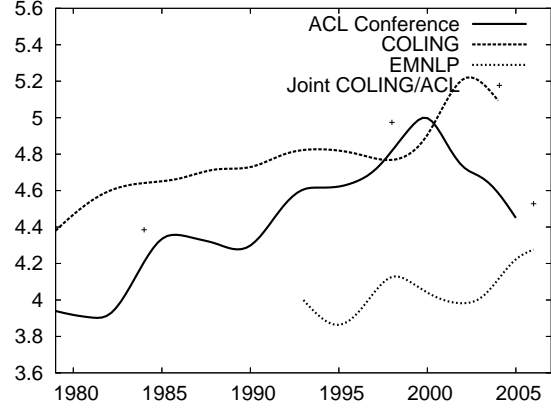


Figure 10: Entropy of the three major conferences per year

and COLING are colocated (1984, 1998, 2006), and marked those colocated years as points separate from either plot. As expected, COLING has been historically the broadest of the three conferences, though perhaps slightly less so in recent years. ACL started with a fairly narrow focus, but became nearly as broad as COLING during the 1990's. However, in the past 8 years it has become more narrow again, with a steeper decline in breadth than COLING. EMNLP, true to its status as a "Special Interest" conference, began as a very narrowly focused conference, but now it seems to be catching up to at least ACL in terms of the breadth of its focus.

Since the three major conferences seem to be converging in terms of breadth, we investigated whether or not the topic distributions of the conferences were also converging. To do this, we plotted the Jensen-Shannon (JS) divergence between each pair of conferences. The Jensen-Shannon divergence is a symmetric measure of the similarity of two pairs of distributions. The measure is 0 only for identical distributions and approaches infinity as the two differ more and more. Formally, it is defined as the average of the KL divergence of each distribution to the average of the two distributions:

$$\begin{aligned}D_{JS}(P||Q) &= \frac{1}{2}D_{KL}(P||R) + \frac{1}{2}D_{KL}(Q||R) \\ R &= \frac{1}{2}(P + Q)\end{aligned}\quad (4)$$

Figure 11 shows the JS divergence between each pair of conferences over time. Note that EMNLP



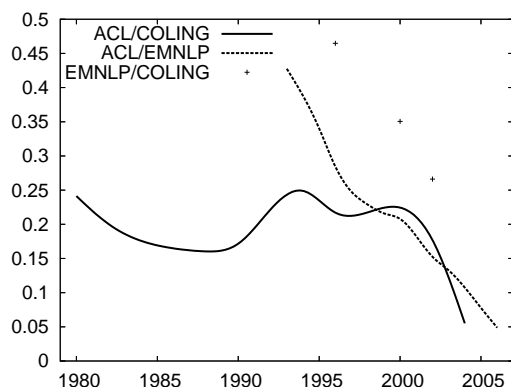


Figure 11: JS Divergence between the three major conferences

and COLING have historically met very infrequently in the same year, so those similarity scores are plotted as points and not smoothed. The trend across all three conferences is clear: each conference is not only increasing in breadth, but also in similarity. In particular, EMNLP and ACL’s differences, once significant, are nearly erased.

## 7 Conclusion

Our method discovers a number of trends in the field, such as the general increase in applications, the steady decline in semantics, and its possible reversal. We also showed a convergence over time in topic coverage of ACL, COLING, and EMNLP as well as an expansion of topic diversity. This growth and convergence of the three conferences, perhaps influenced by the need to increase recall (Church, 2005) seems to be leading toward a tripartite realization of a single new “latent” conference.

## Acknowledgments

Many thanks to Bryan Gibson and Dragomir Radev for providing us with the data behind the ACL Anthology Network. Also to Sharon Goldwater and the other members of the Stanford NLP Group as well as project Mimir for helpful advice. Finally, many thanks to the Office of the President, Stanford University, for partial funding.

## References

- Steven Bird. 2008. Association of Computational Linguists Anthology. <http://www.aclweb.org/anthology-index/>.
- David Blei and John D. Lafferty. 2006. Dynamic topic models. *ICML*.
- David Blei, Andrew Ng, , and Michael Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.
- D. Blei, T. Gri, M. Jordan, and J. Tenenbaum. 2004. Hierarchical topic models and the nested Chinese restaurant process.
- Kenneth Church. 2005. Reviewing the reviewers. *Comput. Linguist.*, 31(4):575–578.
- Laura Dietz, Steffen Bickel, and Tobias Scheffer. 2007. Unsupervised prediction of citation influences. In *ICML*, pages 233–240. ACM.
- Eugene Garfield. 1955. Citation indexes to science: A new dimension in documentation through association of ideas. *Science*, 122:108–111.
- Tom L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *PNAS*, 101 Suppl 1:5228–5235, April.
- Yookyung Jo, Carl Lagoze, and C. Lee Giles. 2007. Detecting research topics via the correlation between graphs and texts. In *KDD*, pages 370–379, New York, NY, USA. ACM.
- Mark T. Joseph and Dragomir R. Radev. 2007. Citation analysis, centrality, and the ACL anthology. Technical Report CSE-TR-535-07, University of Michigan. Department of Electrical Engineering and Computer Science.
- Thomas S. Kuhn. 1962. *The Structure of Scientific Revolutions*. University Of Chicago Press.
- Wei Li and Andrew McCallum. 2006. Pachinko allocation: DAG-structured mixture models of topic correlations. In *ICML*, pages 577–584, New York, NY, USA. ACM.
- Gideon S. Mann, David Mimno, and Andrew McCallum. 2006. Bibliometric impact measures leveraging topic analysis. In *JCDL ’06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 65–74, New York, NY, USA. ACM.
- Xuerui Wang and Andrew McCallum. 2006. Topics over time: a non-Markov continuous-time model of topical trends. In *KDD*, pages 424–433, New York, NY, USA. ACM.