A Study of Academic Collaboration in Computational Linguistics with Latent Mixtures of Authors

Nikhil Johri, Daniel Ramage, Daniel McFarland and Daniel Jurafsky

Stanford University

Motivating Questions

- What is the value added from collaboration?
  - Division of labor?
  - Mixture of individual contributions?
  - New, synergistic ideas?
- Can we model a collaborative publication as a sum of its authors?
- What are the characteristics of influential collaborations?

Dataset

- ACL Corpus
  - 12,000+ papers
  - Ranges from 1965 to 2009
  - Collaborations
    - 8,000+ papers with 2 or more authors

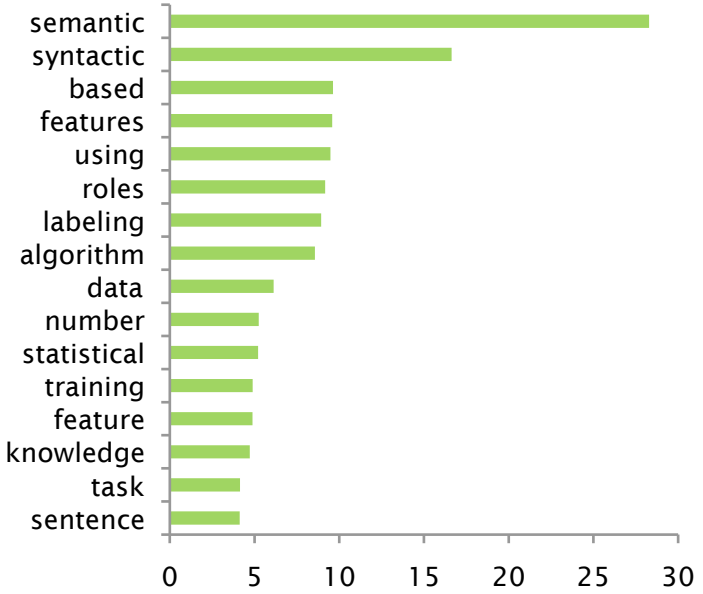
- Citation information
  - Obtained from ACL Anthology Network (AAN)

Morphological features help POS tagging of unknown words across language varieties (2005)

By Huihsin Tseng, Daniel Jurafsky and Christopher Manning

Part-of-speech tagging, like any supervised statistical NLP task, is more difficult when test sets are very different from training sets, for example when tagging across genres or language varieties. We examined the problem of POS tagging of different varieties of Mandarin Chinese. An analytic study first showed that unknown words were a major source of difficulty in cross-variety tagging. Unknown words in English tend to be proper nouns. By contrast, we found that Mandarin unknown words were mostly common nouns and verbs. We showed these results are caused by the high frequency of morphological compounding in Mandarin…

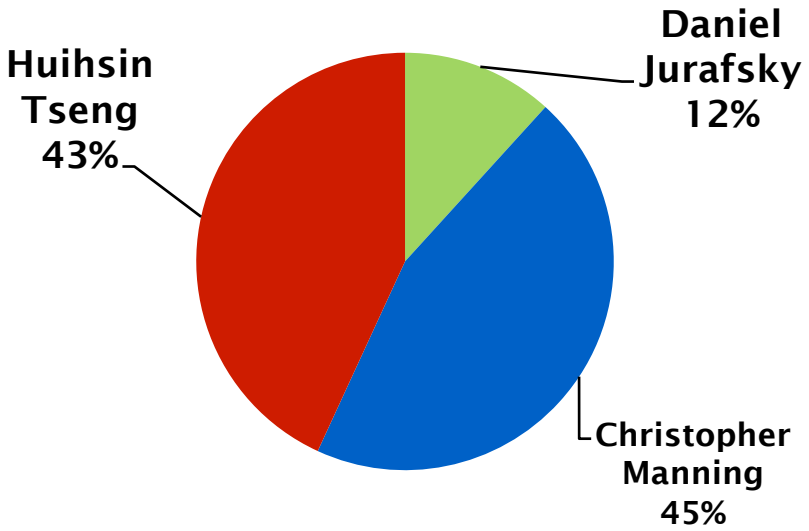
Authorial word distribution for Dan Jurafsky in 2004



Morphological features help POS tagging of unknown words across language varieties (2005)

By Huihsin Tseng, Daniel Jurafsky and Christopher Manning

Part-of-speech tagging, like any supervised statistical NLP task, is more difficult when test sets are very different from training sets, for example when tagging across genres or language varieties. We examined the problem of POS tagging of different varieties of Mandarin Chinese. An analytic study first showed that unknown words were a major source of difficulty in cross-variety tagging. Unknown words in English tend to be proper nouns. By contrast, we found that Mandarin unknown words were mostly common nouns and verbs. We showed these results are caused by the high frequency of morphological compounding in Mandarin…



Methodology

- Latent Mixture of Authors
  - Compute author ‘signatures’ as distributions over terms
  - Compute new signatures for each author for each year
  - Implemented using **Labeled LDA** (Ramage et al.)

- Cosine Similarities Between
  - Document word vector
  - Authors’ signatures on earlier documents

Labeled LDA

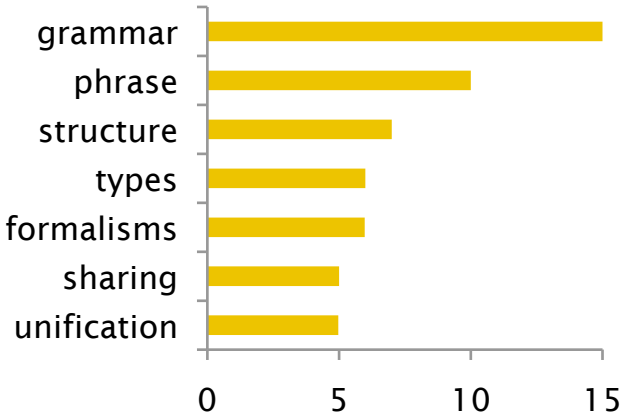
- Variation of Latent Dirichlet Allocation (LDA)
- Topics are constrained to be about specific tags associated with the documents
- In this case, tags = authors
- Result: a probabilistic term 'signature' for each author

Chris Manning	Ronald Kaplan	Martin Kay
model	grammar	phrases
models	lexical	grammars
entailment	functional	reversible
inference	lfg	parsing
local	context	translation
semantic	free	generation
named	formal	formalisms

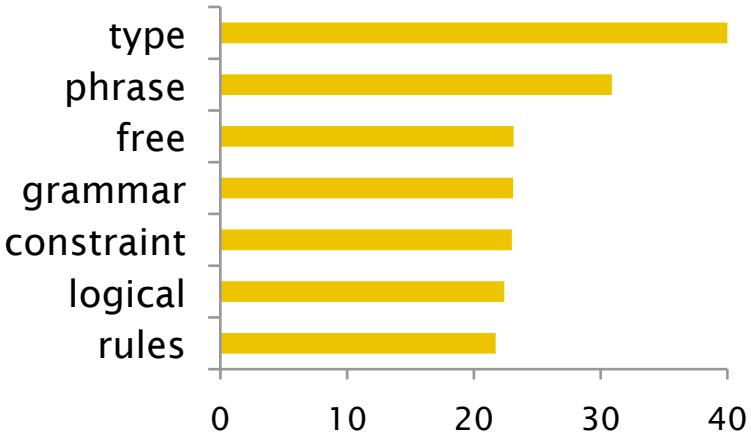
Labeled LDA

- We run a separate model for each year
  - Only papers written up to and including the preceeding year are considered

Fernando Pereira, 1985



Fernando Pereira, 2009



Types of Collaborations

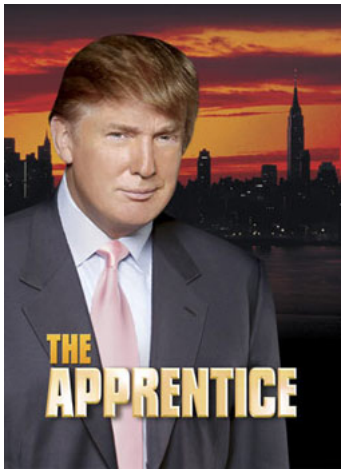
- Identify each document's collaboration type based on:
  - Presence of unestablished authors
  - Number of established authors that the document resembles
  - Similarity of document to established authors' previous work
- Only consider papers with > 0 established authors
  - Not enough track record for new authors

MaxSim Score

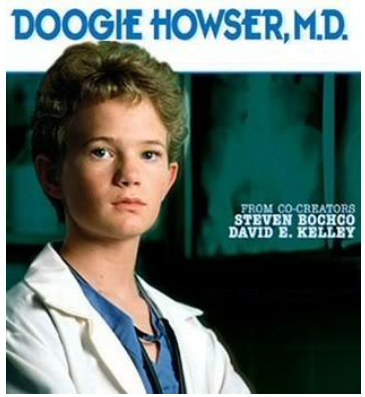
- Highest cosine similarity of paper with any of its established authors' signatures from the preceding year
- Range from 0.01 to 0.83, mean and median ~0.16
- Low MaxSim score ~ new area for all established authors involved

Collaborations with new authors

- One or more established authors
- One or more new authors
- High MaxSim score
  - Indicates similarity to established author's work
  - ***Apprentice*** style paper



- Low MaxSim score
  - Different area from established authors' usual field(s)
  - New ideas brought in by new author or authors
  - ***New blood*** style paper



Apprenticeship Examples

- Some Computational Properties of Tree Adjoining Grammars (1985)
  - K. Vijay-Shankar and Aravind K. Joshi
  - Joshi had already published papers on TAGs
  - Closer to apprenticeship than new blood

- Improvements in Phrase-Based Statistical Machine Translation (2004)
  - Richard Zens and Hermann Ney
  - Looked heavily like Ney, incremental in a sequence of similar papers
  - Had a high MaxSim score

New Blood Example

- Thumbs up? Sentiment Classification using Machine Learning Techniques (2002)
  - Lillian Lee, Bo Pang and Shivakumar Vaithyanathan
    - Lillian Lee established author (professor at Cornell)
    - Bo Pang new grad student
    - Vaithyanathan ML researcher (no prior ACL pubs)
  - This was Lillian Lee's first paper on sentiment
    - Previous work mainly in distributional word clustering
  - Heavy new blood influence from authors Pang and Vaithyanathan

New Blood Papers - Examples

- Finding Parts in Very Large Corpora (1999)
  - Matthew Berland, Eugene Charniak
    - Eugene Charniak established author
    - Matthew Berland undergrad
  - Paper was about semantic extraction of parts from a whole
    - Not a whole lot like Charniak's previous works
  - Berland's senior thesis, and evident new blood influence

Catalyst Example

- Answer Extraction (2000)
  - Steven Abney, Michael Collins, Amit Singhal
    - Abney
      - ***Established*** - worked on Stochastic Attribute-Value Grammars
    - Collins
      - ***Established*** - worked on Parsing
    - Singhal
      - Well known in IR community but considered ***new*** in ACL
  - Paper involves Collins and Abney working in Singhal's area (information retrieval)
    - Singhal was the ***catalyst***

Collaborations without new authors

- Two or more established authors
- High Maxsim score
  - Indicates a ***regular*** collaboration
  - Consistent in at least one author's line of work

- Low Maxsim score
  - Indicates a ***synergistic*** collaboration
  - New area for all authors involved

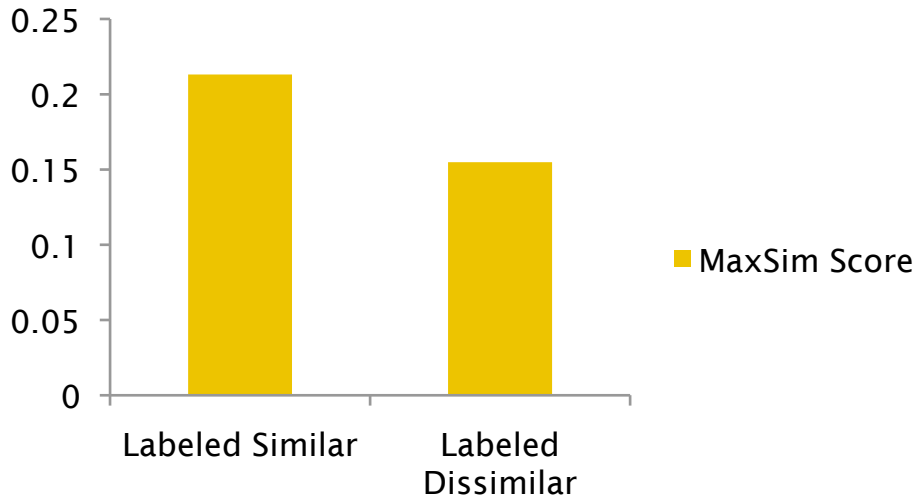
Synergy Spectrum Examples

- Feature Structures Based Tree Adjoining Grammars (1988)
  - K. Vijay-Shankar and Aravind K. Joshi
  - Vijay-Shankar now established, several publications
  - Paper highly resembled previous work
  - ***Low Synergy (or Regular)*** Paper

- Inside-Outside Reestimation From Partially Bracketed Corpora (1992)
  - Fernando Pereira and Yves Schabes
  - Resembles neither author's previous work
  - ***High Synergy*** Paper

Evaluation (I) – Predicting Human Labels

- Expert Annotation
  - 1 expert labeled 120 papers as similar or dissimilar to previous work of the established authors
  - Binary classification accuracy: better than chance results ~ 62.5%
    - Required applying thresholds on MaxSim score



Evaluation (II) – First Author Prediction

- Use similarity scores to predict first author
- Performs better than random chance

Predictor Feature	Accuracy
Random Chance	37.4%
Alphabetical Ordering	43.6%
Author Signature Similarity	45.2%
Frequency Estimator	56.1%

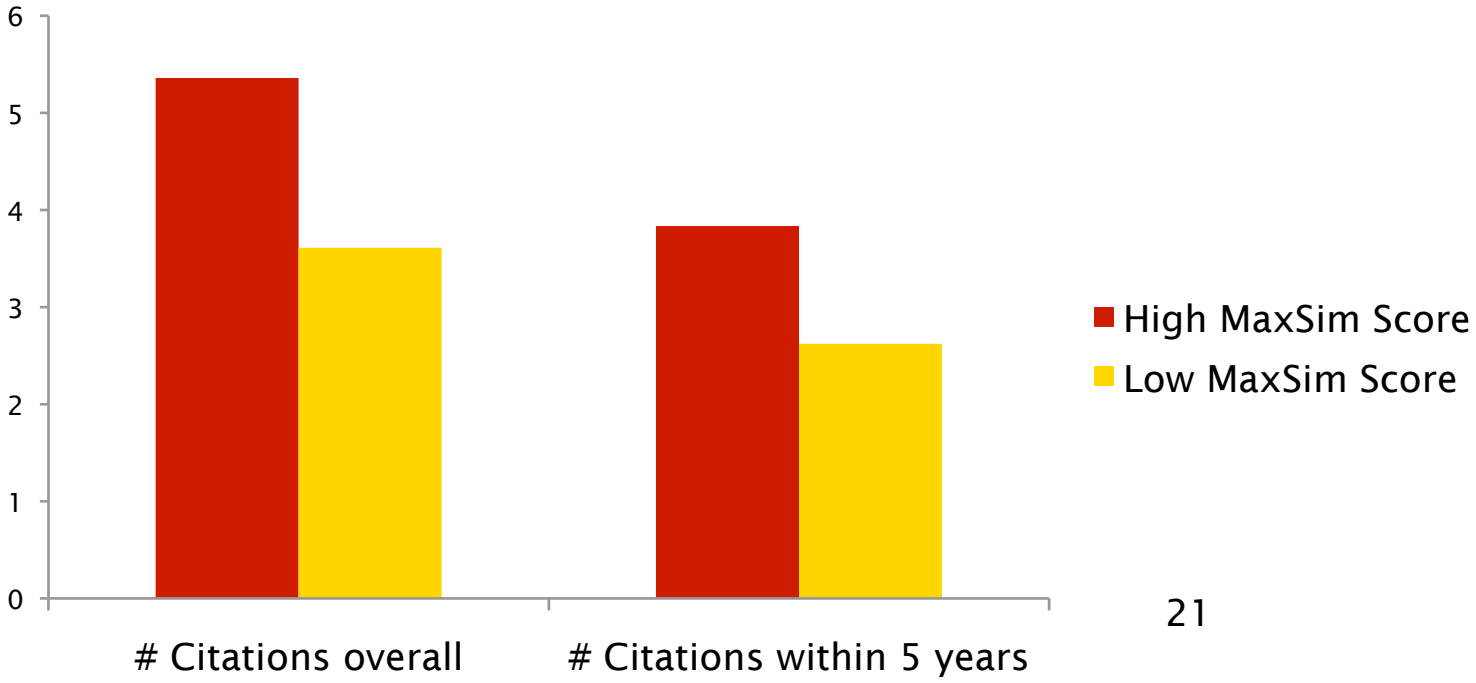
Potential Questions

- What collaboration styles result in higher citation counts?
- Are some subfields more likely to exhibit certain collaboration styles than others?
- Which authors are more like ‘hedgehogs’? Which are more like ‘foxes’?

Citation Analysis

- Collaborations benefit most when at least one author has built a track record in the area

- New graduate students have higher chances of early success working as apprentices



Subfield Analysis

- Average MaxSim scores by subfield
  - High ~ more rigid, formal, requires training, lab-oriented
  - Low ~ more flexible, less defined, open to novelty

Topic	Score
Statistical Machine Translation	0.2695
Prosody	0.2631
Speech Recognition	0.2511
Non-Statistical Machine Translation	0.2471
Word Sense Disambiguation	0.2380

High Similarity Scores

Topic	Score
Question Answering	0.1335
Sentiment Analysis	0.1399
Dialog Systems	0.1417
Spelling Correction	0.1462
Summarization	0.1511

Low Similarity Scores

Hedgehogs vs Foxes (Berlin 1953)

- Identification of 'hedgehogs' and 'foxes'
  - Hedgehogs specialize in a single area
    - Highest average similarity scores to previous work
  - Foxes dabble in several areas
    - Lowest average similarity scores to previous work

Topic	Score
Koehn, Philipp	0.4346
Pedersen, Ted	0.4115
Och, Franz Josef	0.3967
Ney, Hermann	0.3730
Sumita, Eiichiro	0.3671

Top 'Hedgehog' Authors

Topic	Score
Marcus, Mitchell P.	0.0999
Pustejovsky, James D.	0.1047
Pereira, Fernando C. N.	0.1434
Allen, James F.	0.1446
Hahn, Udo	0.1501

Top 'Fox' Authors

Conclusion & Future Work

- Presented a system that analyses collaborations based on authorial deviation
- Modeled publications as a combination of author contributions
- Explored a number of social science questions relating to academic collaboration in ACL
- In the future, we hope to extend this work to larger corpora - PubMed or ISI