Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers

Sonal Gupta Christopher Manning
Natural Language Processing Group
Department of Computer Science
Stanford University

Goals

- Extract key aspects of scientific papers
  - Main contribution
  - Techniques used
  - Domain or task
- Use them to study dynamics of research
- Understand how science is progressing in terms of new *problems*, *techniques* and *applications* in the papers published
  - What influenced statistical machine translation most?
  - Has a field ‘matured’ to be a used as a tool or intermediate subroutine to solve other problems (e.g. POS tagging)?

Key Aspects

Given a paper's abstract

> *We propose a new framework for **predicting links between entities in a graph**. Our system uses a new **ABC algorithm** and it performs better than the XYZ algorithm. We test our system on **Facebook**.*

Predict

- **FOCUS** (main contribution)
  - *predicting links between entities in a graph*
- **TECHNIQUE** (tools or algorithms used)
  - *ABC algorithm*
- **DOMAIN** (problem or task at hand)
  - *Facebook; predicting links between entities in a graph*

Why we need FOCUS?

Abstract 1

*We work on improving the **speech recognition** system using more linguistic features. We use a **discriminative classifier** with our new features and show that our system performs better than state-of-the-art techniques.*

Abstract 2

*We work on a new **regularizer for discriminative classifiers**. Our system performs better than the existing systems on the **speech recognition** task.*

Focus is different even though technique and domain are same!

A DOMAIN for me is a TECHNIQUE for you

.. AND VICE VERSA

- Part-of-speech tagging uses word segmentation and HMM
  - TECHNIQUE: word segmentation; HMM
  - DOMAIN: part-of-speech tagging
- Parsing uses part-of-speech tagging as an intermediate tool
  - TECHNIQUE: part-of-speech tagging
  - DOMAIN: parsing

Why BoW based techniques fail?

- Bag-of-Words techniques assume words are independent
  - Cannot tell whether a phrase is a FOCUS, a TECHNIQUE, or a DOMAIN
- Topic models (e.g. LDA) give higher level topics, like, ‘parsing’, ‘semantics’

- Our approach: Information extraction using dependency patterns

Our approach: Dependency Patterns

- Find patterns in dependency graph of sentences
  - In first iteration, 13 patterns for FOCUS, 7 for TECHNIQUE and 15 for DOMAIN

FOCUS

propose —direct-object→ <phrase tree>

work —prep_on→ <phrase tree>

DOMAIN

algorithm —prep_for→ <phrase tree>

task —prep_of→ <phrase tree>

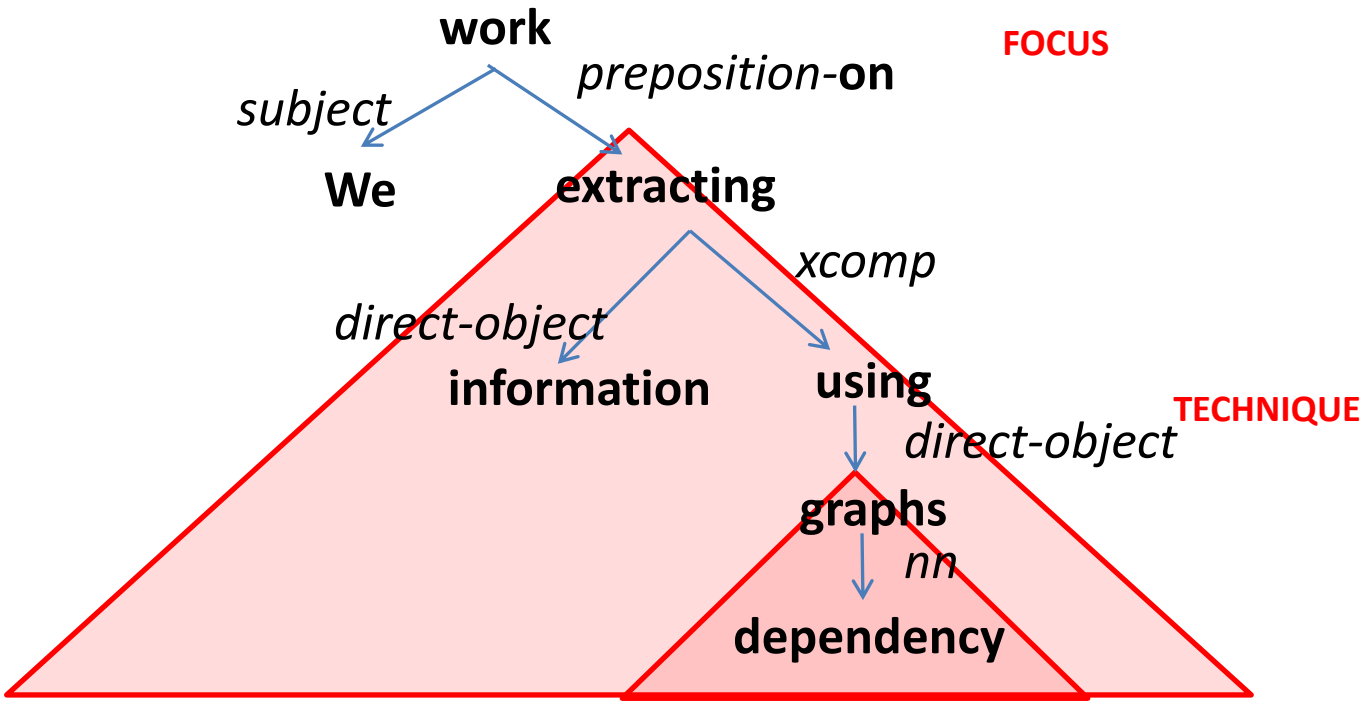
TECHNIQUE

use —direct-object→ <phrase tree>

apply —direct-object→ <phrase tree>

Learn new patterns using bootstrapping!

Example



Our semantic patterns will extract "extracting information using dependency graphs" as FOCUS, and "dependency graphs" as TECHNIQUE.

Learned Patterns using Bootstrapping

TECHNIQUE			DOMAIN		
model	**nn** →	<phrase tree>	**improve**	**direct-object** →	<phrase tree>
rules	**nn** →	<phrase tree>	**used**	**prep_for** →	<phrase tree>
extracting	**direct-object** →	<phrase tree>	**evaluation**	**nn** →	<phrase tree>
identify	**direct-object** →	<phrase tree>	**parsing**	**nn** →	<phrase tree>
constraints	**amod** →	<phrase tree>	**domain**	**nn** →	<phrase tree>
based	**prep_on** →	<phrase tree>	**applied**	**prep_to** →	<phrase tree>
…			…		

nn = any noun that modifies the head noun

Example: Phrases Extracted

- Studying the History of Ideas Using Topic Models
  - **FOCUS**: studying the history of ideas using topic
  - **TECHNIQUE**: latent dirichlet allocation; topic; unsupervised topic; historical trends; that all three conferences are converging in the topics
  - **DOMAIN**: studying the history of ideas; topic; model of the diversity of ideas , topic entropy; probabilistic

Example: Phrases Extracted

- A Bayesian Hybrid Method For Context-Sensitive Spelling Correction
  - **FOCUS**: new hybrid method , based on bayesian classifiers; bayesian hybrid method for context sensitive spelling correction
  - **TECHNIQUE**: decision lists; bayesian; bayesian classifiers; ambiguous; part-of-speech tags; methods using decision lists; single strongest piece of evidence; spelling
  - **DOMAIN :** context-sensitive spelling correction; for context-sensitive spelling correction; spelling

Dataset

- Computational linguistics community using the ACL Anthology dataset (Radev et al. 09, Bird et al. 08)
  - 10,889 abstracts from 1985 to 2009
- Extracted 25,525 phrases for FOCUS, 24,430 for TECHNIQUE, and 33,203 for DOMAIN
- Test set: 462 abstracts labeled by hand
- Inter-annotator agreement: 30 abstracts, each labeled by two PhD candidates in computational linguistics

Extraction Results

Approach	F1	Precision	Recall
FOCUS			
Our system	42.41	31.38	65.39
Inter-annotator agreement	53.33	50.80	56.14
TECHNIQUE			
Seed Patterns	19.72	19.83	19.61
Our system	36.04	27.83	51.14
Inter-annotator agreement	72.02	66.81	78.11
DOMAIN			
Seed Patterns	23.86	23.86	23.87
Our system	37.75	32.23	45.56
Inter-annotator agreement	72.31	75.58	69.32

Challenges in Using Patterns

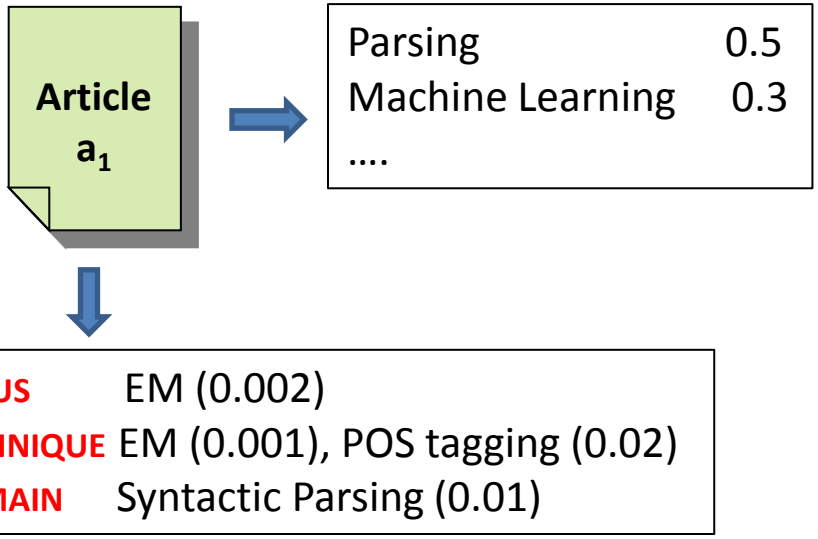
- Intuitions about what their systems can be useful for
  - E.g. " .. we can use our system in parsing, semantic role labeling, and other NLP tasks"
- Previous approaches and techniques listed in the abstracts
- Generic phrases and coreferent phrases
  - "we use a novel algorithm to.."
  - "we use the system to get .."
- Phrases like "the parsing technique we present.." – confusing for patterns

What to do with these key aspects?

- Influence of communities on each other
  - w.r.t. techniques borrowed (e.g. HMM from speech recognition)
  - and adoption of tools produced (e.g. part-of-speech tagging)

Defining Communities from Topics

- Communities: Topics using Latent Dirichlet Allocation (LDA) on full text of the articles
  - LDA gives soft, probabilistic article-to-community scores in an unsupervised manner
  - For each article, LDA gives probabilities over communities/topics
  - Topics “parsing”, “statistical MT”, “probability theory” are treated as communities
- Our case study is on the 74 communities (i.e. topics) of computational linguistics



technique-score(Parsing, EM, a_1) =
*0.001*0.5*

*all-score(*Parsing, EM, a_1) =
*(0.002+0.001+0)*0.5*

score that a community uses a phrase from an article as a TECHNIQUE:

$$technique\text{-}score(community, phrase, article) = 1/z_p \underbrace{count(phrase \in technique \mid article)}_{\text{Tf-idf like score using extraction}} \underbrace{P(community \mid article, \theta)}_{\text{From topic model}}$$

Influence

Influence of community c_1 on community c_2 in year y:

How many phrases in **any of the three classes** from articles in c_1 published in y are used as **TECHNIQUES** in articles in c_2 published at a **later date**?

$$Influence(c_1, c_2, p, a_1)$$

$$= all\text{-}score(c_1, p, a_1) \sum_{a_2, y_{a_2} > y_{a_1}} technique\text{-}score(c_2, p, a_2) C(a_2, a_1)$$

If a_2 cited a_1, 1
Otherwise, 0.5

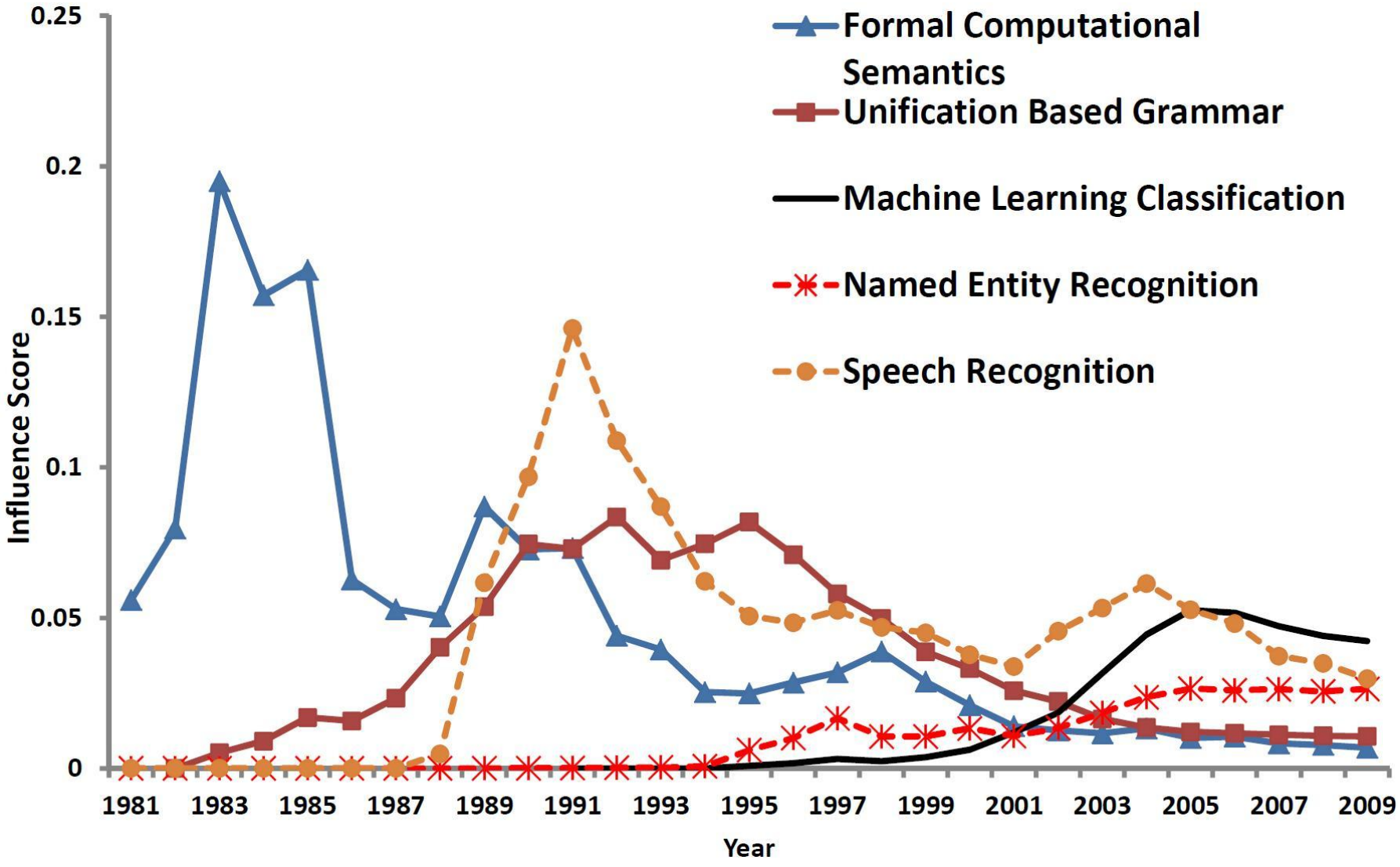
$$Influence(c_1, c_2, y) = \sum_{p, y_a = y} Influence(c_1, c_2, p, a)$$

Communities (decreasing order of influence)	Most influential Phrases
Speech Recognition	EM; HMM; language; contextually; segment; context independent phone; snn hidden markov;
Probability Theory	HMM; maximum entropy; language; EM; merging; EM HMM; natural language; variable memory markov;
Bilingual Word Alignment	HMM; EM; maximum entropy; spectral clustering; statistical alignment; CRFs , a discriminative; statistical word alignment; string to Tree
POS Tagging	maximum entropy; machine learning; EM HMM; POS information; decision tree; hidden markov; transformation based error driven learning; entropy; POS tagging
Machine Learning Classification	SVMs; ensemble; machine learning; gaussian mixture; EM; flat; weak classifiers; statistical machine learning

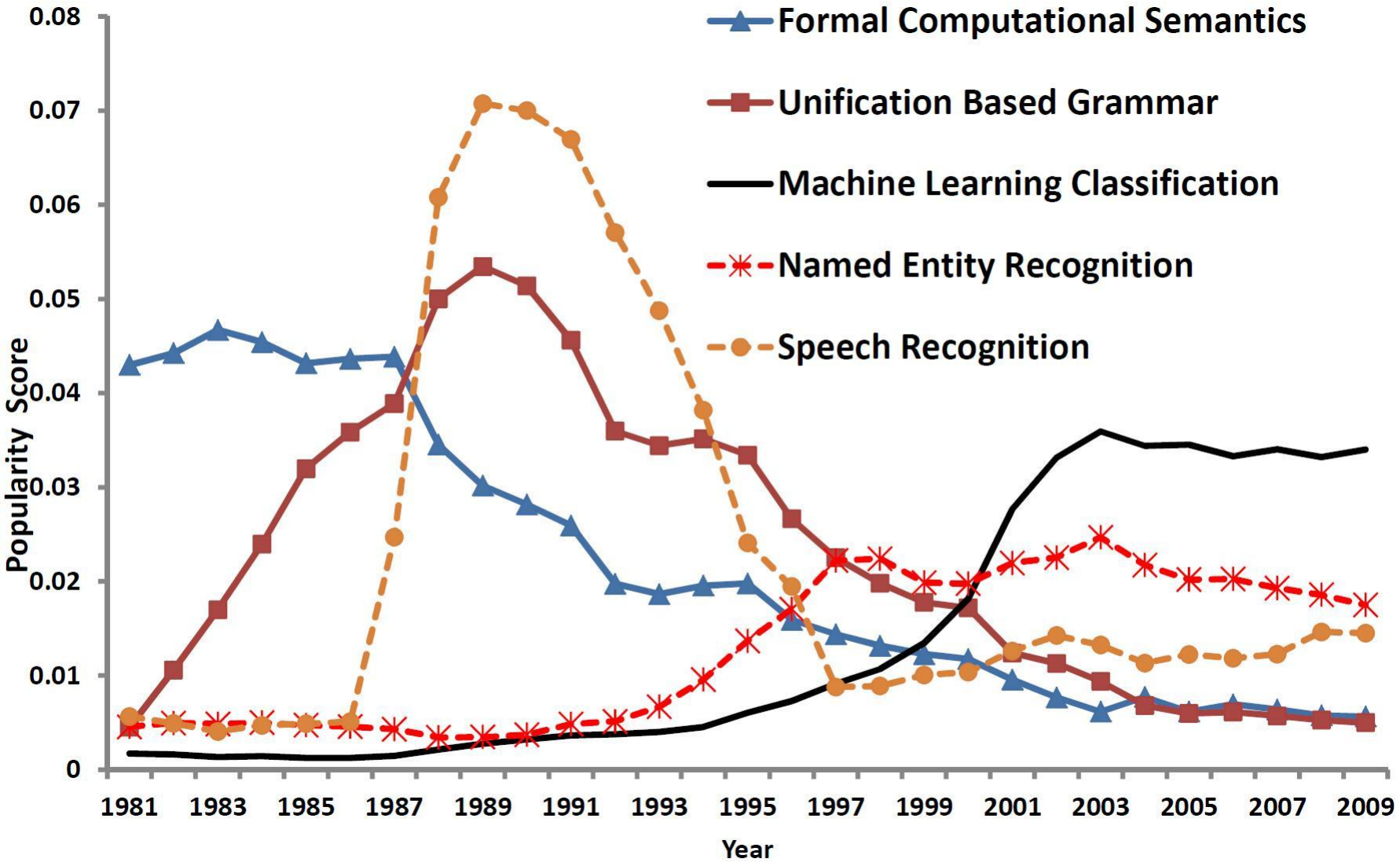
Influence vs. Popularity

- Influence of community c_1 on community c_2
  - How many DOMAIN, TECHNIQUE and FOCUS phrases of papers in c_1 were used as TECHNIQUES by papers published at a later date in c_2
- Related work: Popularity
  - Expected numbers of papers published in year *y*
  - Previous work (Hall et al. 2008, Griffiths and Steyvers 2004, …) have studied this
  - Different from influence!

Influence of Communities

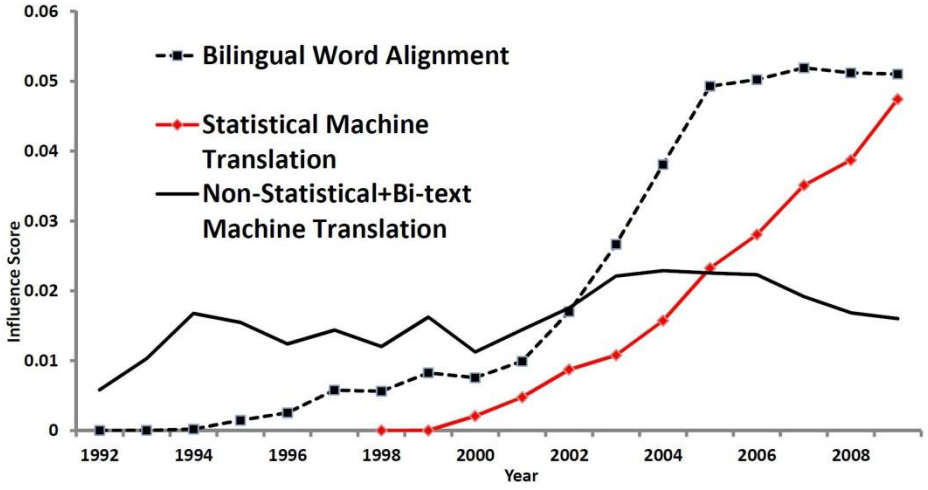


Popularity of Communities

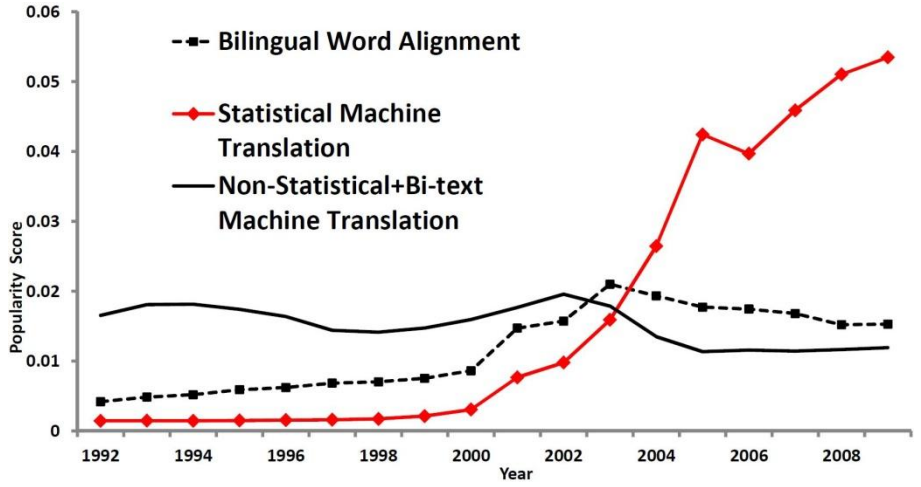


Influence vs. Popularity of MT Communities

Influence



Popularity



Community	Communities that have influenced most (descending order)
Named Entity Recognition	Chunking/Memory Based Models; Discriminative Sequence Models; POS Tagging; Machine Learning Classification; Coherence Relations; Biomedical NER; Bilingual Word Alignment
Statistical Parsing	Probability Theory; POS Tagging; Discriminative Sequence Models; Speech Recognition; Parsing; Syntactic Theory; Clustering+DistributionalSimilarity; Chunking/Memory Based Models
Word Sense Disambiguation	Clustering + DistributionalSimilarity; Machine Learning Classification; Dictionary Lexicons; Collocations/Compounds; Syntax; Speech Recognition; Probability Theory

How about supervised approaches?

- Split the test labeled data (462 abstracts) evenly into training/test for supervised CRF
- Chunk the sentences into phrases
- Features for each chunk
  - n-grams, suffixes, prefixes (and their n-grams)
  - sentence number
  - whether a common word
  - tag for the whole phrase (NP/VP/..)

Results for supervised CRF

TECHNIQUE	F_1	Precision	Recall
Supervised CRF	35.38	**41.55**	31.51
Bootstrapped Patterns	**38.56**	29.37	**56.1**

DOMAIN	F_1	Precision	Recall
Supervised CRF	**53.9**	**52.8**	**55.05**
Bootstrapped Patterns	37.56	30.66	48.45

Conclusions

- We described a novel set of categories to extract key aspects of scientific papers
  - FOCUS, TECHNIQUE, and DOMAIN
- We used dependency patterns to extract the information and learned the patterns using bootstrapping
- We studied influence of communities on each other in terms of techniques used
  - Our case study results: speech recognition and probability theory have been the most influential fields.

Future Work

- Improve extraction accuracy by using semi-supervised approaches like similarity of trigger words
- Study influence in terms of citation graphs
  - Why are you citing a paper?
- Study “residual” effect in co-author graph
  - Did you start using techniques/applications I generally use after our collaboration?
- Study effectiveness of inter-disciplinary research
  - Does inter-disciplinary research lead to innovative techniques specific to the application domain?