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

<table>
<thead>
<tr>
<th>FOCUS</th>
<th>TECHNIQUE</th>
<th>DOMAIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>propose</td>
<td>use</td>
<td>algorithm</td>
</tr>
<tr>
<td>direct-object</td>
<td>direct-object</td>
<td>prep_for</td>
</tr>
<tr>
<td>work</td>
<td>prep_on</td>
<td>prep_of</td>
</tr>
<tr>
<td></td>
<td>prep_for</td>
<td></td>
</tr>
<tr>
<td></td>
<td>prep_of</td>
<td></td>
</tr>
</tbody>
</table>

Learn new patterns using bootstrapping!
Our semantic patterns will extract “extracting information using dependency graphs” as FOCUS, and “dependency graphs” as TECHNIQUE.
Learned Patterns using Bootstrapping

<table>
<thead>
<tr>
<th>TECHNIQUE</th>
<th>RULES</th>
<th>TECHNIQUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>model</td>
<td>nn</td>
<td>&lt;phrase tree&gt;</td>
</tr>
<tr>
<td>rules</td>
<td>nn</td>
<td>&lt;phrase tree&gt;</td>
</tr>
<tr>
<td>extracting</td>
<td>direct-object</td>
<td>&lt;phrase tree&gt;</td>
</tr>
<tr>
<td>identify</td>
<td>direct-object</td>
<td>&lt;phrase tree&gt;</td>
</tr>
<tr>
<td>constraints</td>
<td>amod</td>
<td>&lt;phrase tree&gt;</td>
</tr>
<tr>
<td>based</td>
<td>prep_on</td>
<td>&lt;phrase tree&gt;</td>
</tr>
</tbody>
</table>

...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

<table>
<thead>
<tr>
<th>Approach</th>
<th>F1</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FOCUS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Our system</td>
<td>42.41</td>
<td>31.38</td>
<td>65.39</td>
</tr>
<tr>
<td>Inter-annotator agreement</td>
<td>53.33</td>
<td>50.80</td>
<td>56.14</td>
</tr>
<tr>
<td><strong>TECHNIQUE</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seed Patterns</td>
<td>19.72</td>
<td>19.83</td>
<td>19.61</td>
</tr>
<tr>
<td><strong>Our system</strong></td>
<td>36.04</td>
<td>27.83</td>
<td>51.14</td>
</tr>
<tr>
<td>Inter-annotator agreement</td>
<td>72.02</td>
<td>66.81</td>
<td>78.11</td>
</tr>
<tr>
<td><strong>DOMAIN</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seed Patterns</td>
<td>23.86</td>
<td>23.86</td>
<td>23.87</td>
</tr>
<tr>
<td><strong>Our system</strong></td>
<td>37.75</td>
<td>32.23</td>
<td>45.56</td>
</tr>
<tr>
<td>Inter-annotator agreement</td>
<td>72.31</td>
<td>75.58</td>
<td>69.32</td>
</tr>
</tbody>
</table>
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
<table>
<thead>
<tr>
<th>Article $a_1$</th>
<th>Parsing</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Machine Learning</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>....</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FOCUS</th>
<th>EM (0.002)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TECHNIQUE</td>
<td>EM (0.001), POS tagging (0.02)</td>
</tr>
<tr>
<td>DOMAIN</td>
<td>Syntactic Parsing (0.01)</td>
</tr>
</tbody>
</table>

**technique-score**($\text{Parsing, EM, } a_1$) = $0.001 \times 0.5$

**all-score**($\text{Parsing, EM, } a_1$) = $(0.002 + 0.001 + 0) \times 0.5$

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

\[
\text{technique-score}(\text{community, phrase, article}) = \frac{1}{z_p} \text{count}(\text{phrase} \in \text{technique | article}) \times P(\text{community | article, } \theta)
\]

- **Tf-idf like score using extraction**
- **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?

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

If $a_2$ cited $a_1$, 1
Otherwise, 0.5

$$Influence(c_1, c_2, y) = \sum_{p, y_{a} = y} Influence(c_1, c_2, p, a)$$
<table>
<thead>
<tr>
<th>Communities (decreasing order of influence)</th>
<th>Most influential Phrases</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Speech Recognition</strong></td>
<td>EM; HMM; language; contextually; segment; context independent phone; snn hidden markov;</td>
</tr>
<tr>
<td><strong>Probability Theory</strong></td>
<td>HMM; maximum entropy; language; EM; merging; EM HMM; natural language; variable memory markov;</td>
</tr>
<tr>
<td><strong>Bilingual Word Alignment</strong></td>
<td>HMM; EM; maximum entropy; spectral clustering; statistical alignment; CRFs, a discriminative; statistical word alignment; string to Tree</td>
</tr>
<tr>
<td><strong>POS Tagging</strong></td>
<td>maximum entropy; machine learning; EM HMM; POS information; decision tree; hidden markov; transformation based error driven learning; entropy; POS tagging</td>
</tr>
<tr>
<td><strong>Machine Learning Classification</strong></td>
<td>SVMs; ensemble; machine learning; gaussian mixture; EM; flat; weak classifiers; statistical machine learning</td>
</tr>
</tbody>
</table>
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!
Popularity of Communities

- Formal Computational Semantics
- Unification Based Grammar
- Machine Learning Classification
- Named Entity Recognition
- Speech Recognition
Influence vs. Popularity of MT Communities

Influence

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


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

<table>
<thead>
<tr>
<th>TECHNIQUE</th>
<th>$F_1$</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised CRF</td>
<td>35.38</td>
<td>41.55</td>
<td>31.51</td>
</tr>
<tr>
<td>Bootstrapped Patterns</td>
<td>38.56</td>
<td>29.37</td>
<td>56.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DOMAIN</th>
<th>$F_1$</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised CRF</td>
<td>53.9</td>
<td>52.8</td>
<td>55.05</td>
</tr>
<tr>
<td>Bootstrapped Patterns</td>
<td>37.56</td>
<td>30.66</td>
<td>48.45</td>
</tr>
</tbody>
</table>
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?