Identifying Focus, Techniques and Domain of Scientific Papers

Sonal Gupta
Department of Computer Science
Stanford University
Stanford, CA 94305
sonal@cs.stanford.edu

Christopher D. Manning
Department of Computer Science
Stanford University
Stanford, CA 94305
manning@cs.stanford.edu

Abstract

The dynamics of a research community can be studied by extracting information from its publications. We propose a system for extracting detailed information, such as main contribution, techniques used and the problems addressed, from scientific papers. Such information cannot be extracted using approaches that assume that words are independent of each other in a document. We use dependency trees, which give rich information about structure of a sentence, and extract relevant information from them by matching semantic patterns. We then study how the computational linguistics community and its sub-fields are changing over the years w.r.t. their focus, methods used and domain problems described in the papers. We get sub-fields of the community by using the topics obtained by applying Latent Dirichlet Allocation to text of the papers. We also find "innovative" phrases in each category for each year.

1 Introduction

The evolution of ideas and the dynamics of a research community can be studied using the scientific papers published by the community. But a rich understanding of the development and progress of scientific research requires an understanding of more than just "topics" of discussion or citation links between articles. You need to understand the domain problems of interest, the spread of methods used to approach problem classes, and an understanding of when and why scientists focus on methods versus problems. To get at this level of detail, it is essential to move beyond "bag of words" topical models to be able to connect together how methods and ideas are being pursued. This requires an understanding of sentence and argument structure, and is therefore a form of information extraction, if of a looser form than the relation extraction methods that have typically been studied.

Our work uses information extraction to study evolution of ideas and dynamics of sub-communities in a research community. We match semantic patterns in dependency graphs of sentences to extract information such as a paper's main contribution, techniques used and its domain. We call these categories as FOCUS, TECHNIQUE and DOMAIN, where we define FOCUS as main contribution of a paper, TECHNIQUE as a method used or a solution proposed, and DOMAIN as domain of the problems addressed in the paper. For example, if a paper addresses the problem of regularization in Support Vector Machines and shows improvement in parsing accuracy, then its FOCUS and TECHNIQUE are Support Vector Machines and its DOMAIN is parsing. Extracting such thematic information from scientific papers has wide applications, both from information retrieval and exploratory point of views. Using our approach, we present a study of computational linguistics community by: (i) exploring the dynamics of its sub-communities over time, (ii) studying when certain sub-fields mature and get adapted in the community as tools for solving other problems (for example, we see that parsing text is now getting adapted as a tool for addressing other problems), and (iii) defining and studying 'innovation' in the community in the three categories.

Related Work Extracting FOCUS, TECHNIQUE and DOMAIN phrases using semantic patterns and dependency graphs of sentences is in essence information extraction, and there has been a wide vari-

FOCUS	present → (direct object) [PHRASE-TREE]	work → (preposition_on) [PHRASE-TREE]
TECHNIQUE	use → (direct object) [PHRASE-TREE]	apply → (direct object) [PHRASE-TREE]
DOMAIN	system → (preposition_for) [PHRASE-TREE]	task → (preposition_of) [PHRASE-TREE]

Table 1: Some examples of semantic extraction patterns that extract information from dependency trees of sentences.

ety of work done in the field. A seminal early paper is by Hearst [6], which identifies IS-A relations using hand-written rules. There also has been some work in studying research communities, but as far as we know, we are the first one to use semantic patterns to extract structured information from research papers, and apply them to study dynamics of a research community. Topic models have been previously used to study history of ideas [5] and scholarly impact of papers [4]. However, topic models cannot extract detailed information from text as we do. Instead, we consider topic-to-word distributions calculated from topic models as a way of describing sub-communities.

2 Approach

Information Extraction: We use a few hand written semantic patterns to extract phrases indicative of a paper's FOCUS, TECHNIQUES and DOMAINS from dependency trees of sentences in the paper's abstract. A dependency tree of a sentence is a parse tree that gives dependencies (such as direct-object, subject) between words in the sentence. Some of the semantic patterns we use are shown in table 1. Examples of phrases extracted from some papers are shown in table 2. We use a total of 14 patterns for FOCUS, 7 for TECHNIQUE, and 17 for DOMAIN. For paper titles from which we are not able to extract a FOCUS phrase using the patterns, we label the whole title with the category FOCUS since authors usually include the main focus of the paper in the title. For titles from which we could extract a TECHNIQUE phrase, we labeled rest of the words with DOMAIN (for titles such as 'Studying the history of ideas using topic models'). After extracting the information, we remove common phrases using a stop word list of 10,000 most common phrases upto 3-grams in 100,000 random articles that have an abstract in the ISI web of knowledge database [1]. Next, we explain how to score a sub-field depending on the occurrence of its words in FOCUS, TECHNIQUE and DOMAIN phrases.

Calculating Scores: For getting sub-fields of a research community, we consider each topic generated by Latent Dirichlet Allocation [3] using the publications text as a sub-field. From the LDA model, we get topic-to-word scores ($score(w|T)$) that tell how likely is a word to be generated by a given topic. We then combine the topic-to-word scores with the number of times the words appear in each category phrases in a given year. That is, we compute the unnormalized FOCUS score, $\tilde{F}_T^y$, for a topic T in year y as

$$\tilde{F}_T^y = \sum_{w \in V} score(w|T) \times count(w \in V_{F^y})$$

where V is the word vocabulary, and V_F^y is the focus vocabulary that consists of all words occurring in the FOCUS phrases in year y. We then smooth the scores $\tilde{F}_T^y$ by taking a weighted average of the 2 previous and 2 next years. Similarly, we compute scores for TECHNIQUE and DOMAIN. We explain normalization of the scores in the experiments section.

Innovation: Another application of extracting such detailed information from papers is to study the new domains, techniques and applications emerging in a research community. For each of the categories, we describe "innovation" in each year as the new phrases in the category that have never been used before in the dataset. We rank the phrases by the number of times they are used after that year (as a measure of impact). For this task, we do not include full titles in the FOCUS category. We also extract phrases from all sub-trees of the matched phrase-tree. We deal with abbreviations by counting the number of times a phrase occurs in a bracket after another phrase, and threshold the count to get a list of abbreviations and their canonical names, with high precision. We also remove common words like 'model', 'approach' from the technique phrases for this task.

3 Experiments and Discussion

We studied the computational linguistics community from 1985 to 2009 using the ACL Anthology dataset [2] since it has full text of papers available (note that we cannot use bag-of-word vectors for extracting information using dependency trees). We used title and abstracts of 14,133 papers to extract phrases in the three categories. The total number of phrases extracted were 18,630 for

Paper/Title	FOCUS	TECHNIQUE	DOMAIN
Hall *et al.*[5]	diversity of ideas , topic entropy;	unsupervised topic modeling; historical trends; Latent Dirichlet Allocation; Topic Models	nil
Triplet Lexicon Models for Statistical Machine Translation.	various methods using triplets incorporating long-distance dependencies	triplets; long-distance dependencies; phrases or n-gram based language models	statistical machine translation

Table 2: Extracted phrases for some papers (overlapping phrases). For second example, FOCUS also includes its title.

FOCUS	joint chinese word segmentation; linguistically motivated phrases in parallel texts; scope of negation in biomedical texts
TECHNIQUE	acquired from web; joint srl model; generalized expectation criteria
DOMAIN	semantic dependency parsing; joint chinese word segmentation; conll 2008

Table 3: Top three new phrases in each type in the year 2008

FOCUS, 9,305 for TECHNIQUE, and 5,642 for DOMAIN. We hand labeled 197 abstracts with the three categories to measure precision and recall scores. For each abstract, we compared the unique non-stop-words in each category extracted from our algorithm to the gold labeled dataset. The F-1 scores are: 44.93 for FOCUS, 17.12 for TECHNIQUE, and 21.93 for DOMAIN. The F-1 scores are not high for three reasons: (1) authors many times use generic phrases to describe their work, which are not labeled in the gold labeled dataset (such as 'parallel text', 'faster model', 'computational approach'), (2) the system uses limited number of hand-written patterns, and (3) sometimes the dependency trees of sentences are wrong. We used the Stanford Parser [7] to generate dependency trees. We used the topics generated using the same dataset by Hall *et al.* [5]. They ran LDA with some seeded topics to get a set of 46 topics and labeled them by hand.

Figures 1(a), 1(b), 1(c) and 1(d) compares how some sub-fields are changing w.r.t them being used as methods or worked on as problems. To reduce the effect of different number of seed patterns and their recall ability, we normalize the scores by dividing them by the number of papers and the number of phrases extracted for the *given category* in the given year. The figures show the difference of DOMAIN and TECHNIQUE scores for each year. A positive score means more papers worked on the sub-field as a DOMAIN, and vice-versa. The figures also have top 40 words in the sub-fields from the topic-to-word distribution obtained from LDA. We can see that 'Wordnet' has shifted from being a tool to a domain. The 'Probabilistic Models' sub-field has been widely used as a technique in the computational linguistics community since 1990. The 'Information Extraction' and 'Statistical Parsing' sub-fields, which traditionally have been domains, are increasingly also being used as techniques. This shows that as these domains got mature, they got adapted in the community as techniques for solving more complex problems.

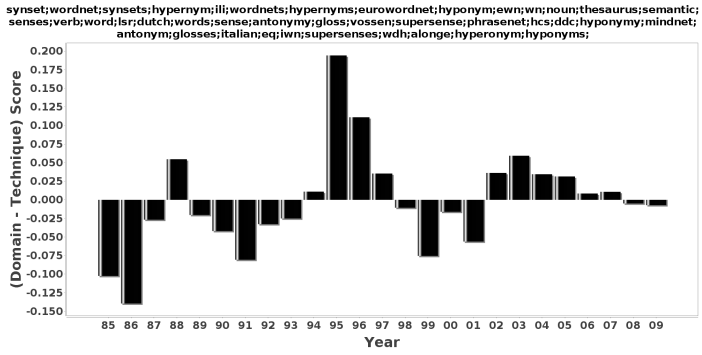
Figures 1(e) and 1(f) compare TECHNIQUE and FOCUS scores, respectively, for some sub-fields. To reduce the effect of different number of phrases extracted every year, we normalize the scores by dividing them by the number of papers and the total number of phrases extracted in the given year. As seen before, 'Probabilistic Models' has been increasingly used as a technique, while 'Syntactic Structure' is showing a declining trend. The latter figure shows that 'Parsing' has been declining as focus in recent years, and more papers now address the problems in 'Information Extraction' and 'Word Sense Disambiguation (WSD)'.[1].

Table 3 shows top "innovative" phrases for each category in the year 2008. We can see that many 'joint' models were suggested and used in the papers in 2008

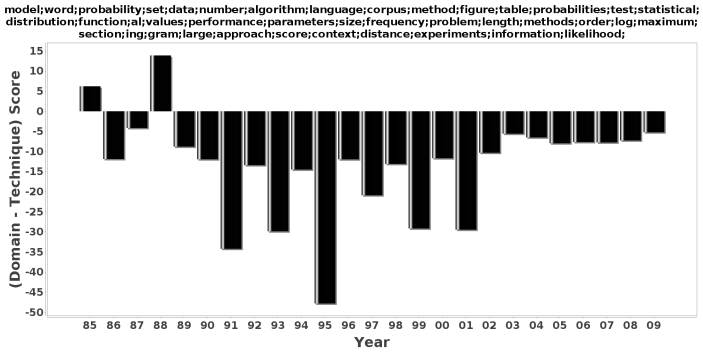
4 Future Directions

Our future work includes using a machine learning framework to extract the information. We are also exploring ways to use our system for studying citation and co-authorship networks.

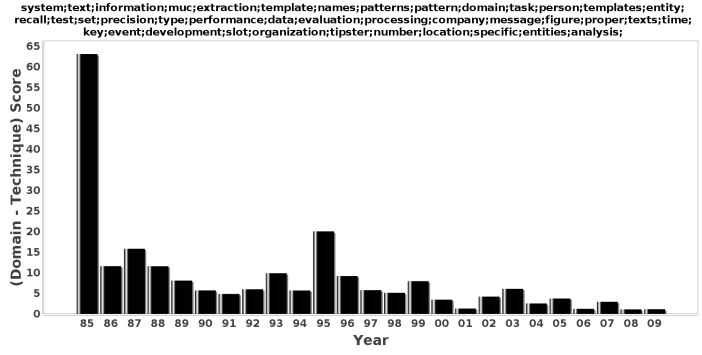
[1]More figures and results are available at the webpage `http://cs.stanford.edu/users/sonal/comparetopics.html`



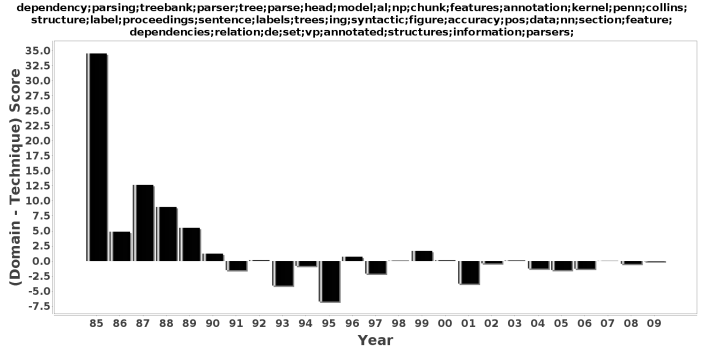
(a) Wordnet



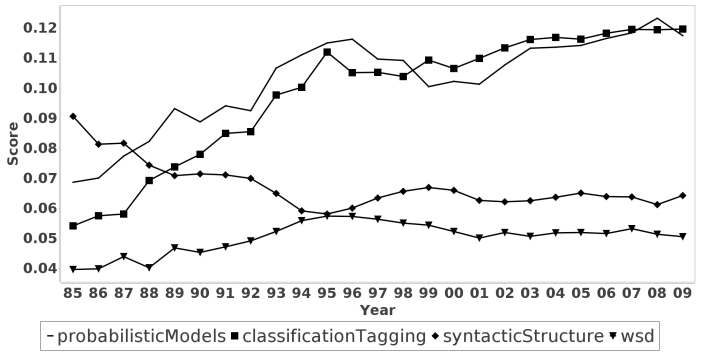
(b) Probabilistic Models



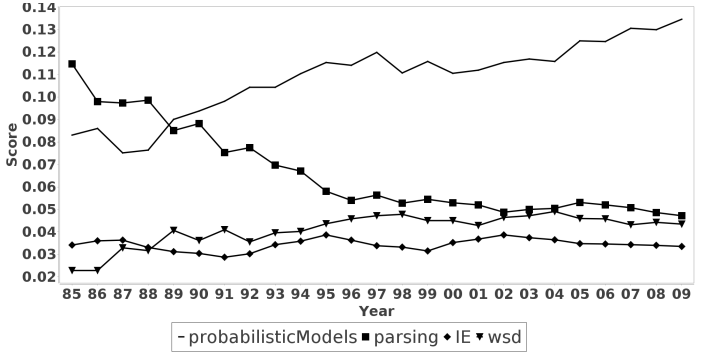
(c) Information Extraction



(d) Statistical Parsing



(e) TECHNIQUE



(f) FOCUS

Figure 1: (a) to (d) plot the difference of DOMAIN and TECHNIQUE scores ($F_D^y - F_T^y$) of some sub-fields, and the last two figures compare different sub-fields for TECHNIQUE and DOMAIN categories, respectively.

Acknowledgments

We are grateful to DARPA grant 27-001342-3 for funding the first author. We are also thankful to NSF grant 0835614 for providing the resources, such as access to the ISI web of knowledge database.

References

[1] ISI web of knowledge. `www.isiknowledge.com`.

[2] S. Bird, R. Dale, B. J. Dorr, B. Gibson, M. T. Joseph, M. yen Kan, D. Lee, B. Powley, D. R. Radev, and Y. F. Tan. The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *LREC*, 2008.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 2003.

[4] S. M. Gerrish and D. M. Blei. A language-based approach to measuring scholarly impact. In *ICML*, 2010.

[5] D. Hall, D. Jurafsky, and C. D. Manning. Studying the history of ideas using topic models. In *EMNLP*, 2008.

[6] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Conference on Computational linguistics*, 1992.

[7] M. D. Marneffe, B. Maccartney, and C. D. Manning. Generating typed dependency parses from phrase structure parses. In *LREC*, 2006.