Towards Extracting Domains from Research Publications

Shilpa Lakhanpal and Ajay Gupta
Department of Computer Science
Western Michigan University, Kalamazoo, MI 49008
shilpa.lakhanpal@wmich.edu; ajay.gupta@wmich.edu

Rajeev Agrawal
Department of Computer Systems Technology
North Carolina A&T State University, Greensboro, NC 27411
ragrawal@ncat.edu

Abstract

Every research paper falls within some specific subject areas called domains of a larger scientific field. In this paper, we present a technique for effectively mining scientific research papers for key domain areas. We combine techniques from natural language processing and machine learning to create a unique method for extracting such domains. Using preposition disambiguation helps us infer the meaning of words or phrases based on their placement within text. Combining this knowledge with supervised learning such as using a Naïve Bayes classifier helps us to classify phrases as domain areas within a scientific field. Thus in essence, our technique derives meaning from text and contributes effectively to the field of text analytics.

Introduction

Text analytics is the process of analyzing unstructured text data with a goal of deriving meaningful information. We narrow our focus into a specific type of information that we may seek from text and pay particular attention to data found in the research sphere in the form of scientific research papers.

A domain refers to a particular branch of scientific knowledge or scientific field. For example, the scientific field of Computer Science has several domains such as data mining, networking, operating systems, etc. The data mining domain in turn has the subdomains such as pattern recognition, machine learning, statistics, etc.

The problem-area addressed in a paper is the focus of research described in that paper. Each research paper or a journal is written to demonstrate the work done by the authors to solve a particular problem, or to achieve a goal.

For solving a problem, the researchers may apply known techniques, or may even devise their own techniques.

Any given research paper is basically just a collection of words. When we read the paper, we might be able to decipher what domains and subdomains it caters to. But the ability to comprehend these topics could be based on our prior knowledge of what constitutes domain or subdomain areas. We will certainly be in error if we presumptuously assume that any and every reader will be pre-equipped with the correct understanding of whether a word or a phrase is a domain, problem-area or technique.

In this work-in-progress paper, we propose an efficient technique for extracting domains from research papers. Our technique uses preposition disambiguation to provide insight into the meaning of text. We validate this meaning by using a supervised learning method. Promising domain extracting techniques can then be easily extended to discover trending domains in a scientific field.

Related Work

A bootstrapping learning technique has been proposed by (Gupta and Manning 2011) to extract items such as domain areas, focus of research and techniques from research papers. Although the work provides key insights, their results are not that encouraging as they themselves claim that their system failed to correctly address patterns which it found to be outside these three pre-defined categories. Analysis of the results indicates that their technique for domain extraction has high recall but suffers from low precision. Our proposed approach obtains good results of high precision and high recall for correctly labelling domains.

Supervised learning for text classification has been widely used in applications of Natural Language Processing (NLP). Hidden Markov Models (HMMs) are widely used statistical tools for modeling generative sequences. HMM has been used for sentence classification (Rong et

al. 2006), where the preferred sequential ordering of sentences in the abstracts of "Randomized Clinical Trial" papers, facilitated its use. The sentences in the abstract are supposed to be ordered in sequence of "background," "objective," "method," "result" and "conclusion" and model states are aligned to these sentence types. Our approach does not depend on a generative process as the "domain", "problem-area" and technique can occur in any random order in a title. Hence our proposed approach targets more generic solutions.

In our previous work (Lakhanpal, Gupta and Agrawal 2014), we extracted the prevalent trends of research using a phrase-based approach. We take our work much further by incorporating intelligent machine learning techniques to extract meaningful domain areas from research papers.

Our Approach

We describe a technique to extract meaning from the titles, keywords and abstracts of a collection of research papers. We extract this inherent meaning, which has been conveyed by the respective authors themselves by making use of the results of an NLP technique of preposition disambiguation. Thereafter our unique methodology succeeds in achieving good results using machine learning techniques. We effectively derive meaning from text without explicitly using the constructs of NLP.

As described above, a problem-area is a current focus of any research, while the domain is the larger subject area into which that and other related research work fall.

But the distinction between a domain and a problem-area is not always well-defined. Sometimes a problem-area that was initially a focus of small amount of research, over time, gains a lot of attention. Researchers begin to zoom in on the minutiae and start generating new problem-areas. Thus what started as a problem-area has now become a domain in its own right.

For the scope of this paper, however, we make no distinction between a domain and problem-area as our goal is to segregate the words / phrases depicting these two from the words depicting techniques or methods.

Although our approach is extendible to any scientific field, we conduct our preliminary experiments in the field of Computer Science.

Definitions

Word
A single and distinct element of language which has a meaning and is used with other words to form a sentence, clause or phrase

Stopword
Word in the language, such as "and", "the", which is very common, but of little value in selecting text meeting a user's need

Sentence
A sequence of words that is complete in itself, containing a subject and predicate, conveying a statement, or question, etc. and consisting of a main clause and, optionally, one or more subordinate clauses

Clause
A unit of grammatical organization said to consist of a subject and predicate

Phrase
A small group of words standing together as a conceptual unit, typically forming a component of a clause

m-gram:
A contiguous sequence of m words

Preposition:
A word governing, and usually preceding, a noun or pronoun and expressing a relation to another word or element in the clause

Preposition with Intention Sense:
The preposition that indicates that the phrase following it specifies the purpose (i.e., a result that is desired, intention or reason for existence) of an event or action

Phrase of Interest (Interesting Phrase):
A phrase that follows a preposition with intention sense and ends before the next preposition in the sentence or ends with the end of the sentence

Derivative:
Keyword or keyword phrase which has one or more words in common with the interesting phrases

Domain Word:
A word that is or has a potential for naming a well-accepted domain area, or is a part of a phrase denoting a well-accepted domain area

Using Preposition Sense Disambiguation

Semantics is a branch of linguistics that deals with the meaning of words and phrases in a particular context. For the computer to understand language as humans do, one of the steps is to elicit this semantic content. And towards achieving this purpose, we need to understand how and in what context, the prepositions are used.

Various prepositions convey various meanings based on the context they are used in. It is the placement and context of prepositions that can provide valuable information towards the meaning of text. The "sense" (Boonthum, Toida, and Levinstein 2006) or the "relation" (Srikumar and Roth 2013) communicated by the presence of various prepositions within different group of words has been investigated. We wish to draw attention to the "intention" sense. For example the intention sense is conveyed by the preposition "for" in the phrase "mining for information". (Boonthum,

Toida, and Levinston 2006) refer to the "complement" of the preposition as conveying the "intention" or "purpose". In English the complement of a preposition refers to a noun phrase, pronoun, a verb, or adverb phrase following the preposition. Technical paper titles generally focus on conveying the gist of the paper, which will be achieved more likely by using technical terminology with less stress on nuances of English language such as adverbs or pronouns. Hence, for simplicity we pick the complement delimited at the other end by the next preposition or end of the title and define it as an "interesting phrase".

We hypothesize that the interesting phrases reflect the "purpose" or the goal of their respective papers as is validated by their very definition and hence in most cases hint upon the larger domains. This hypothesis is supported by the important observation that the authors would probably want to highlight the goal of their research in their titles (Hertzmann 2010).

We would like to emphasize that we want to retrieve the generic part of the interesting phrase. Hence we fetch its part that is common with the keyword section of that paper. The keyword section of a research paper is a section where the authors will enumerate the key phrases or key words of their documents (Sherman 1996). Since titles tend to be unique, their constituents may not by themselves be good representatives of general domain areas. The keywords on the other hand are commonly and widely used, well accepted set of general terms that authors use to label their work. Hence they serve as generic terms that authors might use to mention their domains, problem-areas and techniques.

Grammatically, the title of a paper could be a sentence, clause or phrase. We scan each title to find the prepositions with intention sense. Next, we extract the interesting phrases that follow a preposition that conveys the intention sense. The next step involves finding an intersection between the interesting phrases of each paper and its keyword section. In this step, we retain those keyword or keyword phrases which have one or more words in common with the interesting phrases. This resultant set or the derivative becomes the main element of our analysis.

Supervised Classification

We classify each derivative as a "Domain" or "Not Domain".

We create a repository of domain areas in Computer Science from research and analysis of hot and trending topics across various scientific conferences and journals. This repository consists of a list of unigrams (1-grams). These unigrams either as stand-alone or together with other such members of this list signify well accepted domain areas and serve as domain words.

In analyzing each derivative, if it has any word from the above repository, we label it as a "Domain." The non-appearance of any word from the domain list in a derivative makes it a "Not Domain."

Next we delineate the features of the derivatives that help determine their likelihood of being the domains.

The various sessions of a conference group together the papers that deal with similar goals or topics. The session name or identifier captures each such topic for each group in a synoptic form, logically making it a representative of the domain of its group. While examining each derivative, if it has any word in common with a session identifier, we record its feature as "Found in Session: True" and if not, we record its feature as "Found in Session: False".

Each derivative is a phrase of one or more words. The potential of any word of the derivative to be a domain word can be heightened by its frequent occurrence across different abstracts. We use abstracts because they are written so as to contain an intelligent gist of a paper (Koopman 1997), and hence are likely sections to look for domains. Different abstracts containing the same word can validate the importance of a word, hence count of abstracts becomes a relevant feature. The count of abstracts containing at least one of the words in the derivative phrase is calculated for each derivative.

Training the classifier

We extract the feature sets for the derivative data, and divide them into a training set and a test set in the ratio of 70%-30% respectively. The training set is used to train a new "naive Bayes" classifier.

Our Technique Exemplified

We describe our process through an example. Figures 1 (a) and (b) depict Use Case Diagrams showing the Steps involved in Extracting Derivatives. We use a title of a paper from the ACM SIGKDD 2012 conference.

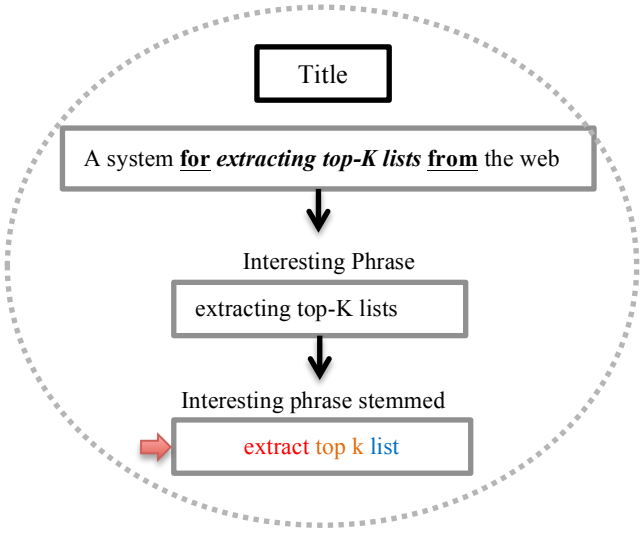

Figure 1(a): Use Case Diagram showing the Steps involved in Extracting Derivatives

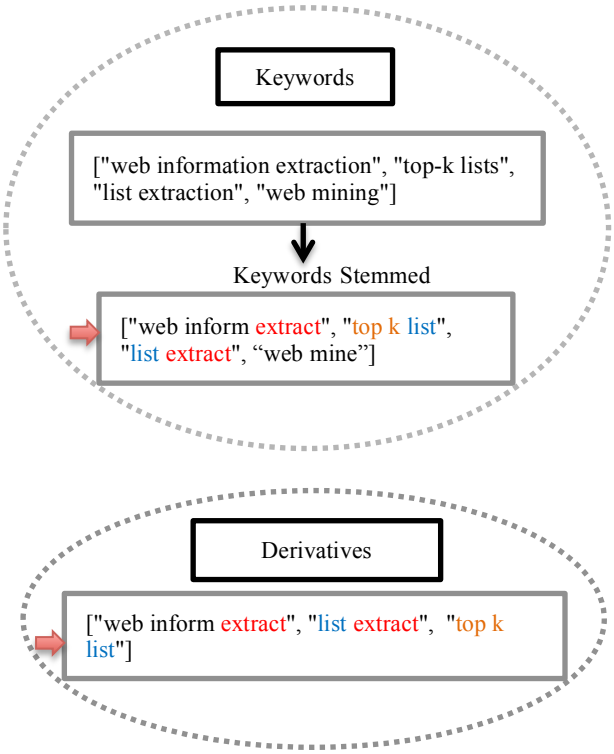

Figure 1(b): Use Case Diagram showing the Steps involved in Extracting Derivatives

Results

We have programmed our technique in python and also have employed some readymade data mining packages. Careful study of the preposition senses narrowed down by (Boonthum, Toida, and Levinston 2006) has allowed us to create our set of prepositions with intention sense namely ["for", "to", "towards", "toward"].

In order to find well-accepted domain areas, we have collected the topics from the Calls for Papers sections from the IEEE International Conference on Data Mining series (ICDM), the IEEE International Conference on Data Engineering (ICDE), and the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) from 2010-2014. Call for papers for any conference contain topics under which papers are sought. Hence they are one of the definitive sources of domains well-accepted by experts in the scientific field.

In a set of experiments, we collected data from the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) from years 2010-2014. This data includes 939 papers from all sessions including keynote, panel, demonstration, poster, industrial and government track apart from the regular research track sessions. Out of the 939 paper titles, 367 have prepositions with intention sense. From the 367, we get 272 non empty derivatives.

Although the final dataset of 272 is small, our results are very encouraging. For 100 iterations, we get an average accuracy of 86.72 % for the classifier. Our point of contention was never the size of the dataset, rather the intelligence we derive from it, based on our technique. Our technique has high precision and high recall as is demonstrated by the values of precision = 0.90 and recall = 0.91from one such iteration.

Conclusions and Future Work

We have obtained encouraging results from our technique, even though the experiments are limited to Computer Science papers following a fixed format. Using preposition disambiguation has helped us in extracting keywords (derivatives) that depict domains.

As future work, we wish to test our technique on a much diverse dataset and evaluate technical robustness when the papers do not have a fixed format. We further wish to extend this *fusion of NLP with supervised classification* and develop methods for extracting techniques from scientific papers. The keywords which were not recognized as derivatives need to be evaluated as potential words denoting techniques.

References

Boonthum, C., Toida, S., and Levinstein, I. (2006) Preposition Senses: Generalized Disambiguation Model. *In Proceedings of the Seventh International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, Lecture Notes in Computer Science, Berlin: Springer, pp. 196-207.

Gupta, S., and Manning, C. D. 2011. Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers. *In Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*.pp 1 – 9.

Hertzmann, A. 2010. Writing Research Papers. http://www.dgp.toronto.edu/~hertzman/courses/gradSkills/2010/writing.pdf.

Koopman, P. (CMU) 1997. How to Write an Abstract. http://users.ece.cmu.edu/~koopman/essays/abstract.html.

Lakhanpal, S., Gupta, A., and Agrawal, R. 2014. On Discovering Most Frequent Research Trends in a Scientific Discipline using a Text Mining Technique. *In Proceedings of the 52nd Annual ACM Southeast Conference*, Kennesaw, GA: ACM, pp. 52:1-52:4.

Rong, X., Supekar, K., Huang, Y., Das, A., and Garber, A. 2006. Combining Text Classification and Hidden Markov Modeling Techniques for Structuring Randomized Clinical Trial Abstracts. *In Proceedings of the American Medical Informatics Association (AMIA) Annual Symposium*, pp. 824 - 828.

Sherman, A 1996. Some Advice on Writing a Technical Report. http://www.csee.umbc.edu/~sherman/Courses/documents/TR_how_to.html.

Srikumar, V., and Roth, D. 2013. Modeling Semantic Relations Expressed by Prepositions. *Transactions of the Association for Computational Linguistics, 1*: 231-242.