Survey: Inducing Taxonomies from Text

Nathan Bartley
University of Southern California
nbartley@usc.edu

Abstract

Taxonomies and facets are of great importance to knowledge-rich domains like science and applications like information retrieval. Given the glut of information available online, it is valuable to be able to quickly sort individual documents by their broader topic areas (e.g., their discipline) and their more granular semantics (e.g., their methodology). In this paper we analyze the current state of facet extraction and automatic taxonomy construction from raw text.

1 Introduction

With the massive digitization effort of scientific articles over the last 15 years, we now have access to huge amounts of information, especially in disciplines related to computer science.[1] To take advantage of such a glut of information, researchers must be able to sift through papers to see if they are applicable or interesting. To be able to search through documents and identify the key *concepts* and *facets* that are discussed would be of great value. For example, it is difficult to instantly understand the *methods* being proposed in a paper, the *domain* the authors are working in, nor the *metrics* used to assess their methods.

Similarly, it would be of great interest to examine on a macroscopic level the direction research is taking by such aggregated concepts and facets. How does the adoption of a particular method spread over time?

As posed by (Siddiqui et al., 2016), the problem of extracting such concepts and facets can be described as the Facet Extraction problem: to extract facets is to label each document in a corpus with a ranked list of concepts for each facet. This means that in a paper about computer vision we may have a number of deep neural network models and preprocessing techniques (the *concepts*) be associated with a *technique* facet for that document.

However, the problem of Facet Extraction as it stands does not allow for much hierarchy in the concepts nor the facets. If we reframe the problem as one where we jointly label each facet with a list of concepts *and* induce a taxonomy over those concepts, then we may recover better concepts and gain macroscopic insight as to what the corpus is concerned about.

To further illustrate this problem, we discuss related work in the realm of extracting facets as well as constructing taxonomies from text.

2 Facet Extraction

An early work in facet extraction is by Gupta et al. (2011), where they characterize a scientific article in terms of its *focus*, *technique*, and *domain*. A focus of an article is its main contribution. A technique is any method or tool used in the article. The domain is the article's application domain. For example, an article that concentrates on regularization in RNNs for speech recognition will have a focus of regularization, techniques of regularization and RNNs, and a domain of speech recognition.

To identify the concepts associated with each of the three facets, the authors match a document's text to semantic patterns built on dependency parse trees. Given a set of seed patterns (e.g, a focus pattern is [$present \rightarrow direct\_object$]),

the authors bootstrap more patterns from the corpus. After re-weighting the discovered patterns, they identify significant facets in each document. They also topic model their corpus, and tie together the topics with the concepts to analyze the influence different communities have on one

[1] In the last three years alone there have been nearly 90,000 pre-print articles published on arxiv.org under computer science.

another.

Something that is interesting about this work is that they utilize semantic patterns and dependency parse trees to directly extract the relations. However, this work does not consider a richer set of facets for analyzing their corpora. Likewise, they do not rely on hierarchical information in the concepts to inform the influence score.

Another important paper in Facet Extraction is by Siddiqui et al. (2016). In this paper they explicitly define the Facet Extraction problem, and present their framework for extracting concepts and assigning them to arbitrary facets (which can be user-specified). The authors treat the assignment of concepts to facets as a joint optimization over four constructed subgraphs: one with links between concept mentions and topical concepts, one with co-occurrence between concept mentions and section names (e.g., Introduction, Methods, Conclusion), concept mentions and relation phrases, and one with concept mentions and suffix phrases (e.g., ”-able” and ”-ition”). They then solve this joint optimization problem as a mixed integer programming problem.

Something interesting in this paper is that they present a framework that does not assume a fixed set of facets, which allows for flexibilty in application across different domains. Similarly, this framework takes advantage of both local sentence-level semantics and global-level corpus statistics to construct its heterogeneous graph rather than relying on one or the other. This allows for multiple levels of granularity in the concepts that get extracted. However, the framework implicitly models the levels of granularity and specificity, which would be explicitly captured in an extraction of the taxonomy of the concepts.

3 Taxonomy Construction

Automatic taxonomy construction has been a problem in computational linguistics for many years, as it has been readily apparent the value in automatically organizing a corpus into a well-structured taxonomy to allow for quick information access (or for instance recommendation of new articles). Early methods rely heavily on pre-defined lexico-syntactic patterns for extraction of straightforward ”is-a” relations (Hearst, 1992), which gives high precision but very low recall given its fixed patterns.

More recently, work has been done at combining insights from neural language models (namely, using word embeddings trained under the Skip-gram model) and an adaptive recursive hierarchical clustering scheme to construct *topic taxonomies* (Zhang et al., 2018). These topic taxonomies are trees that have many semantically coherent concepts assigned to each node, where the concepts are more granular and specific the further down the tree is traversed.

This work by Zhang et al. is interesting because it jointly deals with varying the levels of granularity of each term and relaxes the need for strict patterns to identify terms, ultimately increasing coverage. However, this work is limited in that it requires a fixed number of clusters for its adaptive clustering module. Relaxing this would allow for a more reliable data-driven taxonomy generation. Similarly, this work is implicitly relying on extracting hypernymy ”is-a” relations which limits the possible domain applications of the taxonomies.

This user-specific limitation is addressed by Hi-Expan (Shen et al., 2018). In this work the authors present a framework that takes a domain-specific corpus, and a task-specific seed taxonomy. With this the authors extract new terms from the corpus, and using an iterative process of set expansion and relation expansion fill out the seed taxonomy. This is all carried out as a joint optimization problem that assigns each term to its appropriate parent node in the taxonomy.

This work is interesting because it can expand a taxonomy based on a user-defined relation. However, it is not able to assign multiple terms to the same node in the same way the other frameworks can. In this sense it is difficult to compare specific terms to one another, even those that are children of the same parent node.

References

Sonal Gupta and Christopher Manning. 2011. Analyzing the dynamics of research by extracting key aspects of scientific papers. In *Proceedings of 5th international joint conference on natural language processing*, pages 1–9.

Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics.

Jiaming Shen, Zeqiu Wu, Dongming Lei, Chao Zhang, Xiang Ren, Michelle T Vanni, Brian M Sadler, and

Jiawei Han. 2018. Hiexpan: Task-guided taxonomy construction by hierarchical tree expansion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2180–2189. ACM.

Tarique Siddiqui, Xiang Ren, Aditya Parameswaran, and Jiawei Han. 2016. Facetgist: Collective extraction of document facets in large technical corpora. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 871–880. ACM.

Chao Zhang, Fangbo Tao, Xiusi Chen, Jiaming Shen, Meng Jiang, Brian Sadler, Michelle Vanni, and Jiawei Han. 2018. Taxogen: Unsupervised topic taxonomy construction by adaptive term embedding and clustering. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2701–2709. ACM.