Information Sciences 329 (2016) 125–143

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Generalization of parse trees for iterative taxonomy learning

Boris A. Galitsky*

eBay Inc., San Jose, CA 95125, USA

ARTICLE INFO

Article history:
Received 9 January 2013
Revised 3 July 2015
Accepted 11 September 2015
Available online 21 September 2015

Keywords:
Learning taxonomy
Web mining
Learning constituency parse tree
Search relevance

ABSTRACT

We build a taxonomy of entities which is intended to improve the relevance of search engine in a vertical domain. The taxonomy construction process starts from the seed entities and mines the web for new entities associated with them. To form these new entities, machine learning of syntactic parse trees (their generalization) is applied to the search results for existing entities to form commonalities between them. These commonality expressions then form parameters of existing entities, and are turned into new entities at the next learning iteration.

Taxonomy and paragraph-level syntactic generalization are applied to relevance improvement in search and text similarity assessment. We conduct an evaluation of the search relevance improvement in vertical and horizontal domains and observe significant contribution of the learned taxonomy in the former, and a noticeable contribution of a hybrid system in the latter domain. We also perform industrial evaluation of taxonomy and syntactic generalization-based text relevance assessment and conclude that proposed algorithm for automated taxonomy learning is suitable for integration into industrial systems. Proposed algorithm is implemented as a part of Apache *OpenNLP.Similarity* project.

1. Introduction

Nowadays, in designing search engines and text relevance systems, it is hard to overestimate the role of taxonomies for improving their precisions, especially in vertical domains. Taxonomies, thesauri or concept hierarchies are a crucial component of many applications of Information Retrieval and Natural Language Processing (NLP). However, building, tuning and managing taxonomies and ontologies is rather costly since a lot of manual operations are required. A number of studies proposed automated building of taxonomies based on linguistic resources and/or statistical machine learning (Kerschberg et al. [24], Roth [39], Kozareva et al. [25], Sánchez and Moreno [40]).

A majority of current approaches to taxonomy-supported searches have not been deployed in industry due to the following:

- insufficient accuracy of resultant search, due to a limited coverage of a domain by a taxonomy,
- limited expressiveness in representations of queries of real users,
- inability to disambiguate queries relying on taxonomies,
- high cost associated with manual construction or editing of linguistic resources, and their limited adjustability.

In terms of code reusability, taxonomies remain very labor intensive. Transferring a search engine from one domain to another, the primary component to be rebuilt is a taxonomy.

* Tel.: +1 6502094601.
E-mail address: bgalitsky@hotmail.com

http://dx.doi.org/10.1016/j.ins.2015.09.008
0020-0255/

In this work we will take advantage of web mining based on search engine APIs, deep syntactic parsing as well as learning and matching tree representations of sentences and taxonomies. Proposed industrial-quality taxonomy learning algorithm is oriented at an improvement of vertical search relevance and will be evaluated in a number of search settings and domains. The main challenge in building taxonomy tree is to make it as deep as possible to incorporate longer chains of relationships [22], so that more specific (and more complicated) questions can be answered.

A number of currently available general-purpose and crowd-sourced resources such as DBPedia and Freebase assist with entity-related searches, but are insufficient to filter out irrelevant answers concerning multiple entities, certain action over an entity and a multitude of its parameters. A set of available vertical ontologies, such as genes, bioinformatics, entertainment are also helpful for entity-based searches in vertical domains, however their taxonomy trees are rather shallow, and their utility for filtering out irrelevant answers is rather limited.

A few studies have attempted to learn taxonomies on the basis of textual input (Perrin and Petry [32]). Several researchers explored taxonomic relations explicitly expressed in texts by pattern matching (Hearst [20], Poesio et al. [34]). One limitation of pattern matching is that it involves the predefined choice of semantic relations to be extracted. In this study, to improve the flexibility of keyword-based pattern matching, we use matching of parse trees, which is a higher level of abstraction than sequences of words. We extend the notion of syntactic contexts of (Lin [26]) from a partial case such as noun + modifier and dependency triple toward finding a parse sub-tree in a parse tree. Our approach also extends handling of internal structure of noun phrases used to find taxonomic relations (Buitelaar et al. [4]). Many researchers follow Harris' distributional hypothesis of correlation between semantic similarity of words or terms, and the extent to which they share similar syntactic contexts (Harris [19]).

The contribution of this study is an automated taxonomy building mechanism which is based on initial set of main entities (a seed) for given vertical knowledge domain. This seed is then automatically extended by mining of web documents' abstracts which include a "meaning" of a current taxonomy node. This node is further extended by entities which are the results of inductive learning of commonalities between these documents. These commonalities are extracted using an operation of syntactic generalization, which finds the common parts of syntactic parse trees of a set of documents, obtained for the current taxonomy node. Syntactic generalization has been extensively evaluated commercially to improve text relevance (Galitsky et al. [17], Galitsky et al. [13]), and in this study we also apply it *at the level of paragraphs* for automated building of taxonomies. In (Galitsky [9]) taxonomy construction was considered from the standpoint of transfer learning [35,41], and sentence-level generalization was used.

For industrial search engines, the value of semantically-enabling search engines via taxonomies and ontologies for improving search relevance has been appreciated (Heddon [21]). Once a taxonomy adequately covering all important entities in a vertical domain is available, it can be directly applied to filtering out irrelevant answers. What is worth exploring nowadays is how to apply a real-world taxonomy to search relevance improvement, where this taxonomy is not perfect since it was automatically compiled from the web.

Our taxonomy building algorithm is focused on search relevance improvement, unlike the majority of ontology mining methods which optimize the precision and recall of extracted relations. Therefore evaluation in this study will assess the algorithm performance in terms of search accuracy improvement. Hence we expect the search performance-driven taxonomy learning algorithm to outperform the ones focused on most, or most exact, relations. Our evaluation will be conducted in the vertical and horizontal searches, as well as in industrial environment of text similarity assessment.

We now proceed to a formal problem formulation for a taxonomy-supported search. Let us consider a search query Q and candidate answers $a_1, a_2,\ldots, a_n \in A$ that are produced by a component of a search engine. In this paper we build a relevance verification component based on a taxonomy: it takes Q, A and filters out irrelevant answers to retain a subset of A. We need a mechanism which would determine which keywords in Q must occur in answer a_i; otherwise this answer would be considered irrelevant. To do that, we need to have a hierarchy (ordered sets) of keywords T for a specific vertical domain, so that for Q we can determine "what it is about" in terms of these keywords X which must occur in the relevant answer (then Q is about X and a_i is about X). Selected domain keywords are organized in a taxonomy T which enforces relevant a_i to have keywords X extracted from Q.

This paper elaborates this approach and is organized as follows. We first define the relationships between Q and X. We then explore the relationships between Q, X and A and propose T as a way to enforce relevance and propose an online search algorithm. After that we propose the methodology for how T is constructed, followed by evaluation of search relevance improvement.

The industrial evaluation of a hybrid system reveals that the proposed algorithm is suitable for integration into industrial systems. The algorithm is implemented as a component of Apache OpenNLP project.

2. Improving search relevance by taxonomies

To answer a question, natural language or keyword-based, it is beneficial to 'understand' what this question is about [3,10]. In the sense of current paper this 'understanding' is a preferential treatment of keywords. A Q/A system needs to understand the topic of the questions, what it is about, which keywords are most important. For example, for a query '*sales tax*' one can say that it is about *tax*, but not about *sale*. We denote a relationship between a set of keywords for a question and its subset which expresses the topic of this question as

is-about (*set-of-keywords, subset-of-keyword*).