Entailment-based Text Exploration with Application to the Health-care Domain

Meni Adler
Bar Ilan University
Ramat Gan, Israel
adlerm@cs.bgu.ac.il

Jonathan Berant
Tel Aviv University
Tel Aviv, Israel
jonatha6@post.tau.ac.il

Ido Dagan
Bar Ilan University
Ramat Gan, Israel
dagan@cs.biu.ac.il

Abstract

We present a novel text exploration model, which extends the scope of state-of-the-art technologies by moving from standard *concept*-based exploration to *statement*-based exploration. The proposed scheme utilizes the *textual entailment* relation between statements as the basis of the exploration process. A user of our system can explore the result space of a query by drilling down/up from one statement to another, according to entailment relations specified by an entailment graph and an optional concept taxonomy. As a prominent use case, we apply our exploration system and illustrate its benefit on the health-care domain. To the best of our knowledge this is the first implementation of an exploration system at the statement level that is based on the textual entailment relation.

1 Introduction

Finding information in a large body of text is becoming increasingly more difficult. Standard search engines output a set of documents for a given query, but do not allow any exploration of the thematic structure in the retrieved information. Thus, the need for tools that allow to effectively sift through a target set of documents is becoming ever more important.

Faceted search (Stoica and Hearst, 2007; Käki, 2005) supports a better understanding of a target domain, by allowing exploration of data according to multiple views or *facets*. For example, given a set of documents on Nobel Prize laureates we might have different facets corresponding to the laureate's nationality, the year when the prize was awarded, the field in which it was awarded, etc. However, this type of exploration is still severely limited insofar that it only allows exploration by *topic* rather than *content*. Put differently, we can only explore according to *what a document is about* rather than *what a document actually says*. For instance, the facets for the query *'asthma'* in the faceted search engine Yippy include the concepts *allergy* and *children*, but do not specify what are the exact *relations* between these concepts and the query (*e.g.*, allergy *causes* asthma, and children *suffer from* asthma).

Berant et al. (2010) proposed an exploration scheme that focuses on relations between concepts, which are derived from a graph describing textual entailment relations between *propositions*. In their setting a proposition consists of a predicate with two arguments that are possibly replaced by variables, such as *'X control asthma'*. A graph that specifies an entailment relation *'X control asthma* $\rightarrow$ *X affect asthma'* can help a user, who is browsing documents dealing with substances that affect asthma, drill down and explore only substances that control asthma. This type of exploration can be viewed as an extension of faceted search, where the new facet concentrates on the actual statements expressed in the texts.

In this paper we follow Berant et al.'s proposal, and present a novel entailment-based text exploration system, which we applied to the health-care domain. A user of this system can explore the result space of her query, by drilling down/up from one proposition to another, according to a set of entailment relations described by an *entailment graph*. In Figure 1, for example, the user looks for 'things'

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 79–84,
Jeju, Republic of Korea, 8-14 July 2012.

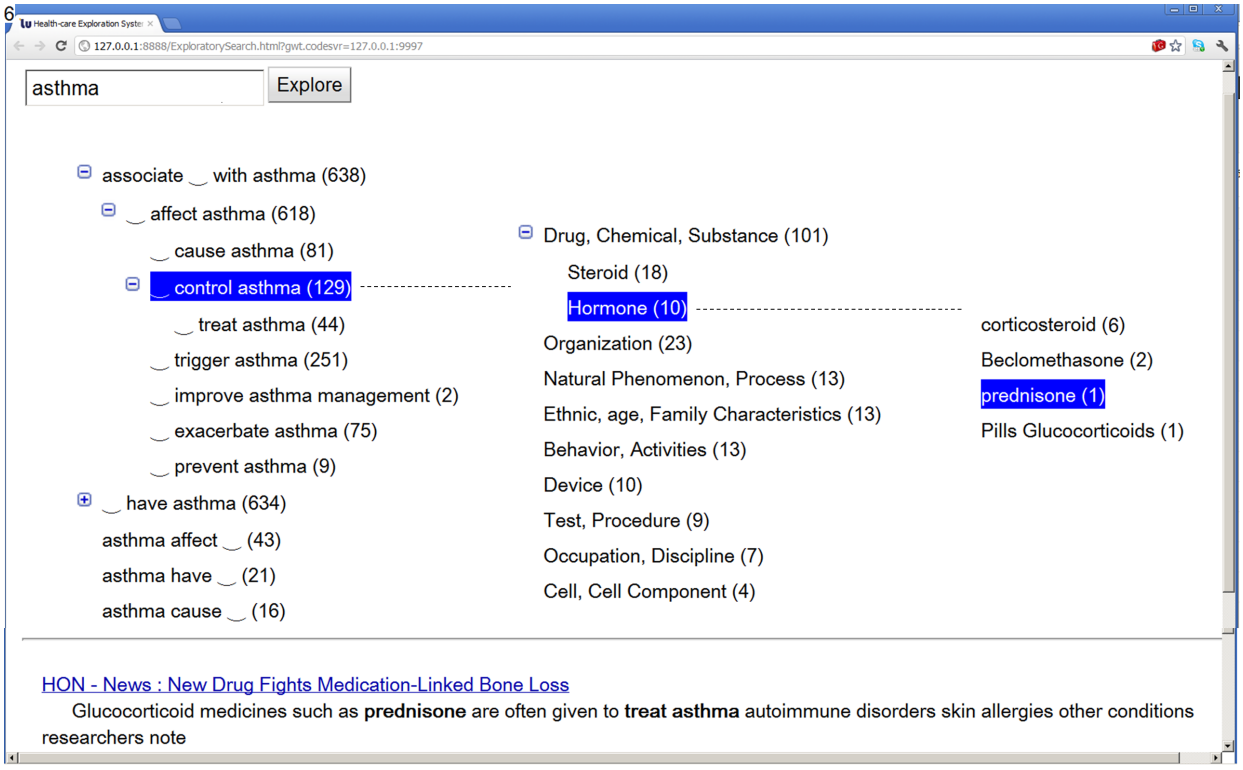

Figure 1: Exploring *asthma* results.

that affect asthma. She invokes an *'asthma'* query and starts drilling down the entailment graph to *'X control asthma'* (left column). In order to examine the arguments of a selected proposition, the user may drill down/up a concept taxonomy that classifies *terms* that occur as arguments. The user in Figure 1, for instance, drills down the concept taxonomy (middle column), in order to focus on *Hormones* that control asthma, such as *'prednisone'* (right column). Each drill down/up induces a subset of the documents that correspond to the aforementioned selections. The retrieved document in Figure 1 (bottom) is highlighted by the relevant proposition, which clearly states that prednisone is often given to treat asthma (and indeed in the entailment graph *'X treat asthma'* entails *'X control asthma'*).

Our system is built over a corpus of documents, a set of propositions extracted from the documents, an entailment graph describing entailment relations between propositions, and, optionally, a concept hierarchy. The system implementation for the health-care domain, for instance, is based on a web-crawled health-care corpus, the propositions automatically extracted from the corpus, entailment graphs borrowed from Berant et al. (2010), and the UMLS[1] taxonomy. To the best of our knowledge this is the first implementation of an exploration system, at the proposition level, based on the textual entailment relation.

2 Background

2.1 Exploratory Search

Exploratory search addresses the need of users to quickly identify the important pieces of information in a target set of documents. In exploratory search, users are presented with a result set and a set of exploratory facets, which are proposals for refinements of the query that can lead to more focused sets of documents. Each facet corresponds to a clustering of the current result set, focused on a more specific topic than the current query. The user proceeds in the exploration of the document set by selecting specific documents (to read them) or by selecting specific facets, to refine the result set.

[1] http://www.nlm.nih.gov/research/umls/

Early exploration technologies were based on a single hierarchical conceptual clustering of information (Hofmann, 1999), enabling the user to drill up and down the concept hierarchies. Hierarchical faceted meta-data (Stoica and Hearst, 2007), or *faceted search*, proposed more sophisticated exploration possibilities by providing multiple facets and a hierarchy per facet or dimension of the domain. These types of exploration techniques were found to be useful for effective access of information (Käki, 2005).

In this work, we suggest proposition-based exploration as an extension to concept-based exploration. Our intuition is that text exploration can profit greatly from representing information not only at the level of individual concepts, but also at the propositional level, where the relations that link concepts to one another are represented effectively in a hierarchical entailment graph.

2.2 Entailment Graph

Recognizing Textual Entailment (RTE) is the task of deciding, given two text fragments, whether the meaning of one text can be inferred from another (Dagan et al., 2009). For example, *'Levalbuterol is used to control various kinds of asthma' entails 'Levalbuterol affects asthma'*. In this paper, we use the notion of *proposition* to denote a specific type of text fragments, composed of a predicate with two arguments (*e.g.*, *Levalbuterol control asthma*).

Textual entailment systems are often based on *entailment rules* which specify a directional inference relation between two fragments. In this work, we focus on leveraging a common type of entailment rules, in which the left-hand-side of the rule (LHS) and the right-hand-side of the rule (RHS) are *propositional templates* - a proposition, where one or both of the arguments are replaced by a variable, *e.g.*, *'X control asthma* $\rightarrow$ *X affect asthma'*.

The entailment relation between propositional templates of a given corpus can be represented by an *entailment graph* (Berant et al., 2010) (see Figure 2, top). The nodes of an entailment graph correspond to propositional templates, and its edges correspond to entailment relations (rules) between them. Entailment graph representation is somewhat analogous to the formation of ontological relations between concepts of a given domain, where in our case the nodes correspond to propositional templates rather than to concepts.

3 Exploration Model

In this section we extend the scope of state-of-the-art exploration technologies by moving from standard concept-based exploration to proposition-based exploration, or equivalently, statement-based exploration. In our model, it is the entailment relation between propositional templates which determines the granularity of the viewed information space. We first describe the inputs to the system and then detail our proposed exploration scheme.

3.1 System Inputs

Corpus A collection of *documents*, which form the search space of the system.

Extracted Propositions A set of propositions, extracted from the corpus document. The propositions are usually produced by an *extraction method*, such as TextRunner (Banko et al., 2007) or ReVerb (Fader et al., 2011). In order to support the exploration process, the documents are indexed by the propositional templates and argument terms of the extracted propositions.

Entailment graph for predicates The nodes of the entailment graph are propositional templates, where edges indicate entailment relations between templates (Section 2.2). In order to avoid circularity in the exploration process, the graph is transformed into a DAG, by merging 'equivalent' nodes that are in the same strong connectivity component (as suggested by Berant et al. (2010)). In addition, for clarity and simplicity, edges that can be inferred by transitivity are omitted from the DAG. Figure 2 illustrates the result of applying this procedure to a fragment of the entailment graph for *'asthma'* (*i.e.*, for propositional templates with *'asthma'* as one of the arguments).

Taxonomy for arguments The optional concept taxonomy maps *terms* to one or more pre-defined concepts, arranged in a hierarchical structure. These terms may appear in the corpus as arguments of predicates. Figure 3, for instance, illustrates a simple medical taxonomy, composed of three concepts (medical, diseases, drugs) and four terms (cancer, asthma, aspirin, flexeril).

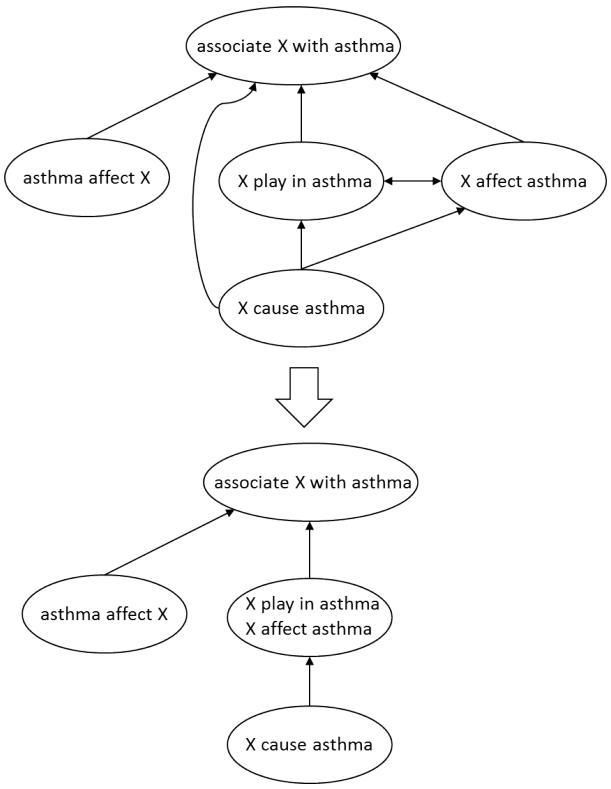

Figure 2: Fragment of the entailment graph for *'asthma'* (top), and its conversion to a DAG (bottom).

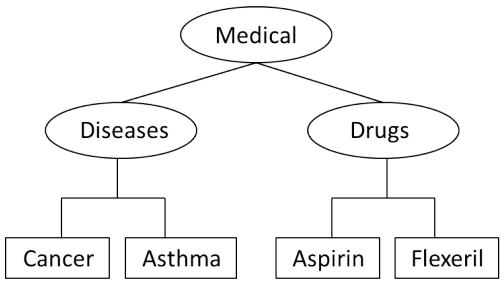

Figure 3: Partial medical taxonomy. Ellipses denote *concepts*, while rectangles denote *terms*.

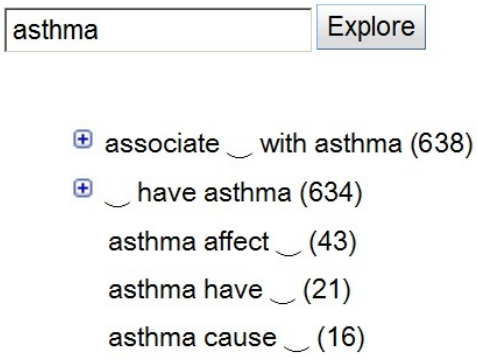

Figure 4: The roots of the entailment graph for the *'asthma'* query.

3.2 Exploration Scheme

The objective of the exploration scheme is to support querying and offer facets for result exploration, in a visual manner. The following components cover the various aspects of this objective, given the above system inputs:

Querying The user enters a search term as a query, *e.g.*, *'asthma'*. The given term induces a subgraph of the entailment graph that contains all propositional templates (graph nodes) with which this term appears as an argument in the extracted propositions (see Figure 2). This subgraph is represented as a DAG, as explained in Section 3.1, where all nodes that have no parent are defined as the *roots* of the DAG. As a starting point, only the roots of the DAG are displayed to the user. Figure 4 shows the five roots for the *'asthma'* query.

Exploration process The user selects one of the entailment graph nodes (*e.g.*, *'associate X with asthma'*). At each exploration step, the user can drill down to a more specific template or drill up to a more general template, by moving along the entailment hierarchy. For example, the user in Figure 5, expands the root *'associate X with asthma'*, in order to drill down through *'X affect asthma'* to *'X control Asthma'*.

Selecting a propositional template (Figure 1, left column) displays a concept taxonomy for the arguments that correspond to the variable in the selected template (Figure 1, middle column). The user can explore these argument concepts by drilling up and down the concept taxonomy. For example, in Figure 1 the user, who selected *'X control Asthma'*, explores the arguments of this template by drilling down the taxonomy to the concept *'Hormone'*.

Selecting a concept opens a third column, which lists the terms mapped to this concept that occurred as arguments of the selected template. For example, in Figure 1, the user is examining the list of arguments for the template *'X control Asthma'*, which are mapped to the concept *'Hormone'*, focusing on the argument *'prednisone'*.

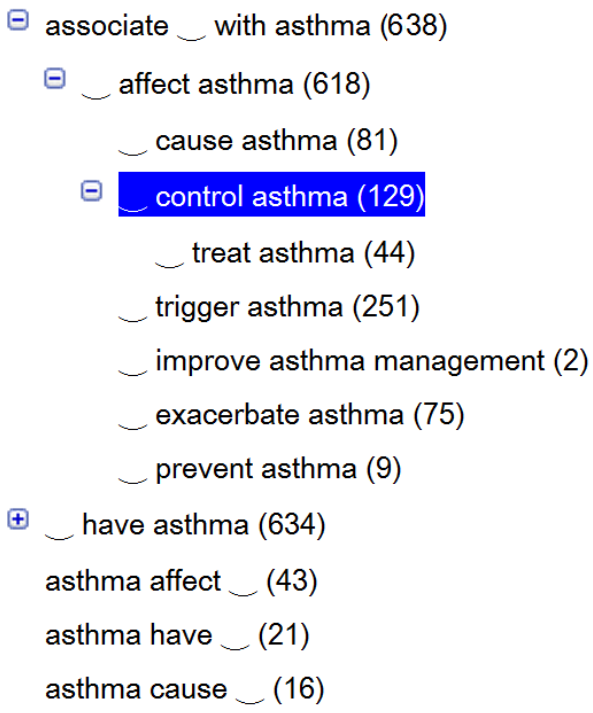

Figure 5: Part of the entailment graph for the *'asthma'* query, after two exploration steps. This corresponds to the left column in Figure 1.

Document retrieval At any stage, the list of documents induced by the current selected template, concept and argument is presented to the user, where in each document snippet the relevant proposition components are highlighted. Figure 1 (bottom) shows such a retrieved document. The highlighted extraction in the snippet, *'prednisone treat asthma'*, entails the proposition selected during exploration, *'prednisone control asthma'*.

4 System Architecture

In this section we briefly describe system components, as illustrated in the block diagram (Figure 6).

The *search service* implements full-text and faceted search, and document indexing. The *data service* handles data (*e.g.*, documents) replication for clients. The *entailment service* handles the logic of the entailment relations (for both the entailment graph and the taxonomy).

The *index server* applies periodic indexing of new texts, and the *exploration server* serves the exploration application on querying, exploration, and data

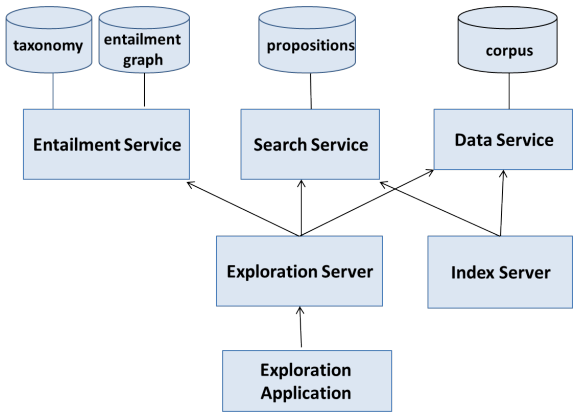

Figure 6: Block diagram of the exploration system.

access. The *exploration application* is the front-end user application for the whole exploration process described above (Section 3.2).

5 Application to the Health-care Domain

As a prominent use case, we applied our exploration system to the health-care domain. With the advent of the internet and social media, patients now have access to new sources of medical information: consumer health articles, forums, and social networks (Boulos and Wheeler, 2007). A typical non-expert health information searcher is uncertain about her exact questions and is unfamiliar with medical terminology (Trivedi, 2009). Exploring relevant information about a given medical issue can be essential and time-critical.

System implementation For the search service, we used SolR servlet, where the data service is built over FTP. The exploration application is implemented as a web application.

Input resources We collected a health-care corpus from the web, which contains more than 2M sentences and about 50M word tokens. The texts deal with various aspects of the health care domain: answers to questions, surveys on diseases, articles on life-style, etc. We extracted propositions from the health-care corpus, by applying the method described by Berant et al. (2010). The corpus was parsed, and propositions were extracted from dependency trees according to the method suggested by Lin and Pantel (2001), where propositions are dependency paths between two arguments of a predi-

cate. We filtered out any proposition where one of the arguments is not a term mapped to a medical concept in the UMLS taxonomy.

For the entailment graph we used the 23 entailment graphs published by Berant et al.[2]. For the argument taxonomy we employed UMLS – a database that maps natural language phrases to over one million unique concept identifiers (CUIs) in the health-care domain. The CUIs are also mapped in UMLS to a concept taxonomy for the health-care domain.

The web application of our system is available at: `http://132.70.6.148:8080/exploration`

6 Conclusion and Future Work

We presented a novel exploration model, which extends the scope of state-of-the-art exploration technologies by moving from standard concept-based exploration to proposition-based exploration. Our model combines the textual entailment paradigm within the exploration process, with application to the health-care domain. According to our model, it is the entailment relation between propositions, encoded by the entailment graph and the taxonomy, which leads the user between more specific and more general statements throughout the search result space. We believe that employing the entailment relation between propositions, which focuses on the statements expressed in the documents, can contribute to the exploration field and improve information access.

Our current application to the health-care domain relies on a small set of entailment graphs for 23 medical concepts. Our ongoing research focuses on the challenging task of learning a larger entailment graph for the health-care domain. We are also investigating methods for evaluating the exploration process (Borlund and Ingwersen, 1997). As noted by Qu and Furnas (2008), the success of an exploratory search system does not depend simply on how many relevant documents will be retrieved for a given query, but more broadly on how well the system helps the user with the exploratory process.

[2] `http://www.cs.tau.ac.il/~jonatha6/homepage_files/resources/HealthcareGraphs.rar`

Acknowledgments

This work was partially supported by the Israel Ministry of Science and Technology, the PASCAL-2 Network of Excellence of the European Community FP7-ICT-2007-1-216886, and the European Communitys Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 287923 (EXCITEMENT).

References

Michele Banko, Michael J Cafarella, Stephen Soderl, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceedings of IJCAI*, pages 2670–2676.

Jonathan Berant, Ido Dagan, and Jacob Goldberger. 2010. Global learning of focused entailment graphs. In *Proceedings of ACL*, Uppsala, Sweden.

Pia Borlund and Peter Ingwersen. 1997. The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 53:225–250.

Maged N. Kamel Boulos and Steve Wheeler. 2007. The emerging web 2.0 social software: an enabling suite of sociable technologies in health and health care education. *Health Information & Libraries*, 24:2–23.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering*, 15(Special Issue 04):i–xvii.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *EMNLP*, pages 1535–1545. ACL.

Thomas Hofmann. 1999. The cluster-abstraction model: Unsupervised learning of topic hierarchies from text data. In *Proceedings of IJCAI*, pages 682–687.

Mika Käki. 2005. Findex: search result categories help users when document ranking fails. In *Proceedings of SIGCHI*, CHI '05, pages 131–140, New York, NY, USA. ACM.

Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules for question answering. *Natural Language Engineering*, 7:343–360.

Yan Qu and George W. Furnas. 2008. Model-driven formative evaluation of exploratory search: A study under a sensemaking framework. *Inf. Process. Manage.*, 44:534–555.

Emilia Stoica and Marti A. Hearst. 2007. Automating creation of hierarchical faceted metadata structures. In *Proceedings of NAACL HLT*.

Mayank Trivedi. 2009. A study of search engines for health sciences. *International Journal of Library and Information Science*, 1(5):69–73.