Towards Acquisition of Taxonomic Inference

Piroska Lendvai

Tilburg centre for Creative Computing
Tilburg University, Netherlands
p.lendvai@uvt.nl

Abstract

On a pilot corpus of Dutch medical encyclopedia texts, we focus on the mechanism of *taxonomic inference* that involves the extraction of co-hyponym terms, and the taxonomic or domain-specific lexico-semantic relation in the form of a textual *hypothesis*, which coordinates these. An initial set of inference elements can be acquired by syntactic and semantic alignment, additionally driven by document structure. From the terms and the related hypothesis we can harvest lexical patterns, which are linked to annotated domain concepts. The aim of the process is to learn inference patterns and apply the system to short as well as unstructured documents where fewer or no discourse-level cues are available, in order to acquire new co-hyponyms linked to their coordinating term via a specified relation.

1 Introduction

It is shown in several studies that thematic coherence has an impact on document structure, hence identifying discourse relations can facilitate access to semantic content; for an overview of relevant issues in the field of discourse parsing see [2]. In the case of stylistically guided documents, pressure for brevity and clarity is superimposed on the creative interplay between form and factual content of narration. An *encyclopedia entry* often provides a definition of an entity of which several subtypes exist; these need to be separately treated in the entry. However, the use of tables, charts, bulleted lists and other visual means that are inherently well suited for representing hierarchically structured data is traditionally not favourable in dominantly textual media that need to remedy this by providing lexical, syntactic, and structural cues to the reader.

Proceedings of the 8th International Conference on Computational Semantics, pages 291–294,
Tilburg, January 2009.

Linguistic constructions that are employed to convey structured information can be macro-propositions, for example 'pre-announcements' of certain subtopics that are going to be addressed in the document, such as *There exists an acute and a chronic form of gallbladder disease*, or recurring lexical and syntactic elements ('chains') across the passages treating the subtopics, as in the excerpt from the encyclopedia entry on *Digestive tract*:

> The first organ is the tongue which is only present in the phylum Chordata. The second organ is the esophagus. ... The third organ is the stomach. ...

Such constructions serve to alleviate the readers' cognitive load required to make semantic inference about hierarchical and semantic relations (e.g., that *tongue*, *esophagus*, and *stomach* are co-hyponyms of the hypernym *digestive tract*, and all three are *organs* of it). Most importantly, many encyclopedia articles treating co-hyponyms do *not* feature any of these linguistic means, but operate by activating real-world knowledge, thus it is an empirical research issue how to trace back such information.

While extracting hyponym-hypernym pairs is a popular topic in the literature, initiated by the seminal paper of [1], identifying specific types of hypernymy relations has only been addressed for a limited set of general concepts [3]. We target the detection of domain-specific lexio-semantic relations (sometimes broadly called 'associative') together with traditional hypernymy and meronymy relations, by modelling it as a taxonomy inference phenomenon. Representing these relations between domain-specific entities can be a first step in the construction of an ontological model, which is an important asset of knowledge-based applications.

Taxonomy inference consists of an n-tuple of terms, two or more of them linked by a coordinate relation, where each of these is linked by either hypernymy or associative relation to a common coordinating term. The relation between mother and child node is governed by a certain semantic property of the parent node, which is in the case of meronymy and hypernymy best expressed by a noun (e.g. in the medical domain *method*, *phase*, *form*), or in the case of domain-specific relatedness by a verb (such as *occurs_in* or *attacks*). The coordinating relation can typically be induced from factual domain knowledge but often remains implicit in the text: while it is non-trivial to retrieve it, especially for the co-hyponymic terms, based on a single document and/or lexico-semantic resources, it is an extremely productive general mechanism of cognitive inference.

2 Data

Our pilot corpus comprises 108 entries from the Dutch Spectrum Medical Encyclopedia[1], these were obtained based on identifying repetitive occurrences of 13 topic labels manually assigned to layout-based sections (`treatment`, `symptoms`, `cause`, etc.) within an entry. Token-level annotation was also manual, using 12 coarse-grained domain concepts such as `body_part`, `duration`, `disease`, `disease_symptom`, `microorganism`, `method_of_diagnosis`.

We illustrate the complexity of phenomena underlying taxonomy inference by the following entry (the document and section titles are set in italics).

> *Jaundice*, or icterus, is a condition whereby a yellowish discoloration of the skin, the mucous membranes, and whites of the eyes appears. This is caused by increased levels of the gall pigment bilirubin in the blood serum. ...
> *Adults*
> Adults can develop jaundice by three means. By an interruption to the drainage of bile in the biliary system (e.g. due to a gallstone or a tumor). ... By diseases of the liver (e.g. hepatitis). The liver's ability to metabolise and excrete bilirubin is reduced, leading to a buildup in the blood. ... By an increased rate of bilirubine production. ...
> *Babies*
> Babies can develop a sort of jaundice (icterus neonatorum) shortly after birth as a consequence of relatively increased breakdown of red blood cells ... .

In this entry, the onsets of the two subsections exhibit some syntactic and lexical parallelism; after establishing their similarity, our aim is to locate the relation between the two main terms (*adults* and *babies*) in these sections. Spotting lexical overlap and using wildcards yields the expression *'X can develop Y'*, and we hypothesise that the verb phrase instantiates the domain-specific lexico-semantic relation shared by the coordinate terms in connection to the term *jaundice*.

Note that this entry does not contain a separate passage of text that can be designated as the hypothesis of the above relation, thus we fall back on the lexical string (i.e., the VP) when extracting the relation. Also note that taxonomy inference can be embedded: we observe syntactic similarity among three sentences in the *Adults* subsection. Their coordinating hypernymy relation is literally expressed by the very first sentence of the section, yielding the pattern *'X can develop Y by [NUM] means'*, this sentence is

[1] `http://www.winklerprins.com/linkpages/Algemenenaslag.html`

thus regarded as the *inference hypothesis* of the three sentences featuring the overlapping onsets.

3 Goals and Evaluation

Our goal is to investigate if we can bootstrap the linking of coordinate terms via a specific type of relation, inferrable from the given document collection. Next to measuring the utility of sentence alignment based on standard, automatically obtained syntactic (dependency, part-of-speech) and semantic (cosine, querying lexical semantic databases) similarity cues as well as discourse markers (anaphore) for this task, our focus is on the role lexical patterns can play in this process. We test the impact of two factors on harvesting lexical patterns as inference elements: (a) syntactic and semantic segmentation: enriching these with dependency as well as semantic roles, and (b) clustering the patterns according to domain concepts.

The results are manually evaluated, preparing training data for further bootstrapping. The system is eventually applied to the full IMIX corpus of Dutch medical encyclopedia texts[2] to process single-section (i.e., unstructured) documents and automatically detect and link coordinate terms by domain-specific relations in these new documents. The obtained terms are matched to available Dutch terminological resources in the medical domain[3].

References

[1] M. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proc. of COLING*, pages 539–545, 1992.

[2] M-P. Péry-Woodley and D. Scott. Introduction to the special issue on computational approaches to discourse and document structure. *Traitment Automatique des Langues*, 47(2), 2006.

[3] M. Thelen and E. Riloff. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proc. of ACL*, 2002.

[2] `http://ilk.uvt.nl/imix`

[3] `http://taalunieversum.org/taal/terminologie/medisch/`