Motivations and Objectives

Background
As the Web becomes easily accessible as a ‘read-write’ medium to more and more users, we can assist to an information overload:
- on the one hand, constantly updated information is available in the form of collection of Web pages, large document corpora, databases, and so on;
- on the other hand, this information is mostly published by humans for humans, it is unstructured and cannot be automatically consumed by a machine.


≡ Knowledge Acquisition Bottleneck
Creating large, usable, expandable and valid representations of semantics about a specific domain of interest represents the most time consuming task of a KM project.

 Ontology Learning from Text
A discipline whose objective is to subsample ontological proposition from collections of natural language propositions, relying on some characteristics which are intrinsic to text:
- the set of descriptive rules represented by the grammar and the syntax of the language (NLP approach)
- the domain specificity of a particular corpus of documents (ontological approach)
- the distribution of terms across all the documents in the corpus (statistical approach)

 Distributional Hypothesis
"Words are similar to the extent that they share similar context" (Harris, 1968)

 Objectives
Being Ontology Learning for Text an approximate activity, users can hardly foresee what is definitely the best solution that will extract the most usable and valid ontology from their document collection.

The main objective of this work is to offer a valid and versatile alternative to other approaches, allowing the user to try different possibilities and find between them the best solution, according to his needs.

This is done by developing a modular approach which:
- analyzes a corpus of documents and extracts its most relevant concepts;
- uses Correspondence Analysis to calculate similarity between these concepts and show how these relation in a 2D space;
- applies different techniques to derive a taxonomy from similarity relationships between concepts.

Our approach

1) Input
Two different corpora are passed as an input to the system: a training set, which is used as a reference to what is considered as common knowledge, and a test set, the one from which we want to extract the concept hierarchy.

2) Indexing
Documents are indexed using Apache Lucene, which takes care of tokenization, indexing, and stopwords filtering. As the indexing process is the more time-consuming, indices (such as the one for the training set) can be saved and reused.

3) Data Filtering
Three different filters can be applied: a Wordnet-based noun detector, an NLP-based one, and a short-terms discarder.

4) Relevant term identification
For every term t in the test corpus we calculate the Information Gain: $IG(t) = H_{theory} - \sum_{p \in P} \frac{|p|}{|P|} \log \frac{|p|}{|P|} H_{p}$, where the entropy $H$ for a generic document set $D_t$ is calculated as: $H_D = -\sum_i p(i) \log p(i)$.
Terms are then ordered by their IG and the top ones are taken as potentially relevant concepts.

5) Similarity computation
Similarity between terms is calculated by applying the framework of Correspondence Analysis. CA allows for a compact representation of term similarities by projecting them in a multidimensional Euclidean space. Term distances in this space are the distances between their m nearest neighbors in the Euclidean space generated with CA.

6) Hierarchy creation
The tool allows one to choose among four different options: Murtagh’s Algorithm [4], Hearst Patterns on Web [1, 2], Maaedche and Staab’s Bootstrapping [3], and an original combination of these last two.

Applications and results

Table 1: Precision measures obtained from the evaluation of the different ontology learning algorithms, obtained by comparing the tool’s results with relations proposed by a human judge.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>AI</th>
<th>Rome</th>
<th>Biology</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Murtagh</td>
<td>6.6%</td>
<td>5.22%</td>
<td>1.87%</td>
<td>4.56%</td>
</tr>
<tr>
<td>Hearst Patterns on Web</td>
<td>60.00%</td>
<td>75.00%</td>
<td>37.90%</td>
<td>57.56%</td>
</tr>
<tr>
<td>Heart Patterns on Web</td>
<td>90.00%</td>
<td>94.52%</td>
<td>85.00%</td>
<td>89.86%</td>
</tr>
</tbody>
</table>


Related Work: