

Project Report

Ontology Extraction

Nainy Sharma, Mitesh Shah

20172008 20172010



Introduction

In this project we try to extract domain-invariant ontologies. The basic methodology is inspired from the paper “Simple Method for Ontology Automatic Extraction from Documents” [Andreia Dal Ponte Novelli, Jose Maria Parente de Oliveira, IJACSA, 2012].

Ontology Extraction helps in Information Retrieval. Information Retrieval has become more and more complex due to growing number and variety of document types. Thus we need ontologies to formally represent concepts. We are refraining from using a rule-based system since such systems are either domain specific or heavily rely on linguistics and need human intervention to form such rules. We try to develop a method which is fast and simple and performs automatic extraction of ontologies from documents or a collection of documents that is independent of the document type and uses the junction of several theories and techniques such Latent Semantic Analysis, Latent Dirichlet Allocation for extraction of initial concepts, WordNet and similarity to obtain correlation of concepts.

Method Used

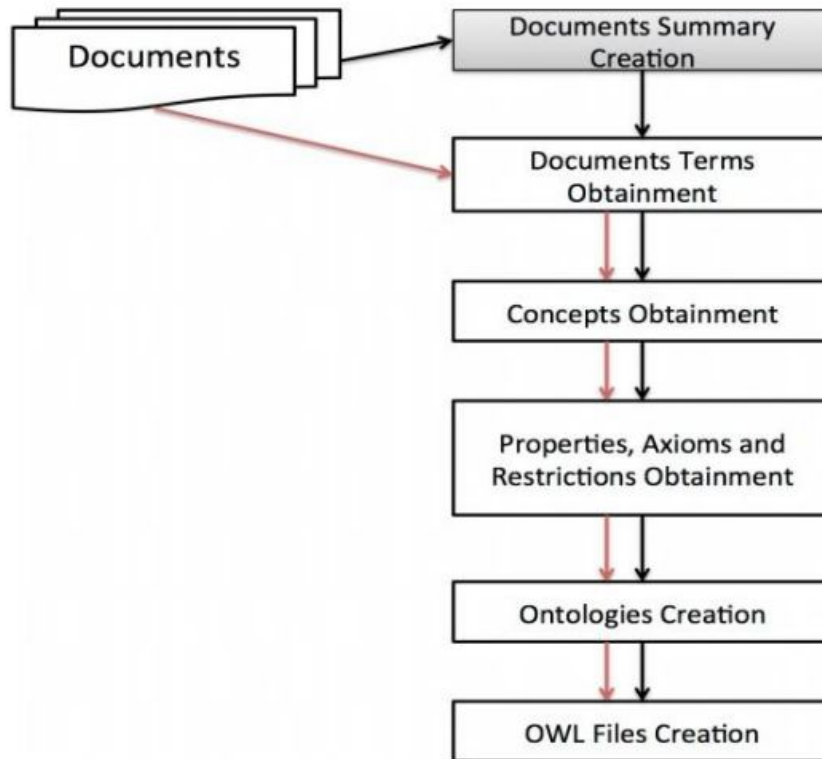



Figure 4. General outline of the proposed method operation.

The general method is outlined in the paper as follows:

1. Perform Summarization on Documents :- This step is avoided and we used datasets that were usually used to train summarization models and we use summaries from those datasets directly for our task.
2. Documents Terms Obtainment :- We clean the documents, perform tokenization, stop words removal, lemmatization to obtain the terms and then form a Term-Document Matrix and create a TF-IDF Matrix u

Summarization of Documents

This step was avoided due to the following reasons:



We wanted to prefer Abstractive Summarization over Extractive Summarization and thus for abstractive summarization we need a dataset that has pre-existing summaries. We implemented both the summarizers and then decided to use the summaries that were originally given for the dataset directly to ease our process.

Document Terms Obtainment

We perform the following tasks on the text data:

1. Case Folding :- Converted the whole text to lowercase.
2. Stop Words Removal :- Removed stopwords (used English corpus so English Stop Words)
3. Punctuation Removal :- Documents like legal case reports etc have a lot of punctuation or special symbols and thus they were removed.
4. Word Normalization :- Used NLTK's WordNet Lemmatizer to normalize the words to their root forms.

After this, we form Term Document Matrix (Bag of Words) on the dataset and then form TF-IDF Matrix on the same.

Concepts Obtainment


Using Latent Semantic Analysis

In the paper mentioned above, the method to obtain concepts was specified as Latent Semantic Analysis (using SVD Decomposition of matrix) and thus we did the same. We created the TF-IDF Matrix and implemented LSA on those matrices. From SVD, we get Term-Concept Matrix. We use this matrix to create concepts containing the respective terms.

As mentioned in the paper, the indexed terms are first level of ontologies and the concepts obtained from this matrix decomposition, are the second level ontologies. For the further level of ontologies, we use Hierarchical Agglomerative Clustering and create a Dendrogram. After analyzing the Dendrogram, we cluster the concepts that seem similar and create clusters of such concepts and then we analyze the clusters.

Using Latent Dirichlet Allocation

Since SVD itself is a computation heavy operation, we also implemented Latent Dirichlet Allocation technique (also commonly used for Topic Modelling with LSA). With LDA, we get two matrices, Terms-Concepts and Concepts-Documents.



Using the same analysis as done above, we will obtain the concepts, our second level ontologies from the terms-concepts matrix. For comparison of results, we implemented LDA on Bag of Words as well as the TF-IDF Matrices.

Again we use Hierarchical Agglomerative Clustering and create a Dendrogram. After analyzing the Dendrogram, we cluster the concepts that seem similar and create clusters of such concepts and then we analyze the clusters.

Following Datasets were used for analysis:

1. Papers on Linguistics
2. Legal Case Reports
3. Book Summaries
4. UN General Debates

We also tried to perform the extraction on corpus containing multiple documents and corpus containing lines of a single documents and analysed the results.

Findings

For better extraction, it was evident that keeping the extraction for corpus containing lines of single document gives better concepts but for that we may need to perform all the LSA/LDA Steps on each document which may be resource consuming. On the flipside, doing the extraction on single documents reduces the complexity of checking whether found concepts belong to same document and creates the task of extracting ontologies from a single document easier.

Here are some concepts obtained from UN General Debates using LSA, LDA with Bag of Words and LDA with TF-IDF values:

Page 10 of 10

Page 10 of 10

A word cloud of terms related to nuclear weapons and proliferation. The words are arranged in a dense, overlapping manner. The most prominent words are 'nuclear', 'proliferation', 'weapon', 'development', 'opportunity', 'relationship', 'cost', and 'proliferation'. Other visible words include 'peaceful', 'stability', 'session', 'office', 'reconciliation', 'prospect', 'regional', 'discharged', 'fuse', 'conventional', 'issue', 'expressing', 'objection', 'grim', 'excellency', 'delegation', 'perception', 'call', 'deployment', 'exemplary', 'fortythird', 'view', 'horizontal', 'duty', 'utilized', 'strong', 'much', 'caputo', 'potential', 'effort', 'peace', 'strongly', 'political', 'sharing', 'negotiation', 'peaceful', 'stability', 'session', 'office', 'reconciliation', 'prospect', 'regional', 'discharged', 'fuse', 'conventional', 'issue', 'expressing', 'objection', 'grim', 'excellency', 'delegation', 'perception', 'call', 'deployment', 'exemplary', 'fortythird', 'view', 'horizontal', 'duty', 'utilized', 'strong', 'much', 'caputo', 'potential', 'effort', 'peace', 'strongly', 'political', 'sharing', 'negotiation'.

LDA with Bag Of Words and TF-IDF work equally well but since we are using the weights of words in LDA as the vectors for terms, the weights in TF-IDF version are very small and thus the clustering for such concepts becomes harder since all concepts seem very close in the space. Thus, even though Bag of Words and TF-IDF approaches are working equally well, after clustering the concepts obtained from Bag of Words seem slightly better.