How to Construct Multilingual Domain Ontologies

Nitsan Chrizman, Alon Itai

Computer Science Department
Technion, Haifa, Israel
nitsan.ch@gmail.com, itai@cs.technion.ac.il

Abstract

The research focuses on automatic construction of multi-lingual domain-ontologies, i.e., creating a DAG (directed acyclic graph) consisting of concepts relating to a specific domain and the relations between them. The domain example on which the research performed is *Organized Crime*. The contribution of the work is the investigation of and comparison between several data sources and methods to create multi-lingual ontologies.

The first subtask was to extract the domain's concepts. The best source turned out to be Wikepedias articles that are under the catgegory. The second task was to create an English ontology, i.e., the relationships between the concepts. Again the relationships between concepts and the hierarchy were derived from Wikipedia. The final task was to create an ontology for a language with far fewer resources (Hebrew). The task was accomplished by deriving the concepts from the Hebrew Wikepedia and assessing their relevance and the relationships between them from the English ontology.

Keywords: Domain ontology, Hebrew ontology, Wikipedia

1. Introduction

Our research focuses on automatic construction of multilingual domain-ontologies, i.e., creating a DAG (directed acyclic graph) consisting of concepts relating to a specific domain and the relations between them. We wished to build a hierarchy such that concepts of the domain are at the bottom level, and related concepts are grouped into (non-overlapping) categories, which in term are grouped into more general categories, until reaching a single root.

Our interests focussed on the containment relation, and did not research into the semantics of other relations. We examined the following subtasks: extracting the domain's concepts; creating an English ontology; and finally how to use the English ontology, which was created for a language with vast resources, to a language with far fewer resources (Hebrew).

Our domain example on which we performed the research is *Organized Crime*. In such a dynamic domain, new concepts (such as criminals) are constantly being added, so it is infeasible to construct the ontology manually. Therefore, we looked only at completely automatic methods and did not investigate methodologies for manually constructing the ontology.

To collect the concepts, we experimented with several data sources: comparing general and domain specific corpora, WordNet and Wikipedia. While the first two sources were promising, the final results were unsatisfactory: we missed too many important concepts, and included many non related concepts. Wikipedia proved to be a better source for finding concepts, but it too yielded many unrelated concepts which we had to remove automatically.

The structure of the ontology was also derived from Wikipedia, but many categories connected unrelated concepts, so they too had to be pruned. The conepts of the Hebrew ontology were constructed from Wikipeia, similarly to the English one. However, since the Hebrew Wikepedia is relatively small, it lacked many connections between concepts. We used the English ontology to find the lacking connections.

Section 2. surveys related work. The data sources are described in Section 3. Section 4. develops methods for collecting the concepts of the English ontology. Methods for establishing the relations between the concepts are described in Section 5. Finally methods for developing the Hebrew ontology are described in Section 6.

2. Related Work

Gómez-Pérez et al. (2010) give a wide survey of works on ontology, with an emphasis on the semantics and how to manually create an ontology. In contrast, we limit our ontology to the subsumption relation and on automatic methods of acquisition. The following works all discuss automatic acquisition.

Wu and Weld (2008) developed the Kylin Ontology Generator (KOG), an autonomous system that builds a rich ontology by combining Wikipedia infoboxes with WordNet using statistical-relational learning. It predicts subsumption relationships between infobox classes while simultaneously mapping the classes to WordNet nodes. KOG also maps attributes between related classes, allowing property inheritance.

Biebow and Szulman (Biebow and Szulman, 1999) presented TERMINAE, a computer-aided knowledge engineering tool which helps building an ontology based on relevant corpus. Its originality is to integrate linguistic and knowledge engineering tools. The linguistic engineering part allows the definition of terminological forms from the study of term occurrences in a corpus. The knowledge engineering part involves knowledge-base management for the ontology.

Velardi, Missikoff and Basili (Velardi et al., 2001) described a text mining technique to aid an Ontology Engineer to identify the important concepts in a Domain Ontology, based on relevant concepts. They used natural language processing tools for two tasks: discovery of terms that are good candidate names for the concepts in the Ontology, and identification of taxonomic relations among these terms.

Navigli and Velardi (2004) presented a method and a tool, OntoLearn, aimed at the extraction of domain ontologies from Web sites. OntoLearn first extracts a domain terminology from available documents. Then, complex domain terms are semantically interpreted and arranged in a hierarchical fashion. Finally, a general-purpose ontology,WordNet, is trimmed and enriched with the detected domain concepts.

Maedche and Staab (2000) developed a general architecture for discovering conceptual structures and engineering ontologies. They used a generalized association rule algorithm that does not only detect relations between concepts, but also determines the appropriate level of abstraction at which to define relations.

Syed, Finin and Joshi (Syed et al., 2008) described the use of Wikipedia and spreading activation to find generalized or common concepts related to a set of documents using the Wikipedia article text and hyperlinks. They started their experiments with the prediction of concepts related to individual documents, and extended them to predict concepts common to a set of related documents.

Suchanek, Kasneci and Weikum (Suchanek et al., 2008) presented YAGO, a large ontology with high coverage and precision. YAGO has been automatically derived from Wikipedia and WordNet. It comprises entities and relations, and contains more than 1.7 million entities and 15 million facts. These include the taxonomic Is-A hierarchy as well as semantic relations between entities. The facts for YAGO have been extracted from the category system and the infoboxes of Wikipedia and have been combined with taxonomic relations from WordNet. A powerful query model facilitates access to YAGO's data.

Hepp et al. (2006) show that standard Wiki technology can be easily used as an ontology development environment for named classes, prove that the URIs of Wikipedia entries are surprisingly reliable identifiers for ontology concepts, and demonstrate the applicability of this approach in a use case.

Declerck, Prez, Vela, Gantner and Manzano-Macho (Declerck et al., 2006) implemented a platform that allows the user to upload a specific ontology, to select labels of the ontology and the language to which this label should be translated. Once the user has made his selections, the systems accesses the EuroWordNet and Wikipedia databases for finding if the selected term is encoded in the resources and displays the results of the search to the user, who can then decide if the suggestions made by EWN or Wikipedia are appropriate.

de Melo and Weikum (2010) investigated how entities from all editions of Wikipedia as well as WordNet can be integrated into a single coherent taxonomic class hierarchy. they relied on linking heuristics to discover potential taxonomic relationships, graph partitioning to form consistent equivalence classes of entities, and a Markov chain-based ranking approach to construct the final taxonomy. This results in MENTA (Multilingual Entity Taxonomy).

Lin and Krizhanovsky (2011) developed automatic translation for obtaining correct matching pairs in multilingual ontology matching. In the case study, the problem entity is a task of multilingual ontology matching based on Wiktionary data accessible. Ontology matching results obtained using Wiktionary were compared with results based on Google Translate API.

In our work we tried to develop a new approach, which uses a foundation ontologies in order to extract a domain ontology in the English language, and based on it, in other language as well. In contrast, the works described above, tried to enrich an existing ontology, or build an ontology from other databases. Furthermore, the works which deal with multilingual ontologies, didn't concentrate on translating a domain ontology based on foundation ontology, which is the basis of our method. Thus, in this work we developed innovative ways to deal with these new aspects, from a different direction.

3. Data

We examined three data sources:

3.1. Contrasting corpora

We collected a pair of corpora: The first consisted of 800 articles belonging to the subject's domain (organized crime). These articles were automatically collected from web-sites dealing with organized-crime, relevant sections in online newspapers, works and articles in this domain, etc. The second corpus was a general topic corpus containing 8304 articles on a wide range of topics and was used as a control group. Its articles were collected from sources similar to the first corpus, apart from not being constrained to the subject topic.

3.2. WordNet

WordNet is a freely and publicly available semantic dictionary of English, developed at Princeton University. We used the JWI (the MIT Java WordNet Interface) which supports access to WordNet versions 1.6 through 3.0, among other related WordNet extensions.

3.3. Wikipedia

Wikipedia is a multilingual, web-based, free-content encyclopedia project. Its vast scope (4,252,811 English articles in June 2013), its diversity, and its data organization structure made it our premier choice. We used the DBpedia knowledgebase project, which contains the information of the English Wikipedia (June 2012 version) in pre-processed files. Furthermore, it contains interlinks for most of the articles, from its English version to the corresponding article in up to 111 languages (including Hebrew).

The Wikipedia database can be viewed as a DAG, whose leaves are Wikipedia's articles and whose internal nodes are Wikipedia's categories.

4. Accumulating the domain's basic concepts

4.1. Contrasting corpora

4.1.1. Method

Using the two corpora (Section 3.1.), our hypothesis was that concepts relevant to the domain will appear more often in the domain corpus than in the general one. Moreover, since many of the domain's concepts consisted of more than one word, we concentrated on collocation (sequences of words that occur frequently), hoping that they will help us identify relevant multi-word expressions (MWE).

We used three kinds of measurements in order to find the MWE most relevant to the domain: TF-IDF – term frequency–inverse document frequency, DC – domain-consensus (Navigli and Velardi, 2004), PMI – Pointwise mutual information. The PMI of a pair of outcomes x and y, measures the discrepancy between the probability of their coincidence given their joint distribution and their individual distributions, assuming independence, i.e., $PMI(x, y) = \log \frac{p(x,y)}{p(x)p(y)}$.

4.1.2. Evaluation

We examined 80 expressions using several combinations of these three measures. Since we don't have a listing of all the relevant concepts, we cannot calculate how many of them weren't found. Thus we were able to measure only the true-positives. The best combination was between TF-IDF and PMI (64.57% true positive) which is still pretty low, so we decided to abandon this data source.

4.2. WordNet

Even though WordNet is a general ontology, it does not contain many multi-word concepts pertaining to the domain (organized crime). Also, it contains very few proper names, which constitute a major part of our domain ontology. Moreover, it misses many connections between related concepts; for example, "organized crime" is not recognized as a subconcept of the concept "crime". Hence this data source was also abandoned.

4.3. Wikipedia's article names

Considering each Wikipedia article as a concept, our goal is to select the ones that belong to the domain based on Wikipedia's structure. In the described methods, we concentrated only on the sub-DAG O_CRIME which consists of all the Wikipedia nodes reachable from the "organized-crime" category, and ignored the remaining articles. To justify this decision, we examined a random set of 500 Wikepedia articles that do not belong to O_CRIME. None was classified as relevant. The large amount of Wikipedia articles prevented us from classifying each article separately, but based on the above results and on the Wikipedia structure, it is reasonable to assume that most of the desired and relevant articles belong to O_CRIME.

We tried four methods as described below. Each relies on the conclusions of the previous ones.

4.3.1. The naïve approach

The motivation: The naïve assumption is that every leaf (article) of O_CRIME should be a relevant concept, because it is an offspring of the category "organized-crime". To evaluate this assumption we manually classified 319 random articles from O_CRIME. Of the 319 articles, only 174 (55%) were relevant. A closer examination revealed that O_CRIME contains many irrelevant articles that deal with topics such as crime-films, thriller-authors, presidents, etc. Thus, this approach didn't provide satisfactory results, but led us to the real challenge: the new goal is to correctly and efficiently separate the relevant leaves of O_CRIME from the irrelevant ones.

4.3.2. Category based

In this step, we tried to classify these articles, based on their parents (categories), i.e., nodes that have a link to the article. An article was deemed *relevant* iff at least λ percent of its parents belonged to O_CRIME. The best results were achieved for $\lambda = 80\%$, for which recall = True-positive/ (True-positive + False-negative) = 64%, precision = True-positive/ (True-positive + False-positive) $=$ 72% and F-measure $= 68\%$.

The results were not yet satisfactory. The large percentage of false-negatives is due to categories like year of birth, year of death, nationality, etc. which appear as parents of relevant articles, and affected the results.

4.3.3. The expanded method

To improve the results we tried first to evaluate the relevance of the categories, based on the percentage of leaves of the category that belongs to O_CRIME. Then a concept was classified as relevant iff most of its parents were relevant. We experimented with various values of λ, but the results were even worse than the naïve approach.

4.3.4. Sub-DAG intersections

In this step we tried to find other sub-DAGs in the Wikipedia DAG to help us correctly classify O_CRIME concepts, using their intersection with O_CRIME. For example, we expected that the intersection of O_CRIME and "Category: Movies" sub-DAG, will contain mainly movies about organized crime. Thus, we hoped that removing the articles in the intersection from the relevant concepts set would bring us closer to our goal. The results didn't meet our expectations, and they were even worse. For example, many crime movies did not belong to the intersection, and this intersection also contained many concepts that were not movies. We suspect that the fact that Wikipedia has a large number of authors affected the consistency of the database's structure.

4.4. Wikipedia's abstract

Due to the unsatisfactory results of the previous methods, we decided to expand the information we relied on in the classification problem, and to look also at the abstracts of the articles. To this purpose, we considered only articles in O_CRIME, as mentioned above.

The method: We constructed a training set L of 290 articles, half of which (R) were classified by a team of human experts as relevant and the rest ($L -$ O_CRIME) were classified as irrelevant. We then examined the words appearing in all the abstracts. For each word we defined its *relevance* as the ratio between the number of its occurrences in R and in $L - R$. The *s-most relevant words* (REL_s^+) are the s words whose ratio is largest, the *s-least relevant words* (REL_s^-) are the s words whose ratio is lowest. The *s-relevance* of a document d is the ratio between the number of words of d which belong to REL_s^+ and the number of words that belong to REL_s^-. Finally, for a parameter $0 < \lambda < 1$, those articles whose s-relevance is greater than

λ	Unclassified	Recall	Precision	F-score
0	0.076	0.97	0.96	0.96
0.3	0.076	0.97	0.96	0.96
0.4	0.083	0.98	0.96	0.97
0.5	0.11	0.98	0.98	0.98
0.6	0.11	0.98	0.98	0.98
0.7	0.134	0.98	0.99	0.98
0.8	0.159	0.99	0.99	0.99
0.9	0.183	0.99	0.99	0.99
1	0.228	0.99	0.99	0.99

Table 1: The results of "Wikipedia's abstract", for several values of the certainty parameter λ.

λ high are deemed *relevant*, and those whose s-relevance was less than λ were deemed *irrelevant*. The remaining articles remained *unclassified*.
For $s = 100$ the method was tested on a test set that was disjoint of the learning set and the results for several values of λ are depicted in Table 1. In order to validate the statistical significance of the results, 5-fold-cross-validation was performed — the results were indeed significant ($p < 5\%$). Experimenting with other values of s did not improve the results.

5. Finding the relationships between the domain's basic concepts

This is the second and final step in building the English ontology. The goal of this step is to find the connections (edges) between the concepts (nodes), that were obtained in the previous step. For this purpose, we decided to rely on a foundation ontology. The general idea was to find the concepts in the foundation-ontology, and then derive the appropriate relationships between them.

5.1. The WordNet foundation-ontology

We tried to use this ontology in order to deduce the relationships between the concepts we accumulated in the previous task, to create a full ontology. We came across several problems when trying to use this ontology:

- Apart for a small number of exceptions, WordNet doesn't contain proper names. So, the insertion of names to the ontology had to be made separately.
- Many relevant concepts are missing in WordNet. Out of the 183 manually-classified relevant Wikipedia's titles, only 18 were found in WordNet. A large part of the remaining 165 articles were proper names, but there are also a lot of relevant MWE we expected to find, for example "drug diversion".
- WordNet misses many connections. For example WordNet doesn't contain a connection between "organized-crime" and "crime". This lack of relevant and important connections in WordNet makes it impossible to create a reasonable domain ontology based on it.

When taking into account these three problems, the best course of action was to use another ontology.

5.2. The Wikipedia database

Wikipedia's large amount of concepts, especially named entities, along with more accurate and various connections between them (compared to WordNet), makes it most suited to our purpose. To find the "best connection" between the concepts in the Wikipedia database, we performed a depth-limited BFS (breadth-first search) on each concept (the depth defined in the initialization point). For any two nodes (articles or concepts) we looked for the most specific concept that generalizes both. Concept c_1 is considered more specific than concept c_2 if c_1 has less descendants and it is further away from the root.
For each two concepts, we look at the BFS-tree, find all the intersections, and choose the best one, based on two criteria:

- Number of children – If their number is large, the category is wider and worse.
- Depth (distance from root) – If this number is large, the category is specific and better.

Assumption: As mentioned above, we treated only article-category and category-category edges, so the connections between the basic concepts (articles) contain only category concepts. Since all the basic concepts are in the organized crime domain, and moreover, they are all in O_CRIME, we assumed that between any two basic concepts there is a (undirected) path which is entirely contained in O_CRIME. This assumption proved to be valid for our domain (organized crime). Thus we limited our exploration to O_CRIME.

To find the connections we employed an iterative method – at each step we perform a single BFS-step for each concept and save, for this concept, the new categories achieved during this step. Then, for each concept, we check if any of these new categories belongs to the list of some other concept. If there is such category, a new connection between these two concepts is established. If the two concepts were already connected, we compare the new connection to the previous one using the criteria of number of children and depth (as explained above), and save only the more favourable connection. Note, a category which is n_1-edges away from the first concepts, and n_2-edges away from the second, will surely be found in $max\{n_1, n_2\}$ steps.

Algorithm (number-of-iterations i, sub-trees-parameter a, depth-parameter b)

Initialization
For each concept create a single node tree, with the concept as the root.

For i iterations, for each concept c:

1. Perform a single BFS step of the concepts trees, and add, the newly discovered concepts to the tree of c. (If a concept which already exists in the tree is rediscovered- ignore it).

```
2. If, as a result of this step, the BFS trees of two concepts intersect:
   For each intersection point:
   (a) If no previous connection is found between those concepts:
       i. Save the intersection point as the "connection concept"
       ii. Save the path from each one of the two basic concepts to the intersection point as the "connection path"
   (b) Else:
       i. Let $n_1$ be the previous "connection concept"
       ii. Let $n_2$ be the new "connection concept"
       iii. If $score(n_1) > score(n_2)$:
            Remove the old connection
            Save the new "connection concept" and the "connection path"
```

```
function $score$(current-node $v$, sub-trees-parameter $a$, depth-parameter $b$)
  1. Let $cMax$ be the max number of children of one node, from all the nodes in $W$, the Wikipedia DAG.
  2. Let $h$ be the height of $W$ (the max distance from node to the root)
  3. Let $c$ be the number of children of $v$.
  4. Let $d$ be the depth of $v$.
  5. Return
     $\max\{(\frac{c}{cMax} \times a) + (1 - (\frac{d}{h} \times b)), score(parent) + 1\}$.
```

Evaluation

We claim, based on the assumption made earlier and on the final results, that the algorithm finds a connection between every pair of concepts, and that these connections are in the domain scope. On the other hand, it is difficult to measure the quality of the found ontology. When trying to do so, several questions coming up:

- What is the definition of quality ontology?
- How to compare two ontologies?
- To which ontologies it will be valuable to compare?

The process resulted in an English domain ontology. It is difficult to measure the quality of the resultant ontology. This is especially hard when we don't have a domain ontology with which to compare.

6. Creating the corresponding Hebrew ontology

At first we intended to use bilingual dictionaries, such as the multi-lingual WordNet. This approach failed since WordNet lacked many important concepts such as proper names. To overcome this problem we tried to add names by transliterating names written in the Latin alphabet to Hebrew. This transliteration is phonetic and since their pronunciation depends on the original language (compare *Charles de Gaulle* and *Prince Charles*) their transliterations depend on the original language which is not always English. Consequently, to perform the transcription we need semantic knowledge, which was beyond out means. Thus dictionary transfer even with transliteration did not yield satisfactory results.

We next tried to use Wikipedia's interlinks between the English articles and their Hebrew counterparts. We examined a set of 145 English articles and even though we expected to find counterparts for most of them, only 27 (19%) had a Hebrew counterpart. On the other hand, many Hebrew Wikipedia articles have no English counterpart (such as articles on Israeli mobsters). We therefore had to add concepts which are unique to the Hebrew Wikipedia.

Our first attempted was to add the neighbors of relevant concepts in the Hebrew Wikipedia DAG: For each relevant article we first inserted all its ancestors and then added all the descendants of the new categories. We conjectured that the relevant concepts that were not found are siblings of found concepts. The result of this method was poor – the vast majority of the added articles were not relevant. It turned out that this method adds many irrelevant concepts, such as year of birth and origin. These general concepts have many descendants, most of which are irrelevant.

Following the construction of the English ontology, we restricted ourselves to subgraphs of H_CRIME, the Hebrew organized-crime DAG, which similar to the English O_CRIME DAG (Section 4.3.), is the sub-DAG rooted at the Hebrew node "Category: Organized crime". As with O_CRIME, it contained many irrelevant concepts, so our task was to weed them out.

Let $A(R)$ be the set of ancestors in O_CRIME of the articles of R, where R is the set of 145 relevant English articles as described in Section 4.4.. A Hebrew article was deemed *relevant* if at least one of its ancestors in H_CRIME had a counterpart in $A(R)$. The algorithm yielded 43 Hebrew articles, which upon manual inspection, all proved to be relevant.

Calculating the false-negative is more complicated. Because of the small size of R, many relevant articles in H_CRIME were not deemed relevant by our algorithm.

Inspecting Hebrew concepts, which were mistakenly classified by the algorithm as irrelevant, revealed that each such concept has at least one ancestor whose corresponding English category is an ancestor of a relevant English article. We therefore conjecture that using all of the English relevant articles during the translation process, would show that all the articles retrieved by our method are relevant. We were unable to prove this conjecture because of the size of Wikipedia in general and of O_CRIME in particular.

7. Acknowledgements

This research has been partially supported by the EU grant CAPER no: 261712.

8. References

Biebow, Brigitte and Szulman, Sylvie. (1999). TERMINAE: a method and a tool to build a domain ontology. In *EKAW '99 Proceedings of the 11th European Workshop on Knowledge Acquisition, Modeling and Management*, pages 49–66, London, UK. Springer-Verlag.

de Melo, Gerard and Weikum, Gerhard. (2010). Menta: inducing multilingual taxonomies from Wikipedia. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1099–1108, New York, NY, USA. ACM.

Declerck, Thierry, Gómez-Pérez, Asunción, Vela, Ovidiu, Gantner, Zeno, and Manzano, David. (2006). Multilingual lexical semantic resources for ontology translation. In *In Proceedings of the 5th International Conference on Language Resources and Evaluation*.

Gómez-Pérez, Asunción, Fernández-Lopez, Mariano, and Corcho, Oscar. (2010). *Ontological Engineering*. Springer.

Hepp, Martin, Siorpaes, Katharina, and Bachlechner, Daniel. (2006). Harvesting Wiki consensus - using Wikipedia entries as ontology elements. In *IEEE Internet Computing*, pages 54–65.

Lin, Feiyu and Krizhanovsky, Andrew. (2011). Multilingual ontology matching based on wiktionary data accessible via sparql endpoint. *CoRR*, abs/1109.0732.

Maedche, A. and Staab, S. (2000). Semi-automatic engineering of ontologies from text. In *In Proceedings of the 12th Internal Conference on Software and Knowledge Engineering*, pages 231–239.

Navigli, Roberto and Velardi, Paola. (2004). Learning domain ontologies from document warehouses and dedicated web sites. *Comput. Linguist.*, 30(2):151–179, June.

Suchanek, Fabian M., Kasneci, Gjergji, and Weikum, Gerhard. (2008). Yago: A large ontology from wikipedia and wordnet. *Web Semant.*, 6(3):203–217, September.

Syed, Zareen, Finin, Tim, and Joshi, Anupam. (2008). Wikipedia as an ontology for describing documents. In *Proceedings of the Second International Conference on Weblogs and Social Media*. AAAI Press, March.

Velardi, Paola, Missikoff, Michele, and Basili, Roberto. (2001). Identification of relevant terms to support the construction of domain ontologies. In *Proceedings of the workshop on Human Language Technology and Knowledge Management - Volume 2001*, HLTKM '01, pages 5:1–5:8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wu, Fei and Weld, Daniel S. (2008). Automatically refining the Wikipedia infobox ontology. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 635–644, New York, NY, USA. ACM.