Okinet: Automatic Extraction of a Medical Ontology From Wikipedia

Vasco Calais Pedro[1], Radu Stefan Niculescu[2], Lucian Vlad Lita[2]

[1] Language Technologies Institute, Carnegie Mellon University

[2] Siemens Medical Solutions

Abstract

The medical domain provides a fertile ground for the application of ontological knowledge. Ontologies are an essential part of many approaches to medical text processing, understanding and reasoning. However, the creation of large, high quality medical ontologies is not trivial, requiring the analysis of domain sources, background knowledge, as well as obtaining consensus among experts. Current methods are labor intensive, prone to generate inconsistencies, and often require expert knowledge. Fortunately, semi structured information repositories, like Wikipedia, provide a valuable resource from which to mine structured information. In this paper we propose a novel framework for automatically creating medical ontologies from semi-structured data. As part of this framework, we present a Directional Feedback Edge Labeling (DFEL) algorithm. We successfully demonstrate the effectiveness of the DFEL algorithm on the task of labeling the relations of Okinet, a Wikipedia based medical ontology. Current results demonstrate the high performance, utility, and flexibility of our approach. We conclude by describing ROSE, an application that combines Okinet with other medical ontologies.

Motivation

In the last decades the number of available medical ontologies has grown considerably. These ontologies enable the use of previous medical knowledge in a structured way. Applications of medical ontologies include: more effective search of patient records, hospital quality improvement programs, (semi)automatic ICD-9 coding for insurance reimbursement, preliminary symptom-based diagnosis, ambiguity reduction when choosing medical tests, and classification of diseases, symptoms, and other medical concepts. For example, when trying to answer whether a patient was prescribed *Aspirin* (for hospital quality improvement measures), one needs to consider similar terms (such as *Ecotrin*, *Bayer pain reliever*, etc). Also, when performing (semi)automatic patient ICD-9 coding, it is useful to map conditions that can be described in various ways (*Heart Attack* can be also stated as *AMI* or *MI* or *Myocardial Infarction* or simply *Infarction*). For preliminary diagnosis at the point of care, ontologies can help by quickly returning diseases that have a given set of symptoms (instances of symptoms and diseases are concepts related by the "symptom of" relationship).

Several proprietary and public efforts such as MESH(Lipscomb 2000) and SNOMED(Spackman, Campbell et al. 1997) have become available and UMLS(Bodenreider and Journals) is rapidly becoming a de facto standard for medical ontologies, containing more than 100 dictionaries. Other medical ontologies include: RadLex (Radiology Information Resource**)**, OMIM (Online Mendelian Inheritance in Man**)**, MEDCIN (medical terminology), LOINC (Logical Observation Identifiers Names and Codes) and ICD-9/ICD-10 Codes.

At the same time, large sources of encyclopedic knowledge are becoming readily available in wiki like form. Resources such as Wikipedia(Denoyer and Gallinari 2006), the largest collaboratively edited source of encyclopedic knowledge(Krotzsch, Vrandecic et al. 2005; Völkel, Krötzsch et al. 2006), Scholarpedia(Izhikevich 2007), Citizendium(Sanger 2007) and the recently launched, incipient Google Knols Project are examples of semi-structured encyclopedic knowledge bases that provide a natural way to collect human knowledge(Lih 2003), with the advantage of naturally solving, to a large degree, the problem of consensus.

These resources represent an intermediate step between unstructured text and structured knowledge and are seen as potential viable sources of knowledge for automatic construction of medical ontologies.

In this paper we propose a general framework to mine structured knowledge from Wikipedia and apply it to the creation of a medical ontology. The paper proceeds as follows: we first discuss related work, and then describe the general framework for building a medical ontology from Wikipedia. We demonstrate our Directional Feedback Edge Labeling algorithm on a task of labeling the relations in Okinet, a Wikipedia based medical ontology. We conclude with a description of the Okinet browser as well as some interesting and promising ideas for future work.

Related Work

Maedche(Maedche and Staab 2002; Maedche 2002) and Navligli et al.(Navigli, Velardi et al. 2003) explored semi-automatic methods for concept and relation extraction, focusing on building ontologies from broad domain documents. Blake and Pratt(Blake and Pratt 2002) worked on extracting relationships between concepts from medical texts. Khoo et al.(Khoo, Chan et al. 2002) matched graphical patterns in syntactic parse trees in order to look for causal relations.

Several pieces of previous work focused on the link structure of Wikipedia to derive structure. Kozlova(Kozlova 2005) mined the link structure in Wikipedia for document classification. Milne et al.(Milne, Medelyan et al. 2006) used the basic link structure to construct domain specific thesauri and applied it to the agriculture domain. Bhole et al.(Bhole, Fortuna et al. 2007) used document classification techniques to determine appropriate documents in Wikipedia that were later mined for social information (people, places, organizations and events).

The Wikipedia Structure

Wikipedia general structure consists of an *article name*, which is unique within the particular wiki structure and thus suitable for a *concept name*, and *links* connecting articles, which are suggestive of semantic relations between them. Each article is typically divided in sections and sometimes contains tables that synthesize information pertinent to the article.

Within the different types of inter-article links, we often find *redirects* (articles that consist solely of a link to another article) and when we find this type of link we can interpret the two concepts described by those articles as synonyms. Each article is normally inserted into one or more categories, thus creating a set of hierarchical relations.

Even though each link seems to carry semantic information between two concepts, only a small percentage is typically used in mining Wikipedia, namely the *redirects* and *categories*. The main challenge of this work is to assign the correct semantic label to the extracted links deemed of interest, when the link is not a redirect.

General Methodology

We propose that we should take an inclusive approach rather than a selective approach to create a medical ontology, where we start by including all the article names as concepts and all the existing links as potential relations. We subsequently rely on extracted features to assign labels, finally discarding links without labels.

The goal is to first create a directed unlabeled graph that mimics the link structure, use the extracted features to generate a small amount of labeled data and run a Directional Feedback Edge Labeling Algorithm to extend the labels to the rest of the links, discarding the links with confidence below a preset threshold.

Feature Extraction

For every link extracted we store a set of features that are associated with that link. The set of features consists of the following:

Document Title

The title of the document where the link was found. This corresponds to the source concept.

Section Header Path

The path composed of the sections up to the section where the link was found. E.g. Diagnosis → Symptoms.

Context

The context surrounding the link. This consists of the 3 words before and after the link.

Link Type

The type of link. This can be *redirect*, *anchor*, *category* or *regular*.

Part of List

Binary feature that is positive if the link was found within a list, such as – *Fatigue*, *Headache*, *Nausea*, *Vomiting*.
In Table 1 we show an example of the information that the extraction of one link generates.

Sample Feature Extraction	
Concept	**fever**
Document Title	**Influenza**
Header Path	**Symptoms and diagnosis > Symptoms**
Context	**Extreme coldness and fever**
Link Type	**regular**
Part of List	**yes**

Table 1. Sample Feature Extraction

Even though we extracted five features, for the purposes of this work, we used only three features. We expect to use the remaining features in future work for the purpose of increasing performance.

Generating Labeled Data

Once we process the entire Wikipedia, we have a directed unlabeled graph where each edge represents a relation between two concepts. For each edge we also have a set of associated features.

After we decide the set of labels we are interested in, we use a combination of heuristics to bootstrap the labeling process. Besides using the redirect anchor and category links to label synonyms and hypernyms, we rely on the following two strategies.

List Based Labeling

Uses articles that list concepts and assigns labels to the instances of those lists that are under corresponding sections. E.g. if we find fever under the section symptoms in article flu and fever is also in the list of medical symptoms article, then we assign symptom as label for the link between flu and fever.

Context Based Labeling

Assigns the section title as label if the context shows that the link is displayed within a list. E.g. If we find *–fever, headache and nausea* under the section *symptoms* under article *flu*, we assign *symptom* as a label for the link between *flu* and *fever*.

After the bootstrapping process, we have a directed graph with a partially labeled relation set. In the next section we introduce the Directional Feedback Edge Labeling Algorithm which starts with a small such set of labeled links and uses graph probability propagation to label the remaining links/relations in the ontology.

Directional Feedback Edge Labeling Algorithm

The Directional Labeling Algorithm relies on neighboring edge trends and directionality to update probabilities of possible labels that can be assigned to an unlabeled relation. The steps of this algorithm are described in Algorithm 1.

Each unlabeled edge starts with equal probability of label assignment. At each iteration, in STEP 1, for each node we update the probabilities of the labels of the outgoing edges by smoothing them with the overall probability distribution of labels over the outgoing edges of that node (essentially multiplying the two probability distributions). This assures we take into account both our current belief about that edge and the overall information contained in the edges going out of that node. To give an intuition why both types of information are important, consider the example in Figure 1. The dashed and the dotted edges represent edges which were labeled during the bootstrapping phase. The dashed edges represent label *SymptomOf* and the dotted edges represent label *Treats*. The solid edges are unlabeled and therefore it is natural to assume that, in the absence of other information, each label is equally likely. However, based on the already labeled outgoing edges at C_1, the unlabeled edge (C_1,C_{10}) has a 2/3 probability to have label *SymptomOf* and 1/3 probability to have label *Treats.* Therefore, our initial belief of the edge (C_1,C_{10}) needs to be updated by incorporating this new information.

For each node *C*, perform Steps 1 and 2, then repeat until convergence.

STEP 1. Let p_{ik} be the probability of the i^{th} outgoing edge (out of *n* possible) from node *C* to have the k^{th} label (out of *m* possible labels). Update the outgoing edge probabilities:

$$P_{ik} \leftarrow \frac{P_{ik} \times \sum_{j=1}^{n} P_{jk}}{\sum_{l=1}^{m} (P_{il} \times \sum_{j=1}^{n} P_{jl})}$$

STEP 2. Update the incoming edge probabilities similar to the previous step.

STEP 3. Once convergence is reached via the above two steps, assign the maximum probability label to an edge as long as this probability is higher than a predefined threshold.

Algorithm 1. Directional Feedback Edge Labeling.

In STEP 2, we then perform the same procedure for each node, but based on incoming edges. Because an edge is an incoming edge for a node and an outgoing edge for another, the label probability distribution for that edge is influenced by the label distributions in both its endpoints. Therefore, after a number of iterations, the label probabilities can be influenced by other label probabilities at any distance in the graph.

Back to the example in Figure 1, the edge (C_1,C_{10}) has a 2/3 probability to be labeled *SymptomOf* if we look only at the outgoing edges from C_1 whereas it has a probability of 1 to be labeled *Treats* if we look only at the incoming edges to C_{10}. This justifies the need to perform the same operation for both incoming and outgoing edges. The need to perform both steps iteratively is twofold: to assure convergence and to allow knowledge to propagate across the network.

After convergence, we select only the edges with labels above a predefined threshold and discard the rest as unreliably labeled.

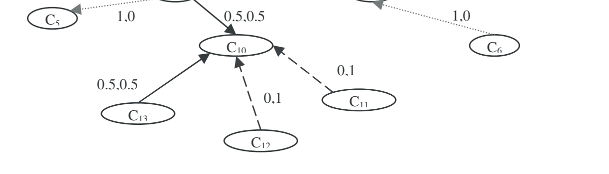


Figure 1. Example of Directional Labeling

UMLS and Okinet

UMLS is perhaps the most important medical ontology currently available. It uses a semantic network to combine the knowledge contained in the set of available dictionaries and allows for easy access to a set of standard ontological relations. The work of mapping the vocabularies demands a large human effort and is very time consuming. Due to the structure of UMLS, certain semantic relations exist only at the semantic network level. This means that in UMLS we are not able to determine symptoms of particular diseases, but rather between classes of concepts. For example, we are not able to find out what are the *symptoms_of flu*, but rather what categories of concepts could represent symptoms of *flu*, which is a problem.

Okinet thus exist as complement to UMLS, allowing the rapid and automatic creation of relations at the instance level, which enables the use of inference processes using both ontologies.

Experimental Setting

In order to test our approach, we used Wikipedia as a test case, even though the methodology could be applied to any other wiki like resource. Our goal is to create an ontology of causes, treatments, symptoms, diagnoses and side effects.

We started by selecting all the concepts contained in the *list of diseases* article, which contains 4000+ diseases and syndromes. We then expanded our article set to include all the articles that linked or were linked to by any of the articles contained in the current set.

Next we performed the feature extraction process followed by the bootstrapping procedure. The results were manually checked to create a gold standard set. This resulted in an ontology with 4308 concepts and 7465 relations divided as depicted in Figure 2.

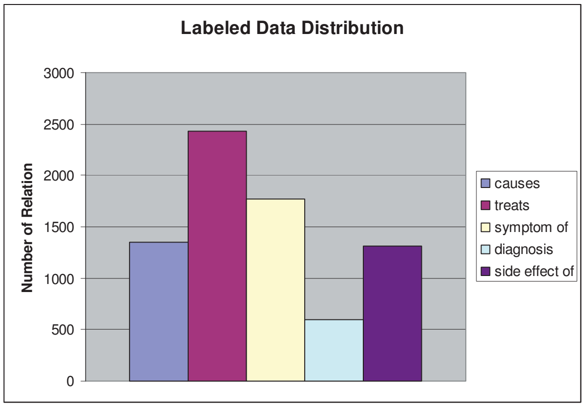


Figure 2. Distribution of relation labels.

Results

We experimented using small percentages of the labeled data as a training seed for the Directional Edge Feedback Labeling algorithm while considering the remaining edges unlabeled. The results of the labeling algorithm were then compared with the original labels. In our experiments, we varied both the percentage of the labeled training data (seed size) as well as the threshold above which we assign a label. We evaluated the results using precision and recall: Precision The percentage of label assignments that were correctly assigned to the proper class.

Recall The percentage of possible labels that were correctly assigned.

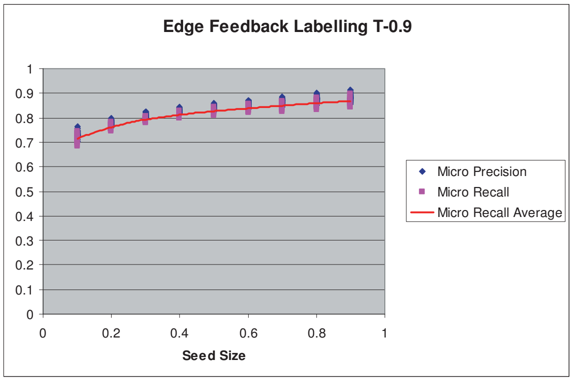


Figure 3. Algorithm performance with threshold 0.9 and variation on the seed size

In Figure 3 we can see the results of varying the size of labeled seed set at a threshold for label assignment at 0.9, which means that we only assign a label with high confidence. Even though we are only showing micro precision and micro recall, the results for macro precision and recall were very similar and were thus not presented for simplicity purposes. Each point in the average line represents 100 hundred runs with a labeled seed size of the indicated value. The precision and recall average vary between 70% and 90% while seeds vary from 10% to 90% of the total labeled set.

Even though the results are very promising, we explored ways to boost the results at small seed sizes. Due to the propagation nature of this algorithm, by stopping after a few iterations, we are in fact preventing long-range labeled edge influences and therefore we can restrict the process of labeling an edge to local neighborhood in the graph. Figure 4 shows the results of stopping the labeling algorithm after two iterations. Given the number of iterations, only edges with a fast convergence rate will update the probabilities distribution enough to get assigned a label. This means that the higher the threshold, the more accuracy we get, even though the recall is sharply reduced. This variation is particularly useful in situations where the precision is more important than recall. Using this technique we would be able to extend the labeled data set with highly accurate labels.

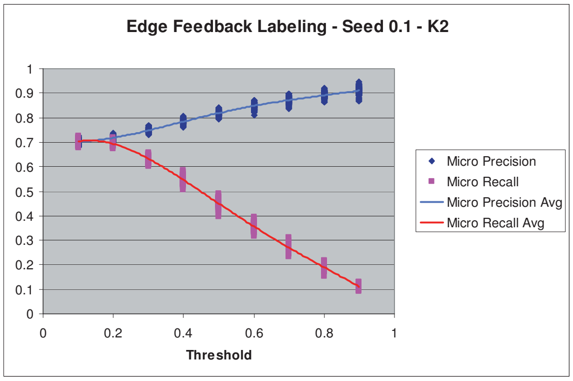


Figure 4. Precision and recall with 10% seed size, algorithm stopped after two iterations and varying the assignment threshold.

Finally we looked at our algorithm as way to reduce uncertainty. Figure 5 shows the results of taking the two highest confidence labels for each edge and considering as correct if either of the assigned labels is correct.

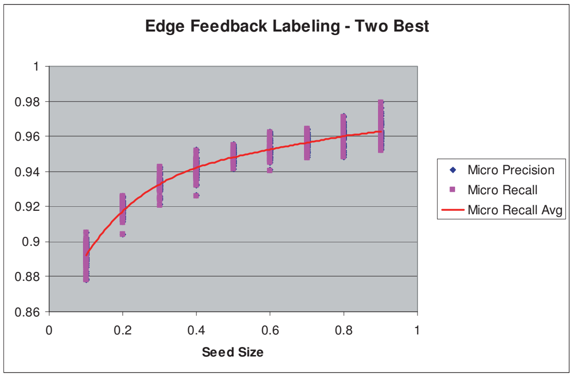


Figure 5. Precision and Recall when considering looking at the two labels with the highest probability

The Rose System

The current focus is on integration of multiple ontologies in the medical domain for immediate use. We currently have integrated UMLS , Okinet and Wordnet (Miller 1995) as a proof of concept application. We have built an ontology search system that uses a federated approach to ontology search as described in (Pedro, Lita et al. 2007) and produced a simple interface for ontology search. We call this integrated system ROSE, the Remind Ontology Search Engine. The goal of ROSE is to allow for rapid access to the ontological knowledge in the ontologies contained therein. In particular, the target domain for ROSE is the medical domain, where there are multiple ontologies, lexicons, and general graph-based resources covering specific parts of the medical knowledge space. In this scenario, a federated ontology search engine is desperately needed.

Web Based Interface

Although ROSE works in a typical server fashion, which can be queried directly using xml, we have also created a visualization tool for ontology browsing. Given the fact that the server is querying several ontologies, possible using resources span several servers, having a localized copy of everything is unfeasible. We therefore opted for the creation of a web interface to visualize the results of ontology queries.

The web interface allows for querying ROSE for synonyms and medical relations from all the available ontologies. It enables the user to visualize the desired relations across ontologies in a graph display and browse subsequent relations with ease and simplicity. It uses AJAX and JSP with JSON as a communication protocol. Below is a snapshot of the result of querying rose for Congestive Heart Failure. We can see the results from Wordnet, UMLS and Okinet represented by different colors.

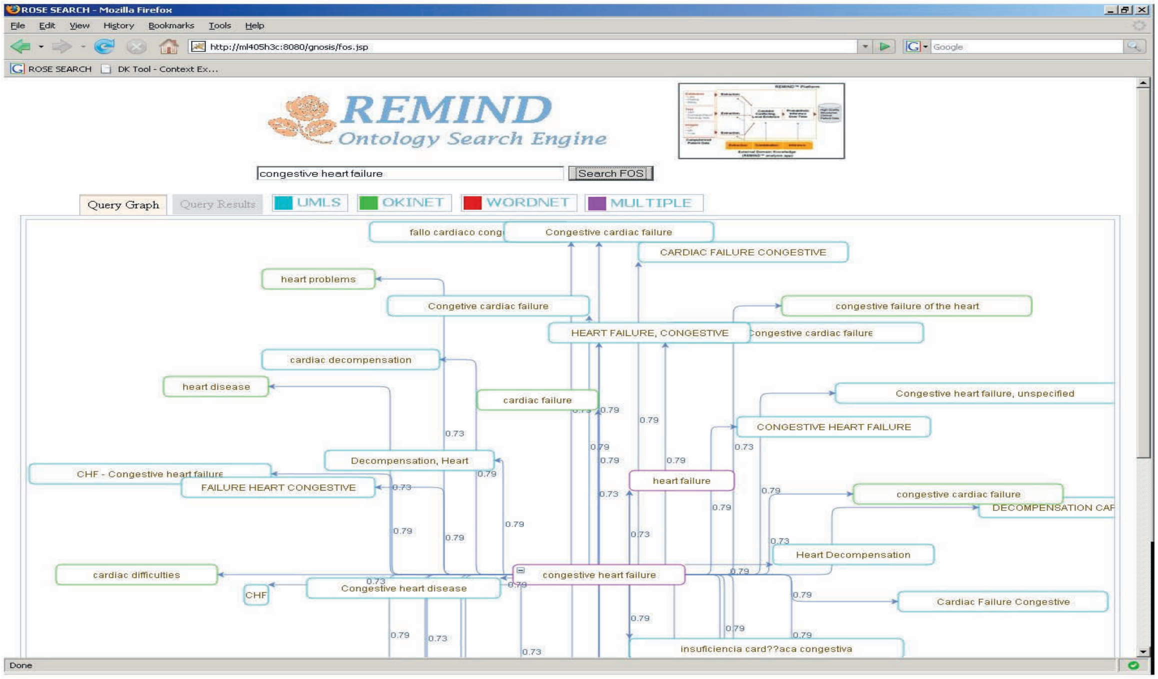
Summary and Future Work

The creation of medical ontologies is a complex and challenging task, in that it requires the analysis of domain sources, background knowledge and obtaining consensus among creators. The current methods are labor intensive and prone to generate inconsistencies. In this paper we propose a novel methodology for creating medical ontologies automatically from Wikipedia. We have measured the precision and recall of our method using a set of experiments and demonstrated the flexibility and utility of our approach. Given that our experiments were performed on a small dataset obtaining high performance, we believe that the described method has the potential to achieve even higher performance on larger datasets, further alleviating the amount of work required for ontology creation and further speeding up the process of knowledge acquisition.

Multiple research avenues remain open for domain-specific ontologies extracted from semi-structured data. In particular for the medical domain, Wikipedia is proving to be a continuously updated source of knowledge. In future work, we would like to incorporate additional features we have already extracted into the edge labeling algorithm. These features have the potential of boosting performance, especially with smaller seed sets.

We also intend to extend our medical ontology to additional concepts and relation types. Complex relations in the medical domain are abundant and higher coverage would allow sophisticated user to explore Okinet beyond frequent, often encountered medical phenomena.

Of particular interest studying how applicable is our method to automatic ontology mapping. Graph-based algorithms are particular well suited to this task, provided sufficient training data or sufficiently large seeds. We are



interested in reducing these seed sets while maintaining high performance.

Another interesting direction for future work is the question of whether seed localization has an impact on the performance of the labeling algorithm.

Finally, since our Okinet generation method does not rely on rule-based components, nor has been infused with medical expert knowledge, a promising research direction consists of extending our approach to non-medical domains such as finance, education, physics, which have at least a moderate wiki presence.

References

Bhole, A., B. Fortuna, et al. 2007. *Mining Wikipedia and relating named entities over time*. Bohanec, M., Gams, M., Rajkovic, V., Urbancic, T., Bernik, M., Mladenic, D., Grobelnik, M., Hericko, M., Kordeš, U., Markic, O. Proceedings of the 10 th International Multiconference on Information Societz IS 8: 12.

Blake, C. and W. Pratt 2002. *Automatically Identifying Candidate Treatments from Existing Medical Literature*. AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases.

Bodenreider, O. and O. Journals. *The Unified Medical Language System (UMLS): integrating biomedical terminology*. Nucleic Acids Research 32(90001): 267-270.

Denoyer, L. and P. Gallinari 2006. *The Wikipedia XML corpus*. ACM SIGIR Forum 40(1): 64-69.

Izhikevich, E. M. 2007. *Scholarpedia, the free peer-reviewed encyclopedia*. from http://www.scholarpedia.org/.

Khoo, C., S. Chan, et al. 2002. *Chapter 4*. The Semantics of Relationships: An Interdisciplinary Perspective.

Kozlova, N. 2005. *Automatic ontology extraction for document classification*, Saarland University.

Krotzsch, M., D. Vrandecic, et al. 2005. *Wikipedia and the Semantic Web-The Missing Links*. Proceedings of Wikimania.

Lih, A. 2003. *Wikipedia as Participatory Journalism: Reliable Sources? Metrics for evaluating collaborative media as a news resource*. Nature 2004.

Lipscomb, C. E. 2000. *Medical Subject Headings (MeSH)*. Bull Med Libr Assoc 88(3): 265-266.

Maedche, A. and S. Staab 2002. *Measuring similarity between ontologies*. Proceedings of the European Conference on Knowledge Acquisition and Management (EKAW): 251–263.

Maedche, A. D. 2002. *Ontology Learning for the Semantic Web*, Kluwer Academic Publishers.

Milne, D., O. Medelyan, et al. 2006. *Mining Domain-Specific Thesauri from Wikipedia: A Case Study*. Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence: 442-448.

Navigli, R., P. Velardi, et al. 2003. *Ontology learning and its application to automated terminology translation*. Intelligent Systems, IEEE [see also IEEE Intelligent Systems and Their Applications] 18(1): 22-31.

Sanger, L. 2007. *Citizendium*. from www.citizendium.org.

Spackman, K. A., K. E. Campbell, et al. 1997. *SNOMED RT: A reference terminology for health care*. Proc AMIA Annu Fall Symp 640(4): 503-512.

Völkel, M., M. Krötzsch, et al. 2006. *Semantic Wikipedia*. Proceedings of the 15th international conference on World Wide Web: 585-594.