Semantic Relationship Extraction and Ontology Building using Wikipedia: A Comprehensive Survey

Nora I. Al- Rajebah
Computer Science Department
College of Computer and Information Sciences
King Saud University

Hend S. Al-Khalifa
Information Technology Department
College of Computer and Information Sciences
King Saud University

ABSTRACT
Semantic web as a vision of Tim Berners-Lee is highly dependable upon the availability of machine readable information. Ontologies are one of the different machine readable formats that have been widely investigated. Several studies focus on how to extract concepts and semantic relations in order to build ontologies. Wikipedia is considered as one of the important knowledge sources that have been used to extract semantic relations due to its characteristics as a semi-structured knowledge source that would facilitate such a challenge. In this paper we will focus on the current state of this challenging field by discussing some of the recent studies about Wikipedia and semantic extraction and highlighting their main contributions and results.

General Terms
Wikipedia, Ontology, RDF, Semantic web.

Keywords
Ontology Building, Semantic Web, Semantic extraction, Wikipedia.

1. INTRODUCTION
Nowadays, the need for ontology models to build semantic web applications is becoming a demand, considering the large number of applications that can benefit from them e.g. information retrieval systems and web search engines. A major step to build such ontologies is to define the concepts and the semantic relationship between them. This can be achieved using an extraction process mechanism. However, the extraction process needs to be done on a huge corpus to assure a full coverage of semantics. Web content, as a huge corpus, is a great source for semantics extraction however applying semantic extraction to the web might seem possible; yet, it would be time and space consuming.

Wikipedia the free online encyclopedia has been well-known as a source for extracting concepts and semantics. The reason behind choosing Wikipedia is that it is a semi-structured knowledge source and it has been organized in a hierarchical manner.

In Wikipedia, articles are the basic unit of information [1]. Usually each article talks about a unique concept, and sometimes the same name can be used for many concepts e.g. KSU can be King Saud University or Kansas State University. In such case a disambiguation page is used to gather all the other possible concepts.

Each article belongs to one or more categories that are grouped according to their relatedness in a hierarchy. For example King Saud University belongs to Universities and colleges in Saudi Arabia category and Educational institutions established in 1957 category. Usually each article begins with a definition statement and a brief overview of the concept. After that, the text is organized into sections that focus on some aspect of the concept. Within the article's text, any reference to other articles is presented as a hyperlink to that article. Also, some articles have infoboxes that provide summarized information about the article in a form of a table where each row provides attribute-value information. Articles belong to the same class e.g. scientists usually share identical or similar infobox template.

In this paper, we will focus on Wikipedia and its usage by semantic extraction methods. We will also discuss a number of recent studies that used Wikipedia in relationship extraction and ontology building.

2. WIKIPEDIA AND SEMANTICS EXTRACTION
Since launching Wikipedia in 2001, this encyclopedia has been a subject of a wide variety of research fields including: Natural Language Processing (NLP), knowledge extraction, ontology building and semantic relatedness measurements. In the following subsections we are going to present a collection of key studies that benefited from Wikipedia's semi-structured nature in relation extraction and ontology building.

2.1 Relation Extraction
As one of the first attempts to extract relations from Wikipedia Ruiz et. al. [2] [3] propose an automatic approach to extend the semantic network of WordNet using relations extracted from Wikipedia. So, for two semantically related WordNet synsets, this method defines lexical patterns from their corresponding Wikipedia articles, in order to extract the missing relations between them in WordNet. Their method successfully extracted 1224 new relations from Wikipedia with a precision ranged from 61%-69% for three different types of relations (Hyponymy, Holonymy and Meronymy).

Ruiz. et. al [4] have also extend their previous work in [2], by proposing an automated method to extract other types of relations besides Hyponymy (isA), Hyperonymy (not-isA) , Holonymy (has part) and Meronymy. In their approach they used several NLP tools such as (Tokenizer, POS Tagger, Stemmer, Named Entity Recognition, Chunker) to extract the lexical patterns from Wikipedia. The approach achieved a human judgement precision

that varied from 90% (for death-year relation) to 7.75% (for player-team relation) this is related to the types of pattern used to extract such relations. In their discussion, they mentioned possible future use of the method to build Ontologies.

Hyperlinks over categories were another major technique used to extract relationships from Wikipedia. For instance, Chernov et. al. [5] discussed the degree in which semantic relations extracted from Wikipedia are correlated. To do so they propose two measures to estimate the strength of the semantic relation between two Wikipedia categories. The first measure is the number of hyperlinks in the pages contained within the categories. The other measure is called Connectivity Ratio which represents a normalization of the number of hyperlinks according to the category size. The authors' main argument is that if there is a strong semantic relation between any two categories then the number of hyperlinks between the pages under them will be large, otherwise, there is no regular semantic relation. In order to evaluate their method, Chenove et. al. introduced Semantic Connection Strength measure (SCS) with the values 0, 1 or 2, where 2 means strong semantic relation and 0 means weak semantic relation. In general, Connectivity Ratio results were better by 25%.

Another widely used technique for relation extraction is infoboxes. Wang et. al. [6] proposed the Positive-Only Relation Extraction (PORE) framework to extract relations from Wikipedia. Instead of using pattern matching, their approach was completely independent of patterns and it was based on B-POL algorithm which is an extension of Positive-Only Learning algorithm, where negative examples are identified and then classified until they reach convergence. In the conducted experiments, they evaluated four types of relations (album-artist, film-director, university-city, band-member). These four relations were extracted from album_infobox, movie_infobox, university_infobox and band_infobox respectively. The evaluation was done by 3 human judges. Album-artist relation achieved the highest F-measure with 79.9% while band-member relation achieved the lowest F-measure that is equal to 47.1%.

Similarly, Wu and Weld [7] proposed the KYLIN system, which automatically extracts information from Wikipedia articles that belongs to the same category in order to create and complete their infoboxes. KYLIN generates infoboxes using Infoboxes Generation Module which consists of three phases: 1) preprocessing, 2) classifying and 3) extracting. Moreover, it can be further used in automatic link generation. Four classes were evaluated during the experiments conducted to test KYLIN's performance: U.S. county, airline, actor, and university. For each one of these classes the authors manually compared KYLIN generated infobox with the one already exist within Wikipedia's articles. Recall results ranged from 60.5% for university, to 95.9% for county. The precision ranged from 87.2 % for airline to 97.3% for county. The authors argued that county achieved the highest recall and precision because it contains a lot of numerical data that can be extracted easily.

Based on KYLIN [7], Wu and Weld [8] proposed KOG (KYLIN Ontology Generator), which generally used to build Ontology that combines Wikipedia infoboxes and WordNet hierarchy. In KOG each infobox template was considered as a class, while the slots belong to the infobox were considered as attributes. KOG has three main components: 1) schema cleaner, 2) subsumption detector, and 3) schema mapper. Subsumption detecting was mainly used to identify isA relations between classes (infobox templates). They used machine learning techniques (SVM and Markov Logic Networks MLN) with a number of classification features to classify the relations between the classes while mapping them to WordNet nodes. Many experiments were conducted to evaluate the system as a whole and to evaluate the performance of each component based on 1269 selected infobox classes. In subsumption detection phase, SVM achieved a 97.2% precision and 88.6% recall. In MLN (basic) the precision decreased to 96.8% while the recall increased to 92.1%. Finally, MLN+ (fully-functional) had increased both, precision to 98.8% and recall to 92.5%.

In their later work, Wu et. al. [9], focused on the improvements in the recall of their system. KYLIN had successfully extracted attributes for many types of classes (infoboxes), unfortunately not all classes had infoboxes contained articles. KYLIN's extraction performance is poor for such classes due to the lack of the possible attributes to extract and the lack of sufficient training examples.

Natural language techniques (NLP) were also utilized to extract relations from Wikipedia via syntactic parsing and tree mining. In this regard, Nguyen et. al. [10][11] proposed a framework to extract relations from Wikipedia using a syntactic parser and several NLP techniques. The process starts by extracting the hyperlinks from each Wikipedia article, then, the whole text is processed by OpenNLP tool that contains a Sentence Splitter, a Tokenizer, and a Phrase Chunker to extract all of the different occurrences of an entity in an article. After that, from each detected sentence's keywords which indicate the type of relation that exist between principals are extracted using Keyword Extractor. Relation Extractor is used to construct a syntactic tree representation of each sentence; those trees are then generalized using machine learning techniques to be applied to other similar sentences. In their experiment they randomly selected 3300 entities (principal and secondary) and manually annotate their relations. 3100 entities are used for training and 200 for testing. The system achieved an average F-measure equals to 38%. Where recall has ranged from 50% to 55% and precision ranged from 20% to 30%.

Cui et. al. [12] focused on extracting concepts from Wikipedia as a requirement in building ontologies. They used infoboxes information, category labels and the first sentence of each article. Definition sentences usually indicate hypernym relation (instanceOf). Verbs-to-be in a definition (e.g. is, are, was) are usually a good indication that the following noun is connected to the title noun with isA relation. A Part-Of-Speech (POS) tagger was used to syntactically parse the sentence. N-gram statistics were used to divide category names into smaller parts. They evaluated the correctness of 50,000 extracted concepts. They compared the three methods individually. Infobox method achieved 90.1% precision with 15% coverage. Definition sentences method achieved 76.7 % precision and 79.2% coverage. Finally, categories method achieved 75.3 % with 100% coverage.

2.2 Ontology Building

Research in the area of ontology building is vast and several studies have investigated many methods to extract concepts and relations for building ontologies.

As one of the early attempts that used pattern matching for ontology building, Herbelot and Copestake [13] used Robust Minimal Recursion Semantics (RMRS) parser to extract Ontologies from biological articles in Wikipedia. RMRS is used to derive a semantic representation of the articles' text. It outputs the sentences into small semantic trees where the roots are the lemmas in each sentence. After parsing, each sentence is systematically processed for pattern matching to extract "isA" relations. The system is evaluated manually and automatically. The initial experiment that extracts the patterns manually resulted in 16% 'rough' recall, 14% manual recall and 92% precision. Herbelot and Copestake have also discussed some factors that might yield the low recall achieved. Further experimentation by extracting the patterns automatically increased the recall to 37% while the precision decreases to 65%.

Another attempt was brought by a project named DBpedia by Auer et. al. [15][14]. DBpedia extracted unstructured knowledge from Wikipedia and converted it into more structured one represented as RDF triples. DBpedia contains about 103 million RDF triples that can be used in several Semantic Web applications. To build its dataset, DBpedia extracted information from Wikipedia's articles infoboxes and stored them as triples, where an article's title was considered as a subject and each attribute became a predicate of the triple while their values were considered as the objects. One of the main obstacles DBpedia has faced was the lack of availability of infoboxes in many Wikipedia articles. Auer et. al. used DBpedia mainly to build a querying system based on the stored triples to answer questions like: What have Innsbruck and Leipzig in common?. On the other hand, no evaluation results were provided to justify their method. In general, the main drawback of DBpedia is that it is poorly structured because there are no semantic links between its content. To overcome this problem, they produce the DBpedia ontology in 2008, which was manually created.

Ponzetto and Strube [16] proposed a method to build a large taxonomy from Wikipedia using its categories hierarchy which is more like a conceptual network without the identification of the semantic relations between its nodes. They improved the semantic aspect of this hierarchy by classifying relation types to be either isA or not isA. They used syntax based methods to match categories' names and label's relations of matched categories as isA. Two types of syntax matching were used: head matching (isA) and modifier matching (not isA). The second method was based upon the connectivity characteristics of the network which had been used to define instanceOf relation between Wikipedia categories using their corresponding article as isA relation if their lemmas match. The final step was to label the rest of unidentified relations to either isA or not-isA relations using lexico syntactic based method that applied 13 predefined patterns to Wikipedia category titles. POS tagger and a SVM-based chunker were needed to identify noun phrases (NP). To evaluate the quality and coverage of the generated taxonomy, a comparison with ResearchCyc yielded recall of 89.1%, precision of 86.6% and F-measure of 87.9%.

Suchanek et. al. in [17] aimed to extend the highly structured WordNet hierarchy by adding concepts and relations from Wikipedia through its categories. The project named YAGO (Yet Another Great Ontology) and it had successfully extracted 1.5 million entities and 5 million facts, some of the facts are represented as isA relation, SubClassOF (hyponymy) relation and the rest are predefined ones (bornInYear, diedInYear, establishedIn, locatedIn, writtenInYear, politicianOf, and hasWonPrize). Another type of relations is Means relation which had been extracted from Wikipedia redirect pages in which it connects synonyms together. YAGO was made with the capabilities of future extension so it had its own data model, which is an extension of the RDF (YAGO Data Model) that was used to represent concepts and relationships between them. They argued that the reason behind defining such model is that the existing Web Ontology Language (OWL and its variants before OWL 2) at that time are either undecidable or lacks in relations expression e.g. unable to represent transitivity. YAGO attached the leaf nodes from Wikipedia categories hierarchy (classes) to the ontology derived from WordNet where each synset is considered as a class in YAGO. All of the Wikipedia articles belong to the same category in Wikipedia (class) that had been mapped into WordNet are added to YAGO ontology as new entities. Many experiments had been conducted to evaluate the accuracy of the ontology generated by YAGO. They manually evaluated the facts produced by the system. The accuracy varied according to the type of relation being evaluated it ranged from 98.72 % for diedInYear relation to 90.84% for establishedIn relation. The average accuracy of the system was 95 %.

In their later work Suchanek et. al. [18] extended their work in [17] by adding new heuristics to extract new entities and facts from Wikipedia and emphasizing on generating a high quality ontology using some quality control mechanisms. The heuristics used yielded a large number of relations which have been evaluated by 13 human judges to compute their precision. 74 heuristics achieved a precision of 95 % or higher and the average precision of YAGO was 95%. YAGO has been employed as a backbone in many other applications including querying, semantic search engines and even in ontologies constructions e.g. Freebase, UMBEL and SUMO.

Pattern matching methods have been exploited in extracting semantic relations from Wikipedia. An example of such usage was introduced by Liu et. al. [19]. In this paper, the main goal was to extract triples from Wikipedia assuring a wide coverage across it with a minimum effort. Wikipedia categories hierarchy played a major role in the proposed method, so for each category the method's job was to extract the property and value that are common among all the articles belonging to a given category. Two types of category pairs were investigated. The first one has explicit property and value e.g. "Songs by artist" and "The Beatles songs", where "artist" is the property and "The Beatles" is the value. The second type has explicit value and implicit property, e.g. "Rock songs" and "British rock songs", where "British" is the value and the property is not provided. Four patterns were used to extract property-contained category names (PCCN) and value-contained category names (VCCN) from

category names. OpenNLP were employed to extract part-of-speech tags for the category names. At the final step, triples are generated for each article where the subject is the article's title. To evaluate the proposed method, 500 Catriple (Category Triples) triples were selected randomly to be judged by human judges, this yielded a precision ranged from 47.0% to 96.4%.

Cyc project has been also utilized in ontology building. As an example of such usage Medelyan and Legg [20] benefited from ResearchCyc by integrating it with Wikipedia in order to generate a folksonomy that combines both Wikipedia and ResearchCyc Their approach mapped Cyc concepts into Wikipedia articles' titles. Two types of mapping were investigated: 1) Exact mapping in which Cyc concepts are mapped directly to Wikipedia article titles or to redirect links within an article as one to one. In case of no matched article found they checked whether its Cyc synonym is available. 2) Ambiguous mapping, this arise when one Cyc concepts can be mapped to many Wikipedia articles. Each Cyc concept would have a list of candidates collected from disambiguation pages. At the end 52,690 Cyc concepts were mapped to Wikipedia articles. To evaluate the method, two datasets were used. The first one was a list of 9,333 manually mapped Cyc synonymous; the precision of the comparison was 93.9%. The second one consists of 100 random pairs of Cyc concepts that were mapped to Wikipedia articles to be judged by six human judges which resulted in a precision of 93%.

The methodology used by Medelyan and Legg [20] has been further improved by the work of Sarjant et. al. [21]. Mappings' methods were improved by adding more conditions, for example in exact mapping they removed "The" from titles. In the evaluation, they evaluated both mapping process and extension process by 22 human judges who evaluated both old (mapping used by [20]) and the current mapping. The newly computed precision for the old mapping was 83% while the new one achieved 91% precision.

De Silva and Jayaratne [22] introduced WikiOnto which extract and model ontology from Wikipedia XML corpus using Natural Language Processing (NLP) and machine learning techniques. WikiOnto composed of three phases, in the first phase concepts and candidate relationships (sub-conceptOf) with each other were extracted using the provided XML document structure. In the next step, a vector-space is constructed for each document which contains the keywords (most relevant concepts) by measuring TFIDF (Term Frequency Inverse Document Frequency) of each concept. After that documents were clustered according to the similarities (cosine similarity) of their corresponding vector-spaces using k-means algorithm. After clustering, ontologies were modeled where each concept is linked to the other concepts that belong to its cluster. To improve the performance of their system they added a syntactic processing method that can extract hyponymic relations between concepts. It is noted that no evaluation results were provided but they mentioned that their project is an ongoing one.

Wikipedia is a multilingual online encyclopedia that is available in many languages. As an example of using other versions of Wikipedia, Farhoodi et. al. [23] used the Persian Wikipedia to build an ontology that facilitated their proposed query expansion method. To generate ontology, a process of three phases was employed. Firstly, a Wikipedia parser extracted the title, keywords, existing links, 'see also' links and list of categories. Then the relationships between the extracted data are defined as "IsRelatedTo" which occur between article's title and keywords article's title and related links, and finally between an article's title and hyperlinks. No specific evaluation results about ontology building were provided.

3. DISCUSSION

The problem of semantic extraction has been addressed using different techniques and throughout the previous sections we presented different studies that employed them. Table 1 shows that most studies have combined one or more methods to accomplish the required task. Also, by combining methods the overall coverage of semantics across Wikipedia articles will be increased accordingly.

Although, Wikipedia categories hierarchy contains duplication and sometimes it is inconsistent compared to other manually created hierarchies e.g. Cyc and WordNet, yet, it has been used in many approaches [5][6][12][16][19][22]. After refinement, it provides a good base hierarchy in which relations are defined as isA. This hierarchy can be extended by other concepts and relations extracted from other sources. On the other hand, other studies used WordNet e.g. YAGO [17][18] and Cyc [20][21] as a base because of their high quality in semantic connections.

Compared to infoboxes, category hierarchy covers a wider area of concepts available within Wikipedia. In fact, the methods that are dependent on infoboxes only e.g. DBpedia [14][15] perform poorly for many reasons, 1) there might be several templates for the same type of articles. 2) Sometimes some infoboexes attributes have no values. 3) Redundancy and inconsistency. Therefore, it is highly preferable to use infoboxes them combined with one or more methods. To the contrary, KYLIN [7][8] used the information available on Wikipedia to complete and create infoboxes that belongs to the same class and sometimes it benefitted from the attributes in the infoboxes of some articles to add more attributes to other articles' infoboxes within the same class. As mentioned in [19] 44.2% of Wikipedia articles have infoboxes while categories covered nearly 81% of Wikipedia information.

Hyperlinks were not used widely in the previous methods, and obviously it is due to the ambiguity that they might lead to. Because, it cannot be guaranteed that the links appeared within Wikipedia's articles may indicate semantic relations between concepts. One reasonable use of hyperlinks was proposed by Chernov et al. [5]. They measured the strength of the semantic relations between categories by counting the number of pages that have hyperlinks to pages in other category.

Parsing and pattern matching methods were used effectively in many approaches. Usually the results achieved were promising. The reason behind this promising result is that when applying patterns to the free text the number of extracted semantic relations would be very large, but for sure the selection of patterns should be done with caution because they should be general enough so that they can yield better performance. Another important issue when dealing with patterns is that the

used language has a great impact on the ability of defining patterns and generalizing them. Also, some approaches used machine learning techniques mainly to train their systems to decide upon relationship types.

To conclude, two characteristics can affect the usability of any relation extraction/ontology building approach:

1) The number of the correctly extracted concepts and relations it produces (the coverage).

2) The quality of the extracted semantic relationships between its extracted concepts (Quality of structure).

4. CONCLUSION

In this paper we focused on Wikipedia and the challenge of semantic extraction. We present some of the key studies that achieved remarkable outcomes. We present their main contribution and results. We also differentiate between the approaches that have been used for such tasks e.g. using Wikipedia hierarchy, infoboxes, hyperlinks between categories and articles, machine learning techniques, pattern matching and NLP tools. We saw that usually one or more approaches are used to achieve better performance. For each of the approaches we mentioned its points of strength and weaknesses. In general, two characteristics affect the usability of any relation extraction/ontology building approach: 1) its coverage based on the number of correct concepts and relations extracted. 2) The quality of structure that has been built which is related to the consistency of the extracted semantics.

5. ACKNOWLEDGMENTS

The authors are thankful to Professor AbdulMalik S. Al-Salman, for his valuable and thoughtful feedback.

6. REFERENCES

[1] Medelyan, O., Milne, D., Legg, C., and Witten, I. H.,"Mining meaning from Wikipedia", *International Journal of Human-Computer Studies*, vol. 67, no. 9, pp. 716-754, Sep. 2009.

[2] Ruiz-Casado, M., Alfonseca, E., Castells, P., "Automatic extraction of semantic relationships for wordnet by means of pattern learning from Wikipedia", *10th International Conference on Applications of Natural Language to Information Systems, NLDB'05*, Alicante, Spain, pp. 67–79, June 2005.

[3] Ruiz-Casado, M., Alfonseca, E., Castells, P., "Automatising the learning of lexical patterns: an application to the enrichment of WordNet by extracting semantic relationships from Wikipedia", *Data Knowledge and Engineering*, vol. 61, no.3, pp. 484–499, Jun. 2007.

[4] Ruiz-Casado, M., Alfonseca, E., Castells, P., "From Wikipedia to semantic relationships: a semi-automated annotation approach", *The First International Workshop: SemWiki'06—From Wiki to Semantics. Co-located with the Third Annual European Semantic Web Conference ESWC'06*, Budva, Montenegro, Jun. 2006.

[5] Chernov, S., Iofciu, T., Nejdl, W., Zhou, X., "Extracting semantic relationships between Wikipedia categories", The First International Workshop: SemWiki'06—From Wiki to Semantics. Co-located with the Third Annual European Semantic Web Con- ference ESWC'06, Budva, Montenegro, Jun. 2006.

[6] Wang, G., Yu, Y., Zhu, H., "PORE: Positive-only relation extraction from Wikipedia text", *The Sixth International Semantic Web Conference and Second Asian Semantic Web Conference, ISWC/ASWC'07*, Busan, South Korea, Nov. 2007

[7] Wu, F., Weld, D., "Autonomously semantifying Wikipedia", *The 16th ACM Conference on Information and Knowledge Management, CIKM'07*, Lisbon, Portugal, pp. 41–50, Nov. 2007.

[8] Wu, F., Weld, D., "Automatically refining the Wikipedia infobox ontology", *The 17th International World Wide Web Conference, WWW'08*, Beijing, China, pp. 635-644, 2008

[9] Wu, F., Hoffmann, R., Weld, D., "Information extraction from Wikipedia: moving down the long tail", *The 14th ACM SigKDD International Conference on Knowledge Discovery and Data Mining (KDD-08)*, Las Vegas, NV, pp. 635–644, Aug. 2008.

[10] Nguyen, D.P.T., Matsuo, Y., Ishizuka, M., "Relation extraction from Wikipedia using subtree mining", *The AAAI'07 Conference*, Vancouver, Canada, pp. 1414–1420, Jul. 2007.

[11] Nguyen, D.P.T., Matsuo, Y., Ishizuka, M., "Exploiting syntactic and semantic information for relation extraction from Wikipedia", *The IJCAI Workshop on Text-Mining and Link- Analysis, TextLink'07*, 2007

[12] Cui, G., Lu, Q., Li, W., and Chen, Y., "Mining Concepts from Wikipedia for Ontology Construction", The 2009 IEEE/WIC/ACM international Joint Conference on Web intelligence and intelligent Agent Technology - Volume 03, Web Intelligence & Intelligent Agent. IEEE Computer Society, Washington, DC, pp. 287-290, Sep. 2009

[13] Herbelot, A., Copestake, A., "Acquiring ontological relationships from Wikipedia using RMRS", *The International Semantic Web Conference 2006 Workshop on Web Content Mining with Human Language Technologies*, Athens, GA, 2006

[14] Auer, S., Bizer, C., Lehmann, J., Kobilarov, G., Cyganiak, R., Ives, Z., "DBpedia: a nucleus for a web of open data", *The Sixth International Semantic Web Conference and Second Asian Semantic Web Conference (ISWC/ASWC2007)*, Busan, South Korea, pp. 715–728, Nov. 2007.

[15] Auer, S., Lehmann, J., "What have Innsbruck and Leipzig in common? Extracting semantics from Wiki content", In: Franconi, et al. (Eds.), *European Semantic Web Conference (ESWC'07)* June 2007. *Lecture Notes in Computer Science*, vol. 4519. Springer-Verlag, pp. 503–517.

[16] Ponzetto, S.P., Strube, M., "Deriving a large scale taxonomy from Wikipedia", *The 22nd national conference on*

Artificial intelligence - Volume 2, AAAI '07, Vancouver, British Columbia, Canada, pp. 1440–1445, Jul. 2007

[18] Suchanek, F.M., Kasneci, G., Weikum, G., "Yago: a core of semantic knowledge", *The 16th World Wide Web Conference, WWW'07.* Banff, Alberta, Canada, *ACM Press, New York*, pp. 697-706, May 2007.

[19] Suchanek F. M., Kasneci G., Weikum G., "YAGO: A Large Ontology from Wikipedia and WordNet", *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 6, no. 3, pp. 203–217, Sep. 2008.

[20] Liu Q., Xu K., Zhang L., Wang H., Yu Y., Pan Y., "Catriple: Extracting Triples from Wikipedia Categories", *Lecture Notes in Computer Science*, vol. 5367. Springer, pp. 330-344, 2008.

[21] Medelyan, O. & Legg, C., "Integrating Cyc and Wikipedia: Folksonomy Meets Rigorously Defined Common-Sense", *The AAAI Workshop on Wikipedia and Artificial Intelligence*, Chicago, Ill pp. 13-18, 2008.

[22] Sarjant, S., Legg, C., Robinson, M., and Medelyan, O., "All You Can Eat" Ontology-Building: Feeding Wikipedia to Cyc", The 2009 IEEE/WIC/ACM international Joint Conference on Web intelligence and intelligent Agent Technology - Volume 01. Web Intelligence & Intelligent Agent. IEEE Computer Society, Washington, DC, pp. 341-348, Sep. 2009.

[23] De Silva, L.; Jayaratne, L.,"Semi-automatic extraction and modeling of ontologies using Wikipedia XML Corpus", *Applications of Digital Information and Web Technologies, 2009. ICADIWT '09. Second International Conference*, pp.446-451, Aug. 2009.

[24] Farhoodi M., Mahmoudi M., Bidoki A. M. Z., Yari A., Azadnia M., "Query Expansion Using Persian Ontology Derived from Wikipedia", *World Applied Sciences Journal* , vol. 7 no. 4, pp. 410-417, 2009.

Table 1. Summary of relation extraction/ontology building studies

Study	Wikipedia Categories Hierarchy	Infoboxes	Hyperlinks	Pattern matching and NLP tools	Knowledge sources (WordNet, Cyc)	Machine learning	Results
Ruiz et. al. 2005, 2006, 2007				*	*		-Precision 61%-69% (2005) -Precision 90% -7.75% (2007)
Chernov et. al. 2006	*		*				NA
Wang et. al. 2007	*	*				*	F-measure 79.9%-47.1%.
Wu and Weld 2007, 2008 (KYLIN,KOG)		*				*	Precision 87.2 % -97.3% (2007) Precision 98.8% (2008)
Nguyen et. al. 2007				*		*	Precision 20% - 30%.
Cui et. al. 2009	*	*		*			Precision 75.3 % - 79.2%
Auer et. al. 2007 (DBpedia)		*					NA
Herbelot and Copestake 2005				*			Precision 92%

Ponzetto and Strube 2007	*			*			Precision 86.6%
Suchanek et. al. 2007, 2008 (YAGO)	*	*			*		Accuracy 95 %. (2007) Precision 95%.
Liu et. al. 2008 (Catriples)				*			Precision 47.0% - 96.4%.
Medelyan and Legg 2008					*		Precision 93%.
Sarjant et. al. 2009					*		Precision 91%
De Silva and Jayaratne 2009	*		*	*		*	NA
Farhoodi et. al. 2009			*				NA