Mining Concepts from Wikipedia for Ontology Construction

Gaoying Cui, Qin Lu, Wenjie Li, Yirong Chen
Department of Computing
The Hong Kong Polytechnic University
Hong Kong, China
{csgycui,csluqin,cswjli,csyrchen}@comp.polyu.edu.hk

***Abstract*—An ontology is a structured knowledgebase of concepts organized by relations among them. But concepts are usually mixed with their instances in the corpora for knowledge extraction. Concepts and their corresponding instances share similar features and are difficult to distinguish. In this paper, a novel approach is proposed to comprehensively obtain concepts with the help of definition sentences and Category Labels in Wikipedia pages. N-gram statistics and other NLP knowledge are used to help extracting appropriate concepts. The proposed method identified nearly 50,000 concepts from about 700,000 Wiki pages. The precision reaching 78.5% makes it an effective approach to mine concepts from Wikipedia for ontology construction.**

Keywords-Concept; Ontology Construction; Wikipedia

I. INTRODUCTION

Ontology construction by linguists or domain experts is time-consuming and difficult to update. For example, the Suggested Upper Merged Ontology (SUMO), was built with collaborators from the fields of engineering, philosophy, and information science [1]. It is generally not feasible to build ontology for different domains manually. Automatic ontology construction attempts to acquire concepts and relations from some domain corpora. In the top-down approach, some upper-level ontology is given and algorithms are developed to expand it from the most general concepts downwards to reach the leaf concepts where their instances can be attached [2]. In the bottom-up approach, concepts and relations are extracted from some domain corpora directly. Most corpus-based ontology construction assumes that concept terms in the domain are known [3]. Thus, the work focuses on identifying relations among the concept terms. In real corpora, concepts and their instances are inter-mixed. In many cases, concept instances appear even more than corresponding concepts. For example, "*Microsoft*" and "*IBM*" are instances of concept "*company*", and such instances may occur more frequently. Thus in a truly corpus-based approach, one important issue is to distinguish concepts from concept instances, which are the intents and extents of the concepts, respectively. It should be pointed out that instances of concepts are normally not considered as a part of ontology. If they are appended in an ontology, they should appear only as leaves. For ontology construction using corpus based approach, there is a natural gap between the ontology as a concept-level structure and the corpus as an instance-rich resource.

Wikipedia (Wiki for short) is the largest online encyclopedia in the world which contains definitions and descriptive information for over 2 million pages [4]. A Wiki page is annotated with a type, such as Article page, Category page, etc., and can contain internal and external hyperlinks such as Category Labels, and other semantic information. The {{Infobox}} Structures declared by contributors is an indicator of instance page. However, only about 15% of Wiki pages contain the {{Infobox}} Structure, which leads to a very low coverage for concept acquisition through {{Infobox}} Structure [5]. In this paper, more Wiki resources are exploited for concept acquisition, such as definition sentences reflecting the *is-a* relation in Article pages and classification information in Category Labels providing topics relevant to Wiki pages. As Categories Labels can be arbitrary, long, and noisy, simple NLP methods can help to acquire more relevant concept terms with lexical processing.

The rest of the paper is organized as follows. Section 2 presents some related works. Section 3 describes the proposed method for acquiring concepts using Wiki resources. Section 4 shows the experiments and evaluation details. Section 5 concludes the paper with future directions.

II. RELATED WORKS

For ontology construction, terms are usually extracted first and relations between them are then mined. If terms are automatically extracted, concepts and instances are difficult to distinguish. Ontology construction methods usually do not distinguish concepts and instances before identifying taxonomic relations [6] unless named entity recognizer or other filtering tools are applied. Mining the web to obtain concepts for ontology construction in a considerable scale without semantic annotation is also not easy. The work in [7] mined concepts and their definitions with the help of search engines. For a term as a topic, its sub-topics or salient concepts are emphasized word phrases by specific html tags with high frequency. They were informative pages returned by some search engines containing definitions extracted according to set of rules. This method is topic oriented and cannot work for general domain concepts. Also, a precision of 61.2% for definitions acquisition is not qualified to mine concepts or relations for ontology construction.

For Wiki mining, most researches take the titles of Wiki pages as the concept terms or concept instances directly. In some research works, concepts are considered as classes corresponding to instances. Some works build taxonomy from Wiki by analyzing the links and other Wiki resources, such as Category Labels. [8] and [9] generate taxonomy by calculating the degree of association between Wiki pages through URLs and Category Labels without discriminating

concepts from instances. The focus of [10] is to associate terms with concepts using the Disambiguation label in Wiki pages and article titles to build a network of terms sharing the same lexical term as a substring in Article pages. This lexical name is then considered the representative for the concept of these title terms. However, there is some abnormality. For example, the lexical term "*Java*" is used in many article titles. So, there is a Disambiguation page named "*Java(disambiguation)*" linking to "*JavaScript*", "*Java(band)*", "*Java Platform*". Consequently, "*Java*" is considered as a concept. However, in IT domain, "*Java*" is in fact an instance of the concept "*programming language*" and it is not a concept term by itself.

Categories in Wiki taxonomy are distinguished into instances and concepts in [11]. The concepts in [11] are classes. Discriminating features are based on observations such as capitalization of Wiki Category titles, presence of plural form in Category titles, etc.. Named entity recognition and other tools are also used. The evaluation for this classification is conducted on the 7,860 extracted concepts in Wiki overlapped with another ontology resource ResearchCyc [12]. The average precision is between 81.6% and 84.5%. However, categories not in the evaluation set tend to have more instances which should be further evaluated. For example, many instance categories such as "*Category: 17th century mathematicians*" and "*Category: 120 mm discs*" are considered as classes. But, they are actually instances and should be classified as such. In fact, in current ontology construction methods, there is no clear way to distinguish concepts from instances.

III. MINE CONCEPTS FROM WIKIPEDIA

This work attempts methods using different annotations and features in Wiki to distinguish concepts from instances. As a large scale on-line encyclopedia, besides {{Infobox}} Structures used in past research [5] which has low coverage, this paper explores definition sentences and Category Labels in Wiki to identify the concept for a given instance Article page with the help of NLP method.

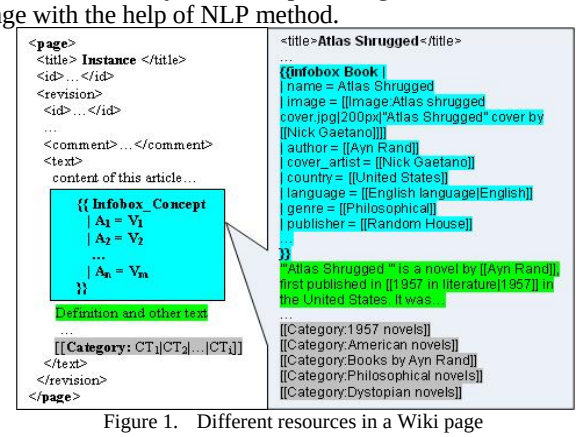

Figure 1. Different resources in a Wiki page

Figure 1 gives an example of a Wiki page containing {{Infobox}} Structure, definition sentence, and Category Labels. The left part of Figure 1 shows the structure of the page and the positions of Wiki labels. The right part shows an example page entitled "*Atlas Shrugged*", which is an instance page of the concept "*book*" because it contains an {{Infobox *book*}} Structure highlighted by the turquoise color. The first sentence highlighted in bright green is the definition sentence which points out that "*Atlas Shrugged*" is a "*novel*". The Category Labels highlighted in gray color indicate the topics of this page as "*Category: 1957 novels*", and "*Category: Books by Ayn Rand*", etc..

A. Using the {{Infobox}} Structures

In this paper, the method using {{Infobox}} Structure will be used as the baseline for the comparison to [5]. An {{Infobox}} Structure is a formatted table present in some Article page *P* and labeled by a common subject $concept_p$ in the form of {{Infobox $concept_p$}}to indicate that it is an instance page with reference to the $concept_p$ it belongs to. Pages with the {{Infobox}} Structures pointing to the same concept are instances of the same concept. More than one {{Infobox}} Structure can be present in an Article page. Thus, an instance can be associated with multiple concepts. For example, the Article page "*Arnold Schwarzenegger*" is an instance of both {{Infobox *Actor*}} and {{Infobox *Governor*}}. {{Infobox}} Structure can be extracted according to the XML and Wiki tags in Wiki pages. In all Article pages, only 15% contains the {{Infobox}} structure. Thus it has quite a low coverage.

B. Using definition sentences

All Article pages contain definition sentences. Often, a definition sentence also provides the hyernym concept associated with the given instance or concept. For example, the definition sentence of "*Atlas Shrugged*" is "*'''Atlas Shrugged'' is a novel by Russian-born writer and philosopher [[Ayn Rand]], first published in 1957 in the [[United States|USA]].*", which states that "*Atlas Shrugged*" is an instance of the concept "*novel*". Before extracting definitions, pre-processing must be conducted to remove unnecessary information such as XML labels and Wiki structures. In principle, the first full stop by the punctuation mark "." (period) signifies the end of the first sentence. In practice, other period marks especially for abbreviations such as in "*Prof.*", "*No.1*", and "*U.S.*" must be preprocessed too.

The main verbs of definition sentences are obtained through syntactic analysis. Nouns or noun phrases following the main verbs are extracted as concept terms. There are two kinds of main verbs, the be-verbs and the non-be-verbs. Be-verbs are "is", "was", "are", etc., which normally indicate the *is-a* relationship. Thus the nouns or noun phrases directly after a be-verb can be considered as the corresponding concepts of the subjects. The relationship indicated by non-be-verbs varies depending on the verbs used. As a full parser for verb extraction and noun phrase extraction can be time consuming and less accurate, only a POS tagger is used in this work. Regular expressions are then applied to identify the noun/noun phrases after verbs. For sentences with verbs followed by "type of", "a kind of", "name of", and "one of", the noun phrases after "of" is extracted as the target.

C. Using Category Labels

Category Labels are used to indicate topic information of Article pages in Wiki. There are also Category pages which

lists all the sub-categories under the current label. If all Wiki pages are organized as a hierarchy by Category pages, the Article pages are always leaf nodes in this hierarchy. Obviously, nodes in the category hierarchy are more likely to be instances and concepts/categories are more likely to be non-leaf nodes. Category pages with no Article page and no sub-categories are also leaf nodes in the hierarchy and likely to be instances. In this paper, these Leaf pages are assumed to serve as instances for which concepts can be mined. No matter they are instances or concepts, the Category Labels contained in them should be concepts. Only about 86% of the Leaf pages are Article pages, which have definition sentences. But all pages contain Category Labels, which means Category labels are contained in more Wiki pages that tha of definition sentences. A Leaf page can be linked to a number of Categories through Category Labels, which usually reflect the relations between instances and concepts. For example, the page "*Atlas Shrugged*" contains several Category Labels such as "*Category: 1957 novels*", "*Category: Novels by Ayn Rand*" and so on. But Category Labels cannot be used directly because they are given by the Wiki page editors. As there is no rule imposed, the labels tend to be arbitrary, long and inconsistent. In the former example, Category labels contain noise such as "*1957*" and "*by Ayn Rand*". Thus, the use of Category Labels as a complete token is not appropriate. Instead, the term "*novels*" which appears in most of the labels is a good candidate concept term.

Based on this observation, Category Labels are split by stop words, such as "in", "of", into smaller components. Only unigrams and bigrams with certain frequencies are calculated and the components with highest product of frequency and length will be considered as the most general concept term to this instance. The algorithm for extracting concepts is listed below:

1 **For** each selected Leaf page ***p***,
2 **If** it is a Wiki Article page,
3 Save the title name ***t*** in ***T***;
4 **Else** if it is a Category page,
5 Delete the prefix "**Category:**";
6 Save the title name ***t*** in ***T***;
7 **End for** (* end of extraction of all titles *)
8 **For** each title name ***t*** in ***T***,
9 Save all the Category Labels ***c*** pointed by ***t*** to C_t (including duplicates);
10 **For** each Category Label ***c*** in C_t,
11 Split ***c*** by stop words from a given stop word list;
12 Save all split component words ***w*** into W_t (including duplicates);
13 **End for** (* extraction of component words *)
14 **For** each ***w*** in W_t,
15 Collect unigrams and bigrams of *w* with their frequencies into G_t;
16 **End for**
(* collection of statistics of component terms *)
17 Select ***g*** from G_t with the highest product of frequency and length (**g* is the concept for ***t*** *)
18 **End for** (* completed for one title *)

Among these steps, steps 1-7 are to collect the titles of Leaf pages. Steps 8-16 are to collect unigrams and bigrams of Category Labels as candidate concepts. Steps 17-18 are to select one concept from candidates for each instance.

IV. PERFORMANCE EVALUATIONS

The evaluation is conducted on the English Wiki corpus containing 1.1 million pages on the cut off date of November 30th, 2006 with about 700,000 Leaf pages. Sampling method is used to select instances and manually evaluate whether their associated concepts are correct. For each set of results from different resources, 400 samples are selected to limit the margin of error within 5% [13]. The evaluation criteria for proposed method are precision and coverage in terms of Leaf pages. The coverage is not the same as recall because it only indicates how many resources are covered, not the number of concepts in these pages. To give a balanced measure of precision and coverage, the F-measure is calculated for precision and coverage, called F'-measure. When using different Wiki information, evaluation must be done separately. For example, the concept for page "*Atlas Shrugged*" is book according to the contained {{Infobox *book*}} Structure. But it is considered as a "*novel*" by using definition sentence and Category information. Both "*book*" and "*novel*" are correct concepts for this instance..

Row 1 of Table I shows the performance of using the {{Infobox}} Structures, labeled as **INF** (INFobox) as the baseline. In INF, only 1,201 concepts are acquired for about 111,623 instances. For example, the concept "*company*" has 2,585 instances, such as "*Microsoft*" and "*Bank of China*". Among the extracted concepts, 90.1% are correct. However, the coverage is about 15% because only 15% of all Leaf pages contain the {{Infobox}} Structure. Thus the F'-measure reaches only 25.5%.

TABLE I. PERFORMANCES OF DIFFERENT RESOURCES

	Infobox (INF)	Definition Sentences			Categories (CAT)
		BDS	NBD	ADS	
Precision	90.1%	76.7%	33.2%	73.2%	75.3%
Coverage	15%	79.2%	7.0%	86.2%	100%
F'-measure	25.5%	77.9%	11.6%	79.2%	85.9%

For definition sentences, the Stanford POS tagger [14] is employed to identify the noun phrases after the verbs. It tags general noun phrases as NN(S) and proper nouns as NNP(S). Our algorithm prefers to select NN(S) as concepts. Only if NN(S) is not present, NNP(S) is taken as the concepts. As be-verb sentences definitely contain the *is-a* relation, the performance of be-verb sentences is separately evaluated in Table 1, labeled as **BDS** (Be-verb Definition Sentences). The evaluation of non-be word sentences and all definition sentences are labeled as **NBD** (Non-be-verb Definitions), and **ADS** (All Definition Sentences), respectively. The precision of BDS is 76.7%, better than that of ADS by 3.5%, which support the assumption that be-verbs are more likely to indicate definitions. The precision of the non-be-verb sentences is only around 33.2%, about 40% less than the be-verb sentences. But, it can help with an additional 7% coverage. So the F'-measure of ADS is 1.3% better than that of BDS. ADS can cover 86.2% of all Leaf pages and gives a much improved coverage compared to that using {{Infobox}} Structure. Errors in precision are caused by two

kinds of problems. The first kind contains the corresponding concepts in the sentences, yet, the extraction of NN and NNP cannot identify them correctly. For example, in the sentence "*Vai Sikahema was an NFL running back who played for 8 seasons from 1986 to 1993.*", "*Running back*" is the correct concept. But, the algorithm only extract "*running*" as the concept. The second kind does not contain the corresponding concepts. For example, in the sentence "*Denham railway station is on the Chiltern Line out of Marylebone towards High Wycombe*.", although "*Chiltern Line*" is a relevant noun phrase to "*Denham railway station*", it is not the right concept. In fact, the corresponding concept should be "*railway station*", which does not even appear after the be-verb. By a rough estimate, 62.5% of errors fall into the first kind, and 37.5% falls into the second kind.

The performance of using Category information is listed in Table I, labeled as **CAT**. The precision is between ADS and BDS. Nearly half of the mistakes are due to place or facility names whose Category Labels are given by other hypernym place names which are also instances rather than concepts. The hypothesis that instance pages should be assigned to Categories that are concepts is obviously incorrect in these cases. For example, the instance "*Orwell, New York*" has two Categories: "*Oswego County, New York*" and "*Towns in New York*". Neither "*county* "nor "*towns*" are selected because the most frequently used phrase is "*New York*" according to the gram-selection method. This problem is not caused by the selection algorithm. In place and organization names etc. where there is a natural hierarchical structure, instances can be linked to other instances of a higher level. By applying a simple pre-processing rule to address this problem, an improvement in precision by 7.6% is already reflected in Table 1. Categories cover 100% Leaf pages. The F'-measure of CAT reaches 85.9%, the best among all three methods.

Table II shows the evaluation of combining definition sentences and the Categories, labeled as **BDS+CAT** and **ADS+CAT**, respectively. BDS+CAT and ADS+CAT methods are combined on the condition that if there is no presence of NN or NNP in BDS/ADS, the corresponding Category information will be used.

TABLE II. PRECISIONS OF COMBINED RESOURCES

	BDS	**BDS+CAT**	**ADS**	**ADS+CAT**
Precision	76.7%	78.5%	73.2%	73.8%

Table II indicates that BDS combined with CAT gives about 1.8% additional performance improvement. Improvement to ADS is only about 0.6%, which makes the combination more worthwhile for BDS. The fact that BDS+CAT outperforms ADS+CAT is because BDS information is more accurate. For example, BDS cannot find a concept for the page "*Domestic violence*" because its definition "*Domestic violence occurs when a family member, partner or ex-partner attempts to physically or psychologically dominate or harm the other.*" does not contain the be-verb. But ADS assigns "*family member*" as the concept which is not correct. However, BDS+CAT can identify "*violence*" as the result according to the Category Labels in the Article page.

V. CONCLUSION AND FUTURE WORKS

This paper proposes a novel approach to mine concepts for instance pages in Wiki using three different resources. The proposed method identified nearly 50,000 concepts for about 700,000 Wiki Leaf pages as concept with a precision of 78.5% to cover much more concepts than existing work with a reasonable precision. The use of unigram and bi-gram statistics effectively eliminate instance information in conceptual descriptions to identify the most appropriate concept terms for instances. Future works can be done on concept extraction with more Wiki resources such as Disambiguation pages. More comprehensive combination of different resources in Wiki can also be tried.

ACKNOWLEDGMENT

This project is partially supported by CERG grants PolyU 5190/04E, PolyU 5225/05E and PolyU Central Grant G-U596.

REFERENCES

[1] I. Niles and A. Pease. "Towards a Standard Upper Ontology", in proceedings of FOIS-2001, available at the following address, http://home.earthlink.net/~adampease/professional/FOIS.pdf, last visited Apr. 1, 2009.

[2] Y. R. Chen, Q. Lu, W. J. Li, W. Y. Li, L. N. Ji, and G. Y. Cui, "Automatic Construction of a Chinese Core Ontology from an English-Chinese Term Bank", *OntoLex07*, Busan, 2007, pp. 78-87.

[3] L. Zhou, "Ontology Learning: State of the Art and Open Issues", Information Technology and Management, 8(3), pp.241-252, 2007.

[4] J. Kazama and K. Torisawa, "Exploiting Wikipedia as External Knowledge for Named Entity Recognition", *EMNLP-CoNLL 2007*, Prague, Jun. 2007, pp. 698-707.

[5] G. Y. Cui, Q. Lu, W. J. Li, and Y. R. Chen, "Automatic Acquisition of Attributes for Ontology Construction", *ICCPOL 2009*, Hong Kong, Mar. 2009, pp. 248-259

[6] R. Navigli and P. Velardi, "Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites", Computational Linguistics, 2004, MIT Press.

[7] B. Liu, C. W. Chin and H. T. Ng, "Mining Topic-Specific Concepts and Definitions on the Web", *WWW 2003*, Budapest, Hungary, May 23-24, 2003.

[8] M. Shirakawa et al., "Concept Vector Extraction from Wikipedia Category Network", *ICUIMC-09*, Suwon, Jan. 2009.

[9] G. Y. Cui, Q. Lu, W. J. Li, and Y. R. Chen, "Corpus Exploitation from Wikipedia for Ontology Construction", *LREC 2008*, Marrakech, 2008, pp. 2125-2132.

[10] A. Gregorowicz and M. A. Kramer, "Mining a Large-Scale Term-Concept Network from Wikipedia", Technical Report #06-1028, The MITRE Corp., Oct. 2006.

[11] C. Zirn, V. Nastase, and M. Strube, "Distinguishing Between Instances and Classes in the Wikipedia Taxonomy", *ESWC2008*, Tenerife, 2008.

[12] Available at http://research.cyc.com/

[13] Margin of error. (2009, May 5). In *Wikipedia, The Free Encyclopedia*. Retrieved 03:46, May 5, 2009, from http://en.wikipedia.org/w/index.php?title=Margin_of_error&oldid=287987112

[14] K. Toutanova and C. D. Manning, "Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger", *EMNLP/VLC-2000*, Hong Kong, Oct. 2000, pp. 63-70.