Loose Method for Pattern Classification in Wikipedia using Duality Theorem for Knowledge Acquisition in Neigbouring Words

Enikuomehin Toyin
Department of Computer Science
Lagos State University
Ojo, Lagos State, Nigeria

Akerele Olubunmi
Department of Computer Science
Lagos State University
Ojo, Lagos State, Nigeria

ABSTRACT

In this paper, we present an approach for structural classification of taxonomies for knowledge acquisition from Wikipedia using standard loose frameworks. Knowledge mapped from WordNet are assigned to corresponding patterns in Wikipedia such that the syse structure are automatically acquired for related patterns and then used for knowledge generation, achievable through Learning. The paper considers the theory of duality principle as posed in Hilbert spaces to describe the operation of two terms related by their linguistic classifications such as hyponyms. Results show that knowledge can be acquired with well formulated pattern, however a lot of gaps still exist which can be solved using manual approaches as that seems to be more efficient based on the experiment conducted.

General Terms

Wikipedia, Taxonomy, Classification, WordNet

Keywords

Structural classification, taxonomy, standard loose framework, POS

1. INTRODUCTION

In [1] a supervised learning method was introduced for natural text to perform learning task after due classification has been carried out. Unlike the earlier approaches that limit their focus on the semantic pattern of the existing categories, the method considers the document in relation to the context at which it is contained. We extend the focus of the paper to Wikipedia articles with interest in relations that exist between concepts as against the links between them. Pattern acquisition from text has been a solid platform for learning automatic frameworks such as knowledge; patterns in ontological domains which have made users rely so much on the ability of the concept to initially generate an appropriate structure before learning takes place. This paper presented herein, uses the knowledge described in Unguided Loose Search (ULS*)* [2]) to present an enhanced model for the acquisition of knowledge based on patterns. The approach involves the use of similar patterns classified under the same subtree or nodes as predetermined on the wikipedia platform. Learning of these tree patterns begin with the generation of a SyntacticSemantic (syse) structure from WordNet. In such cases, the generated structure is mapped unto wikipedia such that the knowledge concepts and structure are automatically extracted when the pattern falls under the same category. The categorization, a form of classification is implemented using a duality principle to create associativity between the terms. The process forms the basis for the generation of the required knowledge.

The rest of the paper is organizational as follows: Section 2 describes the structural mapping of concepts into wikipedia using taxonomies with efficient algorithm, experimental setup and results are evaluated in section 3 while discussions follows in section 4 with a comparison with manual approach in the knowledge extraction process. Recommendations follow to end the paper.

2. STRUCTURAL MAPPING OF CONCEPTS IN WIKIPEDIA

Many earlier approaches has faced tremendous problems resulting from a set of coverage problems. In Ruiz-casado, Machine Learning were used to learn specific semantic patterns whose relation were used to link word (nouns) to wordnet, the expected throughput in the context coverage was not achieved because Wikipedia exist as a restricted content encyclopedia with about 2000 relationships of which 700 exist in wordnet. The accuracy recorded so far in the pattern extraction from wikipedia has largely been due to the consistency in the wikipedia mapping techniques [3], [4]). Automatically refining the wikipedia infobox ontology [5]. If the knowledge patten for a concept, say bird is required the corresponding sense number, say 1 is mapped in WordNet. [6]. The member which has the same sense number as 1 is identified and its taxonomical category is then used for pattern alignment which can further more be used as input for the pattern generation system.

3. OUR APPROACH

Syntactic Semantic SySe patterns are core values in approach of find related knowledge within documents referenced in Wikipedia. Essentially, the focus will use the approach of identifying the first sentence in a document in Wikipedia in other to generate knowledge from the concept presented. The algorithm used in this paper is based on the performance of the hypernym patterns in the example shown in figure 1 below:

3.1 Hypernmy Patterns

Is a

Is any form of

Is typical a

Is a class of

Is defined as a

Example: *Primate ::: Chimpanze*

Musical Instrument ::: Guitar

Thus the meaning patterns also known as Se Patterns are special word patterns that shows the underlining linguistic relation of a certain type of hyponmy. Wikipedia uses this knowledge widely in the description and presentation of their contents. In all cases, two set of word and concepts such as X and Y are connected and this is being done like an operand on operators. Generally, the operations include hyponymy-is a; holonmy-has a; meronmy- part of; synonyms-equal relations and since the is-a is the most used linguistic pattern in Wikipedia, it shall be used for the experimental set up presented in this paper. Simply put, *a goat is a domesticated animal*; is an example of a hyponmy relation relating goat to animals, the SySe pattern can then be built of the form: (*goat, is-a, animal*). The aim of the loose approach is to be able to generate more than one meaning-fetching pattern from concepts. The set of words described above can be transformed into a more useful system if we consider X as A and Y as B, using [7], [8], and [9]. To show that the relationship between the terms. Recall in the duality principle, If {X,R} is a poset , then {X,R(inverse)} is also a poset. Thus, the above can be shown by proving the stated Lemma.

3.2 Lemma 1:

The Dual of $C^p[a,b]$.

Given the interval $[a,b]$ and $1 \le p < \infty$ is the collection of all real valued function that are *p*-times continuously differentiable on $[a,b]$. Clearly, $C^p[a,b]$. is a vector space. We show that for each $x \in C^p[a,b]$, the following

$$\|x\| = \max_{t\in J}|x(t)| + \max_{t\in J}\left|x'(t) + \cdots + \max_{t\in J}\left|x^{(p)}(t)\right|\right| = \sum_{k=0}^{p}\max_{t\in J}\left|x^{(k)}(t)\right|$$

Where $J = [a,b]$, defines a norm.

$$\|\alpha x\| = \max|\alpha x(t)| + \max|\alpha x'(t)| + \cdots + \max\left|\alpha x^{(p)}(t)\right|$$

$$= |\alpha|\left\{\max|x(t)| + \max|x'(t)| + \cdots + \max\left|x^{(p)}(t)\right|\right\}$$

$$= |\alpha|\|x\|$$

If $x, y \in C^p$, then for any $k = 0,1,\ldots, p$

$$\left|x^{(k)}(t) + y^{(k)}(t)\right| \le \left|x^{(k)}(t)\right| + \left|y^{(k)}(t)\right| \le \max\left|x^{(k)}(t)\right| + \max\left|y^{(k)}(t)\right|$$

and also

$$\max\left|x^{(k)}(t) + y^{(k)}(t)\right| \le \max\left|x^{(k)}(t)\right| + \max\left|y^{(k)}(t)\right|$$

Consequently,

$$\|x + y\| = \sum_{k=0}^{p}\max\left|x^{(k)}(t) + y^{(k)}(t)\right|.$$

$$\le \sum_{k=0}^{p}\max\left|x^{(k)}(t)\right| + \sum_{k=0}^{p}\max\left|y^{(k)}(t)\right| = \|x\| + \|y\|$$

[Here $x^{(0)}(t) = x(t)$]

Hence, $C^p[a,b]$. is a normed space.

We next establish the Riesz representation theorem for elements *f* of the dual of the real linear space $C^p[a,b]$.. We define a functional *f* on $C^p[a,b]$. by:

$$f(x) = \int_a^b \left\{x(t) + x'(t) + \cdots + x^{(p)}(t)\right\}dt.$$

Definitely, *f* is linear and bounded with norm

$$\|f\| = b - a$$

$$|f(x)| = \left| \int_a^b \left\{ x(t) + x'(t) + \cdots + x^{(p)}(t) \right\} dt \right|$$

$$\leq (b-a) \left\{ \max_{t\in J} |x(t)| + \max_{t\in J} |x'(t) + \cdots + \max_{t\in J} |x^{(p)}(t)| \right\}$$

$$= (b-a)\|x\|$$

taking the supremum over all *x* of norm 1, we obtain $\| f \| \leq b-a.$

To get $\| f \| \geq b-a$, we choose $x(t) = x'(t) = \cdots = x^{(p)}(t) = 1$

$$\| f \| \geq \frac{|f(x)|}{\|x\|} = |f(x)| = \int_a^b dt = b-a$$

Thus, the neighboring terms of the first sentence can be wrapped up as tokens in the sentence. The initial step is to identify the first sentence of each article and tag the containing tokens. This is necessary in identifying the semantic relation of the context. The proposed processing takes the flow shown in figure 1 below:

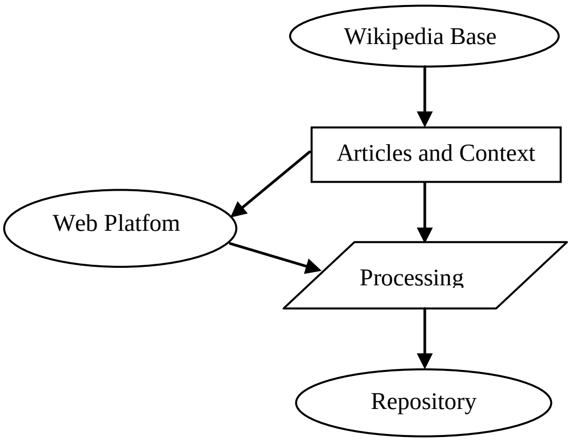

Figure 1: The Process flow

Identifying the concept on the wikipedia server helps to generate the relevance of the document to the set criteria. The system performs a co-document existence resolution on the wikipedia server using a common taxonomy. In achieving this, terms like pronoun which appear at the beginning of a document is suspected to reference the topic of the document [10]. Thereby making it a topical pronoun [11]. The first five terms in any document under the classification is also identified, which is a step performed before the linguistic processor is activated for the Part of Speech (POS) tagging [12] and for lemmatization [13]. A morphological analysis consist of POS tagging and a Lemma (basic form) corresponding to this tag and features combination [14].
Processes such as stemming, lemmatization, term identification, are also performed. Using a generic label, we present a regular expression of the form (N prep, Nverb, …) usable for linguistic instances lemmatization. As an example, if the taxonomy for a goat is considered as being used above, the following properties can be extracted: *a domesticated animal, a subspecies of the wild goat of south Asia, live on land, one of the oldest domesticated anima*l, (DoAn*), belongs to the goat-antelope sub family* and *with a current population of 929 million living*. From the above, a logical knowledge framework can be extracted of the form: *most goat live on land.*

Since knowledge are generally extracted from multidimensional sources which includes databases, web repositories, thesaurus, corporal amongst others. With latest and fastest growing knowledge platform being Wikipedia [15] it is important to understand the structure to which individual source presents its own concept such that encyclopedia like the Wikipedia will not spend much time on structure formatting before knowledge can be built. Wikipedia has shown to be tremendously useful in Information Retrieval, Information Extraction, Ontology, etc. and many other fields. The tokenization can take place once the first sentence of each article has been identified. Thus, the relationship between FS and article is expected to be of the form (1:1).

The advantage point of Wikipedia that interest researchers includes the rate of update of its content by hundreds of thousands of accredited volunteers. Verification of content has also being an issue of concern to web content validity researcher [16].The framework allows for deep search [17]. As each fact is established as content can be surfed for sources. A hypernym extraction algorithm [18] is adapted for the initial concept identification of tokens in Wikipedia and presented below:

- *Large training set with many uncommon patterns "X is a ADJ term that refers to a kind of Y"*
- *Annotated with 4 fields: definiendum (**D**), definitor (**V**) containing the verbal pattern and definiens (**H**) containing the hypernym, and the rest of the sentence (R).*
    - *An <Albedo> (often represented by the generic formula HA)/ is traditionally considered / any **chemical compound** / that, when dissolved in water, gives a solution with a hydrogen ion activity greater than in pure water*

- *The algorithm builds a set of word lattices from the training set. Independent lattices are created for each of the 3 basic fields*

Lattice learning consists of three steps:

1. *each sentence in the training set is pre-processed and each field is generalized to a star pattern*

“[In arts, a chiaroscuro]$_D$ [is]$_V$ [a monochrome picture]$_H$.”

D=“In *, a <TARGET>**”,** V=“is”, H=“a * <HYPER>**”**

1. *Clustering: for each field, the training sentences are then clustered according to the star patterns they belong to;*

In arts, a chiaroscuro is a monochrome picture.

In mathematics, a graph is a data structure that consists of . . .

In computer science, a pixel is a dot that is part of a computer image.

D: In * , a <TARGET>

V: is

H: a * <HYPER>

. Word-Class Lattice construction: for each sentence cluster, a WCL is created by means of a greedy alignment algorithm

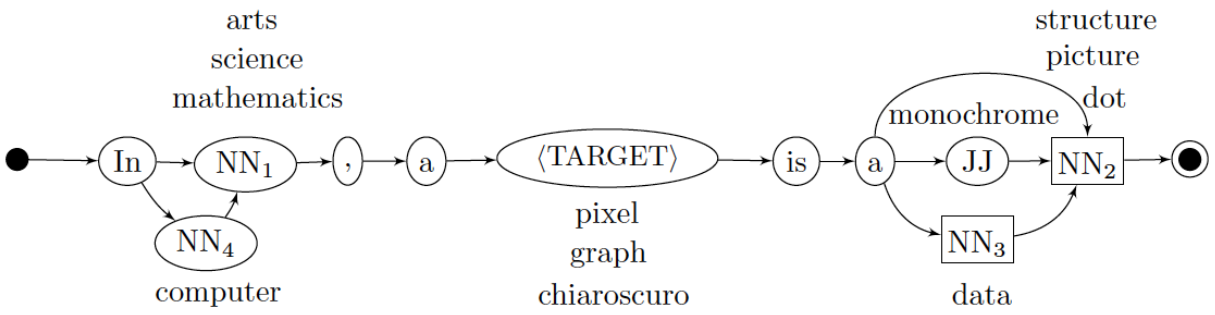

Figure 2: Lattice Decomposition

In this paper, a novel approach for structural acquisition of pattern from Wikipedia is introduced.

4. RESULT AND DISCUSSION

In the experimental setup, WordNet taxonomies were obtained and used as the input class. Essentially, the class animal is defined over a set A (C, G, F) as arguments for Cow, Goat and Fish respectively.

Recall that the term concept used in this paper refers to an item in the classification taxonomy which may correspond to several terms in the form of singular, plural. Synonymy tokens as applicable, the root concept used in the experiment is Animal and machine, and this is because these concepts have large taxonomies in WordNet, which can help to test the validity of our experiment accordingly. The dataset used in the experiment were collected from search engine result. This is done by submitting the generated pattern in figure 1 as query to search engine. For this experiment, google search was used. A total of 2.6g size was retrieved after the algorithm e] was set to truncate at the 15th run. To validate the extracted concepts, the Wikipedia entries relating to each input class in WordNet was mapped to its taxonomical class for POS and lemmatization to be performed by the target tree.

To complete the process, we test for the performance of the following:

Precision: Correctness of retrieved concepts

Recall: The coverage of the retrieved concepts

Learning: Degree of accuracy of learning

The associated scores is retrieved using the following;

$$\text{PrWn} = \frac{\#No\_of\_terms\_in\_WordNet}{\#No\_of\_terms\_returned\_by\_a\lg orithm}$$

$$\text{PrR} = \frac{\#No\_of\_terms\_marked\_correct}{\#No\_of\_terms\_returned\_by\_a\lg orithm}$$

Using the returned data, the output is described in figure 3 below:

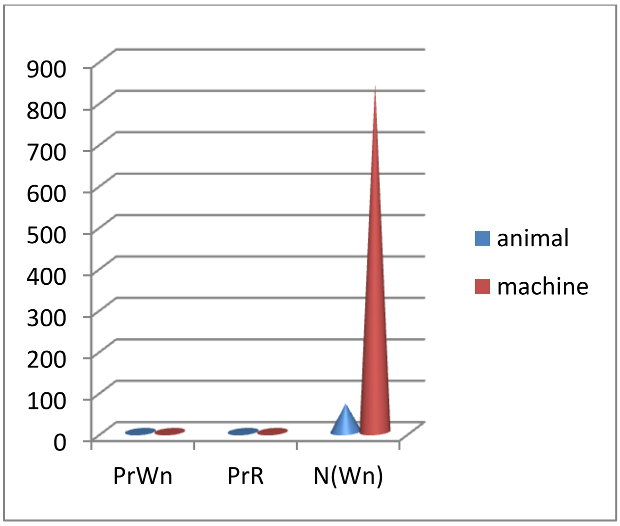

Figure 3: The output using the returned data

The table above shows the retrieved terms from WordNet for term available in the animal class of Wikipedia against term found in the wordnet. N (Wn) stands for terms not available in Wn. The algorithm is set to claim that vid(X, Y) showing the extent to which X is a subconcept of Y based on the platform used in term retrieval. In the case of the set experiment,

Vid(goat, DoAn) would be correct but Vid(goat, Water) will be be incorrect, stressing that the correct knowledge pattern will be vid(Fish,Water). To test the obtained result, a manual comparison is carried out against the automated Wikipedia learning extraction. The result is shown in the table below;

Table 1: The Animal Machine Taxonomical Evaluation

Vid	PrWn	PrP	N(Wn)
Animal	.46	.78	906
Machine	.22	.83	640

The result shows that a possibility of extracting knowledge from derived Wikipedia pattern however further investigation shows that the WordNet does not have does not have about half of the taxonomies generated by the algorithm(906 for animal: 640 for machine0.

5. CONCLUSION

In this paper, we describe a novel method for the extraction of knowledge using the SySe method for pattern generation for similar concepts in WordNet as used within the encyclopedia of Wikipedia, The work shows how the concept can be tested for the validity and furthermore provides a framework for automatic structural learning. Pattern were used to classify concepts thereby enabling assignment of closely related terms as sharing some knowledge.

We conclude that aside the result report above, the experiment shows that it is important to conduct manual evaluation in WordNet results. The terms together with the semantic pattern make up the required knowledge from Wikipedia.

Further research is required in the area of evaluation framework.

6. REFERENCE

[1] Roxana, G., Adriana, B., Oxana, G., and Dan, M. (2006). Automatic Discovery of part-whole relations. Computations Linguistics, 32:1.

[2] Enikuomehin, O., Sadiku, A., & Egbudin, M. (2014). A Critical Review of the Unguided Loose Search (ULS) Process for Natural Language Based Extraction Technique on Relational Databases. Transactions on Machine Learning and Artificial Intelligence, 2(4), 01-11

[3] Wu, F., & Weld, D. (2008). Automatically refining the wikipedia infobox ontology. Proceedings of the 17th international conference on World Wide Web, 635-644

[4] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data, 722-735

[5] Genc, Y., Sakamoto, Y., & Nickerson, J., (2011). Discovering context: Classifying tweets through a semantic transform based on Wikipedia. Foundations of Augmented Cognition: Directing the future of Adaptive Systems, HCL International July 9-14, Orlando, FL, 484-492

[6] Sameh, A. (2013). A Twitter analytic tool to measure opinion, influence and trust. Journal of Industrial and Intelligent Information, 1(1).

[7] Van Rijsbergen, C., (2004). The geometry of information retrieval. Vol. 157. Cambridge: Cambridge University Press, ISBN: 0521838053

[8] Amati, G., & Van Rijsbergen, C. J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Transactions on Information Systems (TOIS), 20(4), 357-389

[9] Van Rijsbergen, C. J. (1986). A non-classical logic for information retrieval. The computer journal, 29(6), 481-485

[10] Klyuev, V., & Oleshchuk, V. (2011). Semantic retrieval: an approach to representing, searching and summarising text documents. International Journal of Information Technology, Communications and Convergence, 1(2), 221-234

[11] Haspelmath, M. (1999). Optimality and diachronic adaptation. Zeitschrift für Sprachwissenschaft, 18(2), 180-205

[12] Voutilainen, A. (2003). Part-of-speech tagging. The Oxford handbook of computational linguistics, 219-232

[13] Korenius, T., Laurikkala, J., Järvelin, K., & Juhola, M. (2004). Stemming and lemmatization in the clustering of finish text documents. Proceedings of the thirteenth ACM international conference on Information and knowledge management, 625-633

[14] Suh, B., Convertino, G., Chi, E. & Pirolli, P., (2009). The singularity is not near: slowing growth of Wikipedia. Proceedings of the 5th International Symposium on Wikis and Open Collaboration Article No. 8

[15] Callahan, E., & Herring, S., (2011). Cultural bias in Wikipedia content on famous persons. Journal of the American society for information science and technology, 62(10), 1899-1915

[16] Bergman, M. K. (2001). White paper: the deep web: surfacing hidden value. Journal of electronic publishing, 7(1).

[17] Navigli, R., Velardi, P., & Faralli, S. (2011). A graph-based algorithm for inducing lexical taxonomies from scratch. In IJCAI, 1872-1877

[18] Kristina, T. & Colin, C. (2009). A global model for joint lemmatization and part-of-speech prediction. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP pp. 486-494