Automatic Extension of a Lexical Ontology Using Web Resources

James Austrow
The Ohio State University
154 W 12th Avenue
Columbus, OH 43210
austrow.1@osu.edu

ABSTRACT

Lexical ontologies, or databases of words and their relationships, are valuable tools for a variety of natural language processing applications. Their costly construction and maintenance times, however, limit their scope and ease of development. This causes them to be difficult to keep up to date with current terminology and concepts. We develop and compare several procedures for automatically updating an existing lexical ontology, focusing on WordNet, based on Wikipedia articles. An appropriate hypernym for each article must be found in order to maintain the hierarchical structure of WordNet, so the different methods we compare focus on different ways to determine this hypernym.

Categories and Subject Descriptors

D.2.8 [**Software**]: Software Engineering—*performance metrics*; H.4 [**Information Systems**]: Applications—*miscellaneous*

General Terms

Experimentation, Performance

Keywords

ontology

1. INTRODUCTION

Lexical ontologies have proven to be very useful in the field of natural language processing. The structure of grouping synonymous word senses together into synsets and arranging them in a graph of relationships such as hypernym-hyponym pairs provides much semantic information that is not apparent from the words themselves. However, one drawback of these systems is that they are typically constructed and maintained by hand, requiring costly effort. One such ontology in common use today is WordNet [1], which will be the focus of this research.

We aim to investigate approaches overcoming the costly maintenance time of lexical ontologies by automatically integrating information found on the web. Wikipedia is a user-edited source of words and concepts, updated more or less in real time. The fact that it is a relatively current source of information makes it attractive for the purpose of keeping a comprehensive ontology such as WordNet updated. Various methods of incorporating word relation data from Wikipedia will be explored.

The remainder of this paper is organized as follows. Section 2.1 describes previous work that is related to this research. Section 2.3 describes the general approach this research is based on and outlines improvements we have made. Section 2.4 details how the methods are implemented. Section 2.5 explains how the results of this research are tested and details the results of the experiment. Section 2.6 shows what future improvements we plan on making. Finally, Section 3 is the conclusion.

1.1 Related Work

The ideas of this research are based heavily on the work of Jiang et al, who propose a method of incorporating articles from Wikipedia into WordNet [3]. They rely on category information from the article to determine its hypernym and achieved encouraging results. However, in a few cases their algorithm detects an incorrect sense of the hypernym within WordNet or identifies a bad category as the most likely hypernym and thus misclassifies the article. Therefore, alternate methods of hypernym extraction are considered. Sang [6], building on the work of Hearst [2] and Snow et al. [7], describes a method of hypernym extraction based on fixed patterns of text. Additionally, Kleigr et al. [4] note that there are several patterns unique to Wikipedia articles that can aid in hypernym identification.

2. EXTENDING WORDNET

2.1 Method

The method of Jiang et. al. is used to automatically extend WordNet with Wikipedia entries. This involves compiling the categories of each article that appear in WordNet as potential hypernyms of the article. The definition of each potential hypernym (from WordNet) and the text of the article are compared for concept similarity to determine the best match. If no category of the article appears in WordNet, the head term of each category is considered instead, and the original category of the chosen hypernym is inserted into WordNet as well.

One of the difficulties of this method is that it is quite slow. A similarity measure must be computed between each pair of concepts appearing in the Wikipedia article and in the WordNet definition of the current candidate, and the total number of concepts in this set is frequently over two thousand. We suggest that the time to run this algorithm can be vastly reduced while maintaining most of the accuracy by only using the concepts from a summarization of the Wikipedia text, such as the first paragraph [5]. As the first paragraph is typically an overview of the topic, most of the more relevant concepts should appear there.

However, the main difficulty of their method is hypernym determination, so additional means of hypernym extraction from text are applied and compared. These new sources of potential hypernyms improve the pool of candidates, leading to an increased likelihood that the correct hypernym can be selected. Looking for fixed patterns in the text of the article should help to identify the hypernym [6]. Furthermore, the first sentence of most articles generally indicates a good hypernym, giving it as "*[title of article]* is a *[hypernym]* which *[elaboration]*" or something similar. This pattern is generally consistent across Wikipedia, especially for larger and more popular articles [2].

2.2 Implementation

The general algorithm for extending WordNet is as follows:

1. Obtain a Wikipedia article which is not already in WordNet
2. Collect potential hypernyms and add all synsets that contain words found this way to a candidate pool
3. For each candidate synset, compute the matching score between its definition and the text of the Wikipedia article [3]
4. Insert the article into WordNet under the synset with the best matching score

This process is graphically illustrated in Figure 1.

We examine improvements that can be made in steps two and three. In step three, we look at improving the speed of the matching score computation by limiting the number of concepts that need to be compared to only those in the first paragraph of Wikipedia text. This has an interesting impact on the accuracy of the resulting hypernyms, which will be further elaborated in the Experiment section.

Furthermore, in step two, we look at ways to improve the quality of the hypernym candidates. The original method uses the categories of the article as hypernym candidates, but the categories do not always make good hypernyms. For example, that method often chooses "birth" or "death" as hypernyms for people because their article has a category such as "1963 Births." Thus, alternate sources for the hypernym must be considered. The second noun phrase in the first sentence of the article tends to be a good hypernym candidate and generally appears in a predictable location in the parse tree of that sentence, as shown in Figure 2. In this example, we would like to extract "tone poem" from the tree; to do

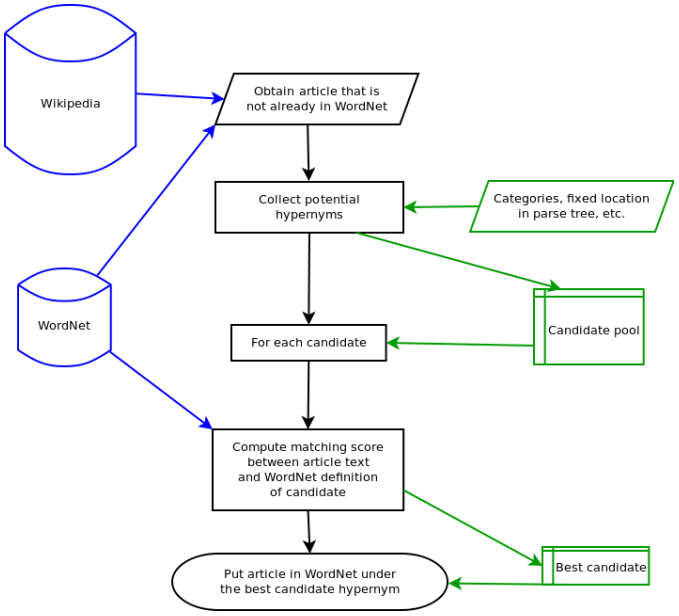

Figure 1: Flowchart for adding Wikipedia articles to WordNet.

this, we can retrieve the head of the noun phrase under the verb phrase under the root of the tree. The tree in Figure 3 illustrates this. Any article whose first sentence follows this "is a" construct will have this kind of structure in its parse tree. After retrieving this word or phrase, we take the best-matching synset that contains it as the hypernym for the article.

To extract the hypothesized hypernym from the parse tree, a specific set of rules is followed for descending the children of each node in the tree, based on the part of speech tag. Those rules are as follows:

1. Start at the S root
2. If there is an S child, select it (to select the first clause of complex sentences)
3. Select the VP child
4. Select the NP child. If successful, stop
5. If no NP child, try VP child, then NP child
6. If still unsuccessful, try S child, then VP, then NP

The goal of this procedure is to find the first noun phrase after the subject of the sentence, which is likely to be a hypernym [2]. It was found experimentally that these patterns cover the majority of sentences that we have some hope of parsing using this method. After obtaining this noun phrase, the following procedure is applied:

1. If this node has a PP child and it is the word "of," select that node
2. Select the NP child. If there is no NP child, select the original NP node

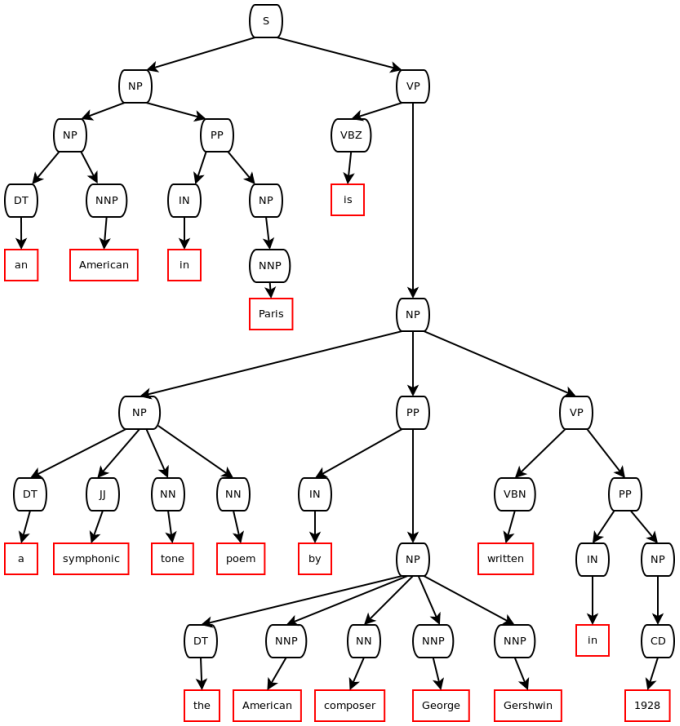

Figure 2: Tree structure of a typical first sentence in a Wikipedia article: "An American in Paris is a symphonic tone poem by the American composer George Gershwin, written in 1928."

3. If a new NP node was found this way, repeat the procedure

The goal of this step is to follow the "[article] is a type of [hypernym]" pattern, which seems to be fairly common. After performing these steps, we end up with a noun phrase node in the parse tree. To obtain the WordNet hypernym from the noun phrase, the following is performed:

1. Retrieve the head word of the noun phrase. This is the working hypernym
2. Find the word just before the head word in the original sentence and add it to the front of the working hypernym
3. If this working hypernym does not appear in WordNet, remove the newly added word and return the working hypernym. Otherwise, repeat from step 2.

This handles cases such as the "An American in Paris" example above, where "tone poem" appears in WordNet but "symphonic tone poem" does not.

Sometimes the original head term does not appear in WordNet. This is corrected for in a hybrid method we also tested, which simply places the output of the parse tree method in the pool of candidates. If the retrieved hypernym is not in WordNet, it will simply use the categories as per the Jiang et. al. method.

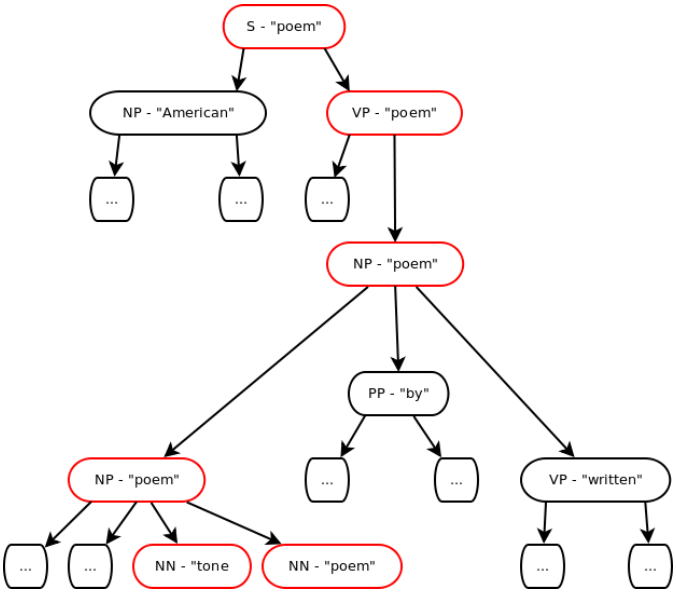

Figure 3: Relevant portion of the same tree with head words annotated and path to hypernym highlighted.

2.3 Experiment

As the method of extending WordNet with Wikipedia is be based on the work of Jiang et. al. [3], the results were tested using the same method as them. A random sample of the new additions were be taken and scored by human readers on a 5-point scale. The original results scored an average of 4.26 (Good). The three methods we tested are: a version of the original method which only uses the first paragraph of the article in concept matching; the parse tree extraction method described previously; and a hybrid method which combines the two approaches by inserting the result of the parse tree method into the candidate pool of the first method.

The version which uses only the first paragraph for concept matching proves to be much faster than using the whole text of the article as shown in table 1. Case studies showing the change in accuracy are given in Tables 2 and 3. Accuracy seems to have fallen for a few articles, but interestingly, accuracy improved for some other articles. We suggest that this improvement is caused by the fact that the body of the article may contain elaboration about a specific feature of the main topic and skews the matching computation towards that facet. This can be seen for the "America the Beautiful" article especially, which was mistakenly put under "Pikes Peak" in the old method. "Barbra Streisand songs" is not the best hypernym, but the new algorithm at least was able to put it under "song, strain" rather than "peak, crown, crest,

Table 1: Speed comparison between different amounts of article text used

Text Used	Articles Processed in Five Hours
Whole Article	7
First Paragraph	3347

Table 2: Case study of the original method of hypernym extraction

Wikipedia Article	Category	WordNet Hypernym
An American in Paris	1928 compositions	musical composition, opus, composition, piece, piece of music
Animalia (book)	Puzzle books	book, volume
Allan Dwan	1981 deaths	death
America the Beautiful	Pikes Peak	peak, crown, crest, top, tip, summit
Albert Sydney Johnston	1862 deaths	death, last

top, tip, summit," so it certainly is an improvement. Using only the first paragraph may have the effect of keeping the concepts used to compute the matching score closer to the overall topic of the article. The new method in which the hypernym is extracted from a parse of the first sentence is showing mixed results. It is able to correctly identify hypernyms that the original method do not, but it also misses many others that the original method identifies correctly. Case studies of this method are shown in Table 4.

The three methods were each run long enough to process 100,000 Wikipedia articles. Of these, a random sample of 300 results were selected from each method. Three volunteers were recruited to grade these samples, using the same scale as used by Jiang et. al. The same 300 entries were used across all three methods in order to better compare results. The results from each volunteer were then averaged to obtain the final scores for each method. The average score over the whole dataset for each method is given in Table 5, while Table 6 shows the individual score breakdown. All of the new methods have a lower average than the original (Jiang et. al.) method. The loss in accuracy for the summarized method is understandable given its dramatic speed increase, but the parse tree search method clearly is underperforming in its current state. The hybrid method improves on it slightly. One issue with the parse tree method is that the rule for following the preposition "of" down to the next noun phrase frequently caused the algorithm to go right past the real hypernym; thus, the rule is not applicable in all situations. This highlights one of the weaknesses of the fixed rule approach that is used here.

2.4 Future Work

The main improvement that could be made to these methods would be to make the parse tree search much more flexible by using machine learning to decide where in the parse tree the hypernym is most likely to be. The fixed rule approach is not nearly flexible enough to handle the large variety of sentence structures that occur in Wikipedia articles, even just in the first sentence. If a good set of features to test could be developed, it seems likely that large improvements could be made to this method.

Table 3: Case study of the faster version of the original method

Wikipedia Article	Category	WordNet Hypernym
An American in Paris	1928 compositions	constitution, composition, physical composition, makeup, make-up
Animalia (book)	Puzzle books	book, rule book
Allan Dwan	Writers from Ontario	writer
America the Beautiful	Barbra Streisand songs	song, strain
Albert Sydney Johnston	United States Military Academy alumni	alumnus, alumna, alum, graduate, grad

Table 4: Case study of the parse tree method

Wikipedia Article	WordNet Hypernym
Allan Dwan	conductor, music director, director
America the Beautiful	song, strain
Albert Sydney Johnston	career, calling, vocation
Citizen Kane	film
Commonwealth of England	republic
Groucho Marx	comedian, comic
Miss Marple	character, reference, character reference

The matching score computation also stands to be improved. In some cases, even when the correct hypernym is chosen out of the candidates, an incorrect sense of that word is chosen from WordNet because of its matching score. For example, as seen in Table 3, the algorithm correctly found "character" as the hypernym word for "Miss Marple," but chose the sense "character, reference, character reference" rather than the more correct "fictional character, fictitious character, character." This suggests that the degree of matching between the Wikipedia text and the candidate's WordNet definition may not be the best measure to use. Furthermore, the use of a true word sense disambiguation algorithm is likely to improve these results.

3. CONCLUSION

The usefulness of lexical ontologies is well documented, their main downside being only the cost and human effort required to create and update them. If such an ontology could be produced automatically, this downside would be all but eliminated. Furthermore, the expansion of user-edited sources of information on the web has greatly incentivised their use as comprehensive lexical ontologies, but so far no technique exists to effectively automatically compile the information they contain. This research has taken steps towards the

Table 5: Average scores for the three methods (and original method)

Method	Average Score
Original	4.26
Category	3.88
Parse Tree	3.42
Hybrid	3.45

Table 6: Evaluation results for the three methods (and original method)

Method	Excellent	Good	Fair	Neutral	Bad
Original	193	47	26	12	22
Category	102	100	64	29	5
Parse Tree	39	134	64	40	23
Hybrid	45	111	91	41	12

automatic extraction of word relations required to build an ontology from web sources like Wikipedia. The testing done here demonstrates a few approaches that are not quite flexible enough for this purpose but hint at improvements to be made in the future.

Acknowledgement

The research reported in this document has been funded partially by NSF grants CNS-0958576 and CNS-0851783.

4. REFERENCES

[1] C Fellbaum. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA, 1998.

[2] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics - Volume 2*, COLING '92, pages 539–545, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics.

[3] S Jiang, L Bing, B Sun, Y Zhang, and W Lam. Enhancing ontology actively and learning the concept granularity agilely: Keeping yourself current.

[4] Tomáš Kleigr, Vojtěch Svátek, Krishna Ch, Jan Nemrava, and Ebroul Izquierdo. Wikipedia as the premiere source for targeted hypernym discovery, 2010.

[5] Aleksander Kolcz, Vidya Prabakarmurthi, and Jugal Kalita. Summarization as feature selection for text categorization, 2001.

[6] Erik Tjong Kim Sang. Extracting hypernym pairs from the web. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 165–168, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[7] Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. Learning syntactic patterns for automatic hypernym discovery. In *Neural Information Processing Systems*, 2004.