Extracting Semantic Relationships
between Wikipedia Articles

Suzette Stoutenburg1,2, Jugal Kalita1, Shayn Hawthorne2
1University of Colorado at Colorado Springs College of Engineering
1420 Austin Bluffs Parkway, Colorado Springs, CO 80918
2The MITRE Corporation, 1155 Academy Park Loop, Colorado Springs, Colorado 80910
suzette@mitre.org, kalita@eas.uccs.edu, shayn@mitre.org

Abstract. Wikipedia is arguably the largest source of collaboratively developed knowledge in the world. However, this valuable set of knowledge remains largely unstructured and therefore unavailable for use by software systems. In recent years, there has been increasing research in the use of Wikipedia as a broadly applicable lexical semantic resource. There has been significant progress toward extracting semantic information from Wikipedia, but only a handful of approaches to date have focused on inferring the relationships between articles and links. We present an approach to extract the meaning of relationships between Wikipedia articles using natural language processing techniques. We use regular expressions to detect linguistic patterns and infer relationships between each linked article pair. Preliminary results are competitive for some relationships, with precision as high as 80%. This suggests that the use of regular expressions over parts of speech is a feasible approach to knowledge extraction. The extracted relationships are expressed in OWL on the Web, forming a semantic network of related concepts that capture, in part, the knowledge contained in Wikipedia.

Keywords: Artificial Intelligence, Knowledge Acquisition, Knowledge Representation, Natural Language Processing, OWL, Semantic Networks, Web Ontology Language, Wikipedia, Wiki.

1 Introduction

Wikipedia1 is a web-based encyclopedia written collaboratively by volunteers around the world. It spans multiple languages and represents one of the largest sources of freely available knowledge in the world. Through 2006, Wikipedia grew exponentially and it continues to grow at a rapid pace today2. Additionally, its model for knowledge capture, the wiki3, allows knowledge to stay current in a rapidly changing world. Case studies have compared professional thesauri with Wikipedia and have shown that the latter contains a substantial proportion of domain-specific concepts and semantic relations [1]. Therefore, there is significantly valuable

1 http://www.wikipedia.org
3 A wiki is a collaborative website which can be directly edited by anyone with access to it [Wikipedia].
information in Wikipedia that is continuously updated. Structuring Wikipedia will allow this vast knowledge base to be exploited by software.

Recently, there has been increasing scientific attention on the goal of the leveraging Wikipedia as a rich online resource of lexical and semantic knowledge. Approaches are emerging very rapidly to generate structured knowledge from Wikipedia. In this paper, we present an approach to construct a network of Wikipedia articles using specified links and we infer relationships between the articles. To do this, we apply natural language processing techniques and regular expressions to detect linguistic patterns. The patterns are used to derive semantic relationships between Wikipedia articles, with the goal of deriving a semantic network of article relationships. For example, given the following set of sentences, we would ideally seek to derive the relationships shown in the notional example in Figure 1. Links are underlined.

“An *automobile* is a *wheeled passenger vehicle* that carries its own *motor*. Most definitions of the term specify that *automobiles* are *constructed for the transport of people*...”

To present our results, we first discuss related work and how our research complements the state of the art. Next, we discuss the structure of Wikipedia in Section 3 and the technical approach in Section 4. We discuss results in Section 5 and conclude with a summary and identification of next steps in Section 6.

## 2 Related Work

The foremost initiative to extract knowledge in Wikipedia is DBpedia [2], [3] by Auer et al. DBpedia is an initiative to extract information from Wikipedia and make the information available on the web using the Resource Description Framework (RDF) [4]. One of the core components of DBpedia is a method to extract knowledge from Infobox templates used within articles. The information extracted is represented in RDF triples that can be downloaded from the DBpedia site. However, DBpedia does not infer relationships between articles that are linked, only that the link exists.

A number of approaches build on the information stores established by DBpedia.

![Notional network of semantic relationships in Wikipedia.](http://dbpedia.org/)
For example, DBpedia Relationship Finder uses the RDF stores in DBpedia to answer queries on the relationships between entities [5]. The interface is not yet machine-to-machine but the interface provides useful relationships to the human user. YAGO [6] extends WordNet using information extracted from Wikipedia category tags. Pei et al. developed a Wikipedia Thesaurus [7], a large scale thesaurus extracted from Wikipedia through exploitation of unique URLs, link structure and link text. From this information, they constructed a global ontology to assist in ontology mapping [7].

Volkel et al. have proposed an extension to Wikipedia, known as “Semantic Wikipedia”, designed to explicitly express the relationships between links [8]. This approach offers a rich extension to Wikipedia but still relies on manual entry. Semantic Media Wiki is a follow-on initiative that allows addition of structured information to Wikis using RDF [9]. In addition, SemanticHacker.com is home of the TextWise SemanticHacker API, advertised as the world's first open API for semantic discovery. This tool tags text with semantics and provides links to related Wikipedia articles.

Earlier approaches extract information from Wikipedia by harnessing implicit semantics in the syntax. Gregorowicz and Kramer built a concept extractor that extracts structure from Wikipedia [10]. They use properties from the Simple Knowledge Organization System (SKOS) to represent relations between components of Wikipedia. Others have used Wikipedia as a resource for natural language processing applications such as question answering [11], text classification [12] and named entity disambiguation [13]. Herbelot and Copestake extracted hyponymic relations and applied a filter to obtain taxonomic relationships in the domain of biology [14]. Strube and Ponzetto used Wikipedia to compute measures of semantic relatedness, which they showed to be as accurate as those from WordNet when applied to co-reference resolution [15]. Suchanek et al. [16] demonstrated strong results with the LEILA system, which applied advanced natural language techniques, such as deep analysis and anaphora resolution, and combined them with machine learning techniques. Other approaches used natural language processing techniques to extract knowledge from the lexical and syntactic structures of Wikipedia [1], [14], [10], [17].

Only recently have researchers focused on extracting the relationships between Wikipedia articles. The work that is most closely related to ours is by Nguyen et al. [18] and the KYLIN system by Wu and Weld [19]. Both utilize semi-structured information in Wikipedia Infoboxes to extract relationships from unstructured text using unsupervised approaches. The f-score for Nguyen’s approach was reported between 32.84% and 62.90% [18]. KYLIN had high accuracy, with precision ranges of 73.9% to 97.3% and recall ranging from 60.5% to 95.9% [19]. Nguyen’s application limits relationships acquired, while our work extracts generic relationships, expressed in the original sentences themselves, as does KYLIN. We do

---

5 http://www.w3.org/2004/02/skos/
6 An infobox on Wikipedia is a consistently-formatted table which is present in articles with a common subject to provide summary information, or navigation to closely related articles in that subject. [From http://en.wikipedia.org/wiki/Help:Infobox]
7 F-score is an accuracy metric that combines Recall and Precision into a single score. Also referred to as F-measure.
limit extraction to relationships between article names, as does Nguyen, while KYLIN extracts relationships between any significant entities in the article. However, both Nguyen’s approach and KYLIN apply syntactic and semantic patterns discovered in Wikipedia Infoboxes and article content while our work employs regular expressions to detect linguistic patterns over the natural language of the articles. We believe that the use of regular expressions offers the advantage of a simple approach to extracting complex relationships in natural language text.

3 About Wikipedia

Wikipedia consists of the following primary components, each of which is described below. These components and the implicit relationships among them are shown in Figure 2.

- Article
- Article redirects
- Article links
- Categories
- Disambiguation pages

An article is the primary construct within Wikipedia, containing detail on the term to be described. The first paragraph usually provides a brief definition of the term, while the full article contains more details, history, relevant links to other information sources, etc. Articles in Wikipedia contain mainly nouns, with only a few verbs and adjectives [20].

To ensure that there is only a single article per term, Wikipedia provides article redirects that link equivalent terms to the main article [1]. Article redirects handle various forms of the term, such as spelling variations, capitalization, abbreviations, synonyms, etc. These redirects provide an excellent source of semantic information, as they express synonymy; they have also been shown to improve named entity recognition [20].

Article links direct the user from one article to another in Wikipedia. Therefore, these links represent relationships, but as yet, these relationships are not typed and the strength of the relationships is not specified. Note that the link labels are not necessarily the same as the title of the linked article; therefore, a link label is another source of synonyms, related terms and spelling variations [20]. For example, the

Fig. 2. Wikipedia components and their relationships.
Fig. 3. Wikipedia articles are linked with tags of different names.

Wikipedia articles Automobile and Internal Combustion Engine are linked by the tag “motor” as shown in Figure 3. Wikipedia also offers a category system [21] that was initially intended as metadata for articles. Tags can be associated to each article collaboratively by authors. Categories can also be used to locate articles that relate to them. Therefore, the category system essentially forms a thesaurus [20]. There is also a hierarchy associated with the category system that provides hyponyms and meronyms for terms [21]. Disambiguation pages provide a set of links to articles on each sense of the term, thus providing a potential rich source for word sense disambiguation.

APIs to access Wikipedia data are starting to emerge. Attempts by Riddle⁸, Strube and Ponzetto [15], and Summers⁹ perform poorly and provide access only to some types of the available use the data extraction capabilities offered by Wikipedia and operate over the data in a local, disconnected fashion [22]. An overview of existing access mechanisms is provided in [21]. There was an early API set proposed in 2007 [20] that offers access to the components of Wikipedia. This is the API we chose to use in this work.

4 Technical Approach

We have developed an approach to automatically extract relationships between Wikipedia articles by processing the natural language of sentences that contain links within an article. The overall application, called WikiSemNet, is shown in Figure 4. In this application, substantial pre-processing is performed, including cleansing and normalizing of the raw data and tokenization of words and sentences. Once tokenized, each word is tagged with part of speech and article names and links to other articles are identified. Next, regular expressions are applied to detect linguistic patterns that suggest the semantics of relationships between Wikipedia articles. The regular expressions are applied over parts of speech, combined with article names and other key words. The resulting relationships between articles form a semantic network that reflects a part of the knowledge captured in Wikipedia. The knowledge extracted is represented using the Web Ontology Language (OWL) [23] for use by software applications on the World Wide Web.

In this section, we discuss each of the processing steps necessary to acquire relationships between Wikipedia articles.

http://search.cpan.org/~triddle/Parse-MediaWikiDump-0.40
http://search.cpan.org/~esummers/WWW-Wikipedia-1.9
4.1 Pre-processing

Significant pre-processing is required in order to apply patterns over the content of articles. To accomplish this, we reuse existing tools wherever possible. First, to access Wikipedia data, we used an emerging API called Java Wikipedia Library (JWPL) [22]. Since real-time access to Wikipedia is discouraged, this API operates on a static set of Wikipedia data. We downloaded the 6 February 2007 version into a local database and implemented an interface using the JWPL library. Second, natural language processing in our solution is based heavily on a tool set called LingPipe

10
http://www.alias-i.com/lingpipe/

11
http://www.alias-i.com/

12
MedPost, Natural Language Processing Resources for Medical Literature,

13
They selected the MedPost12 corpora for part-of-speech tagging [24].

We built software to identify the article name and its references in each sentence. Therefore, we built patterns to detect the article name or any combination of its subcomponents. For example, we found that many times, an article title may contain 3, 4 or more words, but only a subset of those words are used in the article to reference the title. Therefore it was essential to identify the principle entity and its references throughout the text. We also built code to detect the use of pronouns as they relate to the main entity in the article.

In the final pre-processing step, a “tag string” is created dynamically for each sentence and used to match against each regular expression. The tag string is constructed using the part-of-speech tags for the sentence in conjunction with other dynamic information. For example, the article name is dynamically inserted into the tag string depending on its position in the original sentence. Also, in some cases, dynamic information is inserted into the tag string. For example, if the word “born”
appears in the original sentence, it is re-inserted into the tag string, replacing its part of speech. Thus, the tag string is composed of dynamic information along with part-of-speech tags. It is also important to preserve the position of links in the tag string. Therefore, we used brackets to indicate the position of links. We carefully select characters for use in the tag string so as not to conflict with special characters used in regular expressions.

Once constructed, the tag string for each sentence is compared against each set of regular expressions that are designed to detect relationships of interest. For example, consider the following sentence.

"An automobile is a wheeled passenger vehicle."

This sentence is assigned the following tag string. As defined in [24], DD represents an article, VBZ represents is, VVD is an adjective and NN represents nouns.

DD Automobile VBZ DD [[VVD]] [[NN]] [[NN]].

4.2 Relationship Extraction

To automatically extract relationships between Wikipedia articles, we built regular expressions to detect linguistic patterns in sentences in Wikipedia articles that contain links. To analyze the patterns, we developed utilities to extract random articles and associated sentences with links. We studied the sentences and identified a number of patterns to detect semantic relationships between articles and links. Patterns are based on the following key components.

- Article name and position in sentence
- Link name and position in sentence
- Parts of speech of each word in sentence
- Sequence of parts of speech in sentence
- Position of key words in sentence

Figure 5 shows an example. In this case, automobile is the name of the article and this sample sentence contains three links, wheel, passenger and vehicle. We used the parts of speech and positions of articles and links to detect an isA relationship between automobile and vehicle. In this initial effort, we focused on detection of directed relations from article to link; so, the triple formed upon relationship detection is: (article, relation, link). We developed patterns that focus on detection of the following relationships.

- isA
- part of
- locatedIn
- bornOnDate
- diedOnDate
- bornIn

Note that we designed patterns to detect all tenses of the above relations.

In addition to these relationships, we built patterns to extract generic relationships from the original sentences to be used in the derived semantic relationship. For example, from the following sentence fragment,
Fig. 5. Example of derived relationship between article and link using patterns. The most important components in a pattern are the position and part of speech of the article and the links, and the order of parts of speech in the sentence.

"An automobile is a ... vehicle that carries its own motor"

we built a pattern to extract the relationship (Automobile carries motor) expressed in OWL as follows.

```
<owl:ObjectProperty rdf:ID="carries">
  <rdfs:domain rdf:resource="#automobile" />
  <rdfs:range rdf:resource="#motor" />
</owl:ObjectProperty>
```

The same type of approach was used to build relationships using prepositions from the original sentences. These relations are referred to as Generic relation patterns.

To explain the process of pattern detection, we use a simplified example. Consider once again the sentence:

"An automobile is a wheeled passenger vehicle."

This sentence is tagged with part-of-speech tags as follows; double brackets are used to denote links to other Wikipedia articles.

```
DD NN VBZ DD [[VVD]] [[NN]] [[NN]].
```

To detect linguistic patterns in tagged sentences, we define a set of regular expressions. We use part-of-speech patterns along with the position of the article and links, along with key word detection in some cases. To ensure proper detection of the article and its position in the sentence, the article name is inserted into the regular expression dynamically, depending on the relationship that is to be extracted. Link positions are detected by including the link delimiters expressed in Wikipedia directly in the pattern. For example, one of the isA patterns is designed as follows.

```
articleName.*VBZ DD.*[[NN]]+
```

VBZ is the tag for “3rd person singular be, is” and matches the set \{is\}; DD is a determiner tag which matches the set \{a, an, the\}. NN matches any noun that is not a proper noun. (Proper nouns are denoted by NNP. More details on MedPost tags are provided in the appendix.) Therefore, this pattern will match any sentence that includes the article name, followed by any text (denoted by .*), followed by one
member of the set \{ is a, is the, is an\}, followed by any text, followed by one or more
nouns that appear within a link.

Large sets of regular expressions were implemented and tested over small sets of
data then applied over random sets of data in Wikipedia. Some sample regular
expressions are shown in Table 1.

4.3 Experimental Approach

To test the accuracy of our approach, we selected over 600 articles at random and
applied patterns to the text in each random article. We used sampling and semi-
automated annotation to measure precision and recall. For example, to speed the
manual task of generating expected results, utilities were developed to build a file in
which a unique entry for each article-link pair was created for each relationship, along
with a value of “false” indicating that the relationship does not exist. Then, manual
annotation consisted of simply updating only those entries that existed with a value of
“true.” Then, automated techniques were applied to compare experimental results
with “ground truth” values, and precision and recall automatically computed.

5 Results and Discussion

As shown in Table 2, initial results are promising. In particular, patterns to detect
generic relationships proved to be quite powerful, with precision and recall well over
80%. These relationships also provide the most general extraction capability since
relationships are extracted from the natural language text. As can be seen in Table 1,
some of the regular expressions to detect generic relationships became quite complex,
though the overall approach remains quite simple.

The relation isA performed not as well but still with precision of 62.89% and fairly
high recall. The isA relation can be difficult to detect given that there are many
linguistic patterns that express this relation. This is one area that might benefit from
post-processing using additional techniques.

Table 1. Sample Regular Expressions for Relationship Extraction†

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Regular Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generic</td>
<td>.<em>PN&gt;({1,2})((VV&gt;({1,3}),{II}</em>{0,4},(JJ)<em>&lt;NN&gt;({1,2})</em>:0,3)</td>
</tr>
<tr>
<td>isA</td>
<td>PN.<em>VBZ DD.</em>[^II] [^II DD] (&lt; &lt; NN &gt;&gt;&lt;&gt;&lt;)</td>
</tr>
<tr>
<td>partOf</td>
<td>articleName.<em>part II (DD )</em>&lt; &lt;,*&gt; &gt;</td>
</tr>
<tr>
<td>bornOnDate</td>
<td>articleName.*born &lt; &lt; MC &gt; &gt;</td>
</tr>
<tr>
<td>diedOnDate</td>
<td>articleName.*died &lt; &lt; NN NN &gt; &lt; &lt; MC &gt; &gt;</td>
</tr>
<tr>
<td>bornIn</td>
<td>*born.*II &lt; &lt; NN &gt; &gt;</td>
</tr>
<tr>
<td>locatedIn</td>
<td>articleName.*II &lt; &lt; (NN,1,2)&gt; &gt;</td>
</tr>
</tbody>
</table>

† There were dozens of patterns implemented. We only provide a small subset of examples here.
†† Generic relation patterns are designed to extract verbs and prepositions from sentences and used to
express the relationships between articles.
Patterns to detect part of relationships also performed quite well. This was interesting because we used only very simple patterns to detect the relation. For example, we built patterns to detect key words that indicate meronymy, such as part, component, comprise, etc. The strong initial results suggest that application of more complex meronymy detection methods such as those described in [28] would yield even better results.

The patterns to extract information on dates, such as bornOnDate and diedOnDate proved more difficult than expected. While the patterns are fairly simple at first glance, dates are used for many purposes in an article. We were able to constrain the patterns for higher precision but at the cost of recall. We settled on a set of patterns that resulted in an optimal balance between precision and recall.

The patterns to extract bornIn fared well, with precision just under 70% and recall fairly high. The patterns to extract locatedIn relationship performed fairly well, resulting in a set of useful information, with precision and recall over 70%.

To extract useful relationships from Wikipedia, it seems that high recall is preferred at the cost of precision. With a large set of potential relationships returned, post-processing or manual annotation could be applied to finalize the extraction process. In addition, we support the idea of applying a software bot to improve Wikipedia, as suggested by [19]. This type of approach would improve much of the inconsistencies often found in Wikipedia articles.

We found that introducing a process for semi-automating the annotation process was highly valuable, allowing us to increase the size and number of data sets used to test this work. While our overall approach is fairly simple, we found that adjusting the regular expressions for maximum precision and recall can be tedious. Automating the training process would be highly beneficial to extend this work.

### Table 2. Experimental Results

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Precision %</th>
<th>Recall %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generic relation</td>
<td>83.08</td>
<td>87.23</td>
</tr>
<tr>
<td>isA</td>
<td>62.89</td>
<td>78.21</td>
</tr>
<tr>
<td>partOf</td>
<td>70.83</td>
<td>73.91</td>
</tr>
<tr>
<td>bornOnDate</td>
<td>66.67</td>
<td>74.42</td>
</tr>
<tr>
<td>diedOnDate</td>
<td>68.42</td>
<td>81.25</td>
</tr>
<tr>
<td>bornIn</td>
<td>68.75</td>
<td>88.00</td>
</tr>
<tr>
<td>locatedIn</td>
<td>74.58</td>
<td>72.13</td>
</tr>
</tbody>
</table>

5 Conclusions and Next Steps

In this paper, we have shown that it is feasible to apply regular expressions in order to extract semantically rich relationships between Wikipedia articles. Our approach is simple and yet yields robust results, particular for generic relationships, the most powerful of the relationships extracted. Our work is competitive with other emerging approaches and yet offers the simplicity of regular expressions. We also successfully built patterns to extract other common relationships such as isA and partOf.
In the future, we plan to further test and improve the regular expressions by adding a machine learning component. We would like to investigate automatic collection of training data, similar to the approach used in [18]. We would like to improve the patterns to detect meronymy using more of the concepts described in [28]. We would also like to explore adding new pattern types to extract additional knowledge from Wikipedia. In particular, it would be interesting to build patterns to extract biological information from Wikipedia.

References

10. Gregorowicz A., Kramer, M. A.: Mining a Large-Scale Term-Concept Network from Wikipedia [http://www.mitre.org/work/tech_papers/tech_papers_06/06_1028/06_1028.pdf](http://www.mitre.org/work/tech_papers/tech_papers_06/06_1028/06_1028.pdf)