From Wikipedia to Semantic Relationships: a Semi-automated Annotation Approach*

Maria Ruiz-Casado, Enrique Alfonseca and Pablo Castells

Computer Science Department, Universidad Autonoma de Madrid, 28049 Madrid, Spain
{Maria.Ruiz,Enrique.Alfonseca,Pablo.Castells}@uam.es

Abstract. In this paper, an experiment is presented for the automatic annotation of several semantic relationships in the Wikipedia, a collaborative on-line encyclopedia. The procedure is based on a methodology for the automatic discovery and generalisation of lexical patterns that allows the recognition of relationships among concepts. This methodology requires as information source any written, general-domain corpora and applies natural language processing techniques to extract the relationships from the textual corpora. It has been tested with eight different relations from the Wikipedia corpus.

1 Introduction

Wikis are environments where uploading content to the web is extremely simple and does not require the users to have a technical background. Although wikis can be intended for individual use in applications such as Personal Information Management [1], this emergent area is producing many popular collaborative environments where web users are able to share and contrast their knowledge about a certain topic. Wikis have many applications, such as building collaboratively on-line information sites (e.g. dictionaries or encyclopedias) or for coordinating and exchanging information in project management or corporate intranets [2].

The success of many public wikis is due to the interest they have arisen among potential contributors, who are eager to participate due to their particular involvement in the domain under discussion. Dating from 1995, Wiki Wiki Web[1] is dedicated to software development. ProductWiki[2] is an on-line product catalogue. Wikitravel[3] is a world wide travel guide. Other specific domains can be found for instance at Comixpedia[4], a wiki for comics, or Wookipedia[5], that collects information about the Star Wars saga. Some general-domain wiki portals are the on-line encyclopedia Wikipedia[6], the dictionary Wiktionary[7] or the questions-answers portal Answerwiki[8].

* This work has been sponsored by MEC, project number TIN2005-06885

[1] http://c2.com

[2] http://productwiki.com

[3] http://wikitravel.org

[4] http://comixpedia.org

[5] http://starwars.wikicities.com

[6] http://wikipedia.org

[7] http://www.wiktionary.org/

[8] http://answerwiki.com

Its success is mainly due to several factors: the easiness to publish and review content on the web through the web browser, usually avoiding barriers such as logins or technical knowledge; the few restrictions on what a contributor can write; and the user-friendly interface which provides flexible mechanisms to get the content published. The philosophy behind wikis is to sum up the limited effort of each contributor to produce a repository of knowledge, the size of which depends on the number of contributors and the amount of information they are willing to provide. The Wiki platforms take advantage of the synergy of the aggregated work of their individual contributors.

Given that there is usually no authorship (and no responsibilities), concerns have been raised about the quality that can be attained in public wikis. A recent claim that the quality of scientific articles in Wikipedia is equivalent to the quality of the Encyclopaedia Britannica [3], has been reported to be flawed [4]. Anyway, wikis have shown that it is possible to write collaboratively very useful content, available to everyone with little personal cost. This characteristic is specially attractive for the Semantic Web field, where the need to place semantic annotations in existing and upcoming web content constitutes a costly hindrance.

1.1 Blending Semantic Web and Wikis

Most of the world wide web content is written in natural language and is intended for human readers. Due to the vast amount of information contained in the web, nowadays many tasks need a certain degree of automation. When trying to search and process web content automatically, a machine has to cope with language ambiguities and implicit knowledge that it can hardly process. The Semantic Web constitutes an initiative to extend the web with machine readable content and automated services far beyond the original capabilities of the World Wide Web [5], primarily making the web content explicit through semantic annotations. This would allow an easy automated processing when searching and retrieving information. Annotation standards have been developed by the Semantic Web community giving way to the RDF[9] and OWL[10] annotation languages, among other. The annotations refer to ontologies, a knowledge representation formalism that is used to model the underlying knowledge.

But placing semantic tags in the huge amount of existing and upcoming content can also be too costly if it has to be done manually by specialised ontologists. In 2004, the size of the Semantic Web, i.e. the percentage of web pages containing metadata, was estimated to be less than 5% of the total size of the web [6]. Therefore, as some authors have recently pointed out [2, 7, 8], initiatives like Wikis, that allow easy and collaborative information exchange, can be of great help in making the Semantic Web successful. Wiki communities have proved to succeed in collaboratively producing at low cost vast information repositories. The addition of semantic annotations to documents can be achieved following the Wiki philosophy of lowering the technical barriers: the semantic annotations are presented to the user as simply assigning a type to the hyperlinks, and providing internal facilities to transform these annotations into RDF or other Semantic Web annotation language. On the other hand, semantic annotation allows that

[9] http://www.w3.org/RDF/

[10] http://www.w3.org/2001/sw/WebOnt/

Wikis benefit from the automation of management, searching and retrieval pursued by the Semantic Web.

This blend between the Semantic Web and wiki fields is called *Semantic Wikis*. Some projects developing semantic wikis are IkeWiki[11] and Semantic Wikipedia[12].

1.2 Adding ingredients to the blend: Information Extraction

One step further in the creation of semantic annotations through a wiki environment is pointed out by some authors [7] as the possibility of assisting the user in the editing process to facilitate the selection of the typed links. For instance, the authoring environment may generate candidates for link types automatically, by extracting semantic information from the current Wikipedia corpus.

There is a large amount of information already present in the existing Wikipediae: as for March 2006, the English Wikipedia has more than one million articles. German, Spanish, French, Italian, Japanese, Dutch, Polish, Portuguese and Swedish Wikipediae are above one hundred thousand articles each. Using semi-automatic annotation procedures, it should be possible, in a short amount of time, to offer a fully enriched Semantic Wikipedia, which already has a large set of users willing to collaborate. A large-scale annotation may be the trigger to induce many people to switch from a traditional Wikipedia to a semantically annotated one. Therefore, the Semantic Wikipedia may be accelerated in its way to maturity.

The approach that we present in this paper addresses that point. We depart from our previous work in the extraction of semantic relationships to annotate the Wikipedia [9, 10], consisting in disambiguating Wikipedia encyclopedic entries with respect to WordNet, and determining four basic relationships amongst the concepts: hyponymy (*is-a*) and its inverse hyperonymy, and holonymy (*has-part*) and its inverse meronymy. In this paper, we extend the previous work to propose a methodology to automatically extract any other kind of relationship, and describe the results obtained when it is applied to the Wikipedia corpus. This procedure can be integrated in a tool that makes suggestions to users concerning where to place tags in Semantic Wiki documents.

We would like to point out that any automatic tagging procedure will produce some amount of mistakes. However, the work in correcting a few errors will always be smaller than the amount of work needed to create all the semantic tags from scratch. In the same way that current Wikipedia users, when they identify a mistake in an article, are quite content to edit the article and correct it, the users of a semantic Wikipedia could do the same. Rather than departing from documents containing plain text and a few examples of relations, they would start with the full text of the Wikipedia, containing many semantic relationships, and, while reading the articles, they would just have to correct the wrong labels as they find them.

This paper is structured as follows: Section 2 introduces our approach to the automated extraction of patterns from textual sources; Sections 3 details the approach followed to identify new relations; Section 4 describes the experiment conducted with

[11] http://ikewiki.salzburgresearch.at/
[12] http://wiki.ontoworld.org

Wikipedia and the results obtained; finally, Section 6 concludes and points out open lines for future work.

2 Extracting Relationships from text: our approach

Our goal is to find semantic relationships and properties (attributes) in free text automatically. To do so, we have developed an approach based in automatic learning of lexico-syntactic patterns. The procedure starts with a seed list containing pairs of related items. It may be a list containing writers and some of their works; painters and their pictures; or soccer players and the clubs in which they have played. Next, many sentences containing the pairs of terms from the seed list are collected automatically from the web, and they are processed with the following Natural Language Processing (NLP) tools:

- Segmentation (tokeniser and sentence splitter).
- Part-of-speech tagging, i.e. identifying which words are nouns, verbs, adjectives, etc.
- Stemming, to obtain the canonical form of all the words.
- Named Entity Recognition, to identify dates, numbers, people, locations and organisations.
- Chunker (partial syntactic analyser).

The information obtained from the processed sentences can be used to study which words, syntax and entities are typically used in a human language when a particular relation between two concepts is expressed. To do so, we search in the context that surrounds the two concepts in order to find repetitive lexical patterns that appear with the concepts when the relation is present. A pattern models a possible way to convey a semantic relation in natural language, and can be applied to search and extract new pairs of concepts between which the same relation holds.

Figure 1 shows an overview of the process carried out for each type of relationship. The final version of our system is intended to take an XML dump of the English wikipedia and to produce an equivalent file with the semantic annotations added. In this way, the result can be seen directly on a web browser using standard Semantic Wikipedia software.

The following subsections elaborate on the separate steps. A complete technical exposition of the procedure can be found in [10, 11].

2.1 Collecting contextual patterns from the web

Given a pair of related terms appearing in a text, the context of this pair is the text fragment that encloses them. The context boundaries are sometimes expressed as a window of N words to the left and to the right of those two terms, or as a syntactic constituent (e.g. a sentence) containing them both.

For the task that we address in this paper, extracting semantic relationships between words, the context can be very useful. For example, in (1) it can be seen that *tail* is a part of *dog*, because of the possessive pronoun *its*. In the context, the possessive pronoun

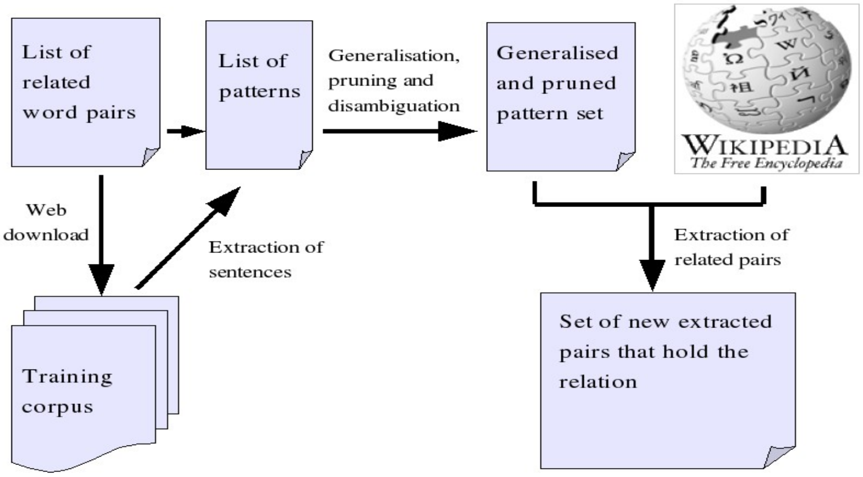

List of related word pairs
List of patterns
Generalisation, pruning and disambiguation
Generalised and pruned pattern set
WIKIPEDIA
The Free Encyclopedia
Web download
Extraction of sentences
Extraction of related pairs
Training corpus
Set of new extracted pairs that hold the relation

Fig. 1. Overview of the procedure.

indicates the existence of a holonymy (part-of) relationship. Also, the verb *composed* in Sentence (2) indicates a relationship *composer-song* between *John Lennon* and *Imagine*.

(1) The happy dog wagged its tail.

(2) The piano on which John Lennon composed “Imagine” is on its way back to the Beatles Story in Liverpool.

In some fields, such as Word-Sense Disambiguation and Information Retrieval, these contexts are usually characterised using a *bag-of-words* approach, where all the words in the context are put together regardless of their relative ordering or syntactic dependencies. In our approach, as we are interested in finding patterns of words that model a relationship, we keep the relative ordering of the words.

A relationship is formed by a triple: the two concepts related and the relationship itself. When the relation is modelled as a pattern, the two concepts participating are usually called *hook* and *target* [12]. So, in the previous example, the relation *composer-song* can be modelled with the pattern (3).

(3) The piano on which *hook* composed *target* is on its way back to the Beatles Story in Liverpool.

As we mentioned before, we start with a seed list containing related pairs. The patterns can be collected, for each relationship, using this list: for each pair of terms,

1. Perform a search on the Internet containing the two terms in the pair.
2. Process with the NLP tools all the sentences that contain the two terms.
3. Substitute the first term by *hook*, and the second term by *target*.

2.2 Pattern Generalisation

The patterns of words with which natural languages can express semantic relationships are usually manifold. Therefore, for each kind of relationship, we need to capture all this lexical variability. The purpose is to group similar patterns to obtain one that is general enough to match different contexts amongst which there are only small differences in paraphrasing. Given that many of the patterns collected in the previous step will have shared parts, that information can be used to guide the generalisation.

The core idea is the following: to generalise two original patterns, the common parts are maintained, and the differences are substituted either by disjunctions or wildcards. For instance, from Sentences (4a,b), the patterns (5a,b) are extracted, and the generalisation (6) can be obtained. The star is wildcard (representing any word), and the vertical bar | means a disjunction, meaning that any of the two adjectives can be found before the target.

(4) a. Alfred Hitchcock directed the famous film Psycho
b. Alfred Hitchcock directed the well known film Psycho

(5) a. *hook* directed the famous film *target*
b. *hook* directed the well known film *target*

(6) *hook* directed the * famous|known film *target*

This pattern can detect the relation director-film and determine the participating concepts in many sentences, e.g. (7a,b,c).

(7) a. Alfred Hitchcock directed the famous film The Birds
b. Bernardo Bertolucci directed the well known film The Last Emperor
c. Woody Allen directed the amusing and famous film Annie Hall

Note that in this example above, pattern (6), has been obtained from two patterns, one containing the word *famous*, and the other containing the words *well known*, but it is not the only possible generalisation. In total, using disjunctions and wildcards, from patterns (5a,b) the following generalisations are plausible:

- Substituting them all with the wildcard, *.
  (8) *hook* directed the * film *target*
- Creating a disjunction *famous|well* followed by the wildcard *.
  (9) *hook* directed the famous|well * film *target*
- Creating the disjunction *famous|known*, preceded by the wildcard *.
  (10) *hook* directed the * famous|known film *target*

We believe that the third one is the most suited to this task, as the terms in the disjunction both undertake the same role in the sentence, as modifiers of the target, while *well* is an adverb modifying the adjective *known*. This can be partly identified automatically if the generalisation procedure takes into account the part-of-speech (PoS) of the words. In this way, during the generalisation step, we only allow for disjunctions of words that share the same part-of-speech.

2.3 Patterns Pruning

In the previous subsection, it is still not clear when to stop generalising. For instance, given that the three sentences in (11) do not have any word in common, their generalisation would be pattern (12), which is clearly unusable because it matches everywhere.

(11) a. *hook* directed the famous film *target*.
b. *hook* won an Oscar with his film *target*.
c. *hook* 's movie *target*.

(12) *hook * target*

In general, very specific and long patterns tend to be very accurate when extracting new relations from text, as they impose many restrictions to the contexts they match. But they are applicable in few cases and hardly extract new relationships. On the contrary, short and very general patterns have a good recall because they match frequently in the texts, but their precision may be low, extracting wrong relationships. It is desirable, therefore, to evaluate the set of patterns as they are generalised, in order to stop as soon as the result of a generalisation produces too many results with a low precision.

A possible way to calculate the precision of the generalised patterns is the one described in [12]. In this approach, after all the patterns have been collected and all their possible generalisations calculated, a development corpus is collected again from the web, searching for pages that contain the hooks from the seed list. The precision of each pattern is estimated as the percentage of times that it extracts, from the development corpus, a pair existing in the seed list.

In our approach [11], we used an improved evaluation methodology that mainly consists in testing the patterns, not only for a development corpus collected with their original hooks, but also for the development corpora collected with the hooks from different relationships, which altogether constitutes what we call the *development super-corpus*. The performance of each individual pattern for each particular relation is tested in the super-corpus.

Based on the scores calculated through this automatic evaluation, the set of patterns can be refined discarding those that do not reach a predefined threshold in precision, or that are not able to match in the super-corpus a minimum number of times.

2.4 Pattern Disambiguation

Up to this point, the patterns have been extracted and generalised separately for each relationship. Hence, we have some patterns that, supposedly, appear between writers and their works, or between people and their birth place. The problem is that it may be the case that the same pattern is extracted for different relations. This is specially relevant in the case of a few short patterns such as the genitive construction (13). As can be seen in Sentences (14a-e), it can be used to convey semantic relationships as dissimilar as *scientist-theory*, *painter-painting*, *writer-work*, *architect-work* or *location-sublocation*, among many others.

(13) *hook* 's *target*

(14) a. Einstein's Theory of General Relativity
b. Bosco's The Garden of Delights
c. Tolkien's Lord of the Rings
d. Gaudi's Sagrada Familia
e. Barcelona's Sagrada Familia

In order to overcome this difficulty, we propose two solutions:

- To take into account the Named Entity (NE) tag of the hook and the target, whenever it has been annotated by the NE recogniser during the NLP processing. In this way, people, locations, organisations and dates would be marked as such in the examples above. If we annotate that a pattern is only applicable if the *hook* is of type *location*, then it will only apply to Sentence (14e) and not to Sentences (14a-d). This can be used to somewhat mitigate the problem.
  Though not completely error-free, patterns including NER are more expressive and present a better ability to differentiate relationships.
- Even so, it may be the case that a pattern still appears in the lists of patterns for different relationships. Currently, many of these patterns are ruled out in the pruning step, because we are taking into consideration the test corpora from all the relationships when we build the *development super-corpus*. Hence, if pattern (13), extracted for the relationship *scientist-theory*, is applied to the test corpus collected for *author-work*, it will erroneously mistag all the book as scientific theories, which will be detected because they do not appear in the seed list for theories, and the precision of the rule will be low.

3 Extraction Procedure

Once we have a set of patterns for each semantic relationship, the extraction procedure is simple. Each pattern will contain:

- A flag indicating whether the hook appears before the target, or vice-versa.
- A left-hand part, consisting of the words at the left of the hook or the target, whichever appears first.
- A middle part, with the words between the hook and the target.
- A right-hand part.

Given a set of patterns for a particular relation, the procedure to obtain new related pairs is as follows:

1. Download the corpus that should be annotated from the web.
2. Clean the HTML code, remove menus and images
3. Process the textual corpus with NLP tools:
   (a) Tokenise the text and split sentences.
   (b) Apply a part-of-speech tagger.
   (c) Stem nouns and verbs.
   (d) Identify Named Entities.
   (e) Chunk Noun Phrases.

4. For every pattern,
    (a) For each sentence in the corpus,
        i. Look for the left-hand-side context of the pattern in the sentence.
        ii. Look for the middle context.
        iii. Look for the right-hand-side context.
        iv. Extract the words in between, and check that either the sequence of PoS tags or the entity type are correct. If so, output the relationship.

The above procedure is repeated for all the relations considered.

4 Experiment and results

4.1 Experimental settings

The experiment carried out consists in extracting several relationships and properties from a test corpus downloaded from the English Wikipedia. The relations and properties considered are:

- Person's birth year
- Person's death year
- Person-birth place
- Actor-film
- Writer-book
- Football player-team
- Country-chief of state
- Country-capital

In order to collect the corpus of the Wikipedia, to ensure that many entries contain the indicated relationships, we have performed a recursive web download starting from the following entries:

- *Prime Minister*, that contains hyperlinks to Prime Ministers from many countries.
- *Lists of authors*, that contains hyperlinks to several lists of writers according to various organising criteria.
- *Lists of actors*, that contains hyperlinks to several lists of actors.
- *List of football (soccer) players*, containing hyperlinks to many entries about players.
- *List of national capitals*, containing the names of national capitals from countries in the world.

From those initial pages, all the hyperlinks have been followed up to a depth of 3–4 hyperlinks, having collected in total 20,075 encyclopedia entries totalling roughly 460 Megabytes after cleaning the HTML files. The NLP toolkit used to process them was the Wraetlic tools v. 2.0[13].

The pruned patterns for the mentioned relations have been produced using the procedure described above. For this experiment, only those patterns showing a precision

[13] Available at http://www.eps.uam.es/∼ealfon/eng/research/wraetlic.html

<table>
<tr><th>Pattern</th></tr>
<tr><td><code>On time_expression TARGET HOOK was baptized|born</code></td></tr>
<tr><td><code>" HOOK (TARGET -</code></td></tr>
<tr><td><code>-|-- HOOK (TARGET -</code></td></tr>
<tr><td><code>AND|and|or HOOK (TARGET -</code></td></tr>
<tr><td><code>By|about|after|by|for|in|of|with HOOK TARGET -</code></td></tr>
</table>

Table 1. Some of the patterns obtained for the relationship *birth year*.

.

Relation	No. of patterns	No. of results	Precision
Birth-year	16	15746	74.14%
Death-year	8	5660	90.20%
Birth-place	3	154	27.27%
Actor-film	11	4	50.00%
Country-Chief of state	109	272	50.00%
Writer-book	176	179	37.29%
Country-capital	150	825	11.45%
Player-team	173	315	7.75%

Table 2. Number of patterns obtained for each relationship, number of results extracted by each pattern set, and precision.

higher than 0.90 in the development super-corpus and that matched at least 3 times are used. Table 1 shows some of the patterns obtained for the property *birth date*. In the second column of Table 2 the total number of patterns that were finally obtained for each of the semantic relationships under consideration is shown.

4.2 Results and discussion

Table 2 shows the number of results (pairs of related terms) that each of the pattern sets has extracted, and the precision attained. This precision has been estimated by correcting manually least 50 results from each relationship.

As can be seen, the precision for birth year and death year is very good, because they are usually expressed with very fixed patterns, and years and dates are entities that are very easily recognised. The few errors are mainly due to the following two cases:

- Named Entity tagging mistakes, e.g. a TV series mistagged as a person, where the years in which it has been shown are taken as birth and death date.
- Names of persons that held a title (e.g. king or president) during a period of time, that is mistakenly considered their life span.

On the other hand, as expected, the other examples have proven more difficult to identify. We have observed the problem, mentioned in the previous sections, that some patterns are applicable for many kinds of relationships at the same time. This phenomenon is specially relevant in the case of the *player-team* relation. The precision of the patterns is 92% when they are applied only to the entries about soccer players, but the figure falls down to 7.75% when applied to the whole Wikipedia corpus collected.

This means that they are patterns that, in the domain of soccer, usually indicate the relationship between the player and its club, but in other contexts they may be conveying a different meaning. One of these patterns is the already mentioned genitive construction. In sports articles, when this construction is found between an organisation and a person is usually expressing the *player-team* relation, as in *Liverpool's Fowler*. But it also extracted many wrong pairs from documents belonging to different topics.

The same also applies to the case of countries and capitals. During training, from phrases such as *Spain's Madrid*, the system extracted the genitive construction as indicating a relationship of capitality, but it is a source of errors because it can also express a part-of relationship between a country and any of its cities.

In the case of *actor-film*, we have observed that in the actors' entries in the Wikipedia, there is usually a section containing all the filmography, expressed as an HTML bullet list. In this way, because the information is already semi-structured, the textual patterns cannot apply. It should be easier to extract that data using other simpler procedures that take benefit of the structure of the entry.

4.3 Automatically generated list pages

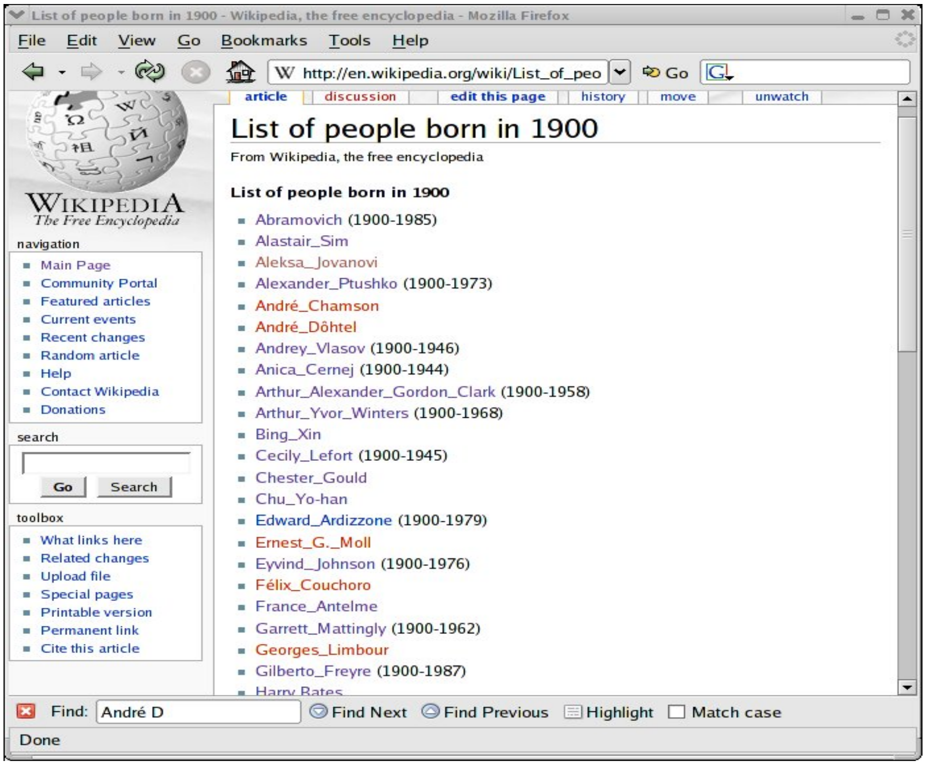
To test a possible application of this procedure in a real case, the algorithm has been used to generate automatically a list page using the data collected from Wikipedia. The example chosen is the list of people born the year 1900. A total of 57 people had been annotated with that birth date. The system was able to obtain the list automatically and, in the cases in which the death date was available as well, it was added as additional information. A manual evaluation of the generated list revealed the following:

- Two out of the 57 people were not people, due to errors in the Named Entity recogniser. These were eliminated by hand.
- One birth date was erroneous, and it was really the date of publication of a book about that person. That entry was also removed.
- One person had been extracted with two different (but correct) spellings: Julian Green and Julien Green, so they were merged in one.
- From the remaining 53 people, 14 did not have an associated Wikipedia entry. They had been extracted from other lists in which they were mentioned, but their biography is not available yet. These were left inside the list, as it is useful information.
- Finally, three people in the list had ambiguous names, so they were directed to disambiguation pages. It was easy to modify the hyperlink so they pointed directly to the particular person with that name that had been born in 1900.

Figure 2 shows a possible layout for the generated page. Note that, in the Wikipedia, there are special categories to group all the people born every year. So, the category called *Category:1900_birth* contains all the people for which someone has categorised their entries as having been born in 1900.

Before our experiment was performed, this category contained 640 people[14] born in that year, all of them with an entry in the Wikipedia. From our list, we have been

[14] Note that, in our experiment, we do not process the whole English wikipedia (more than one million entries) but a small subset containing around 20,000 entries.



List of people born in 1900

From Wikipedia, the free encyclopedia

List of people born in 1900

- Abramovich (1900-1985)
- Alastair_Sim
- Aleksa_Jovanovi
- Alexander_Ptushko (1900-1973)
- André_Chamson
- André_Dôhtel
- Andrey_Vlasov (1900-1946)
- Anica_Cernej (1900-1944)
- Arthur_Alexander_Gordon_Clark (1900-1958)
- Arthur_Yvor_Winters (1900-1968)
- Bing_Xin
- Cecily_Lefort (1900-1945)
- Chester_Gould
- Chu_Yo-han
- Edward_Ardizzone (1900-1979)
- Ernest_G._Moll
- Eyvind_Johnson (1900-1976)
- Félix_Couchoro
- France_Antelme
- Garrett_Mattingly (1900-1962)
- Georges_Limbour
- Gilberto_Freyre (1900-1987)
- Harry Bates

Fig. 2. A possible layout for the page generated automatically containing people that were born in 1900, after a manual correction was performed.

able to identify and categorise four people born in that year that were still not inside the category, and 14 people that are not listed because yet nobody has written their entries.

In the same way that this list has been gathered, it should be easy to create or extend list pages using other criteria, such as famous people born in a certain city, or people that died at a certain age. If the Wikipedia were extended with semantic links, then all these list pages would not need to be stored as static pages inside the Wikipedia server, but ideally they would be generated on-the-fly using the semantic metadata.

5 Related work

To our knowledge, there is no other work reported addressing the task of annotating semi-automatically wiki content for the Semantic Web. However, there is already much research on automatically identifying relationships in unrestricted text. In particular, the use of lexical or lexicosyntactic patterns to discover ontological and non-taxonomic relationships between concepts has been proposed by [13–15], all of whom manually define regular expressions to extract hyponymy and part-of relationships. [16] learns patterns that express company merge relationships. [17] quantifies the error rate of a similar approach for hyponymy relationships at 32%.

Systems that learn these lexical patterns from the web have the advantage that the training corpora can be collected easily and automatically. Several approaches have been proposed recently [21, 22, 12], having various applications in mind: Question-Answering [12], multi-document Named Entity Coreference [23], and the generation of biographical information [24].

Other systems that automatically extract ontological knowledge from text are Text-ToOnto [18], OntoLT [19] and KIM [20].

6 Conclusions and Future Work

In this paper, we propose the use of Natural Language Processing techniques to automatically extract semantic relationships from the Wikipedia. We have shown that, for some relationships, the precision obtained is acceptable, and with a brief manual revision good-quality metadata can be obtained.

We believe that these procedures can contribute in increasing the size of current Semantic Wikis semi-automatically. Even though it will always be necessary a manual revision to identify the wrong results, the work involved in correcting the generated metadata is smaller than creating it all from scratch, as we have shown with the example of generating list pages.

Furthermore, we foresee applications of this work for automatic ontology building and population, as hinted by [7].

Concerning future work, we plan to continue exploring ways to improve the precision of the patterns for the relationships with poorer performance. In particular, the use of patterns that may express different relationships, depending on the context, needs to be further enhanced.

We also plan to try this procedure with yet more relationships, and in other languages.

References

1. Kiesel, M., Sauermann, L.: Towards semantic desktop wiki. UPGRADE special issue on The Semantic Web **6** (2005) 30–34
2. Schaffert, S., Gruber, A., Westenthaler, R.: A semantic wiki for collaborative knowledge formation. In: Proceedings of SEMANTICS 2005 Conference., Vienna, Austria (2005)
3. Giles, J.: Internet encyclopaedias go head to head. Nature, Special Report **438** (2005) 900–901
4. Britannica, E.: Fatally flawed: Refuting the recent study on encyclopedic accuracy by the journal nature (2006)
5. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web - a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American **284** (2001) 34–43
6. Baeza-Yates, R.: Mining the web. El profesional de la información **13** (2004) 4–10
7. Krötzsch, M., Vrandecic, D., Völkel, M.: Wikipedia and the semantic web - the missing links. In: Proceedings of WIKIMANIA 2005, 1st International Wikimedia Conference. (2005)

8. Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., Studer, R.: Semantic wikipedia. In: Proceedings of the 15th international conference on World Wide Web, WWW 2006, Edinburgh, Scotland (2006)
9. Ruiz-Casado, M., Alfonseca, E., Castells, P.: Automatic assignment of wikipedia encyclopedic entries to wordnet synsets. In: Proceedings of the Atlantic Web Intelligence Conference, AWIC-2005. Volume 3528 of Lecture Notes in Computer Science. Springer Verlag (2005) 380–386
10. Ruiz-Casado, M., Alfonseca, E., Castells, P.: Automatic extraction of semantic relationships for wordnet by means of pattern learning from wikipedia. In: Natural Language Processing and Information Systems. Volume 3513 of Lecture Notes in Computer Science. Springer Verlag (2005) 67–79
11. Alfonseca, E., Castells, P., Okumura, M., Ruiz-Casado, M.: A rote extractor with edit distance-based generalisation and multi-corpora precision calculation. In: In Proceedings of the Poster Session of ACL-2006. (2006)
12. Ravichandran, D., Hovy, E.: Learning surface text patterns for a question answering system. In: Proceedings of the ACL-2002. (2002) 41–47
13. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of COLING-92, Nantes, France (1992)
14. Hearst, M.A.: Automated discovery of wordnet relations. In: Christiane Fellbaum (Ed.) WordNet: An Electronic Lexical Database. MIT Press (1998) 132–152
15. Berland, M., Charniak, E.: Finding parts in very large corpora. In: Proceedings of ACL-99. (1999)
16. Finkelstein-Landau, M., Morin, E.: Extracting semantic relationships between terms: supervised vs. unsupervised methods. In: Workshop on Ontologial Engineering on the Global Info. Infrastructure. (1999)
17. Kietz, J., Maedche, A., Volz, R.: A method for semi-automatic ontology acquisition from a corporate intranet. In: Workshop "Ontologies and text", co-located with EKAW'2000, Juan-les-Pins, France (2000)
18. Maedche, A., Staab, S.: Semi-automatic engineering of ontologies from text. In: Proceedings of the 12th Internal Conference on Software and Knowledge Engineering, Chicago, USA (2000)
19. Buitelaar, P., M. Sintek, M.: OntoLT version 1.0: Middleware for ontology extraction from text. In: Proc. of the Demo Session at the International Semantic Web Conference. (2004)
20. Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D., Kirilov, A.: Kim - a semantic platform for information extaction and retrieval. Journal of Natural Language Engineering **10** (2004) 375–392
21. Brin, S.: Extracting patterns and relations from the World Wide Web. In: Proceedings of the WebDB Workshop at the 6th International Conference on Extending Database Technology, EDBT'98. (1998)
22. Agichtein, E., Gravano, L.: Snowball: Extracting relations from large plain-text collections. In: Proceedings of ICDL. (2000) 85–94
23. Mann, G.S., Yarowsky, D.: Unsupervised personal name disambiguation. In: CoNLL-2003. (2003)
24. Mann, G.S., Yarowsky, D.: Multi-field information extraction and cross-document fusion. In: ACL 2005. (2005)