Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets ⋆

Maria Ruiz-Casado, Enrique Alfonseca and Pablo Castells

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

Abstract. We describe an approach taken for automatically associating entries from an on-line encyclopedia with concepts in an ontology or a lexical semantic network. It has been tested with the Simple English Wikipedia and WordNet, although it can be used with other resources. The accuracy in disambiguating the sense of the encyclopedia entries reaches 91.11% (83.89% for polysemous words). It will be applied to enriching ontologies with encyclopedic knowledge.

1 Introduction

The huge availability of data in the World Wide Web (WWW), and its exponential growth from the past few years, has made the search, retrieval and maintenance of the information a hard and time consuming task, specially when these tasks (or part of them) have to be carried out manually. One of the difficulties that prevents the complete automatising of those processes [1] is the fact that the contents in the WWW are presented mainly in natural language, whose meaning ambiguities are hard to be processed by a machine.

The Semantic Web (SW) appears as an effort to extend the web with machine readable contents and automated services far beyond current capabilities [2]. In order to make explicit the meaning underlaying the data, and therefore processable by a machine, a common practise is the annotation of certain words, pages or other web resources using an ontology. Sometimes, the ontologies have to include a high amount of information, or they undergo a rapid evolution. This would be the case of the automatic annotation of news, where the domain is very vast and changing. Therefore, it would be highly desirable to automatise or semi-automatise the acquisition of the ontologies themselves. This problem has been object of recent increasing interest, and new approaches [3] for automatic ontology enrichment and population are being developed, which combine resources and techniques from Natural Language Processing, Information Extraction, Machine Learning and Text Mining.

Text Data Mining, defined as the problem of finding novel pieces of information inside textual data [4], is a research area motivated in part by the large amounts of text available. When the source for mining is the World Wide Web, Text Mining is usually called *web mining*. Text and web mining techniques have been used previously for automatically populating ontologies and lexical semantic networks with concepts [5–8,

⋆ This work has been sponsored by CICYT, project number TIC2002-01948.

3]. In order to construct ontologies semi-automatically, it is necessary to define a similarity metric between concepts that can be used to organise them. A popular procedure is based on the distributional semantics hypothesis, which states that the meaning of two words is highly related to the contexts in which they can appear [9]. In this way, we can assume that the meaning of a word is somehow encoded in the contexts in which we have observed it. A useful formalism for representing contexts is the Vector Space Model [10] (VSM), where a word is described as the bag of the terms which co-occur with it in texts [11, 12] or inside dictionary definitions [13]. There are some possible variations, such as collecting only terms which hold some head-modifier syntactic relationship [14, 15].

Apart from enriching existing ontologies with new concepts, it is also possible to try to discover semi-automatically new relationships between the concepts that already belong to the ontology. To this aim, concept definitions and glosses have been found very useful, as they are usually concise descriptions of the concepts and include the most salient information about them [16]. This has already been applied to the WordNet lexical semantic network [17], which is structured as a directed graph, where nodes represent concepts (called *synsets*, or synonym sets), arcs represent relationships, and each synset is annotated with a gloss. In fact, concept glosses have also been found useful in many other problems, such as Automatic Text Summarisation or Question Answering [18] On the other hand, WordNet glosses have been sometimes criticised, as they do not follow any common pattern and some of them are not very informative. This problem appears, with a higher extent, in the multilingual EuroWordNet [19], where many of the glosses are nonexistent. Therefore, a procedure for automatically extending them would be desirable.

In this paper, we present a procedure for automatically enriching an existing lexical semantic network with on-line encyclopedic information that defines the concepts. The network chosen is WordNet, given that it is currently used in many applications, although the procedure is general enough to be used with other ontologies. The encyclopedia chosen is the Wikipedia, in its Simple English version[1]. The syntactic structures found in Simple English are easier to handle by a parser than those in fully unrestricted text, so the definitions will be easier to process in the future.

2 Procedure

The system built crawls the Simple English Wikipedia collecting definition entries, and associates each entry to a WordNet synset. The processing performed is the following:

1. Retrieve a web page from the encyclopedia.
2. Clean the page from everything except the entry (remove all the menus and navigation links).
3. Analyse the entry with a part-of-speech tagger and a stemmer [20]. Remove all the closed-class words (everything except nouns, verbs, adjectives and adverbs).
4. Attach the definition to the synset in WordNet that it is defining.We may encounter several cases:

[1] http://simple.wikipedia.org/wiki/Main_Page

- There is only one synset in WordNet containing the word described in the entry. This is the case, for instance, of the entry *Abraham Lincoln*. This case is trivial, as the encyclopedia entry can be simply associated with that synset.
- It may also be the case that the term described in the encyclopedia does not appear in WordNet. In this case, the entry is ignored.
- Finally, it may happen that there are several synsets in WordNet containing the word described in the entry. In this case, it is necessary to discover which is the correct sense with which the word is used in the entry.

The last case is a classical problem in Natural Language Processing called *Word Sense Disambiguation* [21] (WSD). It generally uses some metric of similarity between the word to disambiguate (in our case, the Wikipedia entry) and each one of the possibilities (the possible WordNet synsets). Different approaches use co-occurrence information [22], all WordNet relationships [23], or just is-a relations (the *hyperonymy* relationship, which relates a concept with others that are more general) [24], with various success rates. Also, some results indicate that WordNet glosses are useful in calculating the semantic similarity [25].

In our problem, we want to find a similarity metric between encyclopedia entries and WordNet synsets. If they refer to the same concept, we can expect that there will be much in common between the two definitions. This is the reason why the approach followed is mainly a comparison between the two glosses:

1. Represent the Wikipedia entry as a vector e using the Vector Space Model, where each dimension corresponds to a word, and the coordinate for that dimension is the frequency of the word in the entry.
2. Let $S = \{s_1, s_2, ..., s_n\}$ be the set of WordNet synsets containing the term defined in the Wikipedia entry.
3. Represent each synset s_i as the set of words in its gloss: $G_i = \{t_1, t_2, ..., t_{k_i}\}$.
4. Let N = 1
5. Extend the sets G_i with the synonym words in each synset s_i and its hyperonyms to a depth of N levels.
6. Weight each term t in every set G_i by comparing it with the glosses for the other senses. In this way, a numerical vector v_i, containing the term weights, is calculated for each G_i. In the experiments, two weight functions have been tried: tf·idf and χ^2 [22].
7. Choose the sense such that the similarity between e and v_i is the largest. Two similarity metrics between the two vectors have been tested: the dot product [26, pg. 18] and the cosine. If there is a tie between two or more senses, increment N and go back to step 5.

3 Evaluation

The algorithm has been evaluated with a sample of the Simple English Wikipedia entries, as available on November 15, 2004. The version of WordNet used is 1.7. From 1841 Wikipedia terms, 612 did not appear in WordNet, 631 were found in WordNet with only one possible sense (they are monosemous) and 598 Wikipedia terms were found in WordNet with more than one sense (they are polysemous). Figure 1 shows an example of a polysemous term. The following evaluations have been performed:

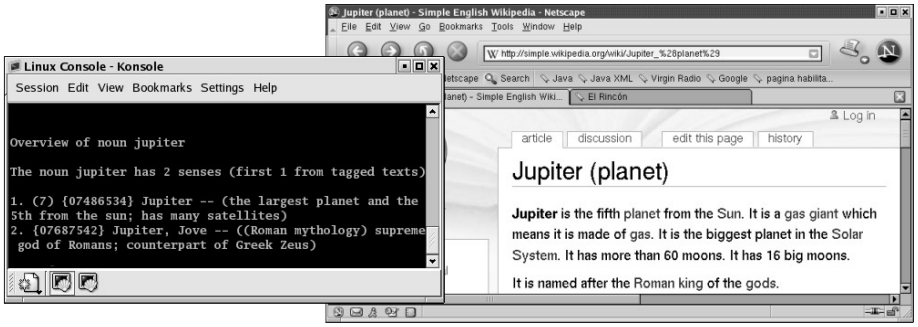

Fig. 1. Entry for *Jupiter (planet)* in the Wikipedia, and WordNet glosses for the synsets that contain the term *Jupiter*.

3.1 Evaluation procedure

Monosemous terms For these terms, the algorithm just associates each Wikipedia entry with the only WordNet synset containing it. A sample, containing the first 180 monosemous terms from Wikipedia, has been manually evaluated, to check whether this assignment is correct.

Polysemous terms In this case, for each Wikipedia entry there were several candidate senses in WordNet, one of which will be chosen by the algorithm. A sample with the first 180 polysemous terms from Wikipedia was manually annotated with the correct sense[2]. In a few cases, the Wikipedia entry included several senses at the same time, because either (a) the wikipedia contained two different definitions in the same entry, or (b) the WordNet senses were so fine-grained that they could be considered the same sense. Regarding this last point, some authors have proposed a previous clustering of the WordNet senses before attempting WSD, to reduce the fine granularity [27]. In these cases, all the right senses are annotated, so the algorithm will be considered correct if it chooses one of them.

The following baseline experiments and configurations have been tested:

- The first baseline consists of a random assignment.
- The second baseline chooses the most common sense of the word in the sense-tagged SEMCOR corpus. This is a set of texts in which every word has been manually annotated with the sense with which it is used, and it can be used to find which is the most common sense of a word in a general text.
- Thirdly, we have implemented Lesk's WSD algorithm [28]. Before applying it, words have been stemmed. Ties between several senses are resolved by choosing SEMCOR's most common sense.
- Our procedure has been tested with three possible variations: two choices for the weight function (tf·idf and χ^2), two possible similarity metrics (cosine and dot product), and either stemming or using the lexical form of the words.

[2] The tagged dataset is available under request at maria.ruiz@uam.es

	Baselines			*Our approach*							
				Dot product				**Cosine**			
				Stemming		No stemming		Stemming		No stemming	
	Random	SEMCOR	Lesk	tf·idf	χ^2	tf·idf	χ^2	tf·idf	χ^2	tf·idf	χ^2
Polysem.	40.10	65.56	72.78	**83.89**	80.56	77.78	77.78	80.56	81.11	78.33	76.67
All	69.22	81.95	85.56	**91.11**	89.45	88.06	88.06	89.45	89.72	88.33	87.50

Table 1. Results obtained for the disambiguation. The first row shows the results only for the polysemous words, and the second one shows the results for all entries in the Wikipedia for which there is at least one synset in WordNet containing the term. The first two columns are the baselines, the third column shows Lesk's algorithm results, and the other eight columns contain the results of the eight configurations tested in our approach.

3.2 Results

With respect to the monosemous terms, 177 out of the 180 assignments were correct, which means an accuracy of 98.33%. Only in three cases the concept defined by the Wikipedia entry was different to the WordNet sense that contained the same term.

Table 1 summarises the accuracy of the different tests for the polysemous terms and for all terms (monosemous and polysemous). These are consistently better than other results reported in WSD, something which may be attributed to the fact that we are comparing two definitions which are supposed to be similar, rather than comparing a definition with an appearance of a term in a generic text. As can be seen, stemming always improves the results; the best score (83.89%) is statistically significantly higher than any of the scores obtained without stemming at 95% confidence. In many cases, also, tf·idf is better than the χ^2 weight function. Regarding the distance metric, the dot product provides the best result overall, although it does not outperform the cosine in all the configurations.

4 Conclusions and future work

In this work we propose a procedure for automatically extending an existing ontology or lexical semantic network with encyclopedic definitions obtained from the web. The approach has been tested with WordNet 1.7 and the Simple English Wikipedia, an Internet encyclopedia built in a collaborative way. We have shown that, for this task, it is possible to reach accuracy rates as high as 91% (83.89% for polysemous words). Interestingly, this result is much higher than the current state-of-the-art for general Word Sense Disambiguation of words inside a text (a more difficult problem), and it shows that current techniques can be applied successfully for automatic disambiguation of encyclopedia entries. We consider this task as a stage previous to knowledge acquisition from a combination of ontologies and encyclopedic knowledge, and opens the following research lines:

1. Analyse the entries that we have associated to WordNet synsets, in order to extract automatically relationships from them, such as *location, instrument, telic* (purpose) or *author*.

2. Generalise the experiment to other ontologies and encyclopedias, and see whether this technique can also be applied to other kinds of texts.
3. Concerning the Wikipedia entries which were not found in WordNet, it would be interesting to explore ways to semi-automatically extend the lexical network with these new terms [5, 6, 8].
4. In the few cases where an entry refers to several synsets in WordNet, divide it distinguishing which fragments of the entry refer to each possible sense.

References

[1] Ding, Y., Fensel, D., Klein, M.C.A., Omelayenko, B.: The semantic web: yet another hip? Data Knowledge Engineering **41** (2002) 205–227
[2] 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
[3] Gómez-Pérez, A., Macho, D.M., Alfonseca, E., nez, R.N., Blascoe, I., Staab, S., Corcho, O., Ding, Y., Paralic, J., Troncy, R.: Ontoweb deliverable 1.5: A survey of ontology learning methods and techniques (2003)
[4] Hearst, M.A. The Oxford Handbook of Computational Linguistics. In: Text Data Mining. Oxford University Press (2003) 616–628
[5] Rigau, G.: Automatic Acquisition of Lexical Knowledge from MRDs. PhD Thesis, Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya (1998)
[6] Hearst, M.A. In: Automated Discovery of WordNet Relations. In Christiane Fellbaum (Ed.) WordNet: An Electronic Lexical Database. MIT Press (1998) 132–152
[7] Agirre, E., Ansa, O., Martínez, D., Hovy, E.: Enriching wordnet concepts with topic signatures. In: Proceedings of the NAACL workshop on WordNet and Other lexical Resources: Applications, Extensions and Customizations, Pittsburg (2001)
[8] Alfonseca, E., Manandhar, S.: Extending a lexical ontology by a combination of distributional semantics signatures. In: Knowledge Engineering and Knowledge Management. Volume 2473 of Lecture Notes in Artificial Intelligence. Springer Verlag (2002) 1–7
[9] Firth, J.: A synopsys of linguistic theory 1930-1955. In F. Palmer (ed.), Selected Papers of J. R. Firth. Longman, London (1957)
[10] Salton, G.: Automatic text processing. Addison-Wesley (1989)
[11] Church, K., Gale, W., Hanks, P., Hindle, D.: 6. In: Using Statistics in Lexical Analysis. In U. Zernik (ed.), Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon. Lawrence Erlbaum Associates, Hillsdale, New Jersey (1991) 115–164
[12] Lin, C.Y.: Robust Automated Topic Identification. Ph.D. Thesis. University of Southern California (1997)
[13] Wilks, Y., Fass, D.C., Guo, C.M., McDonald, J.E., Plate, T., Slator, B.M.: Providing machine tractable dictionary tools. Journal of Computers and Translation (1990)
[14] Lee, L.: Similarity-Based Approaches to Natural Language Processing. Ph.D. thesis. Harvard University Technical Report TR-11-97 (1997)
[15] Faure, D., Nédellec, C.: A corpus-based conceptual clustering method for verb frames and ontology acquisition. In: LREC workshop on Adapting lexical and corpus resources to sublanguages and applications, Granada, Spain (1998)
[16] Harabagiu, S., Moldovan, D.I.: Knowledge processing. In: WordNet: An Electronic Lexical Database. MIT Press (1998) 379–405
[17] Miller, G.A.: WordNet: A lexical database for English. Communications of the ACM **38** (1995) 39–41
[18] Rus, V.: Logic Form For WordNet Glosses and Application to Question Answering. Ph.D. thesis. Computer Science Department, Southern Methodist University (2002)
[19] Vossen, P.: EuroWordNet - A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers (1998)
[20] Alfonseca, E.: Wraetlic user guide version 1.0 (2003)
[21] Ide, N., Véronis, J.: Introduction to the special issue on word sense disambiguation: the state of the art. Computational Linguistics **24** (1998) 1–40
[22] Manning, C.D., Schütze, H.: Foundations of statistical Natural Language Processing. MIT Press (2001)
[23] Hirst, G., St-Onge, D.: Lexical chains as representations of context for the detection and correction of malapropisms. In: WordNet: an electronic lexical database. MIT Press (1998)
[24] Resnik, P.K.: Disambiguating noun groupings with respect to wordnet senses. In: Proceedings of the Third Workshop on Very Large Corpora, Somerset, ACL (1995) 54–68
[25] Mihalcea, R., Moldovan, D.: A method for word sense disambiguation of unrestricted text. In: Proceedings of ACL'99, Maryland, NY (1999)
[26] Kilgarriff, A., Rosenzweig, J.: Framework and results for english SENSEVAL. Computer and the Humanities (2000) 15–48
[27] Agirre, E., de Lacalle, O.L.: Clustering wordnet word senses. In: Recent Advances in Natural Language Processing III. (2004)
[28] Lesk, M.: Automatic sense disambiguation using machine readable dictionaries. In: Proceedings of the 5th International Conference on Systems Documentation. (1986) 24–26