Extending a Lexical Ontology by a Combination of Distributional Semantics Signatures*

Enrique Alfonseca[1] and Suresh Manandhar[2]

[1] Ingeniería Informática, Universidad Autónoma de Madrid, 28049 Madrid, Spain.
Enrique.Alfonseca@ii.uam.es
[2] Computer Science Department, University of York, YO10 5DD York, U.K.
suresh@cs.york.ac.uk

Abstract. Ontologies are a tool for Knowledge Representation that is now widely used, but the effort employed to build an ontology is high. We describe here a procedure to automatically extend an ontology such as WordNet with domain-specific knowledge. The main advantage of our approach is that it is completely unsupervised, so it can be applied to different languages and domains. Our experiments, in which several domain-specific concepts from a book have been introduced, with no human supervision, into WordNet, have been successful.

1 Introduction

Lexical semantic ontologies are now widely used for Natural Language Processing, and several of them are available for English and other languages, such as WordNet [Miller, 1995] and EuroWordNet [Vossen, 1998]. However, they are usually very general, and their enrichment with domain-specific information requires a high degree of supervision. This has motivated the appearance of knowledge acquisition methods for building domain-specific ontologies automatically.

Maedche and Staab [2001] define *Ontology Refinement* (OR) as the adaptation of an ontology to a specific domain or to some user's requirements, without altering its overall structure. An important problem inside OR is the placement of the domain-dependent concepts in the ontology. Applied to lexical ontologies, if we have an ontology $\mathcal{W}$ and a set of domain-specific documents $\mathcal{D}$ containing some unknown concepts and instances $\mathcal{U} = \{u_1, u_2, ..., u_n\}$, we have to find, for every unknown concept or instance u_i, its maximally specific hypernym s_i in the ontology.

This paper reports a system that extended WordNet with new synsets learnt both from Tolkien's *The Lord of the Rings* and Darwin's *The Voyages of the Beagle*. The unknown concepts that were learnt include locations, rivers, seas, animals, races and people. The classification of these concepts in the taxonomy is performed in a fully unsupervised way. The only input it requires is an initial ontology (we use WordNet) and a collection of documents. It will find unknown

* This work has been partially sponsored by CICYT, project number TIC2001-0685-C02-01.

A. Gómez-Pérez and V.R. Benjamins (Eds.): EKAW 2002, LNAI 2473, pp. 1–7, 2002.

concepts in the documents and attach them to WordNet. We show that a Distributional Semantic (DS) model can be a good starting point for locating the right places in the ontology for placing the learnt synsets, and we present a way in which different DS metrics can be combined.

1.1 Related Work

We can group related systems in two groups. *Deterministic systems* are those that provide, for each unknown concept, one or several hypernyms all of which are supposedly correct. One of such systems, described by Hearst [1998], obtains regular expression patterns from free texts by looking at pairs of (hypernym, hyponym) that co-occur in the same sentence, and then uses them to learn new hypernymy relations. However, she notes that these extracted relations contain a high degree of noise. Kietz et al. [2000] quantified the error rate of hand-coded patterns as 32%, so he concludes that the concepts had to be ultimately revised and placed in the hierarchy by the user. A different approach is that described by Grefenstette and Hearst [1992].

Non-deterministic systems are those that provide a set of likely candidates, only some of which are correct. Hastings [1994] describes one such framework, Camille, that learns nouns and verbs. He has beforehand concept ontologies for nouns and verbs about the terrorist domain, and verbs are annotated with selectional preferences, e.g. the object of *arson* is known to be a *building*. If an unknown word was found being the direct object of *arson*, it can thus be classified as a building. Hahn and Schnattinger [1998] describes a similar approach. These systems do not return a single hypernym, but a set of plausible hypotheses.

2 Algorithm

Our aim is the enrichment of WordNet with new concepts learnt from general-purpose texts. Let us suppose that we have found a new term u that is not in WordNet. The aim is to find the place where it should be attached to the ontology. The algorithm we use performs a top-down search along the ontology, and stops at the synset that is most similar to u. The search starts at the most general synset s, and compares u with s and with all its immediate hyponyms. If s is more similar to u than any of s's children, then u is classified as a hyponym of s. Otherwise, we proceed one step downwards to the most similar hyponym. For a detailed description please refer to [Alfonseca and Manandhar, 2002b].

2.1 Similarity Metrics

The tools we used to compute the semantic distance between synsets are all based on the DS model, which assumes that there is a strong correlation between the semantics of a word and the set of contexts in which that word appears, and which has produced good results when applied to fields such as Information Retrieval and summarisation.

We have used the following tools:

- A *topic signature* of a concept c is the list of the words that simply co-occur with c in the same context (we have used the same sentence), and their frequencies.
- A *subject signature* of a nominal concept c is the list of verbs for which c appears as a subject.
- An *object signature* of a nominal concept c is the list of verbs and prepositions for which c appears as an argument.
- A *modifier signature* of a nominal concept c is the list of adjectives and determiners that modify c inside a Noun Phrase.

The intuition behind our procedures is that, if two words are semantically related, their signatures will also be similar. They can be automatically collected for every concept in an ontology, by collecting documents from Internet and collecting the frequencies as described by Agirre et al. [2000]; therefore, the classification procedure can be made fully unsupervised. As an example, Table 1 shows the highest frequency words in the signatures of the concept <*person*>.

In order to compare the topic signature of an unknown concept u and the signatures of a set of WordNet synsets $\{s_1, s_2, ..., s_n\}$, the raw frequencies of the synsets' signatures have be changed into weights, to decide which words do provide support that they are in the context of a synset, and which ones are equally frequent for all synsets [Alfonseca and Manandhar, 2002b]. In a few words, for each decision of the algorithm, the following steps are followed:

- Take the synsets that will be compared to u.
- For each synset, add up the frequencies of the context vectors of all its hyponyms in WordNet, and smooth the frequencies by adding 1 to every frequency value.
- For each one of them, use the rest as a contrast set to change its frequencies into weights. We have obtained the best results with Xi^2 (see [Agirre et al., 2000]).
- To calculate the similarity between u's list of words and frequencies, and a synset s_i's words and weights, perform the dot product of both vectors [Yarowsky, 1992].

2.2 Combining the Similarity Measures

Each signature (topic, subject, object and modifier) provides different similarity values that have to be combined. Let us suppose that we are classifying an unknown concept u, and that we have n choices to take: $\{s_1, ..., s_n\}$ and m signatures. Let us call $P_{sig_j}(s_i)$ the similarity value obtained from the signature sig_j, normalised so all the similarity values obtained from a given signature sum to 1: $\sum_{i=0}^{n} P_{sig_j}(s_i) = 1$. We combine the metrics with a weighted sum, by giving a weight to each of the kind of signatures: $P(s_i) = \sum_{j=0}^{m} weight_j \cdot P_{sig_j}(s_i)$

The baseline experiment was calculated by giving them the same weight $\frac{1}{m}$. In our experiments, we calculated the weights that produce a weight distribution P that is equidistant to the partial distributions P_{sig_j}. The distance metric

Topic signature			Subject signature			Object signature			Modifier signature		
Word	**Freq**	*weight*	**Word**	**Freq**	*weight*	**Word**	**Freq**	*weight*	**Word**	**Freq**	*weight*
Rights	314	23.16	be	23	0.71	of	77	0.38	innocent	16	8.28
Human	162	12.89	have	14	4.24	to	54	3.15	contact	10	9.51
that	161	0.00	use	10	15.09	for	45	3.31	live	8	3.34
Resources	136	19.19	write	6	20.51	in	29	1.11	own	6	5.62
Irights	109	19.94	live	5	4.59	be	23	0.32	indigenous	6	7.28
Department	102	21.77	make	5	6.37	with	15	0.82	other	5	0
Chromosome	96	24.82	kill	4	24.60	on	14	1.30	same	5	6.70
information	65	11.04	work	4	24.60	from	14	2.23	controlling	5	10.25
Center	63	16.04	hold	3	12.65	that	11	0.00	first	3	7.57
Health	63	15.86	produce	3	5.14	as	10	0.06	human	3	2.76
not	56	0.00	suffer	3	11.29	by	9	0.35	right	3	6.36
has	56	3.75	wish	3	16.56	say	6	12.45	whole	3	6.88
have	55	1.98	get	3	12.65	seek	6	13.41	particular	3	5.15

Table 1. Topic, subject, object and modifier signatures of the concept <*person*>. The words shown are the top frequency words and their weights.

chosen to compare distributions is **relative entropy**, also called **Kullback-Leibler distance**. Therefore, we calculated the weights $weight_j$ such that the final distribution P is equidistant (with the minimal distance) to each weight distributions P_{sig_j}, using $D(p||q)$ as the distance metric. These weights are calculated with a simulated annealing procedure. They are initialised as $\frac{1}{m}$, and then we proceeded changing them, slowing down, until the distances $D(P_{sig_j}||P)$ all converge to the same value (if possible). Finally, the synset chosen by the algorithm is $\text{argmax}_i P(s_i)$. Table 2 shows the similarity metrics produced in the first two decisions when classifying the concept <*orc*>, using the topic, subject and object signatures.

3 Experiments and Results

We have calculated the topic, subject, object and modifier signatures for the top 1,200 synsets of the WordNet taxonomy which is rooted by *entity*. This was done automatically by downloading the documents from Internet using the procedure detailed by Agirre et al. [2000].

For learning some new concepts, the domain-dependent texts were processed with our own *ad hoc* shallow parser, and the most frequent unknown words and sequences of words (collocations) were extracted: a total of 46 concepts that appeared 50 or more times in the texts, so we had enough contextual information to classify them. We also hypothesised that every appearance of an unknown common noun (e.g. hobbit, orc, etc.) refers to the same concept, i.e. that they are not polysemous. This did not always occurred (e.g. in Darwin's text *York Minster* was a person; and *St. Yago* was both a person and a place), so this is a shortcoming that should be addressed in the future.

We have used four different metrics. The first one, *accuracy*, is defined as the portion of correctly classified concepts. We have distinguished two measures of accuracy: *strict accuracy* is the percentage of times that the hypernym proposed by the program was the one we expected (as classified by a human), and *lenient*

First decision: entity					
synset	**synset Id**	P_{sig_1}	P_{sig_2}	P_{sig_3}	*total*
being, organism	n00002908	0.40	0.23	0.29	0.3207
causal agency	n00004753	0.38	0.24	0.23	0.3121
location	n00018241	0.11	0.17	0.17	0.1383
body of water	n07411542	0.09	0.12	0.20	0.1112
thing (anything)	n03781420	0.00	0.11	0.02	0.0457
thing (object)	n00002254	0.00	0.11	0.02	0.0442
(16 more)	...	...	...	...	...

Second decision: being					
synset	**synset Id**	P_{sig_1}	P_{sig_2}	P_{sig_3}	*total*
human	n00005145	0.64	0.80	0.40	0.6161
animal	n00010787	0.24	0.18	0.41	0.2790
host	n01015823	0.00	0.01	0.05	0.0243
parasite	n01015154	0.01	0.00	0.04	0.0192
flora	n00011740	0.00	0.00	0.04	0.0169
(34 more)	...	...	...	...	...

Table 2. Similarity values for each of the decisions that have been taken when classifying the unknown concept <*orc*>. The similarities correspond to the topic, subject and object signatures (in that order), and the combination of the three of them. In the first place, when deciding between <*entity*> and its children, the chosen one was <*being, life form*>. In the second decision, the winner was <*human*> (both were correct)

Method	**Accuracy**		**L.A.**	**C.D.**	**C.P.**
	strict	**len.**			
Uniform	13.04%	23.91%	0.34	71.09%	1.98
Entropy	13.04%	28.26%	0.38	73.44%	1.95

Table 3. Comparison of two methods to combine the results provided by the signatures.

accuracy is the percentage of times that the system proposed a hypernym that can be considered valid, although it is not the best one (e.g. if a *man* was classified as *grown man*).

Other metrics that we can measure on our top-down algorithm are **Correct decisions** (C.D.), the percentage of times that a correct decision was chosen at each iteration of the algorithm; and **Correct position** (C.P.), which measures, at each step in the search, when the different children synsets are ordered according to the signatures, the mean position of the correct one. Ideally, this metric has to be as low as possible. Finally, **Learning Accuracy** [Hahn and Schnattinger, 1998] takes in consideration the distance, in the ontology, between the proposed hypernym and the correct one. Please refer to [Alfonseca and Manandhar, 2002a] for a detailed description.

Keeping constant the number of signatures at three: topic, subject and object, we tried the two methods to combine them: the baseline, using a uniform weight, and the simulated annealing, using relative entropy. Results are displayed at Table 3. The procedure that uses relative entropy to find an *intermediate* distribution improved all the metrics when compared to the uniform weighting.

Next, we tested different combination of the signatures. Table 4(a) shows the results. The signature that produced the worst results was the modifier signature: the learning accuracy is the smallest, and the mean position of the correct concept at each decision is very high, nearly 1.5 over the next mark. Also, when used with the others, the modifier signature greatly degrades the results. A manual examination of the signatures seems to indicate that the modified

Method	Accuracy strict	Accuracy len.	L.A.	C.D.	M.P.
Topic	6.52%	17.39%	0.30	68.21%	2.30
Modifiers	16.28%	21.74%	0.29	62.96%	4.47
Subject	10.87%	23.91%	0.30	68.80%	3.06
Object	17.39%	28.26%	0.38	71.43%	2.63
T+S+O	13.04%	28.26%	0.38	73.44%	1.95
TSOM	10.87%	21.74%	0.35	70.31%	2.00

System	Accuracy strict	Accuracy lenient	L.A.
T+S+O	28.26%	36.96%	0.44
	single	**set**	
[Hastings, 1994]	19%	41%	-
Hahn [1998]	21%	22%	0.67
Hahn [1998]-TH	26%	28%	0.73
Hahn [1998]-CB	31%	39%	0.76

Table 4. (a) Results using different signatures. (b) Comparison with related systems.

signature has a low quality, with a high degree of words which are not adjectives. The other three signatures produced acceptable results, and the best mark was attained by combining them all. Most of the errors were produced at the lowest levels of the ontology, when deciding between semantically similar synsets such as <*man*> and <*woman*>, for which the context is not much help.

Some characteristics of WordNet complicate the task, such as the fine-grained senses and the lack of multiple inheritance. For example, geographical locations (e.g. continents) are classified as *object*, while political locations (e.g. countries) are classified as *location*. However, the context of these two kinds of entities are very similar, and they are very different to the context of other *objects* such as *artifacts*. Our algorithm "incorrectly" classified concepts such as mountains or woodlands as *locations*.

Table 4(b) shows results obtained by our system, labelled T+S+O, compared to similar work. There are some important differences, and thence the results are not really comparable. First, the ontologies used in each approach are different, which can have dramatic consequences on the evaluation. Secondly, the other systems are not deterministic. In Table 4(b), the column labelled **single accuracy** shows the percentage of outcomes in which the systems returned a unique hypernym and that one was correct (like our own algorithm); and the **set accuracy** is the percentage of times that the system returned several outcomes among which was the correct one.

4 Conclusions

We have presented here a fully unsupervised method for extending lexical ontologies with unknown concepts taken from domain-specific documents. It can be applied to different domains as it is; and, if we have a shallow parser available, to different languages. It allows the attachment of new concepts to any intermediate level in an ontology, not only at the leaves; and it can tackle large ontologies, such as WordNet. It has been used for generating hypermedia courses using text summarisation as described by Alfonseca and Rodríguez [2002].

Compared to previous approaches (c.f. [Hahn and Schnattinger, 1998, Hastings, 1994]), it requires less resources, as all the signatures are collected automatically, and it does not need a previous encoding of selectional restrictions or article scripts.

The main drawback of our system, as it is now, is that the signatures need to have a certain size in order to provide reliable classifications. An unknown concept that only is cited once will provide a signature with at most a few entries, and the classification will probably be wrong, so an open line for future work is to improve it for low-frequency terms. It has also low precision when classifying concepts that are semantically very related, such as *man* and *woman*. Finally, it still cannot discover whether an unknown word is a synonym of a word already in the ontology.

References

E. Agirre, O. Ansa, E. Hovy, and D. Martinez. Enriching very large ontologies using the www. In *Ontology Learning Workshop, ECAI*, Berlin, Germany, 2000.

E. Alfonseca and S. Manandhar. Proposal for evaluating ontology refinement methods. In *Language Resources and Evaluation (LREC-2002)*, Las Palmas, 2002a.

E. Alfonseca and S. Manandhar. An unsupervised method for general named entity recognition and automated concept discovery. In *1st Conf. on Gen. WordNet*, 2002b.

E. Alfonseca and P. Rodríguez. Automatically generating hypermedia documents depending on the user goal. In *Workshop on Doc. Compression, AH-2002*, 2002.

G. Grefenstette and M.A. Hearst. Method for refining automatically-discovered lexical relations: Combining weak techniques for stronger results. In *Weir (ed.) Statistically based natural language programming techniques, Proc. AAAI Workshop*, 1992.

U. Hahn and K. Schnattinger. Towards text knowledge engineering. In *AAAI/IAAI*, pages 524–531, 1998.

P. M. Hastings. *Automatic acquisition of word meaning from context.* University of Michigan, Ph. D. Dissertation, 1994.

M. A. Hearst. *Automated Discovery of WordNet Relations. In Christiane Fellbaum (Ed.) WordNet: An Electronic Lexical Database*, pages 132–152. MIT Press, 1998.

J. Kietz, A. Maedche, and R. Volz. A method for semi-automatic ontology acquisition from a corporate intranet. In *Workshop "Ontologies and text", EKAW'2000*, 2000.

A. Maedche and S. Staab. Ontology learning for the semantic web. *IEEE Intelligent systems*, 16(2), 2001.

G. A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

P. Vossen. *EuroWordNet - A Multilingual Database with Lexical Semantic Networks.* Kluwer Academic Publishers, 1998.

D. Yarowsky. Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *Proceedings of COLING-92*, pages 454–460, 1992.