Lexicon Generation by Extraction of Context Patterns

Victoria Uren and Enrico Motta

Knowledge Media Institute, The Open University, Milton Keynes, MK7 6AA, UK
v.s.uren@open.ac.uk, e.motta@open.ac.uk

Abstract. Semantic browser technologies such as Magpie require the construction of lexicons to support the identification of terms in Web pages which are linked to a user's chosen ontology. We frame the generation of such lexicons from ontologies as a problem of finding synonyms and hyponyms. Synonym finding using the hypothesis of semantic substitutability relies upon the discovery of patterns in which the target word occurs. Information extraction has the potential to find a range of patterns in text. We present a methodology for finding synonyms for inclusion in lexicons in this way and preliminary tests of the method using standard tools.

Introduction

Browsing the Web and intranets involves finding the right pages and interpreting their contents. The first of these tasks has been far more intensively researched than the second. The Magpie semantic browsing tool helps users to interpret web pages by highlighting terms which are related to entities in a specified ontology and providing relevant contextual services for entities of that type (Dzbor et al. 2003). The ontology provides a perspective shared by a group of users who routinely search the same kinds of data. For instance, a group of financial analysts might be interested in mergers. Magpie would highlight names of companies involved in takeover bids on web pages. A collector service could capture these instances and save them. The result would be a knowledge base in which all the analysts could share the knowledge they have discovered in their browsing sessions.

A key step to providing Magpie services is the generation of the lexicon[1], a resource which allows the system to access, for each instance in the ontology, the various strings which it should recognize and highlight. To date these lexicons have been generated partly by hand and partly by heuristic variation of the strings in the original ontology, e.g. if the ontology contains an instance of the *Researcher* class *"John Domingue"* the lexicon will be automatically populated with both this string and the string *"J. Domingue"*. More general methods are required for producing lexicons for a wider range of domains and potentially in different languages.

The problem of lexicon generation can be generalised as the problem of extending a term, a class or an instance in an ontology, to a list of synonyms or hyponyms. This

[1] This is not a lexicon in the linguistic sense but in the sense defined by Riloff & Jones "a dictionary of words with semantic category labels" (Riloff & Jones 1999).

could be done using resources such as WordNet. However the coverage of WordNet is currently limited both in technical domains and for languages other than English. The problem of identifying synonyms has been explored by the statistical natural language processing community for applications such as expanding queries in information retrieval. They work on the assumption that the semantic similarity between words is related to their contextual similarity, i.e. to the ways in which they are used. This is defined as the notion of semantic substitutability by Miller & Charles (Miller & Charles 1991). Semantic substitutability implies that two words are similar to the extent that one can be exchanged for the other in the context of a sentence without changing the truth values of the sentence. This leads to an experimental approach in which linguistic features are sought which represent context and similarity between words depends on the similarity between their contexts.

Greffenstette (Greffenstette 1992) applies a coarse syntactic analysis to identify three different types of context for nouns: adjectives or nouns which modify a noun, prepositions which modify it, and verbs with which it appears. The overlap between these three lists are used to compute measures of similarity between nouns.

Gauch et al. (Gauch et al. 1999) ignore syntax and concentrate instead on 4 word context vectors which are constructed from the two words before and after each word of interest. Their claim is that the position of a word gives implicit information about its syntactic role. Similarity matrix calculations are used to compute the similarity between words.

Allegrini et al. (Allegrini et al. 2000) present a contextual representation which they call “analogical proportion”, which represents the idea that for two nouns to be similar “one has to be prepared to use them interchangeably in at least two different logical contexts”. An analogical proportion comprises two nouns and two verbs where both nouns have been found in sentences with both verbs.

The above examples give a taste of the work that has been done in the NLP community. In general, experiments have explored particular models for context. Such a model is used to represent patterns in which the target word commonly occurs. Information extraction (IE) systems use a range of approaches to defining patterns for finding entities in text, where entities might be, for instance, dates, people or locations. The question we start to explore in this paper is whether the IE approach to pattern definition could be applied to the synonym finding problem. If so the operational state of IE tools would make them very suitable candidates for building into a lexicon generation system such as the one we envisage for Magpie. It is even possible that the IE approach, in which a system does not commit itself to a particular model of context but instead learns the contexts which are most appropriate for a given problem, may have advantages over approaches where the model is predetermined.

In this paper we first outline in more detail a scenario in which information extraction is used to automatically generate lists of synonyms and hyponyms for terms in an ontology, which are then used to automatically generate a lexicon. Then we present an exploratory experiment using the information extraction tool Amilcare (Ciravegna & Wilks 2003) for the pattern generation task.

Lexicon Generation Approach

We envisage a lexicon generation approach which may be broken down into three stages:

1. Generation of contextual patterns
2. Identification of candidate terms
3. Validation and selection of lexicon terms

Generation of patterns requires seed terms for each class and/or instance in the ontology for which lexicon terms are required. We assume that the ontology will be close enough to natural language to generate these seeds or that the entities in the ontology will have "pretty names". This is reasonable since it is good practice for ontologies to be human-understandable. A collection of texts to learn from is also required. Ideally these should be representative of the domain as a whole. This text collection may need some preprocessing depending on the requirements of the system which is to learn the contextual patterns. This may involve tagging occurrences of the seed terms, parsing the text etc. An IE system will be used to generate the contextual patterns themselves.

Identification of candidate terms is performed by applying the contextual patterns either to the original texts or to new texts. Terms which are tagged by the IE system as examples of the entity will be identified.

Validation and Selection of the final list of terms requires that candidates be ranked according to their similarity to the original entity. Ranking requires numerical scoring. For example, the number of times a string was tagged as a particular entity and the precision of the rules used could be combined in a score. Substring similarity may indicate abbreviations and alternative spellings, e.g. "cycle"/"bicycle" and "colour"/"color". Finally, simple co-occurrence scores have been used successfully to evaluate candidate synonyms (Turney 2001). A combination of methods would be used to rank candidates and select terms for the lexicon.

Experimental Method

In this paper, we present some preliminary tests of the information extraction approach to context finding. In the experiments an information extraction system was trained on a dataset which included sentences containing the term of interest drawn from the British National Corpus (BNC). The resulting rules were applied to further datasets comprising sentences, some of which contained synonyms of the terms used to construct the training set.

The aim was to determine whether the information extraction rules could reliably identify synonyms. The hypotheses were

1. that the extracted context patterns for a particular tag would match synonyms of the tag term more often than synonyms of the other terms on which it was trained,
2. that the patterns would match single sense synonyms more reliably than multisense synonyms,
3. and that the patterns would match nouns with related senses more often than other nouns.

Four cases were considered, each being a vehicle for which the root term had only one main sense in English and for which WordNet listed a number of synonyms and hyponyms. These were “airship”, “bicycle”, “canoe” and “helicopter”. The training set comprised 400 sentences broken down as follows: 4 by 100 sentences randomly selected from BNC containing the terms “airship/s”, “bicycle/s”, “canoe/s” or “helicopter/s” in which the terms were tagged <kwa>, <kwb>, <kwc> and <kwh> respectively, plus 4 by 100 sentences randomly selected from BNC containing the words “a”, “the”, “is” and “are” in which the terms were not tagged. The former 400 sentences acted as negative examples. The tagged terms in the 400 positive examples were replaced with random strings of lowercase characters. This prevented the Amilcare from learning the terms themselves as rules rather than the context.

The testing files comprised untagged sentences drawn from BNC. The terms in the various test sets reflect different degrees of synonymy/hyponymy to the training terms and also include terms with single and multiple common meanings. WordNet was used to guide the selection of terms. The test sets are described below.

SYNAIRSHIP - 207 sentences containing synonyms of “airship”, namely, 100 sentences randomly selected from BNC containing the words “balloon/s”, 100 sentences randomly selected from BNC containing the word/s “zeppelin/s” and the 7 sentences in BNC containing the word “dirigible/s”.

SYNBICYCLE - 116 sentences containing synonyms of “bicycle”, namely 100 sentences randomly selected from BNC containing the word “bike/s”, and
the 16 sentences in BNC containing the words “push-bike/s”.

SYNCANOE - 131 sentences containing synonyms of “canoe”, namely, 100 sentences randomly selected from BNC containing the words “kayak/s”, the 24 sentences in BNC containing the words “outrigger/s”, the 7 sentences in BNC containing the word/s “pirogue/s”.

SYNHELICOPTER - 108 sentences containing synonyms of “helicopter”, namely the 90 sentences in BNC containing the word “chopper” with the part of speech tag NN1 (this excludes cases of “Chopper” used as a nickname), the 17 sentences in BNC containing the word “choppers” and the 1 sentence in BNC containing the word “whirlybird”.

CYCLE - 100 sentences randomly selected from BNC containing the words “cycle/s”. All the alternate meanings of “cycle” were retained in this dataset.

4TRANS - 400 sentences containing terms which are other vehicles, namely 4 by 100 sentences randomly selected from BNC containing the words “skateboard/s”, “bus/es”, “aeroplane/s” and “boat/s”.

ABCH - 400 sentences containing nouns which are not vehicles, namely 4 by 100 sentences randomly selected from BNC containing the words “alcohol”, “banana/s”, “cemetery/ries”, “harmony/ies”.

Relating these datasets to the hypotheses: “bike” and “kayak” are single sense synonyms and we would expect them to match the context patterns more often than the multisense synonyms “chopper” and “cycle”, similarly we would expect the other vehicles in the set 4TRANS, such as “skateboard” to match the context patterns more often than the other nouns in set ABCH, such as “banana”.

Amilcare was selected as the IE system (Ciravegna & Wilks 2003) for three reasons. First, the $(LP)^2$ algorithm, on which it is based, performances well compared with other IE algorithms (Ciravegna 2001). Second, Amilcare is a self contained IE

package with an interface for changing the settings of the algorithm, examining the generated rules etc. Third, we already have positive experience of incorporating the Amilcare API into other systems, e.g. MnM (Vargas-Vera et al. 2003).

Dataset	Tag	Actual	Synonym	Precision
SYNAIRSHIP	kwa	11	3	0.27
	kwb	5	0	0.00
	kwc	4	0	0.00
	kwh	7	2	0.29
SYNBICYCLE	kwa	2	0	0.00
	kwb	9	5	0.56
	kwc	1	0	0.00
	kwh	2	2	1.00
SYNCANOE	kwa	7	0	0.00
	kwb	5	1	0.20
	kwc	7	1	0.14
	kwh	2	0	0.00
SYNHELICOPTER	kwa	3	1	0.33
	kwb	3	0	0.00
	kwc	2	0	0.00
	kwh	5	1	0.20
CYCLE	kwa	3	1	0.33
	kwb	9	5	0.56
	kwc	3	0	0.00
	kwh	2	1	0.50
4TRANS	kwa	5	0	0.00
	kwb	19	6	0.32
	kwc	11	4	0.36
	kwh	15	4	0.27
ABCH	kwa	8	0	0.00
	kwb	11	4	0.36
	kwc	10	1	0.10
	kwh	10	0	0.00

Table 1. Amilcare results for the 7 test sets

Results

Results for the seven test sets are presented in Table 1. Pattern length, i.e. the number of lexical entities in a context, was 4, and all thresholds were set to 1.0, except the error threshold for kwa which was set to 0.8 to increase recall. For each tag

type the *Actual* number of strings that Amilcare tagged as a possible synonym is given, plus the number of these tags that were a *Synonym* of interest in that particular test set. For example in the SYNAIRSHIP test set 27 strings were tagged as synonyms of "airship" of which 2 were one of "balloon/s", "zeppelin/s" or "dirigible/s". *Precision = Synonym/Actual*. This is not always true precision as some of these values are incorrect assignments, e.g., an occurrence of "bike" tagged with kwa is a mistake.

Hypothesis 1 was that the extracted context patterns for a particular tag would match synonyms of the tag term more often than synonyms of the other three terms. This seems to be the case for the tag kwb ("bicycle/s"). There were more than 4 times as many kwb tags applied than any of the other 3 tags for the SYNBICYCLE dataset and more than half the cases were found correctly. However there is no evidence that the kwa and kwc tags are applied preferentially in the datasets SYNAIRSHIP and SYNCANOE.

Hypothesis 2 was that the patterns would match single sense synonyms more reliably than multisense synonyms. This is supported by comparing the results for tag kwb with the SYNBICYLE data set and those with the CYCLE dataset. "cycle/s" has more senses than "bike/s" and although it matched kwb with similar precision it made more assignments to other tags. This hypothesis is also supported by the poor reliability of the kwh patterns when applied to the SYNHELICOPTER dataset which is dominated by the multisense term "chopper/s".

Hypothesis 3 was that the patterns would match nouns of related sense more often than other nouns. There is some evidence to support this as the precisions of kwa, kwc and kwh are higher for the 4TRANS dataset than for ABCH.

Discussion

Overall the evidence to support our hypotheses was weak. However we had one promising case, the <kwb> "bicycle" tag which appeared to behave as expected. Also the patterns do seem to be somewhat better at detecting other transport related nouns than general nouns. When we bear in mind that Amilcare is a general purpose IE system whose rules are aimed at finding entities such as proper nouns, times, dates etc., its performance on this task, though modest, is not discouraging. A special purpose IE system, which concentrated on contextual rules and maybe incorporated some of the methods suggested by the NLP community, such as using information about modifying verbs would be expected to do much better on this task.

Furthermore these tests were limited in scale. The training set had only 800 examples, with 100 examples for each tag, and each "document" was only a single sentence. The training and test sets were generated automatically without any quality checking to ensure, for example, that the negative examples did not contain any examples of vehicles which, being untagged, would have conflicted with positive evidence of patterns. A tool such as Armadillo (Ciravegna et al. 2004) which can learn from a few hand picked, high quality examples may be more suitable for this task.

Finally, the terms used here are in common use. If terms like "bicycle" needed to be included in a lexicon it would be appropriate to get synonyms and hyponyms from

WordNet, as we did to guide the creation of the test sets. The real problem is to generate synonyms for specialist terminology, for which language resources such as WordNet do not exist. It is possible that the language in such technical texts gives stronger contextual patterns than the general purpose examples from BNC that we used here.

The work reported here has given us a clearer understanding of the problems of using IE methods to automatically generate and extend lexicons. Although the initial results are weak we believe that further, better designed experiments now underway may well confirm the suitability of IE methods for this application.

Acknowledgements

Many thanks to Fabio Ciravegna & José Iria for providing Amilcare and for advising us on how to improve the experimental procedures reported here.

References

Allegrini. P., Montemagni, S., Pirrelli, V. (2000), "Learning Word Clusters from Data Types." *COLING 2000*, 8-14.

Ciravegna, F., (2001) Adaptive Information Extraction from Text by Rule Induction and Generalisation, in Proc. of *17th International Joint Conference on Artificial Intelligence (IJCAI 2001)*, Seattle, August 2001.

Ciravegna, F., and Wilks, Y., (2003) Designing Adaptive Information Extraction for the Semantic Web in Amilcare, in S. Handschuh and S. Staab (eds), *Annotation for the Semantic Web*, in the Series Frontiers in Artificial Intelligence and Applications by IOS Press, Amsterdam.

Ciravegna, F., Chapman, S., Dingli, A., Wilks, Y. (2004) "Learning to harvest information for the semantic web". In *Proc. 1st European Semantic Web Symposium*, Heraklion, Greece, May 10-11.

Dzbor, M., Domingue, J. B., Motta, E. (2003) Magpie - towards a semantic web browser , In *Proc. of the 2nd Intl. Semantic Web Conference* , October 2003, Florida US

Gauch, S., Wang, J., Rachakonda, S.M., (1999) A corpus analysis approach for automatic query expansion and its extension to multiple databases, *ACM Transactions on Information Systems (TOIS)*, 17(3), pp. 250 – 269.

Grefenstette G. (1992) Use of syntactic context to produce term association lists for text retrieval, In *Proc. of SIGIR '92*, Denmark.

Riloff E., Jones R., (1999) Learning dictionaries for information extraction by multi-level bootstrapping, In Proc 16th National Conf. on Artificial Intelligence (AAAI-99)

Turney, P. (2001) Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In De Raedt, Luc and Flach, Peter, Eds. In *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)*, pages pp.491-502, Freiburg, Germany.

Vargas-Vera, M., Motta, E., Domingue, J., Lanzoni, M., Stutt, A. and Ciravegna, F. (2002) "MnM: Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup", In *13th International Conference on Knowledge Engineering and Management (EKAW 2002)*, ed Gomez-Perez, A., Springer Verlag.