The Automatic Acquisition of a Broad-Coverage Semantic Lexicon for use in Information Retrieval

Richard F.E. Sutcliffe*, Donie O'Sullivan,
Liam Sheahan and Annette McElligott

Department of Computer Science
and Information Systems
University of Limerick
Limerick, Ireland

Our interest is in developing techniques for constructing a semantic lexicon of broad coverage from machine tractable resources. We work within a paradigm in which each word sense is represented as a vector in an n-dimensional feature space. So far our experiments have encompassed the Merriam-Webster Compact Electronic Dictionary, the Irish An Foclóir Beag and the Princeton WordNet. Our main uses for the results are in full text information retrieval, machine assisted translation and lexical alignment. In this paper we explain the background to the approach, outline the algorithms with which we have been experimenting and report on the results which we have obtained.

1 Introduction

Every natural language processing (NLP) system has a requirement for lexical information. While there has been considerable progress in developing efficient lexical representations of morphological (Koskenniemi, 1983) and syntactic (Hellwig, 1980; Sleator and Temperley, 1991) information, attempts at constructing a wide-coverage lexicon of semantic information have met with considerable difficulty. First, it is very difficult to devise a general yet powerful semantic representation scheme. Meanings are hard to pin down. Second, even if such a scheme exists, it is not easy to create representations conforming to the paradigm for a large set of words. In this paper we present one way in which a broad and shallow lexicon can be developed using as a starting point a machine tractable hierarchical concept taxonomy. The method involves traversal of the taxonomy extracting terms which are then used as semantic features. We first define the kind of representation which we wish to produce — simple distributed patterns where comparison of meaning can be reduced to the dot product operation. Next the basis of the traversal algorithms and post processing is outlined. In the third section we describe a psychometric experiment in which we attempt to measure the efficacy of our representations. Finally, we outline one application of the lexicon, a full-text information retrieval system which operates in the domain of technical software manuals.

*The help of Tony Molloy, Redmond O'Brien and Gemma Ryan is gratefully acknowledged. The first author's email is 'sutcliffe@ul.ie'

2 Distributed Representations

Our goal in this research is to investigate the extent to which linguistic meanings can be captured by distributed patterns. The task of capturing all aspects of lexical meaning is of course very complex (Pustejovsky, 1991). One simple approach such as we describe here can not possibly solve every representational problem. However, a semantic lexicon based around distributed patterns has many advantages over rival approaches and therefore appears to justify careful investigation.

Each lexical entry in our approach consists of a list of <attribute, value> pairs. An attribute is a semantic feature such as 'large' or 'animate'. The

```
[[microscopic,0.24], [plastic,0.24], [magnetic,0.24], [disk,0.24],
[enclose,0.24], [stiff,0.24], [envelope,0.24], [radial,0.24], [incision,0.24],
[use,0.24], [store,0.24], [datum,0.24], [program,0.24],
[personal_computer,0.24], [fabricate,0.16], ['particular(a)',0.16],
[intention,0.16], [artifact,0.14], [embody,0.14], [instrumental,0.14],
[achieve,0.14], [end,0.14], [object,0.12], [entity,0.09], [have,0.07],
[concrete,00.07], [being,0.07], [dwell,0.07]]
```

Figure 1. The semantic representation for floppy_disk produced by the MkII traversal algorithm.

value is a real number between 0 and 1 which reflects the extent to which the feature contributes to the meaning of the concept. Each such pattern can be considered as a vector in an n-dimensional space where n is the total number of features in the system. Thus if each representation is normalised (scaled so that its length is one) a pair of word meanings can be compared using the dot product. If two words word1 and word2 are identical in meaning, this dot product will be 1. On the other hand, if the meaning of word1 has nothing in common with that of word2. then the value will be 0.

The distributed nature of these representations arises because the concept meaning is captured by the set of <attribute,value> pairs as a whole rather than by one piece of information (Hinton, McClelland and Rumelhart, 1986). The most useful ramification of this from our perspective is that the representations can contain a certain amount of incorrect or contradictory data while still showing a good level of performance. This turns out to be an asset when a large lexicon is to be constructed automatically.

A number of projects have constructed distributed lexical representations of the kind we have outlined. Wilks et al. extract features from glosses in the Longmans Dictionary of Contemporary English (Wilks, Fass, Guo, MacDonald, Plate and Slator, 1990). On the other hand Schuetze uses statistical data derived from a parallel text corpus (Schuetze, 1993) while McMahon (1994) has developed a monolingual corpus method. Finally, Nitta and Niwa experiment with both dictionary and corpus methods (Nitta and Niwa, 1994).

3 Constructing the Representations

The family of methods with which we have been working all operate on similar principles. It is necessary to start with a hierarchical concept taxonomy which must either be constructed by hand or be derived from existing machine-tractable resources (Chodorow, Byrd and Heidorn, 1985). The objective is to extract terms from the taxonomy which can be used as features. The algorithm starts at the word to be defined and extracts features from its gloss. It then moves to the next level up in the taxonomy and extracts features from the gloss corresponding to the concept at this level. The procedure is repeated until a finishing condition arises — usually until the top of the taxonomy is reached. The features extracted at a particular level are all accorded the same centrality (strength). This centrality decreases as the taxonomy is ascended.

Once the basic traversal has been performed, a post-processing phase is carried out. This aims to optimise the representations produced in the first phase by collapsing certain sets of features into one feature in order to make the patterns more distributed.

Our initial traversals worked only with nouns and extracted adjectives from a taxonomy constructed by rudimentary analysis of glosses. The dictionaries used were the Merriam-Webster Compact Electronic (Sutcliffe, 1993) and the Irish An Foclóir Beag (Sutcliffe, McElligott and Ó Néill, 1993). Subsequent work has been with the Princeton WordNet (Beckwith, Fellbaum, Gross and Miller, 1992). In the MkI WordNet traversal, there was no post-processing (Sutcliffe, O'Sullivan and Meharg, 1994). The MkII traversal carried out a form of feature reduction which ensured that intrinsic synonyms (members of the same WordNet synset) were only represented by one feature (Sutcliffe, O'Sullivan, Slater and Brehony, 1994). For a sample lexical entry, see Figure 1.

It is important to note that there is a close correlation between a semantic distance measure based on patterns constructed by traversing a taxonomy, and more direct similarity measures based on tax-

onomies. For example Richardson (1994) computes semantic similarity by a form of intersection search in which the hierarchy is ascended from the two concepts being compared until a common node is found. Such methods can be traced back to the work of Quillian (1968). With our approach, all features above the point at which the two taxonomies join will be shared between two concepts for which such a join occurs. The lower down this intersection occurs, the greater the number of shared features. Thus even if the features we extract are entirely erroneous we will still obtain a 'Quillian effect'. Where our approach might gain over such methods is in its ability to obtain correlations which are not directly related to taxonomic relationships.

4 Measurement of Performance

Having constructed a lexicon the next step was to assess its performance. An initial attempt at measuring the performance of the MkI WordNet traversal involved computing the dot products of selected word senses and seeing whether the results conformed to our intuitions. This exercise suggested that one possible way of testing the lexicon was to compare the results of such pairwise associations with those judged by human subjects. An experiment was conducted to investigate this idea and we report on it briefly here.

While the use of the dot product yields a value between zero and one when applied to any pair of semantic representations in the lexicon, there are difficulties when human subjects are asked to make the same kinds of judgement. One particular problem is that while two concepts may not be linked by their literal meanings, there might nevertheless by an indirect connection between then. For example if a person is asked to state how strongly 'lapdog' is related to 'society' they might start to consider the role of dogs in society and thus respond with a high value. In order to restrict subjects to literal meanings, a psychometric questionnaire was designed which asked the respondent to rank a set of ranking words by their degree of match with a particular target word. Figure 2 shows a typical task. The target word is 'server'. The subject must decide which of the ranking words 'document', 'phrase' etc. is closest in meaning to 'server' and give this word a rank of 1. Next, a rank of 2 must be assigned to the remaining ranking word which is closest to 'server'. Ranks 3, 4 and 5 are given to the other words by an analogous process. Example ranks are shown in the figure. The 'lapdog-society' problem is avoided in this approach because we ensure that the set of ranking words contains at least one direct match, relegating indirect matches to lower ranks. For example if 'lapdog' is the target and 'society' and 'hound' are both ranking words, 'hound' is likely to obtain a better rank than 'society'.

An experiment was conducted in which 60 computer terms were extracted from the The *Lotus(R) Ami Pro for Windows User's Guide Release 3.0* and associated with word senses in WordNet. A set of ten ranking tasks was devised, each of the form shown in Figure 2. 30 copies of a questionnaire were run off, each containing the ten tasks. The order of the tasks was randomised for each new questionnaire and the order of the ranking words in each task was also randomised each time. The questionnaires were distributed to staff and postgraduates in the Computer Science and Information Systems Department at Limerick.

The results of the questionnaire were collated and for each <target word, ranking word> pair the mean rank was computed. Next, the lexicon derived from the MkII WordNet traversal 'undertook' the study. In each ranking task the semantic match of the target word with each ranking word was computed. These matches were then used to arrange the ranking words in decreasing order of match, and hence to rank them from 1 to 5.

Finally, the performance of the lexicon was measured by computing the Spearman Rank Correlation of the set of <target word, ranking word> ranks produced by the lexicon with the set of mean ranks obtained from the human subjects. A value of 0.86 was obtained which shows that the mean ranks of the subjects and those of the lexicon were significantly correlated at the $p < 0.01$ level. The essential finding of this study, therefore, is that the semantic representations derived from the lexicon appear to capture some aspects of linguistic meaning as indicated by a set of human subjects. For full details of the study, see Sutcliffe, O'Sullivan, Sharkey, Vossen, Slator, McElligott and Bennis (1994).

```
Target: server (Noun) [filestore accessible from remote machines]

       document [ 3 ] [a text providing information]
         phrase [ 4 ] [group of words]
    workstation [ 2 ] [powerful computer]
         dagger [ 5 ] [character used in printing]
      disk drive [ 1 ] [computer disk drive]
```

Figure 2. An example ranking task from the questionnaire. The numbers are typical figures which the subject might write in when undertaking the question.

5 SIFT - A Sample Application

In this section we briefly outline one application of the lexicon which we have been discussing. It is SIFT-1, a system which takes as input a user query concerning a piece of Microsoft Windows(R) software and attempts to locate parts of a technical user manual which address that query (Sutcliffe, Hellwig, Vossen, O'Sullivan, Slater, Brehony and Sheahan, 1994). Many of the ideas are derived from an earlier project (Sutcliffe, 1991).

The SIFT system consists of two principal components. Firstly, the *document processing component* takes as input an SGML tagged computer manual and associates with the each section, subsection and individual sentence of that manual a representation capturing its meaning. Secondly, the *query processing component* takes in a user query concerning the material covered in the document and produces as output a list of pointers to text portions within it which are ordered by relevance to the query.

The document processing component operates as follows. Each utterance in the input manual is disambiguated and tagged for syntactic category. Each content word (noun or verb) is then converted into a distributed pattern, using the MkII WordNet traversal outlined earlier. The representation stored with the utterance is thus the list of <semantic pattern, syntactic category> pairs for these content words.

When the system is in operation, an input query is first converted into a list of <semantic pattern, syntactic category> pairs exactly as described above. This list is then matched with the representations corresponding to utterances in the text. Portions of the manual which contain good utterance matches are then made available to the user for perusal.

SIFT aims to show the potential of combining tractable semantic representations with conventional retrieval technology. While the utterance representations being used only capture some aspects of sentence meaning, they are a great deal more expressive than the simple keywords of traditional approaches. In particular, it is not necessary to choose exactly the right keyword in order to retrieve the right passage.

6 Conclusions

In this article we have attempted to present the case for a particular kind of semantic lexicon in which each word meaning is represented as a set of <attribute, value> pairs. Such a lexicon can be constructed by one of a family of algorithms which traverse a hierarchical taxonomy of concepts. We have outlined a psychometric experiment which aimed to investigate whether this kind of lexicon reflects people's intuitions about word meaning. In addition we have described one application of the lexicon — full text information retrieval.

In conclusion, we consider that simple distributed patterns of the kind discussed here have considerable potential application in the NLP systems of the future. They are robust, can be constructed automatically and are easy to process. On the other hand, no one would claim that this kind of representation can address all problems in semantics. However, it is naive to think that any paradigm can achieve this. We believe, therefore, that future research should concentrate on broad and shallow approaches of this kind rather than concentrating on the narrow and deep coverage (with no potential for generalisation) which has formed the basis of so much work in the past.

7 References

Beckwith, R., Fellbaum, C., Gross, D., & Miller, G. A. (1992). WordNet: A Lexical Database Organised on Psycholinguistic Principles. In U. Zernik (Ed.) *Using On-line Resources to Build a Lexicon.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Chodorow, M, Byrd, R., & Heidorn, G. (1985). Extracting Semantic Hierarchies from a Large On-Line Dictionary. *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics, Chicago, Illinois*, 299-304.

Hellwig, P. (1980). PLAIN - A Program System for Dependency Analysis and for Simulating Natural Language Inference. In L. Bolc (Ed.) Representation and Processing of Natural Language (271-376). Munich, Germany, Vienna, Austria, London, UK: Hanser & Macmillan.

Hinton, G. E., McClelland, J. L. and Rumelhart, D. E. (1986). Distributed Representations. In D. E. Rumelhart and J. L. McClelland (Eds) Parallel Distributed Processing: Explorations in the microstructure of cognition. Volume I: foundations (pp. 77-109). Cambridge MA: MIT Press.

Koskenniemi, K. (1983). Two-Level Morphology for Morphological Analysis. In *Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI-83*, 683-685.

McMahon, J. G. G. (1994). *Statistical Language Processing based on Self-Organising Word Classification.* Doctoral Dissertation, Department of Computer Science, Queen's University Belfast.

Nitta, Y., & Niwa, Y. (1994). Co-Occurrence Vectors from Corpora vs. Distance Vectors from Dictionaries. *Proceedings of the 15th International Conference on Computational Linguistics, 5-9 August, 1994, kyoto, Japan*, 304-309.

Pustejovsky, J. (1991). The Generative Lexicon. *Computational Linguistics*, 17(4), 409-441.

Quillian, M. R. (1968). Semantic Memory. In M. Minsky (Ed.) *Semantic Information Processing* (pp. 216-270). Cambridge, MA: MIT Press.

Richardson, R., Smeaton, A. F., & Murphy, J. (1994). Using WordNet as a Knowledge Base for Measuring Conceptual Similarity between Words. *Proceedings of the Seventh Annual Irish Conference on Artificial Intelligence and Cognitive Science, Trinity College Dublin, 8-9 September, 1994.*, 179-192.

Sleator, D., D., K., & Temperley, D. (1991). *Parsing English with a Link Grammar* (Technical Report CMU-CS-91-196). Pittsburgh, PA: Carnegie Mellon University, School of Computer Science.

Sutcliffe, R.F.E. (1991). Distributed Representations in a Text Based Information Retrieval System: A New Way of Using the Vector Space Model. In A. Bookstein, Y. Chiaramella, G. Salton & V. V. Raghavan (Eds.) The Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, Chicago, Il., October 13-16, 1991 (pp. 123-132). New York, NY: ACM Press.

Sutcliffe, R. F. E., Hellwig, P., Vossen, P., O'Sullivan, D., Relihan, L., McElligott, A., Slater, B., Brehony, T., & Sheahan, L. (1994). Retrieval from Software Instruction Manuals by Semantic Skimming. Technical Report UL-CSIS-94-17, Department of Computer Science and Information Systems, University of Limerick, December 1994.

Sutcliffe, R. F. E., O'Sullivan, D., & Meharg, F. (1994). A Lexicon of Distributed Noun Representations Constructed by Taxonomic Traversal. *Proceedings of the 15th International Conference on Computational Linguistics, (COLING'94), Kyoto, Japan*, 827-831.

Sutcliffe, R. F. E., O'Sullivan, D., Slater, B. E. A., & Brehony, T. (1994). Traversing WordNet to Create Optimised Semantic Lexical Representations. In Proceedings of the Seventh Annual Irish Conference on Artificial Intelligence and Cognitive Science (AICS'94), Trinity College Dublin, 8-9 September, 1994.

Sutcliffe, R. F. E., O'Sullivan, Sharkey, N. E., D., Vossen, P., Slator, B. E. A., McElligott, A., & Bennis, L. (1994). A Psychometric Performance Metric for Semantic Lexicons. *Proceedings of International Workshop On Directions Of Lexical Research, 15-17th of August, 1994, Beijing, China.*