From SIFT Lexical Knowledge Base to SIFT Lexical Data Base: Creating a Repository for Lexicological Research and Development

Richard F. E. Sutcliffe*, Piek Vossen#, Iskandar Serail#,
Peter Masereeuw#, Peter Hellwig&, Paul Boersma#,
Annelies Bon#, Annette McElligott*,
Donie O'Sullivan* and Liam Sheahan*

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

sutcliffer@ul.ie

+353 61 333644
+353 61 330876 (Fax)

Computer Centrum Letteren#
University of Amsterdam
Amsterdam, The Netherlands

University of Heidelberg,&
Karlstrasse 2,
D-6900 Heidelberg,
Germany

LINGDB'95 Topic areas:

3. Developing (maximally) theory-neutral db schemas for annotation systems.
6. Needs of applications such as lexicography.

What is SIFT?

We have developed a paradigm which allows word meanings to be captured using distributed patterns, that is lists of <feature,centrality> pairs. These representations are generated automatically from machine readable dictionaries by a family of taxonomic traversal algorithms. The meaning of two word senses can be directly compared in this paradigm, yielding a value between zero and one.

In the LRE-2 SIFT project we are developing a system for text retrieval from Technical Manuals. The system takes as input a query relating to the content of the manuals. This is converted to a distributed pattern which is then "swept" over the patterns corresponding to units of the text looking for a good match. Salient portions of text are then shown to the user for perusal. The SIFT-1 prototype is virtually complete. In SIFT-2, representations will capture semantic case information derived from queries and text utterances via PLAIN parse trees (Hellwig, 1980, 1989).

The SIFT LDB

The project has necessitated the development of a database for containing word-based linguistic information required either for the SIFT systems themselves or for the research work which is entailed by their construction. The lexical data is mainly obtained from the Princeton WordNet (Beckwith, Fellbaum, Gross and Miller, 1992) and the Longman Dictionary of Contemporary English (LDOCE) (Proctor, 1978), the latter in the form of lexical data created following experience gained in the ACQUILEX project (Vossen, 1990a, 1991a, 1991b).

The work has raised the following fundamental issues:

* in the context of text retrieval, what exactly is a lexical database?

* what can and can not be represented in such a repository?

* how can such a database be best engineered to allow maximum flexibility in the use of the information it contains?

The activities which we wish to carry out using the LDB are as follows:

* the indexing of documents using word meanings and related data, carried out as an "offline" process,

* the retrieval process itself, involving construction in real time of a meaning representation for the input query,

* the development and testing of taxonomic traversal algorithms used in the construction of the distributed meaning representations,

* the construction and maintenance of terminological data associated with the application domain (Lotus Ami Pro).

* the organisation and maintenance of lexical information required for building the syntagmatic parsing lexicon required for PLAIN.

Progress

So far, a lexical database has been built which provides very efficient access to unstructured data (Masereeuw, 1994). It is engineered to allow use of the same database and access software on the Sun, PC and Macintosh computers. We are currently investigating how best to structure the data in order to support the activities outlined above.

Major issues which have arisen so far are:

* for each type of data, whether to use native binary formats (which are fast) or Prolog-based formats (which are flexible and convenient)

* how to determine and maintain links between material derived from different sources (e.g. LDOCE and WordNet) and intended for widely different purposes (e.g. parsing and lexical semantic representation)

Contribution to the Workshop

The SIFT project raises many questions relating to lexical databases which are relevant to other researchers in computational lexicology, information retrieval and related areas. We would propose to describe our SIFT findings and to explain potential applications of the ideas in other domains.

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.

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.

Hellwig, P. (1989). Parsing natuerlicher Sprachen: Grundlagen, Realisierungen. In S. B. Tori, W. Lenders & W. Putschke (Eds.) Computational Linguistics. An International Handbook on Computer Oriented Language Research and Applications (Handbooks of Linguistics and Communication Science, Vol. 4) (pp. 348-431). Berlin, Germany: de Gruyter.

Masereeuw, P. C. (1994). LTREE and CM: The SIFT LDB Developer's Libraries (SIFT deliverable D10). Amsterdam, The Netherlands: University of Amsterdam, Computer Centrum Letteren.

Proctor, P. (Ed.) (1978). The Longman Dictionary of Contemporary English (LDOCE). London, UK: Longman.

Salton, G. (Ed.) (1971). The SMART Retrieval System - Experiments in Automatic Document Processing. Englewood Cliffs, NJ: Prentice Hall.

Sutcliffe, R. F. E. (1993). Constructing Distributed Semantic Lexical Representations using a Machine Readable Dictionary. In K.T. Ryan and R. F. E. Sutcliffe (eds) Proceedings of AICS-91 - The Fifth Irish Conference on Artificial Intelligence and Cognitive Science, University of Limerick, 10-11 September 1992. London, UK, Berlin, FRG, Heidelberg, FRG, New York, NY: Springer-Verlag

Sutcliffe, R. F. E., McElligott, A., O Neill, G. (1993). Using Distributed Patterns as Language Independent Lexical Representations. In Proceedings of the AAAI-93 Spring Symposium Series: Building Lexicons for Machine Translation, March 23-25, 1993, Stanford, University, CA.

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.

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

Vossen P. (1990a). A Parser-grammar for the Meaning Descriptions of the Longman Dictionary of Contemporary English. (Tech. Rep. NWO, project no. 300-169-007). Amsterdam, Netherlands: University of Amsterdam.

Vossen, P. (1990b). The end of the chain: where does stepwise lexical decomposition lead us eventually? In Proceedings of the 4th Functional Grammar Conference, June 1990, Kopenhagen, Denmark. Also available as Acquilex Working Paper No. 010, July 1990 , Esprit BRA-3030, University of Amsterdam.

Vossen, P. (1991a). Converting data from a lexical database to a knowledge base. Acquilex Working Paper no 027, Esprit BRA-3030, November 1991, Amsterdam.

Vossen, P. (1991b). An empirical approach to automatically construct a knowledge base from dictionaries. Paper to be presented at The Euralex conference, Tampere, 1992 Also available as Acquilex Working Paper no. 25, Esprit BRA-3030, Amsterdam.

Vossen, P. (1991c). Polysemy and vagueness of meaning descriptions in the Longman dictionary of contemporary English. In S. Johansson & A. Stenstrom (Eds.) English Computer Corpora. Selected Papers and Research Guide. Berlin, Germany: Mouton de Gruyter. Also available as Acquilex Working Papers no. 1, Esprit BRA-3030, Amsterdam, 1989.

Vossen, P., Meijs, W., & Broeder, M. den (1989). Meaning and structure in dictionary definitions, In B. Boguraev and E. G. Briscoe (Eds.) Computational Lexicography for Natural Language Processing. London, UK: Longman.

Vossen P., & Copestake, A. (Forthcoming, 1992). Untangling Definition Structure into Knowledge Representation. To appear in E. G. Briscoe, A. Copestake and V. de Paiva (Eds.) Default Inheritance in Unification Based Approaches to the Lexicon. Cambridge, UK: Cambridge University Press. Also in Proceedings of the ACQUILEX workshop on Default Inheritance in the Lexicon, April 1991, Cambridge.

Vossen, P., & Serail, I. (1990). Devil: a taxonomy browser for decomposition via the lexicon. Esprit BRA-3030 Acquilex WP No. 009, University of Amsterdam, Amsterdam.