Discourse Semantics for the Analysis of Change in Language

Jon D. Patrick and Stephen Anthony

Basser Department of Computer Science
The University of Sydney, Australia
{santhony, jonpat}@cs.usyd.edu.au

Abstract

This paper presents a semantic processing framework that offers a new approach to the traditionally problematic knowledge acquisition bottle-neck. The model presented here elucidates the advantages of adopting an interchangeable modular pipeline design of language engineering systems. We argue that a modular design more readily facilitates the automatic acquisition of world knowledge and therefore aids the analysis and comparison of texts. We apply our approach to the identification of semantic change in natural language texts. The domain chosen as a test case is that of discourse in psychotherapy interviews.

1. Introduction

Conservative estimates suggest that the amount of data currently generated by humanity per annum is in the order of exabytes (10^{18}). The most effective and possibly only feasible means by which to deal with such vast ever-expanding quantities of data is to automate the task of knowledge acquisition and reasoning. In light of such motivations, ratification of common representation and content standards for ontologies and lexical resources is becoming an increasingly prominent area of interest. Initiatives such as the World Wide Web Consortium's Resource Description Framework (RDF), the IEEE Standard Upper Ontology Working Group, the Text Encoding Initiative (TEI) (Sperberg-McQueen and Burnard 1994), Conceptual Graphs (CGs) **(**Sowa 1984**)**, and the Knowledge Interchange Format (KIF) (NCITS T2 1998) all aim to facilitate automatic knowledge interchange and aid collaborative research.

In recent years there has been growing interest in the production of digital repositories for human language technologies. A great deal of effort has been directed towards the production of robust digital repositories of structured information designed to document and facilitate natural language interaction. Such efforts include Wordnet **(**Miller 1990**)**, Cyc (Lenat and Guha 1990), VerbNet (Kipper, Dang et al. 2000), Mikrokosmos **(**Beale, Nirenburg et al. 1996**)**, Sensus (Knight, Hovy et al.), Mindnet (Richardson, Dolan et al. 1998), and Protégé-2000 (Noy, Fergeson et al. 2000). Such digital repositories take the form of large-scale lexicons, ontologies and knowledge bases. However, the usefulness of these resources has been significantly diminished predominantly through lack of consensus and consistency with regard to representation and content. The establishment of common formats and content standards for ontologies and lexical resources will allow reuse of results between researchers and better enable the automatic acquisition of lexical knowledge. Robust language processing architectures that utilise such efforts without subscribing to the vagaries of any particular methodology or resource have proven elusive. Such

platforms allow for the facilitation of automated reasoning and inference which may be embedded in applications such as intelligent user agents, information retrieval, decision support systems, knowledge discovery, information extraction, machine translation, speech recognition, and document summarisation.

2. A Model for an Ontolexicon

Our formalism utilises the W3C Resource Description Framework to encode concepts that exist in the world and the relationships that hold between them. WordNet is a lexical database and is the most commonly used source of such relations. WordNet currently contains approximately 120,000 sets of noun, verb, adjective, and adverb synonyms. These sets of synonyms are known as synsets and are interrelated by means of semantic links such as hypernymy, synonymy, antomyny, and meronymy. However, knowledge of such relations between words and concepts are only one aspect of a speaker's lexical knowledge. Speakers also possess knowledge of the syntactic properties of the words in their language. For this reason we also use VerbNet to represent syntactic properties such as predicate-argument structure. VerbNet is a verb lexicon that uses the Levin verb classes (Levin 1993) to systematically construct lexical entries (Kipper, Dang et al. 2000). We see the combination of these two resources as a reasonably powerful building block from which to construct a lexical-semantic processing platform. Deep semantic knowledge is required when processing change in language. Lexical knowledge bases are extremely useful as they support deep, knowledge-intensive processing of language.

Lexical semantics is the study of word meanings and the representation of word meanings in the lexicon. Post-Chomskian semantics focused almost entirely on sentence meaning. More recent research in lexical semantics has restored word meaning as the centre of interest and shifted the theoretical thrust towards generativity. These efforts seek to ground lexical semantics in generative syntax through the use of lexical rules like those used in the Generative Lexicon (Pustejovsky 1995). (Boguraev and Levin 1990; Viegas 1999) expound compelling arguments for the move from static to active lexical data and knowledge bases.

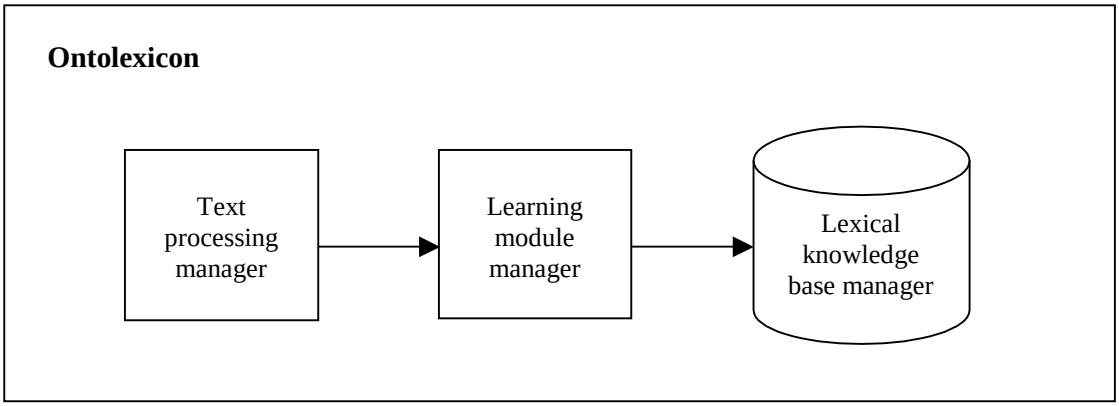

Figure 1: Framework Overview

Our framework incorporates three interrelated processing management modules. Figure 1 illustrates the three overriding processing components. The text processing module begins with unannotated natural language text. This module allows the user to

specify the level and amount of annotation required. Users may choose to include any combination of lexical information as part of the output. An interchangeable machine learning module is used to automatically infer lexico-semantic relations. As shown in Figure 2, the machine learning module manager is tightly integrated with the knowledge base manager that is responsible for the update of the lexical knowledge base. By accommodating a flexible text processing strategy we are able to examine the benefits of a variety of different language processing components.

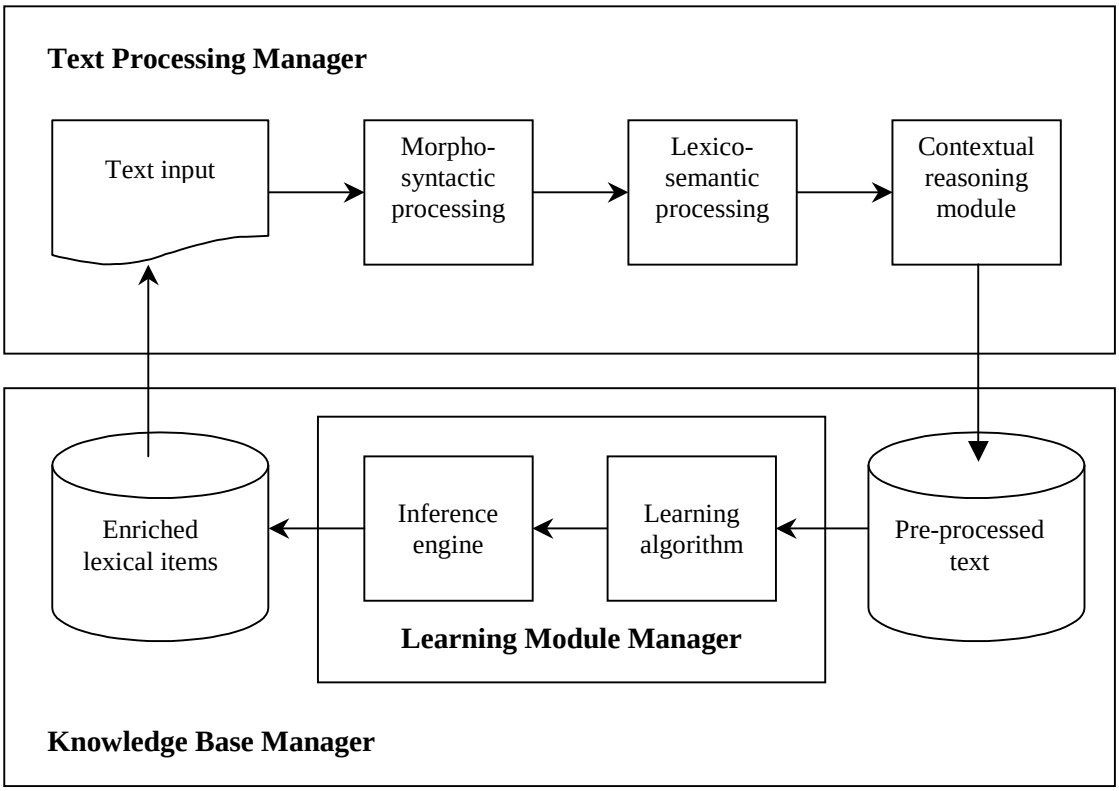

Figure 2: Exchangeable Modules Pipeline Framework

Whilst our system adopts a holistic (depth through breadth) knowledge-intensive approach to language processing the architectural design accommodates exchangeable processing modules. The modular framework allows users to investigate the effects that different processing strategies and theories of meaning have on knowledge acquisition and understanding. We argue that a monolithic system is not possible for language engineering and a pipeline of exchangeable-modules model will produce more tailored and productive systems faster.

Early efforts such as Wordnet and Cyc were developed using time-intensive handcrafted representations. Our strategy adopts a semi-automated approach to knowledge acquisition that endeavours to exploit previous efforts and readily integrate future developments. Initial considerations are strategically geared towards semi-automation in order to accommodate human guidance in the knowledge acquisition process. A semi-automated approach readily facilitates the reintegration of feedback which consequently produces more meaningful representations and more effective knowledge discovery mechanisms.

By initially encoding existing electronic resources we are able to bootstrap our system to a point where automatic knowledge acquisition becomes feasible. Over time the knowledge base will be expanded to include more specialized or less common concepts mined from corpora. The primary goal at this stage is to specify the minimum set of defining concepts, semantic relations, and axioms for the knowledge base to become self-sufficient.

Many polysemous words have vague meanings that are highly dependant on context. For such words the compilation of a comprehensive list of all meanings is not possible. The approach adopted here aims to encode dominant characteristics of word meanings through the combination of thematic roles, semantic predicates, syntactic frames, and selectional restrictions. The aim is to dynamically acquire lexical entries through the application of lexical-sematic rules. Our preferred nomenclature is ontolexical rules. We believe this term more clearly explicates the underlying notion used here; that is the acquisition of lexical knowledge guided by world knowledge, or what we have called the ontolexicon. The interface between the ontolexicon and the processing modules allows for modules to be readily interchanged. Each of the modules effects their operations on the text meaning representation. In turn enriching both the text meaning representation and the ontology. These operations are guided by semantic and world knowledge contained within the ontolexicon.

3. Identifying Language Change to Enhance an Ontolexicon

Our aim is to build a language technology tool that updates its own lexicon and ontology, that is, learns as it analyses texts. We wish to test this proposal in a domain where language change and hence learning about change in language is important. We have assembled a model of a language tool that is based on a highly configurable modular platform that will be used to discover lexico-semantic relationships between items in text which in turn updates the lexical and ontological stores. Hence we have chosen to work with interviews of psychotherapy clients, where it is postulated that change in their language over time is indicative of successful therapy. If our system can identify these changes and incorporate them into its own knowledge stores then we have achieved our goals. The theoretical underpinning of the language change model of successful therapy can be traced back to (Sapir 1927) and has been variously expanded on by (Pittenger, Hockett et al. 1960; Labov and Fanshel 1977; Ferrara 1988, 1994; Lentine 1988; Parker 1995). The most elaborated form of this proposal comes from (Bandler and Grinder 1975).

Our understanding of the world heavily relies upon prior knowledge of causal relationships. The way in which actions change our beliefs of the world is viewed as a fundamental building block to understanding **(**Pearl 2000**)**. Context specific and pragmatic considerations play an indispensable role in the semantic interpretation of communicative acts. Consequently, it is important to identify and represent exactly what effects changes have on our view of the world.

The particular model of therapy used here is the Metamodel (Bandler and Grinder 1975). The Metamodel theorises that a client's language usage presents a distorted version of their experiences held in their deep structure through three primary mechanisms, namely generalisation, deletion, and distortion. Facilitating the presentation of a surface structure that is more closely representative of their deep

structure using the explicitly defined techniques of the Metamodel the therapist aims to assist the client to enrich an otherwise impoverished model of the world.

> The suggestion is that the function of the brain and nervous system and sense organs is in the main eliminative and not productive. Each person is at each moment capable of remembering all that has ever happened to him and of perceiving everything that is happening everywhere in the universe. The function of the brain and the nervous system is to protect us from being overwhelmed and confused by this mass of largely unused and irrelevant knowledge, by shutting out most of what we should otherwise perceive or remember at any moment, and leaving only that very small and special selection which is likely to be practically useful. According to such a theory each one of us is practically Mind at Large … (Bandler and Grinder 1975).

The working hypothesis is that whilst communicating humans present their models of the world in turn projecting any impoverishments that may be contained within their world-view through their language constructs. The process of psychotherapeutic change begins with the detection of impoverishments and then assisting the client by placing them in the position of active self-discovery. The therapist thereby aims to introduce changes into a client's model of the world, this allows them more options in their behaviour.

4. Experiments

We have commenced this work by hand analysis of a corpus. The corpus is a collection of thirty minute psychotherapy interviews taken before, during, and three to six months after therapy for each of the ten clients. The particular language phenomena in which we are interested fall within the three more general categories of generalisation, distortion, and deletion. These are non-specific nouns, non-specific verbs, universal quantifiers, nominalisation, clearly-obviously type, deleted referential index, comparatives, superlatives, modal operators, cause and effect, mind reading of others, mind reading by others, lost performative, and passive voice. Operational definitions, hand tagging criteria, enhanced descriptions based on Systemic Functional Linguistics principles, and potential methods for automatic tagging of these phenomena have been defined. Previous hand tagging of our data has proven support for such claims with seven out of ten clients showing significant positive change, two showing no change and one showing a retrograde change (Patrick 1999).

A statistical study was performed using the results of our hand-tagged data. In this study we compared the significance of counts of the language phenomena of interest taken before, during, and after therapy. The tag counts used in this analysis have been normalised by the length of the client text. As shown in Figure 3 below, the results of our statistical analyses show significant change has been detected in the clients' language. Three out of the four transcripts that were analysed have produced chi-squared values well below the conventional five percent threshold indicating significant change has occurred. These results provide encouraging support and evidence for our empirical study to proceed.

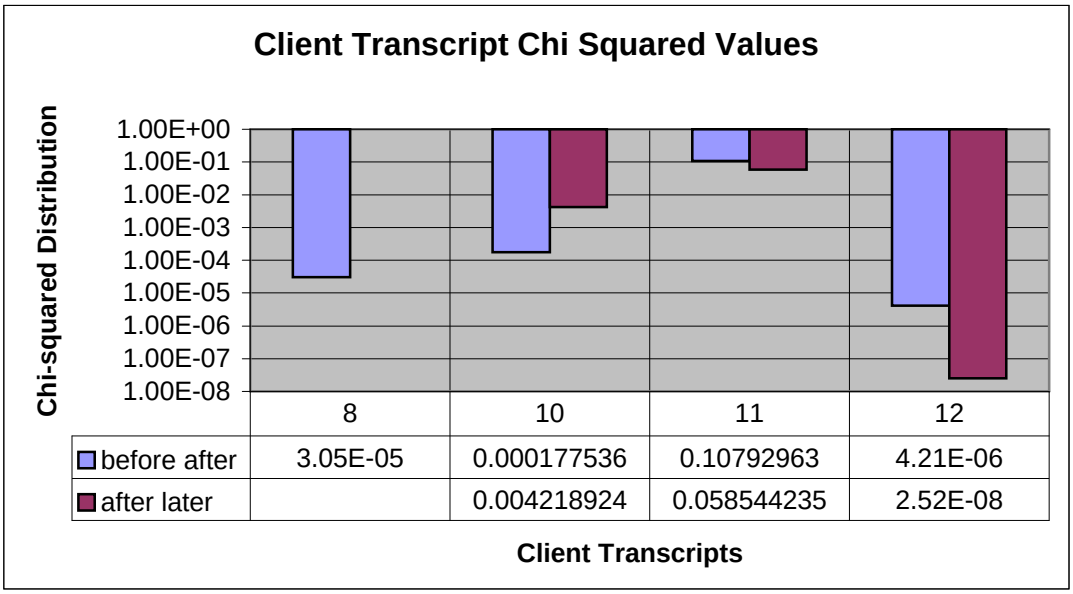

Figure 3: Chi Squared Anaysis

5. Conclusions

We have developed an architectural framework for a language understanding system that is being used to analyse the change in language usage over time. The initial stage has identified a test bed by which to apply and test the framework. The domain which has been chosen as a basis for experimentation is that of change of language usage throughout psychotherapy discourse. The particular linguistic phenomena representing the semantic change of interest have been operationalised. Statistical analysis has proven significant change in discourse may be identified using the defined linguistic phenomena discussed in this paper. Experiments have been initiated in order to automatically identify the linguistic phenomena of interest.

References

Bandler, R. and J. Grinder (1975). The structure of magic: a book about language and therapy. Palo Alto, Calif., Science and Behavior Books.

Beale, S., S. Nirenburg, et al. (1996). Semantic analysis in the mikrokosmos machine translation project. Proceedings of the 2nd Symposium on Natural Language Processing.

Boguraev, B. and B. Levin (1990). Models for lexical knowledge bases. In Proceedings of the 6th Annual Conference of the UW Center for the New OED, Waterloo.

Ferrara, K. (1988). Variation in Narration: Retellings in Therapeutic Discourse. Linguistic Change and Contact. K. Ferrara, B. Brown, K. Walters and J. Baugh. Austin, Texas, University of Texas.

Ferrara, K. (1994). Therapeutic Ways with Words. New York, Oxford University Press.

Kipper, K., T. H. Dang, et al. (2000). Class-Based Construction of a Verb Lexicon. AAAI: 691-696.

Knight, K., E. Hovy, et al. Ontology Creation and Use: SENSUS, USC Information Sciences Institute. http://www.isi.edu/natural-language/resources/sensus.html. Last accessed 2nd of September, 2001.

Labov, W. and D. Fanshel (1977). Therapeutic Discourse: Psychotherapy as conversation. New York, Academic Press.

Lenat, D. B. and R. V. Guha (1990). Building large knowledge-based systems: representation and inference in the Cyc project. Reading, Mass., Addison-Wesley Pub. Co.

Lentine, G. (1988). Metaphor as Cooperation in Therapeutic Discourse. 16th Annual Conference on New Ways of Analyzing Variation: Linguistic Change and Contact. Austin, Texas, University of Texas, Dept of Linguistics.

Levin, B. (1993). English verb classes and alternations: a preliminary investigation. Chicago, University of Chicago Press.

Miller, G. (1990). WordNet: An online lexical database. International Journal of Lexicography 3(4).

NCITS T2 (1998). Knowledge Interchange Format, Working draft of proposed American national standard.

Noy, N. F., R. W. Fergeson, et al. (2000). The Knowledge Model of Protege-2000: Combining Interoperability and Flexibility. In Proceedings of the 12th International Conference on Knowledge Engineering and Knowledge Management (EKAW'2000), Juan-les-Pins, France, Springer Verlag.

Parker, I. (1995). Everyday Behavior(ism) and Therapeutic Discourse: Deconstructing the Ego as Verbal Nucleus in Skinner and Lacan. Therapeutic and Everday Discourse as Behavior Change: Towards a Micro-analysis in Psychotherapy Process Research. J. Siegfried. Norwood, New Jersey, Ablex Publishing Corporation**:** 447-467.

Patrick, J. (1999). Tagging psychotherpeutic interviews for linguisitc analysis. Workshop on Towards Standards and Tools for Discourse Tagging, Association for Computational Linguistics, New Brunswick.

Pearl, J. (2000). Causality: models, reasoning, and inference. Cambridge, Cambridge University Press.

Pittenger, R. E., C. F. Hockett, et al. (1960). The first five minutes: a sample of microscopic interview analysis. Ithaca, N.Y., P. Martineau.

Pustejovsky, J. (1995). The generative lexicon. Cambridge, Mass., MIT Press.

Richardson, S. D., W. B. Dolan, et al. (1998). MindNet: acquiring and structuring semantic information from text. In Proceedings of COLING '98, Université de Montréal, Montréal, Québec, Canada.

Sapir, E. (1927). Speech as a Personality Trait. American Journal of Sociology 32(6): 892-905.

Sowa, J. F. (1984). Conceptual Structures: Information Processing in Mind and Machine, Addison Wesley, Reading, MA.

Sperberg-McQueen, C. M. and L. Burnard, Eds. (1994). Guidelines for Electronic Text Encoding and Interchange. Association for Computers and the Humanities (ACH), Association for Computational Linguistics (ACL), and Association for Literary and Linguistic Computing (ALLC). Chicago, Oxford: Text Encoding Initiative.

Viegas, E. (1999). Opening the world with active words and concept triggers. Breadth and depth of semantic lexicons. E. Viegas. Dordrecht; Boston, Kluwer Academic.