Natural Language Engineering **3** (2/3): 97–101. Printed in the United Kingdom

© 1997 Cambridge University Press

Knowledge representation for natural language processing in implemented systems

SYED S. ALI
Department of Mathematical Sciences, University of Wisconsin–Milwaukee, Milwaukee, WI 53201, USA

ŁUCJA IWAŃSKA
Computer Science Department, Wayne State University, Detroit, MI 48202, USA

(*Received 24 June 1997*)

1 Introduction

This special issue presents the state-of-the-art in implemented, general-purpose Natural Language Processing (NLP) systems that use nontrivial Knowledge Representation and Reasoning (KRR). These systems use full-scale implementations of traditional KRR techniques as well as some newer knowledge-related processing mechanisms that have been developed specifically to meet the needs of natural language processing. The papers cover a wide range of natural language inputs, knowledge and formalisms, application domains and processing tasks, illustrating the key role that knowledge representation plays in all types of NLP systems.

Knowledge Representation and Reasoning (KRR), as a research area, has its origins in the need to represent the meaning of natural language sentences. However, much of the recent work in KRR has not been concerned with the needs of natural language processing. For example, this work has focused on using mathematical, common sense, visual, language-based and logical knowledge. It is by no means clear that these problems have much in common with natural language. Thus, much KRR research today has little to do with the theoretically and practically challenging task of language processing. This special issue attempts to draw together recent work in KRR that has remained true to the original goals of KRR research.

Like KRR, natural language processing, as a research area, has progressed relatively independently of research in other areas. Natural language processing needs KRR. However, the services that are provided by many KRR systems are only marginally motivated by the problems of natural language. Thus, many NLP researchers find that these KRR formalisms are not very useful or outright wrong for their needs. In this special issue, we consider both work that tries to adapt traditional KRR approaches and, to a lesser extent, work that includes new formalisms and techniques that are more linguistically oriented.

The impetus for this special issue was a very successful and well attended sym-

https://doi.org/10.1017/S1351324997001721 Published online by Cambridge University Press

posium on the relationship between the active research areas of KRR and NLP.[1] A conclusion of this meeting was that there is a lack of communication between researchers in the two areas and that this lacunae was detrimental to both fields. In particular, KRR could benefit from more attention to NLP, because grounding KRR in the theoretical and practical problems of NLP will increase the perceived value of KRR research. Conversely, NLP could benefit from more attention to KRR, because the resolution of practical NLP problems requires a great deal of knowledge.

Another important result of the symposium was to identify important issues in the design of NLP systems that use KRR. These issues for language processing include:

- What architecture is most appropriate for NLP systems?
- What pragmatic considerations are important?
- What is the relationship between the KRR and natural language?
- What kind of inference underlies reasoning in language?

The papers in this special issue will propose various solutions to each of these problems.

1.1 The issues

The traditional architecture for NLP systems is the so-called 'pipelined approach' model,[2] which has well-established problems. This model calls for processing information about natural language through four distinct levels: morphology, syntax, semantics and pragmatics. Pipelined implementations have separate components, each with their own representation language. This creates problems for knowledge sharing, because it does not allow tight interplay between levels, or opportunistic, domain-related processing that could skip over processing levels entirely.

Natural language pragmatics includes tasks such as the resolution of anaphoric references, scoping ambiguities, the determination of the speaker's purpose, and the understanding of implicatures. These tasks depend on knowledge of the world and of the extra-linguistic activities in which the speakers are engaged. Typically, NLP systems account for only one or two of these pragmatic considerations. A concern is that dealing with each pragmatic issue in isolation ignores the fact that different kinds of pragmatic information are often used together to process natural language. However, identifying the pragmatic considerations that should be accounted for in a KR language is extremely difficult.

The relationship, if any, between the KRR and the natural language that it represents is not clear. The desiderata include consistency with current theories of language and meaning, ease of syntactic-semantic mapping, accessibility of pragmatic information, efficiency, coverage, and extensibility. In addition, the KRR designer

[1] Details about this symposium (including online papers) are available at the URL: `http://tigger.cs.uwm.edu/~syali/AAAI-Fall-Symp94`

[2] This term was coined by Doug Appelt.

https://doi.org/10.1017/S1351324997001721 Published online by Cambridge University Press

must consider the trade-off between expressiveness and tractability. For example, knowledge representations could be structurally similar to natural language (thus easing the mapping from language into representation), or they could be simililar to logical form (which allows for multiple possible semantic interpretations that must be resolved using context), or they could be based on the traditional knowledge representation, first order logic (which is decontextualized, disambiguated and crisp).

Reasoning in language does not parallel traditional KRR inference methods. One choice is to add new, possibly unsound but plausible, inference methods to KR systems that parallel reasoning in natural language (for example, natural language implicature). The alternative is to use a general-purpose KRR formalism and ignore the natural language. Interestingly, the papers in this special issue (particularly the information extraction papers which are corpus-based) are clearly using natural language as a source of inferential techniques.

1.2 The papers

The seven papers included in this special issue all address the concerns above, in the context of a practical application of NLP that requires KRR. In each case, the authors show in detail how their system processes some utterance, text or dialog. They describe the background information used, how that information is represented and utilized, how the system processes the knowledge to do interesting things and how the information is processed into answers. The papers take very different approaches, but fall naturally into three conceptual groupings:

KRR support for dialog processing

A model for habitable and efficient dialogue management for natural language interaction by Arne Jönsson, and

Uniform knowledge representation for language processing in the B2 system by Susan W. McRoy, Susan M. Haller and Syed S. Ali.

KRR support for processing large corpora

Using a semantic network for information extraction by Robert Gaizauskas and Kevin Humphreys, and

TextNet – A text-based intelligent system by Sanda M. Harabagiu and Dan I. Moldovan.

KRR for general-purpose natural language processing

Recovering nonlinearly distributed knowledge: Computing discourse structure in factual reports by Łucja Iwańska,

Restricted representation of phrase structure grammar for building a tree annotated corpus of Korean by Kong Joo Lee, Jae-Hoon Kim and Young S. Han, and

NRKL, a knowledge representation tool for encoding the 'meaning' of complex narrative texts by Gian Piero Zarri.

https://doi.org/10.1017/S1351324997001721 Published online by Cambridge University Press

1.2.1 KRR support for dialog processing

The papers included in this category – *A model for habitable and efficient dialogue management for natural language interaction* by Arne Jönsson, and *Uniform knowledge representation for language processing in the B2 system* by Susan W. McRoy *et al.* – provide knowledge representation structures suitable for dialog processing in an ongoing interaction. They show sample interactions with users and the associated knowledge representations. In terms of the knowledge representations used, the papers represent two extremes: Jönsson uses shallow, template-based representations that has been extracted automatically from large corpora; Mcroy *et al.* uses traditional deep meaning representations that have been hand-crafted.

The paper by Jönsson describes a dialog model that allows for robust and efficient interaction with natural language interfaces. The model is constructed as a tree of dialog objects. The structure of this model is determined by a dialog grammar that has been developed on the basis of empirical studies of human computer interaction in various simple service applications. The framework has successfully been applied to various systems and interaction modalities, handling written and spoken language. However, these applications all assume that the user has the initiative. Experiments with this model suggest that a wide variety of dialog processing can be done with shallow knowledge representation.

The second paper in this group, by McRoy *et al.*, presents a model of dialog that uses a propositional semantic network to represent and utilize knowledge. The paper discusses the NLP and KRR components of a collaborative system that allows medical students to practice decision-making skills. As a collaborative system, initiative can shift between the system and the user over the course of the dialog. The KRR component supports the NLP component by providing a uniform representation of the content and structure of the interaction at the parser, discourse and domain levels. This uniform representation allows distinct tasks, such as dialog management, domain-specific reasoning and meta-reasoning, to all use the same information source, without requiring mediation. This allows them to handle queries, such as *Why?*, whose interpretation and response requires information from each of these tasks.

1.2.2 KRR support for processing large corpora

The papers included in this category – *Using a semantic network for information extraction* by Gaizauskas and Humphreys, and *TextNet – A text-based intelligent system* by Harabagiu and Moldovan – discuss approaches to KRR support for processing large corpora. The goal of both sets of authors is to build information and knowledge representation structures reflecting some content of the natural language text in the corpus. Work reported by the first paper is an application of NLP to the task of Information Extraction (IE). The research reported in the second paper is motivated the task of automatic knowledge acquisition from large corpora.

The first paper, by Gaizauskas and Humphreys, describes the knowledge representation in a NLP IE system. Their system analyses texts by first translating

https://doi.org/10.1017/S1351324997001721 Published online by Cambridge University Press

individual sentences to a quasi-logical form, and then constructing a weak discourse model of the entire text from which template fills are derived. The authors claim that a deeper processing of texts that requires knowledge representation is necessary for IE systems, if they are to become usable technologies.

The second paper in this group, by Harabagiu and Moldovan, presents a system that maintains a knowledge base structured around an independently developed lexical database. Their system automatically extracts information from texts found in public repositories. The intended application is the automatic collection of common-sense knowledge from texts on the Internet for use in sophisticated Internet interfaces. While the problems of inconsistencies and information distortion are not addressed, the authors present a compelling case for successful knowledge acquisition from natural language.

1.2.3 KRR for general-purpose natural language processing

The three papers in this group – *Recovering nonlinearly distributed knowledge: Computing discourse structure in factual reports* by Łucja Iwańska, *Restricted representation of phrase structure grammar for building a tree annotated corpus of Korean* by Kong Joo Lee *et al*., and *NRKL, a knowledge representation tool for encoding the 'meaning' of complex narrative texts* by Gian Piero Zarri – discuss a variety of knowledge representation approaches for processing unrestricted texts.

The first paper in this group, by Iwańska, presents a heuristics-based algorithm for computing a new type of text structure that reflects nonlinear knowledge presentation. The author discusses quantitative and qualitative performance of the algorithm and presents the results of the extensive tests on large volume of free English texts in two domains. She also shows that the algorithm matches human performance with respect to text segmentation of a complicated dialog.

The second paper, by Lee *et al*., introduces a representation of phrase structure grammars for handling a large corpus of Korean text. Unlike English, Korean word order is partially free and the composition of words can be complex. The differences are significant enough to change the grammar formalism. The authors present a restricted representation of phrase structure grammar to handle the characteristics of Korean more efficiently.

The third paper, by Zarri, describes the Narrative Knowledge Representation Language (NKRL), a language for representing the semantic content of complex narrative texts. NKRL includes four specialized sub-languages and knowledge representations for each of them. The author shows how the NKRL representation retains the main informational elements of the original narrative, including various types of plural expressions.

Acknowledgements

We wish to thank Susan W. McRoy, Susan M. Haller and Deepak Kumar for their helpful comments on earlier drafts of this paper.

https://doi.org/10.1017/S1351324997001721 Published online by Cambridge University Press