http://doi.acm.org/10.1145/1236181.1236182

Reasoning in Natural Language Information Processing – In this Special Issue

DAWEI SONG
Knowledge Media Institute, The Open University, United Kingdom

JIAN-YUN NIE
Department of IRO, University of Montreal, Canada

For any applications related to Natural Language Processing (NLP), reasoning has been recognized as a necessary underlying aspect. Most of the existing work in NLP deals with specific NLP problems in a highly heuristic manner, yet not from an explicit reasoning perspective. Recently, there have been developments on models that allow reasoning in NLP, such as language models, logical models, and so on. The goal of this special issue is to present high-quality contributions that integrate reasoning involved in different areas of natural language processing, both at theoretical and/or practical levels. In this article, we give a brief overview on some major aspects of explicating reasoning in NLP and summarize the articles included in this special issue.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Retrieval models*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing

1. INTRODUCTION

For any applications related to Natural Language Processing (NLP), reasoning has been recognized as a necessary underlying aspect: when we try to retrieve relevant documents in Information Retrieval (IR), to determine the correct answer in a Question-Answering (QA) system, or to determine an appropriate sentence in Machine Translation (MT), etc., some forms of reasoning are often underlying. Most of the existing work in NLP deals with specific NLP problems in a highly heuristic manner, yet not from an explicit reasoning perspective. Recently, there have been developments on models that allow reasoning in NLP, such as language modelling, logical models, models based on Bayesian networks, and so on.

There have been a number of researches considering NLP applications as a reasoning process. As logic is a normalisation of the way we use information to reason and make decisions, the use of appropriate logic allows reasoning to be explicitly modelled. Following the logical uncertainty principle [van Rijsbergen 1986], a number of logic-based models for IR have been proposed in the late 1980s and 1990s to integrate reasoning in IR [Lalmas and Bruza 1998].

Authors' current addresses: Dawei SONG, Knowledge Media Institute, The Open University, Walton Hall, Milton Keynes, MK7 6AA, United Kingdom;, email: d.song@open.ac.uk; Jian-Yun NIE, Department of IRO, University of Montreal, Canada, email: nie@iro.umontreal.ca

Nevertheless, no operational system has been successfully evaluated against IR benchmark collections and applied to large scale IR tasks. Therefore, this trend has not been followed up since then. A functional analysis of logical IR models in [Wong et al. 2001] has uncovered two major difficulties encountered by the logic-based IR:

(1) The lack of automatic means for constructing the background knowledge.
(2) The computational overhead inherent in the symbolic logic when picking up and integrating different types of knowledge to facilitate effective reasoning.

In regard to the first problem, it was argued that when language processing tools have advanced further, the concepts, such as different types of term relationships, under the logic-based models could be applied to IR more easily and more directly so that the problem can be largely alleviated. The second problem has to do with the fact that most logical frameworks exist in the realm of symbolic processing, where reasoning is a sequential process proceeding from assumptions to a conclusion by applying symbolic rules of inference. These logical frameworks often suffer from the high computational complexity when applied to large scale applications. This is partially due to the inherent frame problem of symbolic inference which involves picking up appropriate inference rules to perform reasoning [Gärdenfors 2000]. This can be more serious when different types of inference rules are taken into account.

There are also renewed interests in logical IR. [Song and Bruza 2003] propose to use an information inference mechanism on high dimensional semantic spaces to underpin reasoning at symbolic level for logic based information retrieval. [Lau et al. 2004] is a recent successful attempt in this direction.

The NLP and IR area is currently experiencing a theoretical shift with the strong trend of statistical language modelling approaches. This latter framework is capable of integrating statistical reasoning. We believe that many approaches using language modelling can be easily described from a reasoning perspective. Classical LM approaches [Ponte and Croft 1998] usually assume independence between indexing units, which are unigrams or bigrams. In reality, a word may be related to other word. Such relationships should be properly integrated into LM. Some recent work [Berger and Lafferty 1999; Lafferty and Zhai 2001; Lavrenko and Croft 2001; etc.] has shown that LM framework is capable of integrating statistical reasoning in IR. More recent approaches try to extend the existing LM by incorporating term relationships or dependencies, for example, grammatical links in [Gao et al. 2004], co-occurrence and WordNet relations in [Cao et al. 2005]. [Collins-Thompson and Callan 2005] propose a Random Walk Markov chain model wherein the states are terms and the links are term relationships. [Bai et al. 2005] exploit inferential term relationships extracted by using a more sophisticated approach, i.e., the Information Flow [Song and Bruza 2003].

Cross-language IR (CLIR) is another area in which reasoning is required. Current CLIR approaches consider translation as a separate step. We believe that translation can be better considered by integrating it into a reasoning process. Indeed, as argued in [Nie 2003], CLIR may be considered as a special case of inferential IR, and translation can be considered as a special type of reasoning.

Question Answering is also a typical task that required reasoning. However, reasoning is merely mentioned as such in the current QA approaches, although many specific approaches can be recast as a reasoning process. A logic-based framework "Knowing-aboutness" [Clifton and Teahan 2005] is recently built and demonstrates favourable performance in TREC QA'2004 evaluation.

We believe that the basic techniques necessary for these NLP-related applications are mature enough to think about the problem from a broader point of view. The goal of this special issue is to present high-quality contributions that explicate reasoning involved in different areas of natural language processing, both at theoretical and/or practical levels.

2. ARTICLES IN THIS SPECIAL ISSUE

We received a total number of twenty submissions. Each paper was reviewed by at least three referees. Finally five papers were selected for publication. The selected papers cover the topics on logical, statistical and temporal reasoning for information retrieval, query translation, and topic detection and tracking.

Nie, Bai and Cao: Inferential Language Models for Information Retrieval. This paper presents a novel language modeling framework for incorporating tern relationships and in turn facilitating logical inference in information retrieval. In particular, inferential mechanisms are developed to expand document models and/or query models. Term co-occurrences extracted from relevance feedback documents and the WordNet term relationships are integrated in the language modeling framework via semantic smoothing. The depth of inference is also taken into account by employing the Markov chain model. The proposed language modeling framework opens a door to overcome the scalability problem inherent in traditional symbolic logical models. The framework has been successfully evaluated on large scale English and Chinese collections and demonstrated good performance.

Gao, Nie and Zhou: Statistical Query Translation Models for Cross-Language Information Retrieval. This paper investigates a number of statistical dependency translation models incorporating both statistical co-occurrences and syntactical linguistic structures into cross-language query translation. Special attention is paid to the resolution of query translation ambiguities by taking into account monolingual language model as well as cross-lingual word similarities. Experimental results have shown that the use of linguistic structures in statistical reasoning is more beneficial than the use of co-occurrences and dictionaries. Better results are generated by combining different types of knowledge.

Li, Li and Lu: Topic Tracking with Time Granularity Reasoning. This paper investigates reasoning with the granularity of time in Topic Detection and Tracking (TDT) in order to deal with implicit temporal relatedness between stories and topics. The paper proposes to add temporal expression recognition and time granularity reasoning components to TDT. First, it is argued to take into account not only publishing time but also additional temporal information extracted from the stories, e.g., the actual time when a story occurs. Second, a time granularity model is defined and an algorithm for inferring the strongest co-reference between two time stamps is developed. Experiments on two TDT

datasets show that the time granularity reasoning can improve the performance of TDT tasks.

Xuan-Hieu Phan, Le-Minh Nguyen, Tu-Bao Ho, Susumu Horiguchi: Improving Discriminative Sequential Learning by Discovering Important Associations of Statistics. The main contribution of this paper is a data-driven approach towards mining important association rules with weak statistics hidden in the training data. The authors also show how to incorporate the discovered association rules into the learning model of Conditional Random Fields (CRFs). The proposed approach has been proved effective in experiments in comparison with the traditional CRF approach for two different tasks: phrase segmentation and named entity recognition.

Yi Liu and Rong Jin: A Statistical Framework for Query Translation Disambiguation. This paper proposes a statistical framework for dictionary-based CLIR. Although query translation disambiguation is a long-standing research topic in CLIR, this paper formulates the problem within a statistical framework so that well-known optimization methods can be applied. It estimates the translation probabilities of query words based on the monolingual word co-occurrence statistics. Two variants, namely Maximum Coherence Model and Spectral Query Translation Model are presented. The paper shows that the first method can be considered as a special case of the second one. The experiments also demonstrate that the proposed approaches are more effective than some existing ones.

3. CONCLUSIONS AND FUTURE WORK

We believe that many problems in NLP can be reconsidered from a reasoning perspective in a way more similar to how human beings make inference with information. This could open doors for developing novel and more sensible solutions. Recent advances in statistical processing of natural language, as demonstrated by the articles in this special issue, have resulted in a number of promising methods allowing us to integrate logical and statistical inference in NLP.

Finally, we would like to suggest a number of future research directions including (1) more coherent models for seamlessly integrating logical and statistical inference; (2) incorporation of user and context issues into the reasoning mechanisms, and (3) evaluation paradigms beyond the traditional precision/recall measurement to explore additional benefits (such as explanatory power and compatibility with human reasoning) brought by modelling the NLP problems as reasoning processes.

4. ACKNOLWDGEMENTS

We thank all the authors who submitted papers for their contributions. We are grateful to our dedicated reviewers for their professional reviewing services. The reviewers are Azzah Al-Maskari, Leif Azzopardi, Guihong Cao, Anne De Roeck, George Foster, Jianfeng Gao, Eduard Hoenkamp, Jimmy Huang, Rong Jin, Gareth Jones, Mounia Lalmas, Philippe Langlais, Guy Lapalme, Wai Lam, Wenjie Li, Xue Li, Raymond Lau, Michel Simard, Cheng Niu, François Paradis, Mark Sanderson, Victoria Uren, Maria Vargas-Vera, and Xin Yan.

REFERENCES

BAI, J., SONG, D., BRUZA, P.D., NIE, J.Y., AND CAO, G. 2005. Query Expansion Using Term Relationships in Language Models for Information Retrieval. In *Proceedings of CIKM 2005,* 688-695.

BERGER, A. AND LAFFERTY, J. 1999. Information Retrieval as Statistical Translation. In Proceedings of *ACM/SIGIR'1999,* pp. 222-229.

Cao, G., Nie, J.Y. AND Bai, J. 2005. Integrating Term Relationships into Language Models. *In Proceedings of SIGIR'2005,* 298-305.

CLIFTON, T. AND TEAHAN, W. 2005. Knowing-Aboutness: Question-Answering Using a Logic-based Framework. In *ECIR'2005*.

COLLINS-THOMPSON, K. AND CALLAN, J. 2005. Query Expansion Using Random Walk Models. In *Proceedings of CIKM 2005,* 704-711.

GAO, J., NIE, J., WU, G. AND CAO, G. 2004. Dependence language model for information retrieval, *27th ACM-SIGIR*, pp. 170-177.

GARDENFORS, P. 2000. *Conceptual Spaces: The Geometry of Thought.* MIT Press.

LAFFERTY, J., AND ZHAI, C. 2001. Document Language Models, Query Models, and Risk Minimization for Information Retrieval. In Proceedings of the *24th Annual International Conference on Research and Development in Information Retrieval (SIGIR'01), pp. 111-119.*

LALMAS, M. AND BRUZA, P.D. 1998. The Use of Logic in Information Retrieval Modeling. *Knowledge Engineering Review*, 13(3), 263-295.

LAVRENKO, V., AND CROFT, W.B. 2001. Relevance-Based Language Models. In Proceedings of the *24th Annual International Conference on Research and Development in Information Retrieval (SIGIR'01), pp. 120-127.*

LAU, R., BRUZA, P.D., AND SONG, D. 2004. Belief Revision for Adaptive Information Retrieval. In Proceedings of *ACM/SIGIR'2004,* pp. 130-137.

NIE, J.Y. 2003. Query Expansion and Query Translation as Logical Inference, *Journal of the American Society for Information Science*, 54(4) : 335-346.

PONTE, J. AND CROFT, W.B. 1998. A Language Modelling Approach to Information Retrieval. In Proceedings of *ACM/SIGIR'1998,* pp. 275-281.

SONG, D. AND BRUZA, P.D. 2003. Towards Context-sensitive Information Inference. *Journal of the American Society for Information Science and Technology (JASIST),* 54(4), pp. 321-334.

VAN RIJSBERGEN, C.J. 1986. A Non-classical Logic for Information Retrieval. *The Computer Journal*, 29(6), 481-485.

WONG, K.F., SONG, D., BRUZA, P.D., AND CHENG, C.H. 2001. Application of aboutness to functional benchmarking in information retrieval. *ACM Transactions on Information Systems (TOIS),* 19(4), 337-370.