ISSN 2186-7437

# NII Shonan Meeting Report

No. 2011-4

# Knowledge-leveraged Computational Thinking through Natural Language Processing and Statistical Logic

Sadao Kurohashi
Akihiro Yamamoto
Kentaro Inui
Sebastian Riedel

September 18–21, 2011

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-Ku, Tokyo, Japan

# Knowledge-leveraged Computational Thinking through Natural Language Processing and Statistical Logic

Organizers:
Sadao Kurohashi (Kyoto University)
Akihiro Yamamoto (Kyoto University)
Kentaro Inui (Tohoku University)
Sebastian Riedel (University of Massachusetts Amherst)

September 18–21, 2011

A long-standing problem in Natural Language Processing has been a lack of large-scale knowledge for computers. The emergence of the Web and the rapid increase of information on the Web drastically changed the environment of NLP. The Web is not only a marvelous target for NLP, but also a valuable resource from which knowledge could be extracted for computers, making research and development activities on large-scale text processing and large-scale knowledge acquisition much more popular.

However, beyond the success of large-scale NLP and knowledge acquisition, we are starting to face a new problem: how to manage and use the automatically acquired knowledge. We are still not confident that automatically acquired large-scale knowledge resources will indeed solve NLP problems in real world applications How to incorporate the acquired knowledge into existing NLP frameworks and how to manage them are yet unsolved issues.

Based on this background, the workshop introduces the new research field of Computational Thinking, where computers themselves directly make use of the large-scale knowledge on the Web by combining natural language processing and statistical logical inference. In natural language processing, we need robust structuring of text, making advances in discourse analysis, zero pronoun resolution, and syntactic and semantic parsing. In addition, to handle the uncertainty and ambiguity inherent in language and in automatically acquired knowledge, we also have to construct a theory and system of inductive knowledge derivation, allowing us to carry out statistical inductive logic computation on a massive scale.

Based on these advances, we can develop a high-dimension statistical induction framework that finds links in the useful knowledge represented in language, and by doing so, construct a knowledge analysis engine that goes beyond mere search, effectively harnessing the entire Web knowledge space.

The aim of the workshop is to bring researchers and practitioners together in order to discuss Knowledge-leveraged Computational Thinking.

# Overview of Talks

## Toward Knowledge-leveraged Computational Thinking

Sadao Kurohashi, Kyoto University

A long-standing problem in Natural Language Processing has been a lack of large-scale knowledge for computers. The emergence of the Web and the rapid increase of information on the Web drastically changed the environment of NLP. The Web is not only a marvelous target for NLP, but also a valuable resource from which knowledge could be extracted for computers. We have been constructing an open search engine, TSUBAKI, as a large-scale NLP infrastructure, which provides neatly-prepared huge Web corpus and a semantic search facility exploiting a high-accuracy parser and automatically acquired synonymous expressions. We also automatically acquired Japanese caseframes, predicate-argument patterns for 40 thousand verbs/adjectives from 15 billion sentences, and showed their impacts on case analysis, anaphora resolution and synonymy recognition. What we need next is an inference engine, as a core module for knowledge-leveraged computational thinking, which manages and uses the automatically acquired huge knowledge, coping with uncertainty of NLP analysis and language itself.

## Symbolic and Statistical Modelling: Time for the Twain to Meet

Timothy Baldwin, University of Melbourne

There has historically been a divide in NLP between symbolic (or rule-based)and statistical methods, and the pendulum swing is currently very much in the direction of statistical methods. In practice, however, statistical methods stand to gain much from symbolic methods in terms of constraints, and symbolic methods stand to gain much from statistical methods in terms of disambiguation, scalability and robustness. In this talk, I will outline notable successes at the boundary between symbolic and statistical modelling, and speculate on possible future directions in the space.

## Linked data for NLP or by NLP?

Key-Sun Choi, KAIST

If we call Wikipedia or Wiktionary as "web knowledge resource", the question is about whether they can contribute to NLP itself and furthermore to the knowledge resource for knowledge-leveraged computational thinking. Comparing with the structure inside WordNet from the view of its human-encoded precise classification scheme, such web knowledge resource has category structure based on collectively generated tags and structures like infobox. They are called also as "Collectively Generated Content" and its structuralized content based on collective intelligence. It is heavily based on linking among terms and we also say that it is one member of linked data. The problem is in whether such collectively generated knowledge resource can contribute to NLP and how much it can be effective.

The more clean primitives of linked terms in web knowledge resources will be assumed, based on the essential property of Guarino (2000) or intrinsic property of Mizoguchi (2004). The number of entries in web knowledge resources increases very fast but their inter-relationships are indirectly calculated by their link structure. We can imagine that their entries could be mapped to one of instances under some structure of primitive concepts, like synset of WordNet. Let's name such primitives to be "intrinsic tokens" that are derived from collectively generated knowledge resource under the principles of intrinsic properties. The procedure could be approximately proven and it will be a kind of statistical logic. We then go to the issues about what area of NLP can be solved by the so-called intrinsic tokens and their relations, a resultant approximately generated primitives.

Can NLP contribute to the user generation process of content? Consider the structure of infobox in Wikipedia more closely. It will be discussed about how NLP can help the population of relevant entries, like the social network mechanism for multi-lingual environment and information extraction purpose.

The traditional NLP starts from words in text but now also works have been undergoing on the web corpus with hyperlinks and html markups. In web knowledge resources, the words and chunks have underlying URIs, a kind of annotation. It signals a new paradigm of NLP.

## Scalable Abduction for Deep NLP (ILP-based Reasoning for Weighted Abduction)

Kentaro Inui, Tohoku University

Abduction is inference to the best hypothesis to explain observations. Hobbs et al. (1993) demonstrate that abduction gives a reasonable formalization of the process of discourse understanding, and a variety of NLP subtasks can be resolved with a single abduction-based framework. However, abductive reasoning quickly becomes intractable as the amount of background knowledge is increased to cover the millions of axioms necessary for robust discourse processing. This computational bottleneck is preventing abductive reasoning from benefiting from the recent advances in computational resources for commonsense reasoning. In this talk, we propose an efficient implementation of Hobbs et al.'s abductive discourse interpretation framework, weighted abduction. Our framework transforms the problem of explanation finding in weighted abduction into a linear programming problem. Our experiments showed that our approach efficiently solved problems of plan recognition and textual entailment recognition, outperforming an existing system for weighted abduction.

## Logic, Natural Language and the Language of Thought

Robert Kowalski, Imperial College London

In the philosophy of language, there are three main schools of thought:

- An agent's thoughts are represented in a private language of thought (LOT), which is independent of public, natural languages.

- The LOT is a form of the public, natural language that the agent speaks.

- Human thinking does not have a language-like structure.

I will argue for the first of these three alternatives, and argue moreover that human natural language processing involves both translating natural language into the LOT, and connecting translations with existing thoughts represented in the LOT.

I will also argue that the LOT can be viewed as a simplified form of logic, in which goals and beliefs are both represented as conditionals in clausal form. Goals are represented as general clauses, with the expressive power of full first-order logic (FOL), and beliefs are represented more simply as clauses in logic programming form. This clausal representation is a canonical form, which reduces most thinking in the language of thought to forward and backward reasoning. Moreover, the clausal representation has a connectionist structure and implementation, which can give the impression that thinking does not have a language-like structure at all.

## Toward Inference Rules for First-order Clauses with Uncertainty

Akihiro Yamamoto, Kyoto University

Recently knowledge can be extracted from large data like the Web but it usually has uncertainty. First-order logic is useful for knowledge representation and is now extended so that it can treat uncertain knowledge. Uncertainty could be well combined with semantics of first-order formulae, but even if semantics including uncertainty is provided, we should develop some proof theory. In this talk we provide a inference rule for clauses, formulae of a special type, based on resolution principle.

## POSTECH approaches to computer assisted language learning

Gary Geunbae Lee, POSTECH

Although there have been enormous investments into English education all around the world, not many differences have been made to change the English instruction style. Considering the shortcomings for the current teaching-learning methodology, we have been investigating advanced computer-assisted language learning (CALL) systems. This paper aims at summarizing a set of POSTECH approaches including theories, technologies, systems, and field studies. On top of the state-of-the-art technologies of spoken dialog system, a variety of adaptations have been applied to overcome some problems caused by numerous errors and variations naturally produced by non-native speakers. Furthermore, a number of methods have been developed for generating educational feedback and mining educational data from Internet. Integrating these efforts resulted in intelligent educational robots - Engkey - and virtual 3D language learning games, Pomy. To verify the effects of our approaches on students' communicative abilities, we have conducted a field study at an elementary school in Korea. The results showed that our CALL approaches can be enjoyable and fruitful activities for students. Although the results of this study bring us a step closer to understanding computer-based education, more studies are needed to consolidate the findings.

## From Information to Knowledge: Memory vs. Inference

Rafael E. Banchs, Institute for Infocomm Research

During the last decades we have seen the development of a large number of techniques and applications which allow distilling useful and valuable information from vast amounts of raw data. An interesting, and relatively new, challenge for artificial intelligence is to pursue further and develop methods for leveraging knowledge from information. In this presentation we enter into a debate about the unprecedented possibilities that the World Wide Web phenomenon is providing for the discovery and generation of knowledge. Indeed, the huge amounts of information available in the Internet, along with the currently available computational power, define an interesting framework for the creation of a collective knowledge repository that computational agents can exploit for a diverse scope of tasks and applications. We focus our discussion on the specific roles that both memory and inference plays in the conceptualization, extraction, generation and representation of knowledge. Some different approaches to memory an inference (including some commonly used cognitive, geometrical and statistical models) are revisited under the optics of knowledge representation. Rather than providing definite answers and solutions, the main objective of this presentation is to post new questions, as well as to reformulate old ones under a new perspective, on the specific problems related to extracting knowledge from the World Wide Web and representing it in a meaningful way such that it can be exploited and enriched by computational agents.

## Semantic Transliteration of Personal Names

Haizhou Li, Institute for Infocomm Research

Machine transliteration is the process of automatically rewriting the script of a word from one language to another, while preserving pronunciation. The last decade has seen a tremendous progress and a growth of interests from theory to practice of machine transliteration. In this talk, I will present a brief overview of the fundamentals, algorithms and applications, in particular, transliteration of personal names. I will also report the findings in the most recent transliteration evaluation campaigns - NEWS 2009 and NEWS 2010 Machine Transliteration Shared Tasks.

## A Linguistic Approach for Understanding and Utilization of Mathematical Knowledge

Akiko Aizawa, National Institute of informatics

Mathematical expressions play an important role in scientific knowledge dissemination through research papers or textbooks. With their abstract and structured representations, mathematical expressions have long been considered to be out of the scope of natural language processing. However, in order to understand and utilize mathematical knowledge, mathematical expressions often should be disambiguated and complemented by accompanying natural language text, and the interpretation sometimes requires domain knowledge such as conventions of notations and well-known theorems. In this talk, we explore and discuss how the formalism of mathematical expressions can be identified and further utilized using conventional natural language processing techniques.

## Developing a textual entailment resource towards identification of propositions in a text

Yusuke Miyao, National Institute of informatics

Abstract: This talk describes a recent effort on the development of a textual entailment data set. Rather than assuming a sub-component of applications like question answering and multi-document summarization, we focus on a real-world task to judge whether a natural language proposition is true or false according to a given text. I will describe the design of resource development and features of the obtained resource.

## Unsupervised Semantic Parsing

Pedro Domingos, Washington University

Extracting knowledge from text has long been a goal of AI and NLP, but progress has been difficult. Manual approaches are too brittle, and supervised learning ones require an unrealistic quantity and quality of labeled data. To address this problem, we have recently developed the first unsupervised approach to semantic parsing (i.e., translating raw text into a formal meaning representation). It is based on Markov logic, which combines Markov networks and first-order logic. Our USP system transforms dependency trees into quasi-logical forms, recursively induces lambda forms from these, and clusters them to abstract away syntactic variations of the same meaning. The most probable semantic parse of a sentence is obtained by recursively assigning its parts to lambda-form clusters and composing them. We have evaluated our approach by using it to extract a knowledge base from biomedical abstracts and answer questions. USP greatly outperforms previous approaches on this task. (Joint work with Hoifung Poon.)

## A general MCMC method for logic-based probabilistic modeling

Taisuke Sato, Titech

I present a general MCMC method for PRISM, a logic-based probabilistic modeling language. It is a generalization of an MCMC method for PCFGs to the one for PRISM that covers from Bayesian networks to probabilistic grammars. I describe how to estimate the marginal probability of data from MCMC samples and how to perform Bayesian Viterbi inference using an example of Naive Bayes model augmented with a hidden variable.

## Comparative Analysis of Concerns in Blogs across Languages with Wikipedia as a Multilingual Knowledge Source

Takehito Utsuro, Tsukuba University

This presentation studies how to compare concerns in blogs across languages. To solve this task, we first introduce our framework of categorizing blog posts collected with a certain search query into sub-topics. In this framework, the sub-topic of each blog post is identified by utilizing Wikipedia entries as a knowledge source and each Wikipedia entry title is considered as a sub-topic label. We next focus on results of applying this framework to Japanese and English blogospheres. Through those analyses, we show that it becomes much easier to quickly overview the distribution of sub-topics over the whole blog posts collected with a certain search query. We then compare the results of analyzing concerns in blogs across languages. In this comparative analysis between Japanese and English, we show that differences in concerns can be easily recognized by simply comparing observed sub-topics across languages. Furthermore, we show that, even when we observe certain sub-topics in both languages, still we can recognize differences between languages through much more detailed analysis.

## Multilingual Word Sense Disambiguation under Resource Constraint

Pushpak Bhattacharyya, IIT Bombay

Word Sense Disambiguation (WSD) is a fundamental problem in Natural Language Processing (NLP). Amongst various approaches to WSD, it is the supervised machine learning (ML) based approach that is the dominant paradigm today. However, ML based techniques need significant amount of resource in terms of sense annotated corpora which takes time, energy and manpower to create. Not all languages have this resource, and many of the languages cannot afford it.

In the current presentation, we discuss ways of making use of whatever resource is created for WSD. First we describe a novel scoring function and an iterative algorithm based on this function to do WSD. This function separates the influence of the annotated corpus (corpus parameters) from the influence of wordnet (wordnet parameters), in deciding the sense. Next we describe how the corpus of one language can help WSD of another language, i.e., LANGUAGE ADAPTATION. This is presented in three setting of "complete", "some" and "no" annotation. From this we move on to DOMAIN ADAPTATION where the notion of active learning and injection are pursued to do WSD in a domain with little or no annotated corpora. The extensive evaluation and good accuracy figures lend credence to the viability of our approach which points to the possibility of expanding from one language-domain combination to all language-domain combinations for WSD, i.e., multilingual general domain WSD, a long standing dream of NLP.

The talk is presented in a multilingual setting of Indian languages. There are 22 official languages in India with strong requirements of machine translation and cross lingual search. Our languages of focus in this talk are Hindi and

Marathi along with English and the domains of focus are Tourism and Health which are important to India.

The presentation is based on work done with PhD and Masters students Mitesh, Salil, Saurabh, Anup, Sapan and Piyush, published ACL10, COLING10, EMNLP09 and GWC10.

## Word Sense Disambiguation in Multiple Languages using Wordnets

Francis Bond, Nanyang Technological University

In this talk I look at some of the issues in disambiguating multiple languages using wordnets. Some preliminary analysis is given of Chinese, English and Japanese data.

## Acquiring Knowledge from the Web and its Application to Predicate-Argument Structure Analysis

Daisuke Kawahara, Kyoto University

The biggest issue for artificial intelligence (AI) and natural language processing (NLP) systems has been how to acquire world knowledge. In this talk, we first introduce our approach to acquire case frames from the web, each of which represents an event that consists of a predicate and its arguments. This process is gradually performed to obtain highly reliable and rich relations. Secondly, we describe a method for automatically acquiring strongly-related events (kinds of script knowledge) from the web using the case frames. Both of the acquired knowledge is used in an analyzer of predicate-argument structures, and it achieves cutting-edge performance on various domains of text.

## From Question Answering to Information Acquisition

Xiaoyan Zhu, Tsinghua University

Most web search engines are based on key word search techniques, which do not formally capture the explicit semantic meaning of a query but provide a relatively comfortable way for the user to specify information needs on the basis of keywords. However, there are still some problems caused by the lost of semantic meaning of the query and the lack of users' intention prediction, and the complicated form of returned information. What is the ideal way of internet information acquisition? In this talk, we present a prototype system, QAnswer, which returns sophisticated and succinct answers to users' questions. The platform is supported by key techniques such as semantic relatedness measure, information similarity measure, question typing, concept extension, authority/expert modeling, emotion analysis, opinion extraction, text summarization. The key problems will be discussed as challenges in the future: how to integrate large scale, heterogeneous and multi-modality data; how to build an appropriate knowledge representation structure; as well as how to achieve user intention prediction and answer evaluation based on the semantic content and context understanding.

## Extracting high-level semantic relations from unstructured web text: How much knowledge is needed?

Stijn De Saeger, NICT

In this talk we discuss some of our experiences regarding the themes of this workshop in the context of knowledge acquisition, more specifically the semi-supervised extraction of high-level semantic relations like causation and prevention, which is often assumed to require knowledge-intensive semantic processing like coreference resolution, paraphrase recognition, entailment and discourse analysis. We show how semantic word classes obtained through large-scale clustering of nouns in Web text can be used to learn class dependent extraction templates: lexico-syntactic patterns that select for relation instances from certain semantic class combinations. Class dependent patterns address some long-standing problems due to pattern ambiguity, and enable accurate and efficient relation extraction at Web scale. We also explore ways in which the relation instances acquired this way can further be leveraged to extract other instances that are outside the reach of pattern-based methods – instances mentioned in complex and/or infrequent linguistic contexts for which no simple extraction templates can be learned, and which are traditionally assumed to require complex solutions involving coreference resolution, discourse analysis and inference mechanisms. Throughout the talk, the idea of leveraging previously acquired knowledge to solve more difficult extraction problems is a recurring theme, and we discuss some implications for the idea of knowledge-leveraged computational thinking.

## Thinking Web: Hypothesis generation based on the information in the Web

Kentaro Torisawa, NICT

NLP research on the Web has focused on retrieval, extraction and discovery of information explicitly written on the Web. Evidently this line of research has received much attention and still has many interesting problems left to solve. Yet explicitly written knowledge, even given a huge corpus such as the Web, does not cover and explain all the phenomena observed in the real world, and human decision making based on such limited knowledge may miss important clues, with potentially grave consequences. To increase the amount of knowledge that can help us in our decision making and to reduce the risk of overlooking relevant information, it is important to try another research direction, namely the "discovery" of information that is not explicitly written on the Web. We expect that such "dormant" knowledge includes many innovative ideas or unexpected risks, and proper consideration of this information can lead to better decision making. In this talk, I would like to introduce our work on knowledge acquisition through inference, using "seed" knowledge and inference rules automatically acquired from the Web. After giving real-world examples demonstrating the need for such inference methods I will discuss the methodology, some experimental results, some unexpected problems we encountered in our experiments, and our future research direction.