Introduction to Special Section on Paraphrasing

Paraphrasing, conveying the same meaning in different ways, is an intrinsic part of natural languages. The research field of Automatic Paraphrasing encompasses the tasks of collecting, identifying, and generating paraphrases in an automatic or a computer-aided manner. In addition, researchers have investigated the contribution of automatic paraphrasing techniques to many natural language applications, such as question answering (QA), information extraction (IE), multi-document summarization (MDS), and machine translation (MT). For example, in Machine Translation, paraphrases have been used for rewriting and simplifying input sentences, enlarging translation phrase tables, expanding human references for automatic evaluation, and so forth.

This special section of ACM TIST is intended to cover state-of-the-art research in automatic paraphrasing. Especially, we highlight the applications of paraphrasing techniques in real-world systems, such as MT systems and search engines. Seven articles are included in the special section. One of them is about paraphrase extraction from monolingual corpora, while the other six discuss the applications of paraphrases, including paraphrasing for machine translation, sentence compression, word meaning computing, and plagiarism detection.

There are three articles that focus on applying paraphrasing techniques for MT. These articles cover the three main research directions mentioned, namely, source sentence rewriting, phrase table enlargement, and human reference expansion.

In “Using Targeted Paraphrasing and Monolingual Crowdsourcing to Improve Translation” by Philip Resnik, Olivia Buzek, Yakov Kronrod, Chang Hu, Alexander J. Quinn, and Benjamin B. Bederson, the authors propose enhancing the translation quality of an SMT system based on crowdsourcing. A remarkable advantage of the proposed method is that it involves only monolingual workers to identify target-side translation errors and supply source-side paraphrase, rather than relying on workers with bilingual expertise. The proposed solution has the potential of providing a more cost-effective approach to translation in scenarios where machine translation would be considered acceptable to use if only it were generally of high enough quality. It also has the potential to vastly reduce the burden of human effort for cases in which bilingual translators postedit machine translation output.

In the article “Distributional Phrasal Paraphrase Generation for Statistical Machine Translation” by Yuval Marton, the author focuses on extracting paraphrases to improve the coverage of the translation model. The proposed method extracts paraphrases from large-scale monolingual corpora based on distributional similarity. The extracted paraphrases are then used to augment a translation phrase table with pairs not covered by the initial table. The novelty of the proposed method lies in it being language-independent, and hence it does not rely on bitexts for generating paraphrases or new phrase pairs.

In “Generating Targeted Paraphrases for Improved Translation” by Nitin Madnani and Bonnie Dorr, the authors adopt an approach that uses automatic paraphrase generation to tune parameters for an SMT system. Specifically, given a single reference translation, they build a paraphrase generation system that can produce several different semantically equivalent variants that can then be used as additional reference translations. Experimental results on several language pairs have demonstrated that the proposed approach can improve translation quality. Furthermore, this article presents

DOI: http://dx.doi.org/10.1145/2483669.2483670

a novel way to generate "targeted" paraphrases that yields substantially larger gains in translation quality.

The other articles in this special section are the following.

In "An Abstractive Approach to Sentence Compression" by Trevor Cohn and Mirella Lapata, the authors show that sentence abstraction is a meaningful task in which humans can compress sentences by employing several rewrite operations in addition to deletion. They propose a discriminative tree-to-tree transduction model for the abstraction task that can account for structural and lexical mismatches. The model incorporates a synchronous tree substitution grammar, which encodes a large space of paraphrasing rules extracted from bilingual corpora. Evaluation results show that the approach yields shorter target sentences that are grammatical and mostly preserve the meaning of the longer source sentences while using rewrite rules.

In "An Inference-Based Model of Word Meaning in Context as a Paraphrase Distribution" by Taesun Moon and Katrin Erk, the authors address the problem of word-meaning computing. In particular, the article introduces a usage model of word meaning that frames the task of contextualization as a probabilistic inference problem. This model characterizes a word's meaning as a distribution over potential paraphrases, which is inferred using an undirected graphical model. Evaluated on a paraphrasing task, the proposed model outperforms the state-of-the-art usage vector model on all parts of speech except verbs.

In "Paraphrase Acquisition via Crowdsourcing and Machine Learning" by Steven Burrows, Martin Potthast, and Benno Stein, the authors regard plagiarism detection as a paraphrase recognition task. They acquire passage-level paraphrases via crowdsourcing and use the corpus in the plagiarism detection competition. This article describes in detail how to improve the quality of the crowd paraphrases with a classification framework in which different machine learning models and state-of-the-art features are investigated. Moreover, the article gives a comprehensive cost and time analysis showing excellent savings.

In "Multi-Technique Paraphrase Alignment: A Contribution to Pinpointing Sub-Sentential Paraphrases" by Houda Bouamor, Aurélien Max, and Ann eVilnat, the article addresses the task of sub-sentential paraphrase acquisition from monolingual parallel corpora. The authors propose an exploration of different techniques for the task of paraphrase acquisition. Specifically, five techniques are selected for their complementarity in terms of resources and algorithms, which are applied to two corpora for two languages and four corpus types of various origins. The authors report detailed results on the performance of each individual technique as well as two different technique combinations. An important part of this article is its exhibition and classification of difficult-to-acquire paraphrase pairs.

We hope that this special section will help readers access the state-of-the-art research in the field of automatic paraphrasing. We also hope that more researchers will enter this interesting field to promote the research. We would like to thank the authors submitting articles to this special section. We are also grateful to the reviewers for their great work.

Haifeng Wang
Bill Dolan
Idan Szpektor
Shiqi Zhao
Guest Editors