Coling 2010

23rd International Conference on Computational Linguistics

Tutorial notes

Paraphrases and Applications

Shiqi Zhao and Haifeng Wang

Baidu Inc.

August 22, 2010

To order the CD of Coling 2010 and its Workshop Proceedings, please contact:

Chinese Information Processing Society of China
No.4, Southern Fourth Street
Haidian District, Beijing, 100190
China
Tel: +86-010-62562916
Fax: +86-010-62562916
cips@iscas.ac.cn

Tutorial Instuctor

Shiqi Zhao: *Baidu Inc., Baidu Campus, No. 10, Shangdi 10th Street, Beijing, 100085, China. +86-10-59926892, zhaoshiqi@baidu.com, http://ir.hit.edu.cn/~zhaosq/*

Shiqi Zhao is a postdoctoral researcher in Baidu Inc. (www.baidu.com). He received his PhD in computer science from Harbin Institute of Technology in 2010. Shiqi has studied paraphrasing for several years. The research topics include paraphrase acquisition, generation, and applications. He has published more than 10 papers on paraphrasing at several major conferences and journals, including IJCAI-2007, ACL-08: HLT, ACL-IJCNLP 2009, Journal of Natural Language Engineering, etc.

Haifeng Wang: *Baidu Inc., Baidu Campus, No. 10, Shangdi 10th Street, Beijing, 100085, China, +86-10-59928072, wanghaifeng@baidu.com, http://ir.hit.edu.cn/~wanghaifeng/*

Haifeng Wang is a senior scientist at Baidu Inc. He is also a visiting professor at Harbin Institute of Technology. He received his PhD in Computer Science from Harbin Institute of Technology in 1999. He was an associate researcher at Microsoft Research China from 1999 to 2000, a research scientist at iSilk.com between 2000 and 2002, and the chief research scientist and deputy director at Toshiba (China) R&D Center till Jan. 2010. He has authored more than 60 scientific papers on natural language processing. He served as area chair, tutorial chair, workshop chair, session chair and PC member at several major conferences, including area co-chair and session chair at ACL-IJCNLP 2009, tutorial co-chair for ACL 2010, workshop co-chair for COLING 2010, etc.

Outline

Paraphrases are various expressions that convey the same meaning. Research of paraphrasing is critical in many related NLP research areas, such as machine translation (MT), question answering (QA), information retrieval (IR), information extraction (IE), natural language generation (NLG), etc.

This tutorial is intended to provide the attendees with an in-depth look at the identification, generation, application, and evaluation of paraphrases. The tutorial first reviews studies on paraphrase identification (or extraction), which aims to acquire paraphrases from various data sources, such as large-scale web corpora, monolingual parallel corpora, monolingual comparable corpora, bilingual parallel corpora, as well as some other resources.

It then surveys methods on paraphrase generation, in which the MT-based method will be highlighted, while the other kinds of methods, including thesaurus-based, pattern-based, and NLG-based methods, will also be introduced.

We then discuss the applications of paraphrases in related research areas, especially in MT. We will show how paraphrases can help to alleviate data sparseness problem, simplify input sentences, tune parameters, and improve automatic evaluation in statistical MT systems.

The last part of the tutorial is about the evaluation of paraphrases. Till now, no approach has been widely accepted on paraphrase evaluation, which leaves it as an open issue. This tutorial will summarize existing approaches to paraphrase evaluation, which include human evaluation, automatic evaluation, and application-driven evaluation.

The target audience will be NLP researchers, practitioners, and students. But participants do not need prior knowledge of paraphrasing.