Chinese Textual Sentiment Analysis: Datasets, Resources and Tools

Lun-Wei Ku
Institute of Information Science
Academia Sinica
128 Academia Road, Section 2
Nankang, Taipei 11529, Taiwan
lwku@iis.sinica.edu.tw

Wei-Fan Chen
Institute of Information Science
Academia Sinica
128 Academia Road, Section 2
Nankang, Taipei 11529, Taiwan
viericwf@iis.sinica.edu.tw

1 Description

The rapid accumulation of data in social media (in million and billion scales) has imposed great challenges in information extraction, knowledge discovery, and data mining, and texts bearing sentiment and opinions are one of the major categories of user generated data in social media. Sentiment analysis is the main technology to quickly capture what people think from these text data, and is a research direction with immediate practical value in big data era. Learning such techniques will allow data miners to perform advanced mining tasks considering real sentiment and opinions expressed by users in additional to the statistics calculated from the physical actions (such as viewing or purchasing records) user perform, which facilitates the development of real-world applications. However, the situation that most tools are limited to the English language might stop academic or industrial people from doing research or products which cover a wider scope of data, retrieving information from people who speak different languages, or developing applications for worldwide users.

More specifically, sentiment analysis determines the polarities and strength of the sentiment-bearing expressions, and it has been an important and attractive research area. In the past decade, resources and tools have been developed for sentiment analysis in order to provide subsequent vital applications, such as product reviews, reputation management, call center robots, automatic public survey, etc. However, most of these resources are for the English language. Being the key to the understanding of business and government issues, sentiment analysis resources and tools are required for other major languages, e.g., Chinese.

In this tutorial, audience can learn the skills for retrieving sentiment from texts in another major language, Chinese, to overcome this obstacle. The goal of this tutorial is to introduce the proposed sentiment analysis technologies and datasets in the literature, and give the audience the opportunities to use resources and tools to process Chinese texts from the very basic preprocessing, i.e., word segmentation and part of speech tagging, to sentiment analysis, i.e., applying sentiment dictionaries and obtaining sentiment scores, through step-by-step instructions and a hand-on practice. The basic processing tools are from CKIP Participants can download these resources, use them and solve the problems they encounter in this tutorial.

This tutorial will begin from some background knowledge of sentiment analysis, such as how sentiment are categorized, where to find available corpora and which models are commonly applied, especially for the Chinese language. Then a set of basic Chinese text processing tools for word segmentation, tagging and parsing will be introduced for the preparation of mining sentiment and opinions. After bringing the idea of how to pre-process the Chinese language to the audience, I will describe our work on compositional Chinese sentiment analysis from words to sentences, and an application on social media text (Facebook) as an example. All our involved and recently developed related resources, including Chinese Morphological Dataset, Augmented NTU Sentiment Dictionary (ANTUSD), E-hownet with sentiment information, Chinese Opinion Treebank, and the CopeOpi Sentiment Scorer, will also be introduced and distributed in this tutorial. The tutorial will end by a hands-on session of how to use these materials and tools to process Chinese sentiment.

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Tutorial Abstracts,
pages 5–8, Osaka, Japan, December 11-17 2016.

Tutorial Web Site: http://www.lunweiku.com/

2 Materials

Below is the summary of the materials that will be covered in this tutorial:

Resources: please see (Ku et al., 2011; Yohei et al., 2010; Ku et al., 2010; Ku et al., 2009; Ku et al., 2007; Ku et al., 2006; Wang and Ku, 2016; Chen and Ku, 2016; Chen et al., 2016).

Tools: please see (Chen et al., 2015; Ku et al., 2011; Ku et al., 2009; Ku and Chen, 2007).

3 Prerequisites

From which areas do we expect potential participants to come? Natural Language Processing, Web Mining, Machine Learning, Statistics, and Social Media Analytics

What prior knowledge, if any, do we expect from the audience? We do not require the audiences to have any background knowledge on the Chinese language. However, we expect the audience already understand some basic concepts and terminologies on natural language processing and sentiment analysis, such as part of speech tagging and opinion polarity.

What will the participants learn? The goal of this tutorial is to introduce the materials, resources and tool for Chinese sentiment Analysis. We will also highlight the main research challenges and unsolved issues in these areas, as there are still some room for improvement. Therefore, participants will not only acquire the knowledge and recent advances on Chinese sentiment analysis, but can also get ready for the basic Chinese text processing after this tutorial.

4 Lecturers

Lun-Wei Ku (lwku@iis.sinica.edu.tw) is now an Assistant Research Fellow in Institute of Information Science, Academia Sinica. She received her M.S. and Ph.D. degrees from Department of Computer Science and Information Engineering, National Taiwan University. Previously she worked as an assistant professor in the Department of Computer Science and Information Engineering, National Yunlin University of Science and Technology (Yuntech), Taiwan. Her research interests include natural language processing, information retrieval, and computational linguistics, especially on sentiment analysis. She has been working on Chinese sentiment analysis since year 2005 and was the co-organizer of NTCIR MOAT Task (Multilingual Opinion Analysis Task, traditional Chinese side) from year 2006 to 2010. Her international recognition includes CyberLink Technical Elite Fellowship (2007), IBM Ph.D. Fellowship (2008), ROCLING Doctorial Dissertation Distinction Award (2009), and Good Design Award Selected (2011). Other professional international activities she involved include: Member-at-Large, AFNLP (2016); Information Officier, ACM SIGHAN; Sentiment Analysis and Opinion Mining, Area Co-Chair, ACL-IJCNLP 2015 and EMNLP 2015; Publication Co-Chair, The 6th International Joint Conference on Natural Language Processing (IJCNLP 2013); Publicity Chair, The Twenty-fourth Conference on Computational Linguistics and Speech Processing (Rocling 2012); and Finance Chair, The Sixth Asia Information Retrieval Societies Conference (AIRS 2010).

Wei-Fan Chen (viericwf@iis.sinica.edu.tw) received the BS and MS degrees in communication engineering from National Chiao Tung University, in 2010 and 2012, respectively. He is a research assistant in the Institute of Information Science at Academia Sinica in Taipei, Taiwan. He has published papers in top, well recognized journals and conferences like IEEE TKDE (2016), COLING (2016), HCII (2015), AAAI symposium (2015), and ISCSLP (2012) and joined professional activities actively in NLP and AI domains. His research interests span a broad range of topics focusing on sentiment analysis, deep learning, computer-assisted language learning and speech processing.

5 Tentative Program

---

1. Overall Introduction (50 min)
Lecturer: Lun-Wei Ku
—-Definition, motivation, and challenge of the Chinese sentiment analysis
—-Introduction to related work and our previous results
—-Introduction to the Chinese language, mostly from the aspect of text processing

2. Introduction to the Resources and Tools (30 min)
Lecturer: Lun-Wei Ku
—-Available datasets
—-Available resources

Coffee Break: 20 min

---

3. Introduction to the Sentiment Analysis Tool: CopeOpi (20 min)
Lecturer: Lun-Wei Ku
4. The concept and design of CopeOpi Hands on: Real data (40 min)
Lecturer: Wei-Fan Chen
—-Getting data and environment ready
—-Preprocessing of the Chinese text: segmentation, par-of-speech tagging, parsing
—-Using NTUSD, ANTUSD and CopeOpi
—-Linking the sentiment and the lexical knowledge ontology

5. Final Wrap-up, Conclusion and Q/A (20 min)

Acknowledgements

Research of this paper was partially supported by Ministry of Science and Technology, Taiwan, under the contract MOST 104-2221-E-001-024-MY2.

References

Wei-Fan Chen and Lun-Wei Ku. 2016. UTCNN: a deep learning model of stance classification on social media text. In *COLING (to appear)*.

Wei-Fan Chen, Lun-Wei Ku, and Yann-Hui Lee. 2015. Mining supportive and unsupportive evidence from facebook using anti-reconstruction of the nuclear power plant as an example. In *Spring Symposium on Socio-Technical Behavior Mining: From Data to Decisions (2015 AAAI Symposium Series)*, pages 10–15.

Wei-Fan Chen, Fang-Yu Lin, and Lun-Wei Ku. 2016. WordForce: Visualizing controversial words in debates. In *COLING (to appear)*.

Lun-Wei Ku and Hsin-Hsi Chen. 2007. Mining opinions from the web: Beyond relevance retrieval. *Journal of the American Society for Information Science and Technology*, 58(12):1838–1850.

Lun-Wei Ku, Yu-Ting Liang, and Hsin-Hsi Chen. 2006. Opinion extraction, summarization and tracking in news and blog corpora. In *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs, AAAI Technical Report SS-06-03*, pages 100–107.

Lun-Wei Ku, Yong-Shen Lo, and Hsin-Hsi Chen. 2007. Test collection selection and gold standard generation for a multiply-annotated opinion corpus. In *Proceedings of 45th Annual Meeting of Association for Computational Linguistics (ACL)*, pages 89–92. Association for Computational Linguistics.

Lun-Wei Ku, Ting-Hao Huang, and Hsin-Hsi Chen. 2009. Using morphological and syntactic structures for chinese opinion analysis. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1260–1269. Association for Computational Linguistics.

Lun-Wei Ku, Ting-Hao (Kenneth) Huang, and Hsin-Hsi Chen. 2010. Construction of chinese opinion treebank. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 1315–1319.

Lun-Wei Ku, Ting-Hao (Kenneth) Huang, and Hsin-Hsi Chen. 2011. Predicting opinion dependency relations for opinion analysis. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 345–353.

Shih-Ming Wang and Lun-Wei Ku. 2016. ANTUSD: A large chinese sentiment dictionary.

Seki Yohei, Ku Lun-Wei, Sun Le, Chen Hsin-Hsi, and Kando Noriko. 2010. Overview of multilingual opinion analysis task at ntcir-8-a step toward cross lingual opinion analysis. In *Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access (NTCIR)*, pages 209–220.