XIAODAN ZHU

zhu2048@gmail.com
+1 (613) 762-0386

1200 Montreal Road, M-50
Ottawa, Canada

Education

Dec., 2009	PhD, Department of Computer Science, University of Toronto, Canada
Jul., 2000	M.E., Department of Computer Science, Tsinghua University, China

Research Experience

Jan., 2010–present

National Research Council (NRC) Canada; Researcher

Deep neural networks

I am working on deep high-order neural networks for structured output[2], and deep learning models for semantic (e.g., sentiment) embedding and composition [1]. (More work is in submission.)

Web text (e.g., social media) analysis; sentiment analysis

My recent research is also on analyzing social media text. In Semeval-2014 Task 9: Sentiment Analysis in Twitter, our systems (NRC-Canada) ranked first in five of the ten subtask-domain combinations among about 40 teams [4]. In Semeval-2014 Task 4: Aspect Based Sentiment Analysis, our models ranked first in three of the six subtasks among about 30 teams [8].

Earlier in SemEval-2013, among 40+ teams, we ranked first in Tweet sentiment analysis challenge[13] for both the tweet-level and phrase-level competition. Particularly, in this challenge I led our efforts on detecting term-level sentiment. An extension to the work is described in this recently accepted JAIR paper [3]. I am also working on modeling negation of sentiment [1] and the impact of sentiment in automatic machine translation [6].

Semantic structure analysis; spoken document understanding

To understand the document-level semantic/rhetorical structures, we investigated a less ambitious semantic-structure-alignment problem [16][18][19] that aims to find the correspondence between an existing hierarchical semantic structure and the text in which the structure is embedded (consider aligning a tree structure of electronic slides with the corresponding oral presentations as a special case). The ultimate goal of this project is to understand a more general problem, i.e., document-level semantic structure analysis. The problem is in general much more difficult but also more feasible when domain semantic and structural knowledge are available as a priori. For example, we hope to better analyze the structures of clinical discharge or progress

reports. All the work has been summarized in our recent IEEE TASLP paper [10].

Medical Informatics

We study information extraction (IE) for biomedical texts. Our models ranked at the top place in the i2b2-2010 international competition, among 40+ teams from the world [17]. Specifically, I led NRC's effort on relation detection, which was one of the three subtasks in the challenge and in general a core IE problem. Our relation model ranked at the top among the submitted systems (statistically tied with another top system). After the challenge, we further showed that incorporating syntactic kernels into our domain-semantics-abundant model can further improve it to achieve the best-ever results on the task [11].

We have also studied the problem of mining temporal relations between medical events. The model we built for the most recent i2b2-2012 Challenge was again top-ranked [12]. Indeed, our semantic relation models are based on a general framework that is designed to handle different types of semantic relations.

Citation analysis

Through investigating a number of language and non-language features[5], we tried to automatically identify the subset of references in a bibliography that have a central academic influence on the citing paper.

Sep., 2003–Dec., 2009

University of Toronto; PhD candidate

Extracting salient utterances from spoken documents

My thesis studied salient utterance extraction and redundancy removal in spoken documents. To this end, we reexamined a wide variety of textual and speech-related features, in addition to the impact of speech recognition errors [22][28]. Based on that, we further proposed an acoustics-based model that directly leverages repeated acoustic patterns to estimate both similarity and redundancy [21]. In addition, we also utilized additional semantic knowledge available in relevant written text to boost our performance [23][25].

Biomedical-text polarity analysis; question-answering

In a joint project, we studied the polarity detection of clinic outcomes through combining linguistic and domain knowledge [29]. We used the results to boost the performance of question-answering [26] conducted on medical text.

Jul. –Oct., 2009

May –Sep.,

IBM T.J. Watson Research Center (Yorktown Heights, NY); Research Intern

Machine translation

In two summers, I worked with IBM's T.J. Watson Research Lab in a project aiming to build a speech-to-speech machine translation system. My research manager (Dr.

2007

Bowen Zhou) and I built from scratch the core hierarchical (formally syntax-based) model, which was later used to compete for the DARPA Transtac challenge and was a top-ranked model. A later version that considers some soft syntactic constraints from statistical constituent parsers is described in [24]. In addition, I also worked on pruning the models to render a fast translation speed and reduce memory consumption.

Jun.–Sep., 2006

Google Inc. (New York, NY); Intern

Text mining on big data

I joined a project that aims to mine entities/objects and their attributes from the Web (the whole Google repository). I designed and implemented an unsupervised method to extend the coverage of the existing models in extracting such information, e.g., <*Bill Gates*, *Microsoft*, *Harvard*>, based on any given seeds of attributes. The model learns reliable semi-structured patterns of HTML tags to achieve this goal. The whole model was parallelized and could process the entire Google Web repository within several hours. In principle, this is the same type of problems as finding relational similarities in lexical semantics, although the HTML-tag sequences are utilized instead, which helps circumvent the difficulties of processing plain-text patterns and still collects a huge collection of reliable objects and attributes.

Jul.–Aug.,2005

Avaya Labs Research (Denver, CO); Research Intern

Call-type classification

My duty was to conduct call-type classification for customer-service call-centers. By using the agent-side audio only through speaker adaptive training, we found that the overall classification performance was comparable to that obtained by using both agents' and customers' audio, which suggests that there is no need to recognize customers' speech, given the inevitable higher word error rates on it.

Nov. 2001–Sep., 2003

Microsoft Research (Beijing, China); Visiting Scholar

Information extraction

I applied log-linear reranking methods to improve the performance of a state-of-the-art source-channel model for Chinese named-entity identification. Transformation-based learning was used afterwards to adapt the system for different standards used in SIGHAN evaluation. I also studied the identification of single-character Chinese named entities [31]. I built a finite-state-transducer toolkit to extract factoids (many types of named entities other than persons, locations, and organizations), which includes the functionalities of determinization (for sequentiable or subsequentiable FSTs) and building bi-machines (for unambiguous FSTs).

Jul., 2000–
Nov., 2001

Intel Research (Beijing, China); Full-time Researcher

Information extraction

I studied the indexing and search of closed caption of Chinese broadcast news [33]. I also studied Chinese query processing to match users' natural-language (non-keyword) queries with the frequently-asked-question (FAQ) database. New queries were detected and tracked in order to update the FAQ database of Intel's corporate-wide technical call-center recordings.

Sep., 1997–
Jul., 2000

Tsinghua University (Beijing, China); Master student

Information extraction

My Master's thesis studied two typical information extraction tasks: named-entity identification and template filling, in a specific domain: financial news. I adapted general-purpose models for Chinese word segmentation and shallow parsing to this specific domain. The named-entity extraction was based on a source-channel model and the template filling also used the parsing output. I studied general Chinese temporal representation with the aim to understand the temporal relationship between financial events [32][34]. I also joined a side project that studied the combinatorial regulations of Chinese semantic classes [36].

Publications

[1] Xiaodan Zhu, Hongyu Guo, Svetlana Kiritchenko, Saif Mohammad. An Empirical Study on the Effect of Negation Words on Sentiment. The 52th Annual Meeting of the Association for Computational Linguistics (**ACL-2014**). Baltimore, USA.

[2] Hongyu Guo, Xiaodan Zhu, and Renqiang Min (co-first authors). 2014. A Deep Learning Model for Structured Outputs with High-order Interaction, In Proceedings of NIPS Workshop on Representation and Learning Methods for Complex Outputs. Montreal, Canada.

[3] Svetlana Kiritchenko, Xiaodan Zhu, Saif Mohammad. State-of-the-Art in Sentiment Analysis of Short Informal Texts. (Accepted) Journal of Artificial Intelligence Research (**JAIR**).

[4] Xiaodan Zhu, Svetlana Kiritchenko, and Saif Mohammad. NRC-Canada-2014: Recent Improvements in the Sentiment Analysis of Tweets. Proceedings of the 2014 International Workshop on Semantic Evaluation. 2014, Dublin, Ireland **(describing our top-performing models in Semeval-2014 Task-9: Sentiment Analysis in Twitter)**

[5] Xiaodan Zhu, Peter Turney, Daniel Lemire, and Andre Vellino. Measuring academic influence: Not all citations are equal. Journal of the Association for Information Science and Technology, (**JASIST**), 66 (2), 408--427, 2015.

[6] Boxing Chen and Xiaodan Zhu. Bilingual Sentiment Consistency for Statistical Machine Translation. (Accepted to) Conference of the European Chapter of the Association for Computational Linguistics (**EACL-2014**).

[7] Saif Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. Sentiment, Emotion, Purpose, and Style in Electoral Tweets. (Accepted) Information Processing & Management (**IPM**).

[8] Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif M. Mohammad. NRC-Canada-2014: Detecting Aspects and Sentiment in Customer Reviews Proceedings of the 2014 International Workshop on Semantic Evaluation. 2014, Dublin, Ireland **(describing our top-performing models in Semeval-2014 Task-4: Aspect Based Sentiment Analysis)**

[9] Saif M. Mohammad, Xiaodan Zhu, and Joel Martin. Semantic Role Labeling of Emotions in Tweets. ACL Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA). 2014. Baltimore, USA.

[10] Xiaodan Zhu, Colin Cherry, and Gerald Penn. A Graph-partitioning Framework for Aligning Hierarchical Topic Structures to Presentations. IEEE Trans. on Audio, Speech, and Language Processing (**TASLP**) 21(5): 1102-1112 (2013).

[11] Xiaodan Zhu, Colin Cherry, Svetlana Kiritchenko, Joel Martin, and Berry de Bruijn. Detecting Concept Relations in Clinical Text: Insights from A State-of-The-Art Model. (Accepted). Journal of Biomedical Informatics (**JBI**) 20(5): 843-848 (2013).

[12] Colin Cherry, Xiaodan Zhu, Joel Martin, and Berry De Bruijn. A la Recherche du Temps Perdu - Extracting Temporal Relations from Medical Text in the 2012 i2b2 NLP Challenge. Journal of the American Medical Informatics Association (**JAMIA**). 20(5): 843-848 (2013) **(describing our top-performing models in i2b2-2012 Challenge on detecting temporal relations among medical events.)**

[13] Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu (co-first authors). NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. Accepted to Proceedings of the 2013 International Workshop on Semantic Evaluation. Atlanta, USA. June 2013**. (describing our top-performing systems in Semeval-2013 Task-2: Sentiment Analysis in Twitter)**

[14] Xiaodan Zhu. Spotting keywords and sensing topic changes in speech. Computational Intelligence for Security and Defence Applications. July, 2012.

[15] Anthony McCallum, Cosmin Munteanu, Gerald Penn, Xiaodan Zhu. Ecological Validity and the Evaluation of Speech Summarization Quality. NAACL Workshop on Evaluation Metrics and System Comparison for Automatic Summarization. June, 2012.

[16] Xiaodan Zhu, A Normalized-Cut Alignment Model for Mapping Hierarchical Semantic Structures onto Spoken Documents, The Fifteenth Conference on Computational Natural Language Learning (**CONLL-2011**), Portland, Oregon, USA

[17] Berry de Bruijn, Colin Cherry, Svetlana Kiritchenko, Joel Martin, and Xiaodan Zhu, Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010, Journal of the American Medical Informatics Association (JAMIA), May 2011 **(describing our top-performing systems in i2b2-2010 Challenge on identifying medical entities, their relations, and assertion of medical problems.)**

[18] Xiaodan Zhu, Colin Cherry and Gerald Penn, Indexing Spoken Documents with Hierarchical Semantic Structures: Semantic Tree-to-string Alignment Models, IJCNLP-2011, Changmai, Thailand

[19] Xiaodan Zhu, Colin Cherry, Gerald Penn: Imposing Hierarchical Browsing Structures onto Spoken Documents. **COLING-2010**, Beijing, China.

[20] Cosmin Munteanu, Gerald Penn, and Xiaodan Zhu. Improving Automatic Speech Recognition for Lectures through Transformation-based Rules Learned from Minimal Data. The 47th Annual Meeting of the Association for Computational Linguistics (**ACL-2009**), Singapore.

[21] Xiaodan Zhu, Gerald Penn and Frank Rudzicz. Summarizing multiple spoken documents: finding evidence from untranscribed audio. The 47th Annual Meeting of the Association for Computational Linguistics (**ACL-2009**), Singapore.

[22] Gerald Penn and Xiaodan Zhu. A critical reassessment of evaluation baselines for speech summarization of spontaneous conversations. The 46th Annual Meeting of the Association for Computational Linguistics (**ACL-2008**), Columbus, USA.

[23] Xiaodan Zhu, Xuming He, Cosmin Munteanu, and Gerald Penn. Using latent Dirichlet allocation to incorporate domain knowledge for topic transition detection. Proceedings of the International Conference on Spoken Language Processing (**Interspeech-2008**), Brisbane, Australia.

[24] Bowen Zhou, Bing Xiang, Xiaodan Zhu, and Yuqing Gao. Prior derivation models for formally syntax-based translation using linguistically syntactic parsing and tree kernels. ACL-2008 Second Workshop on Syntax and Structure in Statistical Translation, Columbus, USA.

[25] Xiaodan Zhu, Siavash Kazemian, and Gerald Penn. Identifying salient utterances from Web spoken documents using descriptive hypertext. IEEE Workshop on Spoken Language Technology (SLT-2008), Goa, India.

[26] Yun Niu, Xiaodan Zhu, and Graeme Hirst. Question answering in the medical domain: the role of clinical outcome and polarity. Proceedings of the American Medical Informatics Association 2006 Annual Symposium (**AMIA-2006**), Washington, D.C., USA.

[27] Xiaodan Zhu and Gerald Penn. Summarization of spontaneous conversations. Proceedings of the 9th International Conference on Spoken Language Processing (**Interspeech**-2006), Pittsburgh, Pennsylvania, USA.

[28] Xiaodan Zhu and Gerald Penn. Comparing the roles of textual, acoustic and spoken-language features on spontaneous-conversation summarization. Proceedings of the 11th Human Language Technology Conference / 5th Meeting of the North American Chapter of the Association for Computational Linguistics (**NAACL-2006**) (short), New York, USA.

[29] Yun Niu, Xiaodan Zhu, Jianhua Li, and Graeme Hirst. Analysis of polarity information in medical text. Proceedings of the American Medical Informatics Association 2005 Annual Symposium (**AMIA-2005**), Washington, D.C.

[30] Xiaodan Zhu and Gerald Penn. Evaluation of sentence selection for speech summarization. RANLP-2005 Crossing Barriers in Text Summarization Research Workshop, Borovets, Bulgaria.

[31] Xiaodan Zhu, Mu Li, Jianfeng Gao, and Chang-Ning Huang. Single character Chinese named entity recognition, ACL-2003 Second SIGHAN Workshop on Chinese Language Processing, Sapporo, Japan.

[32] Kam-Fai Wong, Wenjie Li, Chunfa Yuan, and Xiaodan Zhu. Temporal representation and classification in Chinese. International Journal of Computer Processing of Oriental Languages. 15(2): 211-230 (2002).

[33] Xiaodan Zhu, Qian Diao, and Joe F. Zhou. A two-character hash function for Chinese word indexing. In Proceedings of the 6th Joint Conference of Computational Linguistics (in Chinese), 2001.

[34] Xiaodan Zhu and Chunfa Yuan. An algorithm for situation classification of Chinese verbs, ACL-2000 Second Workshop on Chinese Language Processing, Hong Kong, China.

[35] Xiaodan Zhu. Information extraction from financial news and the related temporal information analysis. Masters thesis (in Chinese), Tsinghua University, 2000.

[36] Wei Xu, Chunfa Yuan, Changning Huang,and Xiaodan Zhu. A study on the combinatorial regulation of Chinese semantic classes, Communication of COLIPS.