Understanding Short Texts*

Zhongyuan Wang
Microsoft Research
zhy.wang@microsoft.com

Haixun Wang
Facebook Inc.
haixun@fb.com

1 Intended audience

- Researchers in natural language processing, data management, knowledge engineering, text mining, and information retrieval;
- Industrial practitioners in search, ads, semantic query processing, and other knowledge-powered applications.
- Junior researchers and graduate students in text analysis, who are potentially interested in large-scale knowledge representation and acquisition, machine learning, graph algorithms.

2 Introduction

Everyday, billions of short texts are being produced, including search queries, ad keywords, tags, tweets, conversations in messengers, social network posts, etc. Unlike documents, short texts have some unique characteristics which make them difficult to handle.

- First, short texts, especially search queries, do not always observe the syntax of a written language. This means traditional NLP techniques, such as syntactic parsing, do not always apply to short texts with good results.
- Second, short texts contain limited context. The majority of search queries contain less than 5 words, and tweets can have no more than 140 characters.

Because of the above reasons, short texts give rise to a significant amount of ambiguity, which makes them extremely difficult to handle. On the other hand, search engines, seem to be able to handle short texts (queries) quite well. This however does not mean search engines have an organic understanding of short texts, rather, the semantic gap is filled in by human click-through signals.

Human beings can understand short texts with ease, although many of them are ambiguous. How does the mind get so much out of so little, especially when the input data is sparse, noisy, and ambiguous? In 2011, a Science paper called "How to grow a mind: statistics, structure, and abstraction" (Tenenbaum et al., 2011) pointed out that "If the mind goes beyond the data given, another source of information must make up the difference." Lots of efforts have been put into filling this gap.

A growing number of approaches leverage external knowledge to address the issue of inadequate contextual information that accompanies the short texts. These approaches can be classified into two categories:

- **Explicit Representation Model (ERM)** Explicit approaches try to analyze and model short texts by following the traditional natural language processing steps, including *segmentation* (Hua et al., 2015), *labeling (sense disambiguation)* (Wang et al., 2015a; Song et al., 2011; Kim et al., 2013; Hua et al., 2015; Wang et al., 2015b), and *syntax structure (dependency parsing)* (Wang et al., 2014b). Take *Conceptualization* as an example, it maps the short text to concepts defined in a certain taxonomy or knowledge base. Bayesian rules (Song et al., 2011) and co-occurrence network (Hua et al., 2015) are leveraged for concept inference. A holistic

* You can find more materials from our tutorial website: http://www.wangzhongyuan.com/tutorial/ACL2016/Understanding-Short-Texts/

model (Wang et al., 2015b) is also proposed to allows all available signals to fully interplay in these subtasks to better understand short texts. Then based on explicit representation, which can be treated as both human understandable and machine understandable vectors, more advanced functions such as *term similarity* (Li et al., 2013), *short text similarity* (Song et al., 2011), *short text classifier* (Wang et al., 2014a) are proposed based on vector similarity measures. These functions can be further used for various applications such as *ads and search relevance* (Wang et al., 2015a), *query recommendation* (Wang et al., 2014a), and *web table understanding* (Wang et al., 2012). The most important advantage of explicit models is that its representation results are easily understood by human beings. Therefore, these models can be customized for different cases.

- **Implicit Representation Model (IRM)** On the other hand, implicit approaches tries to leverage big data and learning based techniques (especially deep learning) to model latent semantic representation for short texts. Lots of work focuses on mapping texts to semantic space, which is called *embedding*, ranging from *word embedding* (Mikolov et al., 2013c; Mikolov et al., 2013a), *phrase embedding* (Cho et al., 2014; Socher et al., 2010; Mikolov et al., 2013b; Yu and Dredze, 2015), to *sentence embedding* (Le and Mikolov, 2014; Palangi et al., 2015; Kiros et al., 2015). The well-known approaches includes: using surrounding context to predicate the central word/phrase/sentence, or vice versa. Usually, implicit approaches are designed for specific scenarios, such as *short text conversation* (Shang et al., 2015; Sordoni et al., 2015; Vinyals and Le, 2015) and *question answering* (Severyn and Moschitti, 2015; Qiu and Huang, 2015). In these given scenarios, it is more easily to get large training data. Then recurrent neural network (RNN), long short-term memory (LSTM), and their variants are widely used to train the model. Encoder-decoder framework is also frequently adopted with these models to capture the semantics of texts.

The purpose of this tutorial is to survey recent advances on the topic of short text understanding, and discuss fundamental problems, techniques as well as open issues in this vibrant area.

3 Tutorial Overview

This tutorial aims at presenting a comprehensive overview of short text understanding based on explicit semantics (knowledge graph representation, acquisition, and reasoning) and implicit semantics (embedding and deep learning). We note that no tutorial on the topic yet exist across NLP, web, IR, or databases conferences, and we believe that this tutorial is timely for both surveying the field, and educating both application developers and aspiring researchers.

3.1 Central theme

The tutorial is going to survey many applications, including search engines, ads, automatic question-answering, online advertising, recommendation systems, etc., that may benefit from short text understanding.

The central theme of the tutorial is representation, as in all these applications, the necessary first step is to transform an input text into a machine-interpretable representation, namely to "understand" the short text. We will go over various techniques in knowledge acquisition, representation, and inferencing has been proposed for text understanding, and we will describe massive structured and semi-structured data that have been made available in the recent decade that directly or indirectly encode human knowledge, turning the knowledge representation problems into a computational grand challenge with feasible solutions in sight.

3.2 Tutorial outline

Following is the outline of the tutorial. The total length is about 3 hours.

- **Part I. Introduction and foundations (20 min)** We will introduce the challenge of short text understanding, and its various applications, in order to motivate and inspire the audience of this problem area. This section will also provide a quick overview for the rest of the tutorial.

- **Part II. Explicit short text understanding (80 min)** We will introduce current popular knowledge base systems which are used for building explicit models. Then we will

introduce the explicit representation such as conceptualization for segmentation, labeling, syntax structure analysis, and applications.

- **Part III. Implicit short text understanding (60 min)** We will introduce the major approaches used for building word embedding, phrase embedding, and sentence embedding. Then we will introduce how deep neural networks are built on top of these embedding for short text related applications.

- **Conclusion** We will introduce open research and application challenges (10 min)

4 Related Tutorials

Part of this tutorial (learning the knowledge-base for text understanding) was presented at ACM Multimedia 2014 & 2015 entitled "Learning knowledge bases for text and multimedia" (Xie and Wang, 2014), which was the most attended tutorial at the conference (by attendee counts). The "Inferencing in Information Extraction: Techniques and Applications" (Barbosa et al., 2015) at ICDE 2015 is also related. But our proposal is the first that comprehensively study on short text understanding.

The estimate of the audience size: 100.

5 Proposer bios

Zhongyuan Wang is a Researcher at Microsoft Research Asia (MSRA). He leads two projects at MSRA: Enterprise Dictionary (knowledge mining from Enterprise) and Probase (knowledge mining from Web). He got his Ph.D. degree in computer science from Renmin University of China, and his PhD thesis is "Short Text Understanding". Zhongyuan Wang has published 20+ papers (including ICDE 2015 Best Paper Award on short text understanding) in the leading international conferences, such as VLDB, ICDE, IJCAI, CIKM, etc. He is also the co-author of the book "Web Data Management: Concepts and Techniques", published in 2014. His research interests include knowledge base, natural language processing, semantic network, machine learning, and web data mining. Homepage: `http://wangzhongyuan.com/en/`.

Haixun Wang is a research scientist / Engineering manager at Facebook. Before Facebook, he is with Google Research, working on natural language processing. He led research in semantic search, graph data processing systems, and distributed query processing at Microsoft Research Asia. He had been a research staff member at IBM T. J. Watson Research Center from 2000 - 2009. He was Technical Assistant to Stuart Feldman (Vice President of Computer Science of IBM Research) from 2006 to 2007, and Technical Assistant to Mark Wegman (Head of Computer Science of IBM Research) from 2007 to 2009. He received the Ph.D. degree in computer science from the University of California, Los Angeles in 2000. He has published more than 150 research papers in referred international journals and conference proceedings. He served PC Chair of conferences such as CIKM12 and he is on the editorial board of IEEE Transactions of Knowledge and Data Engineering (TKDE), and Journal of Computer Science and Technology (JCST). He won the best paper award in ICDE 2015, 10 year best paper award in ICDM 2013, and best paper award of ER 2009. Homepage: `http://haixun.olidu.com/`.

6 Conference

This proposal is submitted to ACL 2016.

References

Denilson Barbosa, Haixun Wang, and Cong Yu. 2015. Inferencing in information extraction: Techniques and applications. In *International Conference on Data Engineering (ICDE)*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*.

Wen Hua, Zhongyuan Wang, Haixun Wang, Kai Zheng, and Xiaofang Zhou. 2015. Short text understanding through lexical-semantic analysis. In *International Conference on Data Engineering (ICDE)*.

Dongwoo Kim, Haixun Wang, and Alice Oh. 2013. Context-dependent conceptualization. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence (IJCAI)*, pages 2654–2661. AAAI Press.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *NIPS*, pages 3276–3284.

Quoc V Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *ICML*.

Peipei Li, Haixun Wang, Kenny Zhu, Zhongyuan Wang, and Xindong Wu. 2013. Computing term similarity by large probabilistic isa knowledge. In *ACM International Conference on Information and Knowledge Management (CIKM)*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.

Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, and Rabab Ward. 2015. Deep sentence embedding using the long short term memory network: Analysis and application to information retrieval. *arXiv preprint arXiv:1502.06922*.

Xipeng Qiu and Xuanjing Huang. 2015. Convolutional neural tensor network architecture for community-based question answering. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1305–1311.

Aliaksei Severyn and Alessandro Moschitti. 2015. Learning to rank short text pairs with convolutional deep neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 373–382. ACM.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53th Annual Meeting of Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP'15)*.

Richard Socher, Christopher D Manning, and Andrew Y Ng. 2010. Learning continuous phrase representations and syntactic parsing with recursive neural networks. In *Proceedings of the NIPS-2010 Deep Learning and Unsupervised Feature Learning Workshop*, pages 1–9.

Yangqiu Song, Haixun Wang, Zhongyuan Wang, Hongsong Li, and Weizhu Chen. 2011. Short text conceptualization using a probabilistic knowledgebase. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence (IJCAI)*, pages 2330–2336. AAAI Press.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proc. of NAACL-HLT*.

Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. 2011. How to grow a mind: statistics, structure, and abstraction. *Science*, 331(6022):1279–85.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. In *ICML Deep Learning Workshop*.

Jingjing Wang, Haixun Wang, Zhongyuan Wang, and Kenny Zhu. 2012. Understanding tables on the web. In *International Conference on Conceptual Modeling*, October.

Fang Wang, Zhongyuan Wang, Zhoujun Li, and Ji-Rong Wen. 2014a. Concept-based short text classification and ranking. In *ACM International Conference on Information and Knowledge Management (CIKM)*, October.

Zhongyuan Wang, Haixun Wang, and Zhirui Hu. 2014b. Head, modifier, and constraint detection in short texts. In *IEEE 30th International Conference on Data Engineering (ICDE)*, pages 280–291. IEEE.

Zhongyuan Wang, Haixun Wang, Ji-Rong Wen, and Yanghua Xiao. 2015a. An inference approach to basic level of categorization. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 653–662. ACM.

Zhongyuan Wang, Kejun Zhao, Haixun Wang, Xiaofeng Meng, and Ji-Rong Wen. 2015b. Query understanding through knowledge-based conceptualization. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*.

Lexing Xie and Haixun Wang. 2014. Learning knowledge bases for text and multimedia. In *Proceedings of the ACM International Conference on Multimedia*, pages 1235–1236.

Mo Yu and Mark Dredze. 2015. Learning composition models for phrase embeddings. *Transactions of the Association for Computational Linguistics*, 3:227–242.