Deep Natural Language Processing for Search Systems

Weiwei Guo, Huiji Gao, Jun Shi, Bo Long
LinkedIn
Mountain View, California
{wguo,hgao,jshi,blong}@linkedin.com

ABSTRACT

Deep learning models have been very successful in many natural language processing tasks. Search engine works with rich natural language data, e.g., queries and documents, which implies great potential of applying deep natural language processing on such data to improve search performance. Furthermore, it opens an unprecedented opportunity to explore more advanced search experience, such as conversational search and chatbot. This tutorial offers an overview on deep learning based natural language processing for search systems from an industry perspective. We focus on how deep natural language processing powers search systems in practice. The tutorial introduces basic concepts, elaborates associated challenges, reviews the state-of-the-art approaches, covers end-to-end tasks in search systems with examples, and discusses the future trend.

KEYWORDS

Deep Learning, Natural Language Processing, Search Engine

ACM Reference Format:
Weiwei Guo, Huiji Gao, Jun Shi, Bo Long. 2019. Deep Natural Language Processing for Search Systems. In *42nd Int'l ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3331184.3331381

1 INTRODUCTION

Search engines deal with rich natural language data such as user queries and documents. Improving search quality requires processing and understanding such information effectively and efficiently, where natural language processing technologies are generally leveraged. As the representative data format in search systems, query or document data are represented as a sequence of words. Understanding such sequential information is generally a nontrivial task with traditional methods, with challenges from both data sparsity and data generalization. Deep learning models provide an opportunity to effectively extract the representative relevant information, thus better understand complicated semantics and underlying searcher intention. Recent years have seen the significant improvements brought by deep learning in various natural language processing tasks, indicating its great potential in promoting search systems.

SIGIR'19, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07.
https://doi.org/10.1145/3331184.3331381

However, developing deep learning models for natural language processing in search systems inevitably needs to meet the requirements of the complicated ecosystem of search engine. For example, some systems need frequent model updates, which excludes lengthy model training. In addition, low serving latency constraint precludes complex models from being used. How to keep model quality with relatively low complexity is a constant challenge faced by deep learning practitioners.

In this tutorial, we summarize the current effort of deep learning for natural language processing in search systems, as shown in Section 2. We first give an overview of search systems and natural language processing in search, followed by basic concepts of deep learning for natural language processing [8, 12, 13, 22]. Then we introduce how to apply deep natural language processing in search systems in practice: (a) **query/document understanding** that focuses on extracting and inferring relevant information from text data, specifically, classification [11, 16], entity tagging [10] and entity disambiguation [20]; (b) **retrieval and ranking** for semantically matching relevant documents [9, 14], where the strong latency restrictions can be alleviated by various methods [5, 7]; (c) **language generation** techniques designed to proactively guide/interact with users to further resolve ambiguity in the original search. Three representative generation tasks [3, 6, 15] are introduced that heavily rely on neural language modeling [1], sequence-to-sequence [21],or generative adversarial networks [4], etc. At last, we share our hands-on experience with LinkedIn search in **real-world scenarios**.

This tutorial gives a comprehensive overview of applying deep natural language processing techniques in above components through an end-to-end search system. In addition to traditional search engine, we include several use cases of advanced search systems such as conversational search [16–19] and task oriented chatbots [23]. We also highlight several important future trends, such as interacting with users via query generation, and latency reduction to meet the industry standard [2].

Targeted audience
The target audience is very broad. It is suitable for academic and industrial researchers, graduate students, and practitioners. After the tutorial, we expect the audience have learnt concepts and principles of applying state-of-the-art deep learning for natural language processing in search systems, and gained real-world experiences in an end-to-end search engine. No specific prerequisite is required, but some basic knowledge of deep neural networks will help.

2 TUTORIAL OUTLINE

1. Introduction (15 mins)

- Overview of Search Systems
- Natural Language Processing in Search

2. Deep Learning for Natural Language Processing (30 mins)

- Preliminaries
- Language Understanding
- Language Generation

3. Deep Natural Language Processing in Search Systems (75 mins)

(1) Query/Document Understanding
  - Entity Tagging: word level prediction
  - Entity Disambiguation: knowledge base entry prediction
  - Intent Classification: sentence level prediction

(2) Document Retrieval and Ranking
  - Efficient Candidate Retrieval
  - Deep Ranking Models

(3) Language Generation for Search Assistance
  - Query Suggestion: word-level sequence to sequence
  - Spell Correction: character-level sequence to sequence
  - Auto Complete: partial sequence to sequence

4. Real-world Examples (50 mins)

(1) An End-to-end Example of Deep Natural Language Processing in LinkedIn Search System

(2) Conversational AI
  - Chatbot for LinkedIn Help Center
  - Natural Language Search at LinkedIn

5. Future Trends and Conclusions (10 mins)

3 PRESENTERS' BIOGRAPHY

Dr. Weiwei Guo is a senior software engineer at LinkedIn where he leads several efforts to apply the deep learning models into search productions. Previously, he was a research scientist in Yahoo! Labs working on query understanding. He obtained his Ph.D. in Computer Science from Columbia University in 2015, and B.S from Sun Yat-sen University. Weiwei has published over 20 papers in top conferences including ACL, EMNLP, NAACL, SIGIR with 1000+ citations. He has served Program Committee for many conferences including ACL, EMNLP, NAACL, AAAI.

Dr. Huiji Gao leads the AI Algorithms Foundation team at LinkedIn. He has broad interests in machine learning/AI and its applications, including recommender systems, computational advertising, search ranking, and natural language processing. He received Ph.D. in Computer Science from Arizona State University, and B.S./M.S. from Beijing University of Posts and Telecommunications. He has filed over 10 U.S. patents and published 40 publications in top journals and conferences including KDD, AAAI, CIKM, WWW, ICDM, SDM, DMKD with thousands of citations.

Dr. Jun Shi is a staff software engineer at LinkedIn, where he leads various efforts on promoting natural language processing in search with deep learning technologies. His research interest lies in the area of machine learning with emphasis on natural language processing. Previously he worked at Yahoo!, Broadcom, Qualcomm and Intel. While at Yahoo!, he was an author of CaffeOnSpark and TensorflowOnSpark. He was a contributor to Tensorflow and created verbs interface for Tensorflow. Jun holds a doctoral degree in Electrical Engineering from UCLA. He was a co-recipient of 2009 IEEE Communications Society & Information Theory Society Joint Paper Award.

Dr. Bo Long leads LinkedIn's AI Foundation team. He also worked at Particle Media, Yahoo! Labs, IBM Watson and Google Lab. He has 15 years of experience in data mining and machine learning with applications to web search, recommendation, and social network analysis. He holds dozens of innovations and has published peer-reviewed papers in top conferences and journals including ICML, KDD, ICDM, AAAI, SDM, CIKM, and KAIS. He has served as reviewers, workshops co-organizers, conference organizer committee members, and area chairs for multiple conferences, including KDD, NIPS, SIGIR, ICML, SDM, CIKM, JSM etc.

REFERENCES

[1] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *JMLR* (2003).

[2] W. Bruce Croft, Donald Metzler, and Trevor Strohman. 2010. *Search engines: Information retrieval in practice.* Reading: Addison-Wesley.

[3] Shaona Ghosh and Per Ola Kristensson. 2017. Neural Networks for Text Correction and Completion in Keyboard Decoding. In *arXiv preprint.*

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS.*

[5] Song Han, Huizi Mao, and William J Dally. 2016. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *ICLR.*

[6] Yunlong He, Jiliang Tang, Hua Ouyang, Changsung Kang, Dawei Yin, , and Yi Chang. 2016. Personalized Language Model for Query Auto-Completion. In *CIKM.*

[7] Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. In *arXiv preprint arXiv:1705.00652.*

[8] Sepp Hochreiter and JÃijrgen Schmidhuber. 1997. Long short-term memory. *Neural computation* (1997).

[9] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *CIKM.*

[10] Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. In *arXiv preprint arXiv:1508.01991.*

[11] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *EMNLP.*

[12] Yann LeCun and Yoshua Bengio. 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks* (1995).

[13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS.*

[14] Bhaskar Mitra and Nick Craswell. 2018. An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval* (2018).

[15] Dae Hoon Park and Rikio Chiba. 2017. A neural language model for query auto-completion. In *SIGIR.*

[16] Chen Qu, Liu Yang, W. Bruce Croft, Johanne R. Trippas, Yongfeng Zhang, and Minghui Qiu. 2018. Analyzing and Characterizing User Intent in Information-seeking Conversations. In *SIGIR.*

[17] Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *CHIIR.*

[18] Sudha Rao and Hal DaumÃĺ III. 2018. Learning to Ask Good Questions: Ranking Clarification Questions using Neural Expected Value of Perfect Information. In *ACL.*

[19] Gary Ren, Xiaochuan Ni, Manish Malik, and Qifa Ke. 2018. Conversational Query Understanding Using Sequence to Sequence Modeling. In *WWW.*

[20] Wei Shen, Jianyong Wang, and Jiawei Han. 2014. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering* (2014).

[21] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *NIPS.*

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, ÅĄukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS.*

[23] Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2018. Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-Based Chatbots. In *ACL.*