Guest editorial

On information discovery with machine intelligence for language

Background

Information discovery concerns application under human control using "algorithms designed to analyze data or to extract patterns in specific categories from data" (Fayyad *et al.*, 1996; Klösgen and Żytkow, 1996). Natural language is one of the essential sources of information discovery applications (Sun *et al.*, 2017; Yim *et al.*, 2016; Young *et al.*, 2019). According to Mariani *et al.* (2019), the number of NLP papers increased from 24 in 1965 to 3,314 in 2015.

The average rejection rate of top ACL conferences, such as ACL, EMNLP, COLING and NAACL-HLT, has stayed above 70% in the past three years. At the same time, the number of submissions has exploded. As an example, ACL 2020 received 3,429 valid paper submissions, smashing previous records. This number represents a 15% increase over the number of submissions received for ACL 2019 (*2020Q1 Reports: ACL 2020 – Admin Wiki*, 2020).

In recent years, because of the challenge of "Big Data," the capability to process, analyze, store and understand these data sets is overwhelmed (Fan and Bifet, 2013). As Madden mentioned, it was "too big, too fast, and too hard" for the existing technology to handle (Che *et al.*, 2013). For instance, nowadays, to make rules by human beings or generate features manually from terabytes of unstructured tweets could be a nightmare for the researchers.

Fortunately, the development of machine intelligence empowered the researchers in the information discovery field. Machine learning has been applied in various kinds of tasks, such as multimedia concept retrieval, image classification, video recommendation, social network analysis and text mining (Pouyanfar *et al.*, 2019).

Machine learning, especially deep learning, played a significant role in the NLP field. Transfer learning on NLP has seen a series of breakthroughs since the year 2018. Fast.ai's ULMFiT was submitted to arXiv on January 18, 2018 (Howard and Ruder, 2018) and has already been cited 192 times (as of May 2, 2020). Google's BERT paper was submitted on October 11, 2018 and has 5,057 citations to date (Devlin *et al.*, 2019).

Applications

With the help of new transformer models and transfer learning technology, the NLP applications benefit a lot. Here are some typical examples.

Question answering and reading comprehension

Raffel *et al.* (2019) created a new big natural language data set called "Colossal Clean Crawled Corpus" and achieved state-of-the-art results on many benchmarks covering summarization. The most significant difference in performance the authors observed was that denoising objectives outperformed language modeling and deshuffling for pretraining. They found that the BERT-style objective performs best, though the prefix language-modeling objective attains similar performance on the translation tasks.

In study by Garg *et al.* (2019), the authors provide empirical evidence on the benefits of using TANDA on two commonly used benchmarks for AS2: WikiQA and TREC-QA, which enables a direct comparison with previous work. The authors have presented a novel approach for fine-tuning pretrained transformer models and tested it on a general natural language inference task, namely, answer sentence selection. It can be more effectively used for fine-tuning on the target NLP application, being more stable and easier to adapt to other tasks. Meanwhile, it is robust to noise. This is an essential advantage in terms of scalability, as the data of different target domains can be typically smaller than the data set for the transfer step, thereby causing the main computation to be factorized on the initial transfer step.

Online Q&A websites have attracted many users and are considered as reliable sources by experts from various fields. Annamoradnejad *et al.* (2020) experimented with the BERT model to solve a part of the problem of moderation actions in QA websites. Results confirm that by simple fine-tuning pretrained BERT model, the authors can achieve high accuracy, in little time and on less amount of data.

Yang *et al.* (2019) reviewed and compared the conventional AR language modeling and BERT for language pretraining. The authors observe that both BERT and XLNet perform partial prediction, i.e. only predicting a subset of tokens in the sequence. XLNet achieves substantial improvement over previous pretraining objectives on various reading comprehension tasks such as SQuAD and RACE.

Named entity recognition and text summarization

Li *et al.* (2020b) evaluated the proposed method on four NLP tasks, part-of-speech tagging, named entity recognition, machine reading comprehension and paraphrase identification. Hyperparameters are tuned on the corresponding development set of each data set. The authors propose the dice-based loss to narrow down the gap between training objective and evaluation metrics. Experimental results show that the proposed loss function helps to achieve a significant performance boost without changing model architectures.

Li *et al.* (2020a) reformalized the NER task as an MRC question-answering task. The proposed method obtains SOTA results on both nested and flat NER data sets, which indicates its effectiveness. The authors observe colossal performance boost on the nested NER data sets over previous state-of-the-art models, achieving F1 scores of 85.98%, 86.88%, 83.75% and 80.97% on ACE04, ACE05, GENIA and KBP-2017 data sets, respectively.

The current issue and full text archive of this journal is available on Emerald Insight at: **https://www.emerald.com/insight/2398-6247.htm**

Information Discovery and Delivery
48/3 (2020) 113–116
© Emerald Publishing Limited [ISSN 2398-6247]
[DOI 10.1108/IDD-08-2020-090]

Lee *et al.* (2019) used additional corpora of different sizes for pretraining and investigated their effect on performance. Results show that the performance of BioBERT on three NER data sets (NCBI Disease, BC2GM and BC4CHEMD) changes with the size of the PubMed corpus.

Yan *et al.* (2020) introduced ProphetNet, a sequence-to-sequence pretraining model that learns to predict future n-gram at each time step. ProphetNet achieves the best performance on both abstractive summarization and question generation tasks compared to the models using the same base scale pretraining data set. The new model also achieved state-of-the-art results on *CNN/DailyMail* and Gigaword using only about 1/3 of the pretraining epochs of the previous model.

Takase and Okazaki (2019) proposed length-dependent positional encodings, LDP E and LRP E that can control the output sequence length in transformer. The proposed method significantly improved the quality of headlines on the Japanese headline generation task while preserving the given length constraint. For English, the proposed method generated headlines with the desired length precisely and achieved the top ROUGE scores on the DUC-2004 test set.

Sentiment analysis

Munikar *et al.* (2019) used the pretrained BERT model and fine-tuned it for the fine-grained sentiment classification task on the SST data set. The authors compare the results with the standard RNN and the more sophisticated RNTN. Both of them were trained on SST from scratch, without pretraining. Even with such a simple downstream architecture, the model was able to outperform complicated architectures such as recursive, recurrent and convolutional neural networks.

Xie *et al.* (2019) studied the effect of TSA on two tasks with different amounts of unlabeled data, namely, Yelp-5 where the authors have only 2.5k labeled examples and 6 m unlabeled examples. Their new approach UDA uses state-of-the-art data augmentation found in supervised learning to generate diverse and realistic noise and enforces the model to be consistent with respect to these noises. The authors showed that UDA leads to consistent improvements across all labeled data sizes on IMDb and Yelp-2.

Machine translation

Edunov *et al.* (2018) trained on all available bitext excluding the ParaCrawl corpus and removed long sentences as well as sentence-pairs with a high source/target length ratio. The authors found that synthetic data can achieve up to 83% of the performance attainable with real bitext and achieved a new state-of-the-art result of 35 BLEU on the WMT'14 English-German test set by using publicly available benchmark data only.

For the low-resource neural machine translation, Imankulova *et al.* (2019) proposed a multilingual multistage fine-tuning approach and observed that it substantially improves Japanese and Russian translation by over 3.7 BLEU points compared to a strong baseline.

Topics covered

In addition to the academic attention garnered by these fields in recent years, these topics have moved into the public imagination, and we regularly see news and magazine articles about artificial intelligence, machine learning and natural language processing. For example, in 2007, the whole world witnessed the AlphaGo's ability to defeat the world Champion Go player (AlphaGo, 2020).

So, we decided to make *this special issue*, for a special issue concerning the application of Natural Language Processing with Machine Intelligence to Information Discovery very attractive.

In our Call For Paper (CFP), we invited authors worldwide to submit papers that address the questions listed further, as well as related questions not outlined in this proposal:

- language modeling for information retrieval;
- transfer learning for text classification;
- word and character representations for cross-lingual analysis;
- information extraction and knowledge graph building;
- discourse analysis at sentence level and beyond;
- synthetic text data for machine learning purposes;
- user modeling and information recommendation based on text analysis;
- semantic analysis with machine learning; and
- other topics related to this proposal.

To draft the CFP, we gathered and browsed 1,393 CFPs from "NLP" section of the website wikicfp.com. The technical terms in our sub-topics are not unique, yet we have found no CFP with the same focus on real-world application of these methods and techniques, especially in the information discovery area (NLP Call for Papers for Conferences, Workshops and Journals at WikiCFP, 2020).

Judging from the question list, one may tell we do care about the development of new techniques, but we are not especially interested in results only for the sake of outperforming an existing system. Sometimes, researchers are just "torturing the data until it confesses." We also care about reproducibility; far too many papers have been published without corresponding code release, making it difficult (or even impossible) for others to reproduce the reported results. Also, many papers have claimed to achieve a "state-of-the-art" result, yet the only improve performance from (e.g.) 96.1% to 96.2%. Is this a significant evidence of a better model or was the system just lucky from randomness?

We prefer work that brings new data sets into concern, rather than building more complex structures to solve problems on well-studied data sets such as the Yelp or IMDb review corpora. Even a "good enough" model can benefit millions of people who use them every day, as there are thousands of underrepresented languages in the world.

We care in particular about how to use state-of-the-art technology on real-world data, to meet information users' demands and to provide value to society.

Articles in this special issues

Eventually, we received many more submissions than we had ever expected. Because of the high-quality criteria and volume limitation of the *Journal of Information Discovery and Delivery*, we had to reject a lot of them. Here are the four submissions we finally accepted after several rounds of peer reviewing and revisions.

The paper "Aspect Context-Aware Sentiment Classification of Online Consumer Reviews" tried to solve the classification problem of online consumer reviews (OCR) with a hybrid machine learning approach. While the most related works suffered from the issue of ambiguous sentiment bearing words, the authors used the skip-gram model to grasp the context. Domain-independent seed words are used instead of expensive lexicons to make the trained model more scalable. Distributed word vectors were used in both cosine similarity calculation and feature generation Process. Experiments were carried out on Amazon mobile phone review and hotel reviews from Tripadvisor and the results showed not only the new approach can improve the accuracy but also low down the time complexity as well.

The research "Optimization of Hierarchical Reinforcement Learning Relationship Extraction Model" proposes a novel model BERT-HRL for entity and relation extraction from the unstructured data set. Different from previous studies, the proposed model combines both entity and relation extraction tasks into a unified framework and trains the parameters of two tasks collaboratively; the research also integrates BERT into the model to optimize the word embedding and encoding process. Some strategies, such as punctuation marks and positional information, are also introduced to optimize the model's performance. Experimental results show that the proposed model outperforms the baseline models with a 13% improvement in the NYT10 data set. The proposed BERT-HRL model is further verified based on extensive experiments and case studies. It also illustrates that it is an efficient solution for researchers in different academic domains to make use of the methodologies to process their unstructured text more accurately.

The article "Tracing the evolution of AI: Conceptualization of artificial intelligence in mass media discourse" tried to answer the big question, "What is artificial intelligence?" The research analyzed the public's perception of artificial intelligence (AI) from the perspective of media construction. The authors used five major news media reports on AI in the past 30 years to analyze the evolution of AI concepts from seven aspects: scientific subject, keyword, country, institution, people, topic and opinion polarity. Using text-mining methods such as named entity recognition and topic modeling, this research reveals that the concept of Ai is gradually fragmented because of the seizure of AI by the parties. The "artificial intelligence" defined by the news media is a fusion concept of science and business.

Furthermore, adverse reports mainly focus on various issues related to AI ethics. The research results can help us understand various discussions around AI and provide more perspectives on AI's functions, prospects and traps. At the same time, this research can promote a broader dialogue and expectations of various industries on the future development of AI.

In work "Emotional Communication Analysis of Emergency Microblog Based on the Evolution Life Cycle of Public Opinion," the authors explored the relationship between the vitality of a tweet, as measure by its diffusion quantity, the sentiment expressed by that tweet and its occurrence within the lifecycle of public opinion dissemination about an event in which that tweet occurred. The tweets about the emergency event of Hurricane Irma are used to explore the factors influencing information diffusion within social media during emergencies. The study uses the emotion classification scheme from Ekman's (1992) emotional theory to analyze how different sentiments expressed by tweets impact the information diffusion on the social network. Besides, the paper uses the concept of the lifecycle of public opinion dissemination to understand the frequency of tweets of particular sentiments during a particular phase of the lifecycle. The findings show that the emotion expressed by a tweet and lifecycle of public opinion dissemination influence the target retweeting behavior of the audience. Specifically, the results indicate negative emotions, such as "sadness" and "fear," are most influential during the initial and outbreak stages of an emergency event such as Hurricane Irma. The conclusion made in the research can shed light on the real-time management of public opinion transmission, efficient information gathering and emergency management plans.

Now, for our dear readers and colleagues, we do hope you can enjoy this Special Issue.

Shuyi Wang
School of Management, Tianjin Normal University, Tianjin, China and School of Information Resource Management, Tianjin Normal University, Tianjin, China

Chengzhi Zhang
School of Information Management, Nanjing University of Science and Technology, Nanjing, China and Nanjing University of Science and Technology, Nanjing, China, and

Alexis Palmer
Department of Linguistics, University of North Texas, Denton, Texas, USA

References

2020Q1 Reports: ACL 2020 – Admin Wiki (2020), 2020Q1 Reports: ACL 2020 – Admin Wiki, available at: www.aclweb.org/adminwiki/index.php?title=2020Q1_Reports:_ACL_2020 (accessed 2 May 2020).

AlphaGo (2020), "AlphaGo: the story so far", available at: from/research/case-studies/alphago-the-story-so-far (accessed 6 May 2020).

Annamoradnejad, I. Fazli, M. and Habibi, J. (2020), "Predicting subjective features from questions on QA websites using BERT", arXiv:2002.10107 [Cs], available at: http://arxiv.org/abs/2002.10107

Che, D., Safran, M. and Peng, Z. (2013), "From big data to big data mining: challenges, issues, and opportunities", in Hong, B., Meng, X., Chen, L., Winiwarter, W. and Song, W. (Eds), *Database Systems for Advanced Applications*, Vol. 7827, Springer Berlin Heidelberg, pp. 1-15, doi: 10.1007/978-3-642-40270-8_1

Devlin, J. Chang, M.-W. Lee, K. and Toutanova, K. (2019), "BERT: pre-training of deep bidirectional transformers for language understanding", arXiv:1810.04805 [Cs], available at: http://arxiv.org/abs/1810.04805

Edunov, S. Ott, M. Auli, M. and Grangier, D. (2018), "Understanding back-translation at scale", arXiv:1808.09381 [Cs], available at: http://arxiv.org/abs/1808.09381

Ekman, P. (1992), "An argument for basic emotions", *Cognition and Emotion*, Vol. 6 Nos 3/4, pp. 169-200.

Fan, W. and Bifet, A. (2013), "Mining big data: current status, and forecast to the future", *ACM SIGKDD Explorations Newsletter*, Vol. 14 No. 2, pp. 1-5, doi: 10.1145/2481244.2481246.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996), "The KDD process for extracting useful knowledge from volumes of data", *Communications of the ACM*, Vol. 39 No. 11, pp. 27-34.

Garg, S. Vu, T. and Moschitti, A. (2019), "TANDA: transfer and adapt pre-trained transformer models for answer sentence selection", arXiv:1911.04118 [Cs], available at: http://arxiv.org/abs/1911.04118

Howard, J. and Ruder, S. (2018), "Universal language model fine-tuning for text classification", arXiv:1801.06146 [Cs, Stat], available at: http://arxiv.org/abs/1801.06146

Imankulova, A. Dabre, R. Fujita, A. and Imamura, K. (2019), "Exploiting out-of-domain parallel data through multilingual transfer learning for low-resource neural machine translation", arXiv:1907.03060 [Cs], available at: http://arxiv.org/abs/1907.03060

Klösgen, W. and Żytkow, J.M. (1996), "Knowledge discovery in databases terminology", *Advances in Knowledge Discovery and Data Mining*, ACM, pp. 573-592.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H. and Kang, J. (2019), "BioBERT: a pre-trained biomedical language representation model for biomedical text mining", *Bioinformatics*, doi: 10.1093/bioinformatics/btz682.

Li, X. Feng, J. Meng, Y. Han, Q. Wu, F. and Li, J. (2020a), "A unified MRC framework for named entity recognition", arXiv:1910.11476 [Cs], available at: http://arxiv.org/abs/1910.11476

Li, X. Sun, X. Meng, Y. Liang, J. Wu, F. and Li, J. (2020b), "Dice loss for data-imbalanced NLP tasks", arXiv:1911.02855 [Cs], available at: http://arxiv.org/abs/1911.02855

Mariani, J., Francopoulo, G. and Paroubek, P. (2019), "The NLP4NLP corpus (I): 50 years of publication, collaboration and citation in speech and language processing", *Frontiers in Research Metrics and Analytics*, Vol. 3, pp. 36, doi: 10.3389/frma.2018.00036.

Munikar, M. Shakya, S. and Shrestha, A. (2019), "Fine-grained sentiment classification using BERT", arXiv:1910.03474 [Cs, Stat], available at: http://arxiv.org/abs/1910.03474

NLP Call For Papers for Conferences, Workshops and Journals at WikiCFP (2020), "NLP call for papers for conferences, workshops and journals at WikiCFP", available at: http://wikicfp.com/cfp/call?conference=NLP (accessed 2 May 2020).

Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., Shyu, M.-L., Chen, S.-C. and Iyengar, S.S. (2019), "A survey on deep learning: algorithms, techniques, and applications", *ACM Computing Surveys*, Vol. 51 No. 5, pp. 1-36., doi: 10.1145/3234150.

Raffel, C. Shazeer, N. Roberts, A. Lee, K. Narang, S. Matena, M. Zhou, Y. Li, W. and Liu, P.J. (2019), "Exploring the limits of transfer learning with a unified text-to-text transformer", arXiv:1910.10683 [Cs, Stat], available at: http://arxiv.org/abs/1910.10683

Sun, S., Luo, C. and Chen, J. (2017), "A review of natural language processing techniques for opinion mining systems", *Information Fusion*, Vol. 36, pp. 10-25.

Takase, S. and Okazaki, N. (2019), "Positional encoding to control output sequence length", arXiv:1904.07418 [Cs], available at: http://arxiv.org/abs/1904.07418

Xie, Q. Dai, Z. Hovy, E. Luong, M.-T. and Le, Q.V. (2019), "Unsupervised data augmentation for consistency training", arXiv:1904.12848 [Cs, Stat], available at: http://arxiv.org/abs/1904.12848

Yan, Y. Qi, W. Gong, Y. Liu, D. Duan, N. Chen, J. Zhang, R. and Zhou, M. (2020), "ProphetNet: predicting future N-gram for sequence-to-sequence pre-training", arXiv:2001.04063 [Cs], available at: http://arxiv.org/abs/2001.04063

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R. and Le, Q.V. (2019), "XLNet: generalized autoregressive pretraining for language understanding", in Wallach, H., Larochelle, H. Beygelzimer, A., d Alché-Buc, F., Fox, E. and Garnett, R. (Eds), *Advances in Neural Information Processing Systems*, Vol. 32, Curran Associates, pp. 5753-5763, available at: http://papers.nips.cc/paper/8812-xlnet-generalized-autoregressive-pretraining-for-language-understanding.pdf

Yim, W-W., Yetisgen, M., Harris, W.P. and Kwan, S.W. (2016), "Natural language processing in oncology: a review", *JAMA Oncology*, Vol. 2 No. 6, pp. 797-804.

Young, I.J.B., Luz, S. and Lone, N. (2019), "A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis", *International Journal of Medical Informatics*, Vol. 132, p. 103971.