Research Statement

Wei Xu, Assistant Professor, Department of Computer Science and Engineering, The Ohio State University

I want to tackle a grand challenge in AI and natural language processing (NLP) to make computers understand the complex meaning that is expressed in different human language forms. I design **machine learning** models for **natural language generation** with stylistic variations (e.g., generating variations of news texts to express the same meaning, but that are simpler so that even a 10-year old could read them) and learning semantics from large data for **natural language understanding** (e.g., recognizing texts that have the same meaning to track down the spread of misinformation). These two sides of my research are very connected from the data point of view, but explore very different learning algorithms. More generally, I am interested in creating robust technologies to process not only well-edited text, such as news, but also noisy **user-generated data**, such as that found in social media or biomedical lab instructions.

1 Natural Language Generation / Stylistics

The core of most text-to-text generation problems is sentential paraphrasing with stylistic constraints, which can be thought of as monolingual machine translation (e.g., English→Simple English). While automatic bilingual machine translation (e.g., German→English) has become better and better, natural language generation remains one of the most challenging research problems in NLP due to the complexity of the human editing process, the limited amount of high-quality data, and difficulties in evaluation.

Take **text simplification** (Figure 1), my favorite generation task, for example, it involves a delicate mixture of lexical and syntactic paraphrasing, compression, and sentence splitting in order to make text easier to read and understand. It is practically useful for children[1] and people with disabilities (e.g., deaf, dyslexia, autism) as well as many others to read medical or legal documents, etc. Simplification was popularly studied between 1997 and 2004 from the cognitive science perspective with rule-based methods, and then revived in 2010 because of the availability of Simple English Wikipedia and the development of statistical methods. However, in Xu et al. 2015 [2] and Xu et al. 2016 [3], I uncovered and analyzed two severe problems of the falsely assumed quality of the then-standard dataset derived from Wikipedia data and the evaluation setup using the BLEU [4] metric.[2] To address these issues, I created a new tunable metric, SARI [3], which is effective as a learning objective for training both statistical and neural machine learning models, and al-

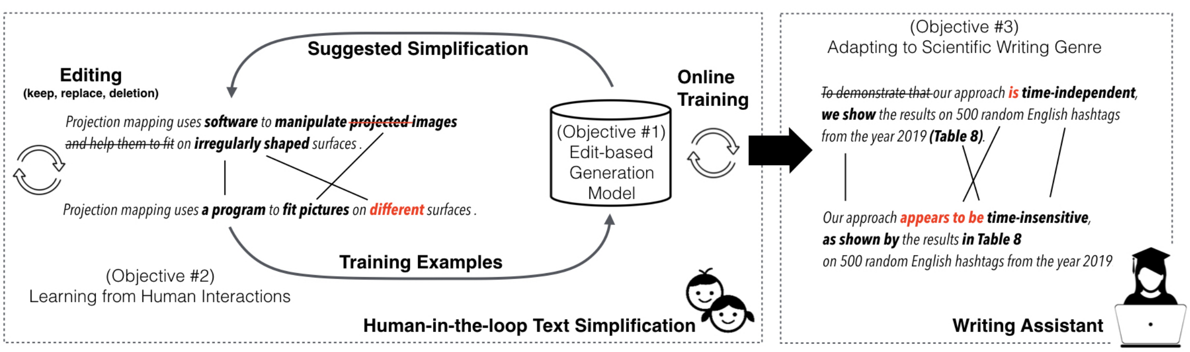

Figure 1. Illustration of my research plan on natural language generation, including three research objectives and two educational applications: (a) text simplification for children and (b) scientific writing for college students. The methodology is general and applicable for other generation tasks and text genres. The example editing adjustments shown in red can be collected from user interactions or simulated from static text corpora.

[1] According to a report released by U.S. Department of Education [1], more than 65% of 8th graders in American public schools are not proficient in reading and writing.

[2] BLEU is the most commonly used evaluation metric for automatic machine translation.

lows researchers to quickly iterate model designs. SARI is now implemented by the Google AI group in TensorFlow, a popular open-source library for training deep learning models. SARI has also been used for different natural language generation tasks, including sentence fusion, style transfer, and text-to-animation generation. We also introduced a high-quality crowdsourced corpus and a professionally edited simplification corpus [2] in collaboration with an education startup company, Newsela, which are now the standard benchmark used by research groups worldwide [5, 6, 7, 8, and others].[3] The automatic simplification system I designed [3] by optimizing syntactic machine translation models toward SARI is currently one of the best performing methods. More recently, my PhD student Mounica and I proposed a unified neural ranking model [9] with a Gaussian-based vectorization layer. It achieves state-of-the-art performance on three lexical simplification tasks, which ensemble a fast lightweight pipeline and will enable the development of a real-time interactive system (see future work below).

Besides simplification, I have worked on a variety of other natural language generation problems, ranging from **error correction** [10, 11], to **user's stylistic preferences** [12] and **paraphrase generation** [13, 14]. Each touches on different learning algorithms with different constraints and data sources. Currently, I am working on: (i) data augmentation for sequence-to-sequence models to combine different generation operations as part of a collaborative research project with HCI and deaf education experts, supported by NSF's Cyberlearning program; (ii) multilingual text simplification that simplifies inputs while retaining important information to help machine translation [15] and information extraction [16, 17, 18, 19, 20, 21, 22], as one of our innovations for the new IARPA's BETTER program; (iii) document-level generation to address the lack of research on how neural network models handle cross-sentence information and document structure [23]; and (iv) better semantic models and crowdsourcing methodology to create larger and higher-quality training data for generation (more details in Section 2). In particular, I found these research directions exciting for **future work**: (i) interactive and human-in-the-loop natural language generation (prototyped in [14]) that can effectively learn and incorporate user representations into generation for quality control and personalization; (ii) more controllable and interpretable neural generation models that can mimic human editing actions explicitly with specific paraphrasing modules, instead of the black-box sequence-to-sequence models; (iii) more complex natural language generation scenarios, including unsupervised and semi-supervised generation which are especially useful for multilingual cases, and new task of scientific writing and revisions. My **long-term goal** is to develop reading and writing assistant technologies, and to find more human-computer interaction (HCI) and education experts to collaborate with.

2 Natural Language Understanding / Semantics

My approach to natural language understanding is learning and modeling very-large-scale paraphrases [24]. Intellectually, I think that paraphrases are fascinating, allowing me to focus on elegant and scalable machine learning models for inferring semantic relationships between words and sentences. This research direction sprouted from my early work on automatic text summarization [25, 16], and is in the same vein as other great PhD theses and research on paraphrases, including from Regina Baziley [26], Chris Callison-Burch [27], and Percy Liang [28]. But, I take an entirely different approach to obtain paraphrases on a much larger scale and with a much broader range than any previous work, essentially by developing more robust machine learning models and leveraging social media data. These paraphrases can enable natural language processing systems, such as machine translation or automatic knowledge base construction, to handle rare words (e.g., `NetsBulls series` ↔ `Nets and Bulls games`), idiomatic expressions (e.g., `gets the boot from` ↔ `has been sacked by`), abbreviations (e.g., `Man City` ↔ `Manchester City`), language shifts (e.g., `is bananas`

[3] See NLP Progress for the datasets and the current state-of-the-art for the most common NLP tasks, including text simplification research progress written by Fernando Alva Manchego: `http://nlpprogress.com/english/simplification.html`

↔ `is great`), and other lexical variations (e.g., `oscar nom'd doc` ↔ `Oscar-nominated documentary`).

One of my ambitious plans is to construct **LanguageNet**[4] (Figure 2), a large-scale database for human language that continuously updates with paraphrases extracted from timely social media and news streams. To this end, we designed a series of unsupervised [11] and supervised learning methods for paraphrase identification from social media data (also applicable to question/answer pairs [29] for QA systems), ranging from multi-instance learning [30], to subword neural network models [31] and tree LSTM [29]. Our recent work [32] demonstrated the feasibility of obtaining paraphrases continuously from tweets and news that contain the same web links and trending topics. Our newest model using neural semi-Markov conditional random fields can align semantically similar text units across sentences more accurately [33]. We have also been working on improving crowdsourcing techniques to collect human-labeled data for training deep learning models to identify semantic relations. We are currently in the third iteration to refine our methods (first version released as a shared task at SemEval 2015 [34]), and we just collected the largest sentential paraphrase corpus to date of over 130,000 English sentence pairs this summer. While researchers have long recognized the importance of modeling semantic relations between sentences and developed various datasets, none of the existing datasets (e.g., MSRP [35], STS [36], SNLI [37]) have the same quantity, quality, and naturalness. I believe this **multi-year ongoing project** will fill the void and lead to significant impacts in natural language processing (NLP) research, analogous to how the ImageNet[5] project of labeled images, led by Feifei Li, has transformed computer vision research. I plan to extend the LanguageNet to include more different languages, and to support tracking information spread across multiple social media platforms. This line of my research is supported by a NSF CRII award and a research subcontract from the DARPA SocialSim program. It is also closely related to paraphrase generation, a popularly studied natural language generation (Section 1) task but with no existing large high-quality data. Other **future work** includes utilizing this new continuously updating semantic resource to improve **information extraction** [16, 17, 18, 19, 20, 21, 22] systems, as well as other NLP systems, to dynamically adapt to new emerging or low-frequency words, event phrases, and entities by pivoting through alternative paraphrases (e.g., `poshtel` ↔ `posh hostel` ↔ `luxury hostel`).

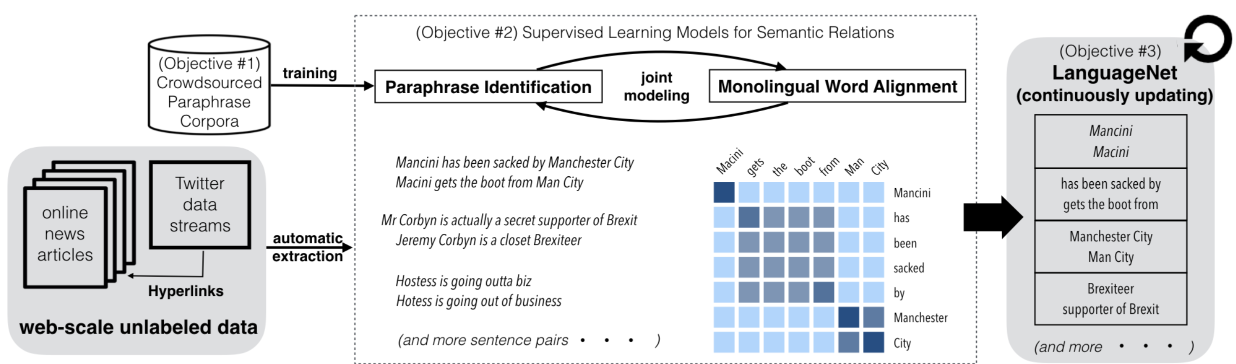

Figure 2. Illustration of my approach for learning large-scale paraphrases, designing automatic semantic models, and constructing a timely semantic resource, LanguageNet. Paraphrases are crucial for natural language understanding, such as in IBM's Watson QA system [38, 39] to connect questions and answers (e.g., `Who is the CEO that is stepping down from Boeing?` → `after Boeing Co. Chief Executive Harry Stonecipher was ousted from`).

[4] LanguageNet is one of the eight winning projects of the AI for Everyone Award in 2018.

[5] `https://en.wikipedia.org/wiki/ImageNet` *"AI researcher Fei-Fei Li began working on the idea for ImageNet in 2006. At a time when most AI research focused on models and algorithms, Li wanted to expand and improve the data available to train AI algorithms."*

3 Noisy User-generated Data / Social Media

For AI to truly understand human language and help people (e.g., instructing a robot), I think, we ought to study the language people actually use in their daily life (e.g., posting on social media), besides the formally written texts that are well supported by existing NLP software. I thus focus on specially designed learning algorithms and the data for training these algorithms to develop tools to process and analyze noisy user-generated data. I work a lot with Twitter data [25, 11, 40, 41, 42] because it is publicly available in large quantity and as it is easy for other researchers to access and reproduce results. Social media also contains very diverse languages for studying stylistics (Section 1) [12] and semantics (Section 2) [30, 34, 32], carrying information that is important for both people's everyday lives and national security. In the past three years, with my students, I have expanded my scope to cover a wider range of user-generated data, including biology lab protocols [43], Quora [29], Urban Dictionary, StackOverflow, and GitHub [44].

For example, in our ACL 2019 paper [45], we presented a novel neural ranking model (Figure 3) to analyze the semantic meaning of hashtags, which are often used in social media and carry important information. The open-source tool we released, **HashtagMaster**, can break hashtags into meaningful word sequences (`#IRGCOUTOFIRAQ` $\rightarrow$ `IRGC out of Iraq`[6]) with 92–95% accuracy, several points better than the previously available Microsoft Word Breaker API. Our proposed pairwise ranking model with an adaptive multi-task learning objective can capture the subtle differences between possible word sequences and better handle non-standard spelling variations, in contrast to the standard approaches used in previous work. This is an improvement and extension of the original neural ranking algorithm we proposed to measure the complexity of words/phrases [9] for text simplification (Section 1). Since the model is not language dependent, I plan to extend HashtagMaster to handle hashtags written in other languages, such as Arabic, and to introduce it to other research communities that often analyze hashtags, such as communication and political science.

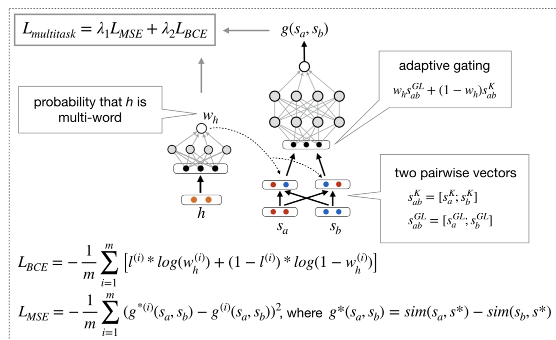

Figure 3. Pairwise neural ranking models with adaptive multi-task learning. Given two candidate segmentations s_a and s_b of hashtag h, the goal is to predict the segmentation's relative quality score g.

Another recent work of ours was on **Biomedical Lab Protocols**, published at NAACL 2018 [43], where we looked into automatically extracting machine-readable representations from biomedical lab procedures written in unstructured natural language text. Our methodology has since been picked up by researchers at UMass/MIT and extended to materials science procedural texts. I am currently working with students and collaborators to enlarge the dataset and plan to organize a shared task at the **Workshop on Noisy User-Generated Text (WNUT)** to facilitate research and development for scientific experiment automation. The WNUT Workshop is an annual one-day event I co-founded and have co-organized since 2015. It collocates with top NLP conferences (ACL/EMNLP/NAACL/COLING) and gathers around 100$\sim$150 researchers worldwide each year.

I am interested in instructional or procedural language, as it is understudied in the NLP research literature, yet is very important for future research in human-computer and human-robot interactions. Another exciting aspect of research on user-generated data to me is that I can work not only with human languages, but also with metadata and non-language data (e.g., programming code). This opens up a lot of opportunities to develop novel machine learning models, and to collaborate with people on interdisciplinary projects.

[6]IRGC, the Islamic Revolutionary Guard Corps, is a branch of the Iranian Armed Forces and designated as a terrorist organization by the government of the United States.

References

[1] *The 2017 National Assessment of Educational Progress (NAEP) Reading Report Card*, Retrieved on June 20, 2019. https://www.nationsreportcard.gov/reading_2017/nation/achievement?grade=8.

[2] **Wei Xu**, Chris Callison-Burch, and Courtney Napoles. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics (TACL)*, 3:283–297, 2015.

[3] **Wei Xu**, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics (TACL)*, 4:401–415, 2016.

[4] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, Pennsylvania, USA, 2002.

[5] Xingxing Zhang and Mirella Lapata. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 584–594, Copenhagen, Denmark, 2017.

[6] Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 85–91, 2017.

[7] Sanqiang Zhao, Rui Meng, Daqing He, Andi Saptono, and Bambang Parmanto. Integrating transformer and paraphrase rules for sentence simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3164–3173, Brussels, Belgium, 2018.

[8] Carolina Scarton and Lucia Specia. Learning simplifications for specific target audiences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 712–718, Melbourne, Australia, 2018.

[9] Mounica Maddela and **Wei Xu**. A word-complexity lexicon and a neural readability ranking model for lexical simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3749–3760, Brussels, Belgium, 2018.

[10] **Wei Xu**, Joel Tetreault, Martin Chodorow, Ralph Grishman, and Le Zhao. Exploiting syntactic and distributional information for spelling correction with web-scale n-gram models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1291–1300, Edinburgh, Scotland, UK, 2011.

[11] **Wei Xu**, Alan Ritter, and Ralph Grishman. Gathering and generating paraphrases from Twitter with application to normalization. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 121–128, 2013.

[12] Daniel Preoţiuc-Pietro, **Wei Xu**, and Lyle Ungar. Discovering user attribute stylistic differences via paraphrasing. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI)*, pages 3030–3037, Phoenix, Arizona, 2016.

[13] **Wei Xu**, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. Paraphrasing for style. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pages 2899–2914, Mumbai, India, 2012.

[14] Quanze Chen, Chenyang Lei, **Wei Xu**, Ellie Pavlick, and Chris Callison-Burch. Poetry of the crowd: A human computation algorithm to convert prose into rhyming verse. In *Proceedings of the 2nd AAAI Conference on Human Computation and Crowdsourcing (HCOMP): work-in-progress*, pages 10–11, Pittsburgh, Pennsylvania, 2014.

[15] Mingkun Gao, **Wei Xu**, and Chris Callison-Burch. Cost optimization in crowdsourcing translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 705–713, Denver, Colorado, 2015.

[16] Wenjie Li, **Wei Xu**, Mingli Wu, Chunfa Yuan, and Qin Lu. Extractive summarization using inter- and intra-event relevance. In *Proceedings of the 44th annual meeting of the Association for Computational Linguistics (ACL)*, pages 369–376, Sydney, Australia, 2006.

[17] Kristen Parton, Kathleen R. McKeown, Bob Coyne, Mona T. Diab, Ralph Grishman, Dilek Hakkani-Tür, Mary Harper, Heng Ji, Wei Yun Ma, Adam Meyers, Sara Stolbach, Ang Sun, Gokhan Tur, **Wei Xu**, and Sibel Yaman. Who, what, when, where, why? Comparing multiple approaches to the cross-lingual 5W task. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 423–431, Suntec, Singapore, 2009.

[18] **Wei Xu** and Ralph Grishman. A parse-and-trim approach with information significance for chinese sentence compression. In *Proceedings of the ACL Workshop on Language Generation and Summarisation*, pages 48–55, Suntec, Singapore, 2009.

[19] **Wei Xu**, Ralph Grishman, and Le Zhao. Passage retrieval for information extraction using distant supervision. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 1046–1054, Chiang Mai, Thailand, 2011.

[20] **Wei Xu**, Zhao Le, Raphael Hoffmann, and Ralph Grishman. Filling knowledge base gaps for distant supervision of relation extraction. In *Proceedings of the 2013 Conference of the Association for Computational Linguistics (ACL)*, pages 665–670, Sofia, Bulgaria, 2013.

[21] Maria Pershina, Bonan Min, **Wei Xu**, and Ralph Grishman. Infusion of labeled data into distant supervision for relation extraction. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 732–738, Baltimore, Maryland, 2014.

[22] Zhengbao Jiang, **Wei Xu**, Jun Araki, and Graham Neubig. Generalizing natural language analysis through span-relation representations. In *Submission*, 2019.

[23] Yang Zhong, **Xu, Wei**, and Junyi Jessy Li. Document and discourse factors for sentence deletion in text simplification. In *Submission*, 2019.

[24] **Wei Xu**. *Data-Drive Approaches for Paraphrasing Across Language Variations*. PhD thesis, New York University, New York, NY, USA, 2014.

[25] **Wei Xu**, Ralph Grishman, Adam Meyers, and Alan Ritter. A preliminary study of Tweet summarization using information extraction. In *Proceedings of the NAACL Workshop on Language Analysis in Social Media*, 2013.

[26] Regina Barzilay. *Information Fusion for Multidocument Summarization: Paraphrasing and Generation*. PhD thesis, Columbia University, New York, NY, USA, 2003.

[27] Chris Callison-Burch. *Paraphrasing and Translation*. PhD thesis, University of Edinburgh, Edinburgh, Scotland, 2007.

[28] Percy Liang. *CAREER: Interactive Training of Semantic Parsers via Paraphrasing*, 2016. `https://www.nsf.gov/awardsearch/showAward?AWD_ID=1552635`.

[29] Wuwei Lan and Wei Xu. Neural network models for paraphrase identification, semantic textual similarity, natural language inference, and question answering. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 3890–3902, Santa Fe, New Mexico, 2018.

[30] **Wei Xu**, Alan Ritter, Chris Callison-Burch, William B. Dolan, and Yangfeng Ji. Extracting lexically divergent paraphrases from Twitter. *Transactions of the Association for Computational Linguistics (TACL)*, 2:435–448, 2014.

[31] Wuwei Lan and **Wei Xu**. Character-based neural networks for sentence pair modeling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 157–163, New Orleans, Louisiana, 2018.

[32] Wuwei Lan, Siyu Qiu, Hua He, and **Wei Xu**. A continuously growing dataset of sentential paraphrases. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1224–1234, Copenhagen, Denmark, 2017.

[33] Lan Wuwei and **Xu, Wei**. Neural semi-markov conditional random fields for word alignment. In *Submission*, 2019.

[34] **Wei Xu**, Chris Callison-Burch, and William B. Dolan. SemEval-2015 Task 1: Paraphrase and semantic similarity in Twitter (PIT). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval)*, pages 1–11, 2015.

[35] Bill Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the 3rd International Workshop on Paraphrasing*, 2005.

[36] Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval)*, 2014.

[37] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 632–642, Lisbon, Portugal, 2015.

[38] Salim Roukos. Cognitive computing: an NLP renaissance! *Invited talk on IBM Watson at University of Pennsylvania*, 2015.

[39] Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. Paraphrase-driven learning for open question answering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sofia, Bulgaria, 2013.

[40] Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and **Wei Xu**. Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter lexical normalization and named entity recognition. In *Proceedings of the ACL 2015 Workshop on Noisy User-generated Text (WNUT)*, pages 126–135, Beijing, China, 2015.

[41] Benjamin Strauss, Bethany E Toma, Alan Ritter, Marie-Catherine de Marneffe, and **Wei Xu**. Results of the WNUT16 named entity recognition shared task. In *Proceedings of COLING 2016 Workshop on Noisy User-generated Text (WNUT)*, pages 138–144, Osaka, Japan, 2016.

[42] Jeniya Tabassum, Alan Ritter, and **Wei Xu**. TweeTime: A minimally supervised method for recognizing and normalizing time expressions in Twitter. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 307–318, Austin, Texas, 2016.

[43] Chaitanya Kulkarni, **Wei Xu**, Alan Ritter, and Raghu Machiraju. An annotated corpus for machine reading of instructions in wet lab protocols. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 97–106, New Orleans, Louisiana, 2018.

[44] Jeniya Tabassum, Mounica Maddela, **Wei Xu**, and Alan Ritter. An empirical study of named entity recognition in the computer programming domain. In *Submission*, 2019.

[45] Mounica Maddela, **Wei Xu**, and Daniel Preoţiuc-Pietro. Multi-task pairwise neural ranking for hashtag segmentation. In *The 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2538–2549, Florence, Italy, 2019.

Teaching Statement

Wei Xu, Assistant Professor, Department of Computer Science and Engineering, The Ohio State University

I teach three courses at the Ohio State University: an introductory AI course, an advanced AI course on natural language processing and speech, and a new course I designed that is related to my research. All these classes are open to both undergraduate and graduate students. I take different teaching approaches for different classes and audiences (see below for more details). My general teaching philosophy is that the lectures need to be interesting and engaging to keep student's attention for them to learn, while AI algorithms need to be taught in a serious way to be understood beyond just how to use them. I would share my experience working at Amazon and Microsoft to illustrate when and where I used a specific algorithm, while I use blackboard to walk through some key equations and algorithms besides using slides. There are programming-based homework assignments in all three classes, including some with open-ended questions to challenge students (e.g. use Twitter API to obtain data and analyze its language mix).

There are more than 1,200 CS undergraduates at OSU. I found many are very talented, eager to learn and take AI as their specialization. I am actively involved in both the undergraduate study and the curriculum committees in my department to shape and improve the undergraduate curriculum. XXX add about undergraduate research XXX

CSE 5539-0010 Social Media and Text Analysis (http://socialmedia-class.org)

I developed this new course to teach undergraduate and graduate students practical skills as well as the most timely research topics. The course website with open-sourced educational material is publicly accessible and actively updated. The *Twitter API Tutorial* I made has been used by many researchers and teachers in different universities and disciplines, including linguistics and business analytics (e.g., University of Utah CS1060, Georgia Tech CS6452, Iowa State University MIS546).

For the new class I designed, I also have taught a simplified version at the North American Summer School on Logic, Language, and Information (NASSLLI) in 2016, and a mini version with hands-on examples on Google Colab for high school students.

CSE 5522 Artificial Intelligence II

For this introductory AI course, I assign take-home quizzes almost weekly to help students keep up, and for me to see if there is anything many students missed so I can do a quick recap if necessary. I generally encourage students ask me questions in the class and take time to answer them.