Wei Xu

Research Statement

My research enables computers to understand and generate human language by learning from billions of Internet users. I develop new machine learning and probabilistic models for automatically gleaning semantic knowledge about language from massive social media and web data, with a variety of applications.

In particular, I advance research on paraphrase which concerns phrases or sentences that have similar meaning (i.e. synonyms). Numerous researchers need and have used paraphrases for a wide range of applications, including question answering, information extraction, machine translation, summarization, textual entailment, etc. For example, it is crucial in IBM's Watson system to connect questions and answer clues (e.g. "whose arrival in ... ?" ↔ "... landed in ..."). Paraphrase is a very challenging research topic because of the huge search space and the diverse wordings — a few researchers have made major contributions in paraphrase acquisition and they are renowned in the field of natural language processing. Intellectually I think that paraphrase is fascinating, allowing me to focus on elegant and scalable models for inferring relationship between words and sentences. My unique contributions include:

- Designing the first successful models to extract paraphrases from Twitter's massive data stream [1, 2, 3, 4]. My models scale to billions of sentences and provide much broader coverage than previous paraphrase research of only several thousands sentences. These web-scale paraphrases enable natural language systems to handle errors (e.g. "everytime" ↔ "every time"), lexical variations (e.g. "oscar nom'd doc" ↔ "Oscar-nominated documentary"), rare words (e.g "NetsBulls series" ↔ "Nets and Bulls games"), and language shifts (e.g. "is bananas" ↔ "is great"). But it is difficult to capture such lexically divergent paraphrases by the conventional similarity-based approaches. I invented joint word-sentence models that use multiple instance learning to infer latent word relations and relax the reliance on sentence similarity. My new probabilistic graphical model is the current state-of-the-art, outperforming deep learning and latent space methods. Using social media data also gives me a unique edge to make broader impacts on social science, political science, security, etc [5].

- Advancing natural language generation via paraphrasing. Generating sentential paraphrases faces a large search space as bilingual translation, but a much smaller optimal solution space due to specific task requirements. To address these nuances of monolingual generation, I advocate for informed adaptations of statistical machine translation framework [6, 7, 8, 2, 5]. I established a new line of research that targets different language styles and handles user-generated or historic or complex texts. I also uncovered multiple serious problems in the state-of-the-art results reported between 2010 and 2014 in text simplification research [9].

I am looking forward to developing higher performance models and open source paraphrase data resources, which are currently scarce but in high demand, to benefit both the research community and industry. In what follows, I will discuss two major research thrusts (paraphrase extraction and generation) and my research agenda in detail.

1 Extraction: Joint Word-Sentence Graphical Models

Probabilistic graphical models provide a principled and interpretable way to model semantic relationship in language. I design new graphical models using multiple instance learning to reduce reliance on expensive human annotation and — in a novel way — to allow more flexible word-sentence relations. My models are scalable for extracting both open-domain and event-specific paraphrases from large volumes of data:

Joint Open-domain Paraphrase Models Twitter engages millions of users, who naturally talk about the same topics simultaneously and frequently convey similar meaning using diverse linguistic expressions. It provides paraphrases at web-scale that did not exist before [3], but also presents serious challenges including a huge pairwise search space and great lexical divergence. The popular similarity-based methods cannot handle paraphrases like "Mancini has been sacked by Manchester City" $\leftrightarrow$ "Mancini gets the boot from Man City" that have limited overlap in surface or low-dimensional space; nor sentence pairs like "Somebody took the Marlins to 20 innings" $\not\leftrightarrow$ "Anyone who stayed 20 innings for the Marlins" that share many words in common but have different meanings. To address these challenges, I proposed a novel joint word-sentence approach based on the assumption that: two sentences under the same topic are paraphrases if and only if they contain at least one word pair as paraphrase anchor [1]. The topics are phrases that have a sharp frequency increase in the data stream (e.g. "Mancini" in the above example). The anchor is a pair of two identical or different words that are indicative of sentential paraphrase. The topical constraint narrows the search space and ensures accuracy, while the at-least-one assumption relaxes the reliance on lexical matching to increase coverage and diversity.

I created a log-linear latent variable model to operationalize this intuition through multiple-instance learning, where labels are only observed on sentence-level paraphrase candidates instead of labels on each individual word pair [1]. Figure 1 shows the model structure and an example instantiation. The word and sentence layers are joined by a deterministic-or factor to integrate the assumption. It is a supervised model that offers state-of-the-art performance and uses training data I created by crowdsourcing. The word pair anchors, which are difficult to annotate, are latent in the model and are inferred from the sentence-level labels during learning. The model is very scalable, using non-lexicalized features, online learning and the Viterbi approximation that replaces costly

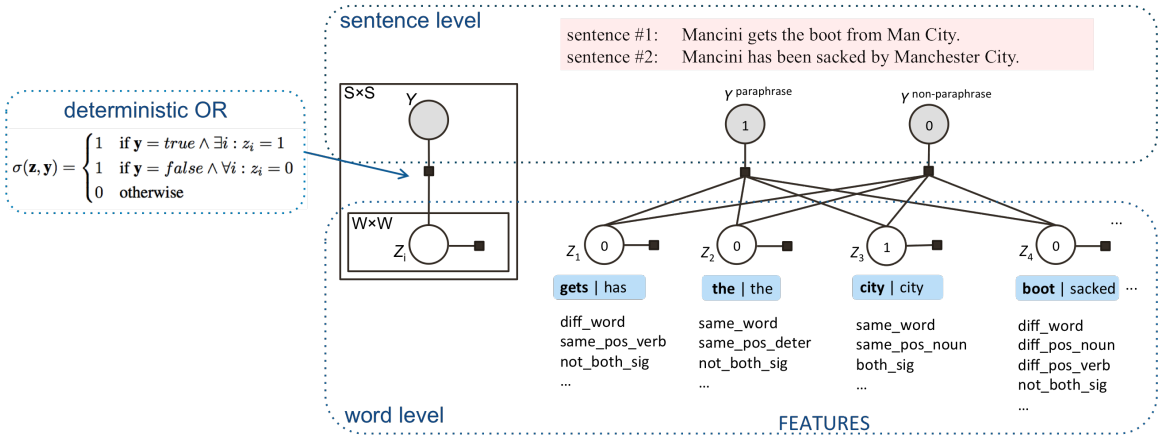

Figure 1: Multi-instance Learning Paraphrase Model is a latent variable model that jointly reasons between words and sentences and determines whether sentence pairs convey similar meaning.

expectation calculations with maximizations during parameter learning.

My work [1, 2, 3] has sparked wide interests from both academia and industry. In 2015, I organized a shared-task on Paraphrase and Semantic Similarity in Twitter [4] at SemEval, the major venue for evaluation of computational semantic systems. 19 teams including Stanford, Columbia, University of Maryland, MITRE participated in the competition. More than 140 research groups and companies have requested my open-source code and data since. I was invited to present my work at a DARPA PI meeting.

Joint Event-specific Paraphrase Models (for Information Extraction) My methods using multi-instance learning are also applicable for extracting structured relational data (e.g. founded[Steve Jobs, Apple Inc.]) from raw text (e.g. "Steve Jobs created Apple Inc. in the ..." $\leftrightarrow$ "... the life of Steve Jobs, the Apple Inc. co-founder ...") [10, 11, 12]. The conditional probability models allow joint reasoning about aggregated cross-sentence and individual sentence-level decisions. They leverage large-scale knowledge bases (e.g. Wikipedia Inbox or Freebase) to heuristically label millions of sentences for training, instead of relying on costly human annotation. I fixed a major issue of inherited noise in the automatically created training data by adding constraints [12] or feedback layers [10] into the graphical models. I also showed that event-based approaches can improve automatic summarization of news articles [13] and social media posts [14].

2 Generation: A Statistical Text-to-Text Framework

The core of most text-to-text (T2T) generation problems is sentential paraphrasing, which can be thought of as monolingual machine translation. I advocate for a statistical generation framework by adapting and building on top of statistical machine translation (SMT) technology.

Comparing to bilingual machine translation, T2T generation looks for not any meaning-equivalent sentence of an input, but one with different and specified wordings. For example, text simplification aims for outputs that are more readable by people with limited language skills, such as children and people with disabilities, dyslexia or autism (Figure 2). In other words, T2T requires an exponential search space larger than SMT, but has a smaller optimal solution space. To address the first challenge, I use phrasal and syntactic paraphrase rules extracted from web-scale data [3, 8, 6]. This enables construction of a much more diverse set of possible sentence outputs than previous work that uses monolingual parallel corpus of limited size. Specially, paraphrases I extracted from Twitter contain both ill- and well-formed variations. This allows me to build a phrase-based text normalization (e.g. "outta biz" $\leftrightarrow$ "out of business") system using additional error controls in SMT decoding [3]. It overcomes previous work's limitation to only word-to-word normalization. At the same time, larger search space aggravates the difficulty to find outputs that satisfy the task-specific requirements. In fact, I had to design new light-weight objective functions and

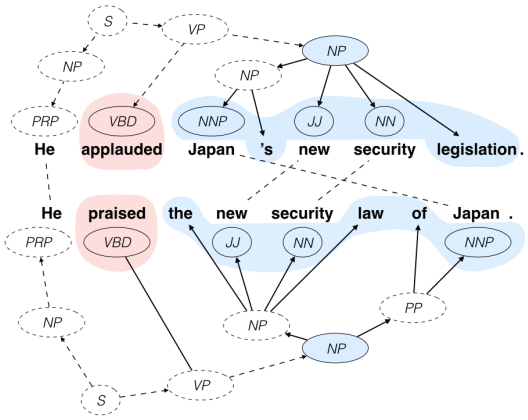

Figure 2: Syntax-based statistical machine translation machinery is adapted to simply input (top) using both lexical and syntactic paraphrases.

rich rule-level feature functions to work collectively for discriminative parameter tuning. My most recent work sets the new state-of-the-art for text simplification using an efficient pairwise ranking optimization that is scalable with regard to the number of features [6]. It is implemented into the open source syntax-based SMT framework, Joshua, to perform complex structural transformations (Figure 2). I also showed that syntactic information can improve spelling error correction [8]. Most importantly, I discovered that the state-of-the-art approaches in text simplification published between 2010 and 2014 were based on false assumptions of data quality and biased evaluation setup. My arguments were published in 2015 by TACL and well received by the research community [9].

Besides interesting technical challenges, paraphrase generation also has interesting real-world applications, such as for education and digital humanities:

Text Simplification I collaborate with an education tech start-up to simply news articles for school curriculum under the U.S. Government's Common Core education initiative. My work is supported by a NSF EAGER grant, which I co-authored [6, 9].

Error Correction I improved noisy text normalization [3] and error correction for language education [8] in an collaboration with researchers at Educational Testing Service, the organization that provides SAT, TOEFL, GRE tests, etc.

Stylistic Rewriting I was the first to generate paraphrases targeting a specific writing style, demonstrating the capability to automatically transform Shakespeare's plays into contemporary modern English and vice versa [7]. This research makes historic books more accessible and shows potential in writing aid systems (e.g. more feminine or sophisticated sounding) [5]. I also mentored two undergraduate students to do a fun project that converts prose into sonnets by combining automatic and human computing algorithms [15]. It is one of my many works [1, 2, 6, 16] that use **crowdsourcing** for natural language processing.

3 Future Work

My research interests lie at the intersection of machine learning, natural language processing, and social media. To facilitate related research, in 2015 I founded and co-organized a global Workshop on Noisy User-generated Text (`http://noisy-text.github.io/`) that was collocated with ACL, the flagship conference in natural language processing. I plan to deepen my research in three areas and expand to more interdisciplinary research:

Big Data Paraphrase Since this is a very new topic, there are many directions to explore. I think the key research challenge continues to be how to capture richer structure in the data — interactions between words, sentences, cross-sentence topics, social (user) graph, etc. Another key emphasis is developing scalable algorithms that can make use of large and streaming data. Some ideas I have been thinking of and excited about are: 1) aligning words/phrases between sentences of same meaning but very different wordings; 2) learning phrasal and sentential paraphrases jointly; 3) improving concurrently lexical paraphrases and distributional word representations, which are essential in deep learning NLP approaches; 4) studying structural paraphrases that can help text simplification or machine translation between languages of different word orders; 5) incorporating social context from social network platforms such as author, time, geographic and social connection metadata; 6) streaming paraphrase models that can capture newly-coined words/phrases as

they are invented. I also plan to extend to multilingual paraphrases which will allow evidence-sharing across languages, most importantly from resource-rich languages like English to resource-poor languages like Arabic or Turkish.

Social Media Text Analysis Current social media analytics often use word-based approaches. I am taking it to the next level — deeper analysis of text meaning. Paraphrase, as a fundamental semantic relation in natural language, is my secret weapon. Knowing whether two sentences have similar meaning can directly help event detection in social media and automatic summarization. I plan to 1) capture stylistic differences across different demographic groups that will complement my research on generation (our preliminary study was recently accepted at AAAI 2016 [5]); 2) mine paraphrases for special domains or purposes such as politics, health, sports, terrorism, cybersecurity, etc; 3) extract slang terms and their meaning explanations (a special form of paraphrases). I am advising one undergraduate and one masters student conducting pilot studies. I also have an ambitious long-term plan that is to create digital textbooks for different languages, especially colloquialisms, using paraphrases, translations and images learned from social media or other user-generated data. It can potentially help computers to better model human language and vision at the same time.

Neural Language Generation I am interested in developing deep learning approaches for text generation. The neural network encoder-decoder framework was recently proposed [17, 18] for bilingual machine translation in 2014. Preliminary studies by Facebook, Google Research and Microsoft Research in 2015 show promising results on sentence compression [19, 20]. I am working to create the first neural model for text simplification and plan to extend it to other language generation tasks. I am especially interested in exploring four new ideas for modeling: 1) incorporating paraphrases and other linguistic knowledge; 2) relying on less parallel text; 3) handling unseen words such as proper names; and 4) paraphrasing to mimic specific styles or personality traits (e.g. formal vs. informal, man vs. woman, polite vs. impolite, professional vs layman).

Interdisciplinary Research (Education, Bioinformatics, Digital Humanities, Political Science) Since machine learning techniques are widely applicable and text data are widely available in many domains, for my long-term research plan, I will not limit myself to a single domain. I plan to search for good collaboration opportunities that can have substantial impact. My current work has already touched on topics involving education [6, 8, 9], social science [1, 3, 4, 5] and digital humanities [7, 15]. My expertise in text analytics for noisy user-generated content can directly extend to electronic health records, political debate transcripts, discussion boards in massive open online course (MOOC), etc. I am also interested in further advancing the multiple instance learning and paraphrase techniques, which have been applied to the biomedical informatics field for drug activity prediction and medical terminology discovery.

References

[1] **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)*, 2014.

[2] **Wei Xu**. *Data-Driven Approaches for Paraphrasing Across Language Variations*. PhD thesis, Department of Computer Science, New York University, New York, 2014.

[3] **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 (BUCC)*, 2013.

[4] **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)*, 2015.

[5] 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*, 2016.

[6] **Wei Xu**, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing statistical machine translation for simplification. *Transactions of the Association for Computational Linguistics (TACL)*, accepted.

[7] **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)*, 2012.

[8] **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)*, 2011.

[9] **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)*, 2015.

[10] **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)*, 2013.

[11] **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)*, 2011.

[12] Maria Pershina, Bonan Min, **Wei Xu**, and Ralph Grishman. Infusion of labeled data into distant supervision for relation extraction. In *Proceedings of the 2014 Conference of the Association for Computational Linguistics (ACL)*, 2014.

[13] Wenjie Li, **Wei Xu**, Mingli Wu, Chunfa Yuan, and Qin Lu. Extractive summarization using inter- and intra-event relevance. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (COLING-ACL)*, 2006.

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

[15] 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)*, 2014.

[16] 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)*, 2015.

[17] Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. Sequence to sequence learning with neural networks. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2014.

[18] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *arXiv preprint arXiv:1409.0473*, 2014.

[19] Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015.

[20] Katja Filippova, Enrique Alfonseca, Carlos A Colmenares, Lukasz Kaiser, and Oriol Vinyals. Sentence compression by deletion with LSTMs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015.