Research Description

Wei-Yun Ma
Computer Science at Columbia University
wm2174@columbia.edu
http://www.cs.columbia.edu/~ma/

My research is in the area of Natural Language Processing (NLP). I am fascinated by the idea of giving computers the ability to process human language. I like to develop formal, computational models from a statistical perspective; but I also enjoy making greater use of linguistic knowledge in the model design. For me, integrating the two fields of statistics and linguistics to solve practical problems is one of the greatest challenges in NLP and is also one of the most interesting parts.

Over the past few years, my research has focused primarily on Multi-Engine Machine Translation (MEMT), which attempts to achieve better translation performance by fusing or selecting the output of multiple translation engines. I have also worked on many fundamental problems of Chinese NLP, including unknown word identification and automatic grammatical information acquisition.

Past and Current Research

Multi-Engine Machine Translation

Given the wide range of successful statistical MT approaches that have emerged recently, including phrase-based MT [1], hierarchical phrase-based MT [2] and syntax-oriented MT [3,4], it would be beneficial to take advantage of their individual strengths and avoid their individual weaknesses. MEMT attempts to do so by either fusing the output of multiple translation engines or selecting the best one among them, aiming to improve the overall translation quality. The most popular fusion approach is through a word-level fusion framework, i.e, Confusing Network decoding [5,6]. It is difficult, however, to consider syntax and semantics in a word-level fusion framework because the minimum unit of syntactic and semantic analysis is a phrase or a sentence rather than a word.

To address the problem, I proposed to use a phrase as the fusion unit and thus present a novel phrase-level fusion framework based on the idea of paraphrasing [7,8]: my system first selects the best translated sentence from multiple MT systems, named the “backbone”, and then paraphrases the backbone using information about consensus across sentences and other features. Another dimension of my MEMT research is to select the best translation sentence among the output of multiple translation engines. I exploited some complex syntactic features to evaluate translation quality in a log linear model, including a supertag-based structural language model [9] and syntactic error analysis using a feature-based lexicalized tree adjoining grammar [10].

My current work in MEMT is to continue with the paraphrasing framework, focusing on how to enhance the paraphrasing process to model word reordering better through a formal and effective method. I am designing a paraphrasing grammar based on synchronous

context-free grammar to paraphrase the backbone using the consensus about the reordering of words.

Chinese NLP - Unknown Word Identification

Before working on MEMT, I had also worked on several problems of Chinese NLP. One of them is unknown word identification [11,12,13]. It is well known that there is no space to mark word boundaries in Chinese text. As a result, identifying words is difficult, because of segmentation ambiguities and occurrences of unknown words. Conventionally unknown words were extracted by statistical methods because they are simple and efficient. However the statistical methods that do not use linguistic knowledge suffer the drawbacks of low precision and low recall; that is because low frequency new words are rarely identifiable by statistic methods.

To address the problem, in addition to statistical information, I tried to use as much information as possible, such as morphology, syntax, semantics, and world knowledge. The identification system fully utilizes the context and content information of unknown words in the steps of detection, extraction and verification. A practical, online unknown word extraction system was implemented which identifies new words, including low frequency new words, with high precision and high recall rates. The system ranked top1 in a segmentation contest held by ACL SIGHAN workshop in 2003.

The online system: http://ckipsvr.iis.sinica.edu.tw/

Chinese NLP - Automatic Grammatical Information Acquisition

Besides unknown word identification, I also worked on another task of Chinese NLP - automatically acquiring grammatical information from a corpus [14]. The study is based on Word Sketch Engine (WSE) [15], in which the original claims are two fold: that linguistic generalizations can be automatically extracted from a corpus with collocation information provided that the corpus is large enough; and that such a methodology is easily adaptable for a new language.

Based on observation and study of Chinese syntax, I designed dozens of fundamental Chinese syntactic rules for WSE to extract collocation information from a given Chinese corpus. Using the collocation information, WSE is able to generate a one-page grammatical summary for every word. The results attest to the claim the WSE is adaptable for a new language. More critically, I show that the quality of collocation information provided has a direct bearing on the result of grammatical information acquisition; when provided with rich, precise collocation information, both the quantity and quality of the extracted grammatical information improves substantially over simple handcrafted grammatical rules.

The online system: http://wordsketch.ling.sinica.edu.tw/

Future Research

MEMT for Semantic-based MT

Continuing the line of my MEMT research, in the future, I plan to design a MEMT model to fuse outputs of semantic-based MT and phrase-based MT engines, and investigate when and where to use the output of either engine. The motivation of this direction is because I believe the two kinds of engines reflect the two major brain operations a human uses to translate sentences - "understand (semantics)" and "memorize (phrase translations)"; people use the two kinds of operations to complete a translation process simultaneously.

Chinese Social Media Understanding

Following the line of my Chinese NLP research, in the future, I plan to extend my past NLP experiences in Chinese syntactic analysis to Chinese semantic analysis in social media. In recent years, social media has been pervasive on the web, such as blogs and forums. People write their experiences and opinions on the web and share them with each other. To better understand the tremendous, informal written materials on the web, a special Chinese parser for social media is required. I plan to work on the development of a Chinese parser for social media. In addition, I will study the semantic composition of individual sentences for understanding the meaning of discourse or documents in social media.

References

[1] Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. *In Proceedings of Human Language Technology Conference and Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*

[2] David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

[3] Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, Ignacio Thayer. 2006. Scalable Inference and Training of Context-Rich Syntactic Translation Models. *In Proceedings of the Annual Meeting of the Association for Computational Linguistics*

[4] Steve DeNeefe and Kevin Knight. 2009. Synchronous Tree Adjoining Machine Translation. *In Proceedings of EMNLP*

[5] Evgeny Matusov, Nicola Ueffing, and Hermann Ney. 2006. Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. *In Proceedings of EACL*

[6] Antti-Veikko I. Rosti, Spyros Matsoukas, and Richard Schwartz. 2007. Improved word-level system combination for machine translation. *In Proceedings of ACL*

[7] Wei-Yun Ma and Kathleen McKeown. 2012. Phrase-level System Combination for Machine Translation Based on Target-to-Target Decoding. *In Proceedings of the 10th*

Biennial Conference of the Association for Machine Translation in the Americas (AMTA), San Diego, CA.

[8] Wei-Yun Ma, Kathleen McKeown. 2011. System Combination for Machine Translation Based on Text-to-Text Generation. *In Proceedings of Machine Translation Summit XIII*

[9] Wei-Yun Ma, Kathleen McKeown. 2013. Using a Supertagged Dependency Model to Select a Good Translation in System Combination. *In Proceedings of NAACL-HLT*

[10] Wei-Yun Ma and Kathleen McKeown. 2012. Detecting and Correcting Syntactic Errors in Machine Translation Using Feature-Based Lexicalized Tree Adjoining Grammars. *In International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP)*, Vol 17, No. 4, pp. 1-14.

[11] Keh-Jiann Chen and Wei-Yun Ma. 2002. Unknown Word Extraction for Chinese Documents. *In Proceedings of COLING*

[12] Wei-Yun Ma and Keh-Jiann Chen. 2003. A Bottom-up Merging Algorithm for Chinese Unknown Word Extraction. *In Proceedings of the second SIGHAN Workshop on Chinese Language Processing*

[13] Wei-Yun Ma and Keh-Jiann Chen. 2003. Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff. *In Proceedings of the second SIGHAN Workshop on Chinese Language Processing*

[14] Chu-Ren Huang, Wei-Yun Ma, Yi-Ching Wu, and Chih-Ming Chiu. 2006. Knowledge-Rich Approach to Automatic Grammatical Information Acquisition:Enriching Chinese Sketch Engine *In Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation (PACLIC)*with a Lexical Grammar.

[15] Kilgarriff, Adam and Tugwell, David. 2002. Sketching Words. *In Marie-Hélène Corréard (ed.): Lexicography and Natural Language Processing. A Festschrift in Honour of B.T.S. Atkins. Euralex.*

[16] Chen Keh-Jiann, Shu-Ling Huang, Yueh-Yin Shih, Yi-Jun Chen. 2005. Extended-HowNet- A Representational Framework for Concepts. *OntoLex 2005 - Ontologies and Lexical Resources IJCNLP-05 Workshop, Jeju Island, South Korea*

[17] Yu-Ming Hsieh, Ming-Hong Bai, Jason S. Chang and Keh-Jiann Chen. 2012. Improving PCFG Chinese Parsing with Context-Dependent Probability Re-estimation. *CLP-2012.*