Bing Zhao

407 South Craig Street (InterAct Lab)
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Tel: (412)268-4546 (Office)
(412)320-0377 (Cell)
Fax:(412)268-6298
Email: bzhao@cs.cmu.edu
http://www.cs.cmu.edu/~bzhao

EDUCATION

CARNEGIE MELLON UNIVERSITY — Pittsburgh, PA, USA

Ph.D. , Computer Science, (expected in March-May. 2007)
Advisors: Stephan Vogel, Eric P. Xing, and Alex Waibel

CARNEGIE MELLON UNIVERSITY — Pittsburgh, PA, USA

M.S. , Language Technologies, 2003

INST. OF AUTOMATION, CHINESE ACADEMY OF SCIENCES — Beijing, China

M.S. , Pattern Recognition and Artificial Intelligence, 2001

UNIVERSITY OF SCIENCE & TECHNOLOGY OF CHINA — Hefei, Anhui, China

B.S. , Electronic Engineering and Information Science, 1998

RESEARCH INTERESTS

Natural language processing, Statistical machine translation, Speech recognition, Information extraction, Machine learning, Text mining.

PROFESSIONAL EXPERIENCE

2001–2006 **Research Assistant** — Language Technologies Institute, SCS, CMU

Statistical Alignment Models for Translational Equivalence

- Bilingual Topic AdMixture Models including BiTAMs and HM-BiTAM
- Unsupervised models for word alignment, Inner-outer Bracketing Models
- Sentence alignment for comparable corpus, data aligned: LDC2002E18
- Major Projects include GALE (05-06), TIDES (02-05) and IWSLT (05-06), STEAM, etc..

Fall 2005 **Research Intern** — IBM T.J. Watson Research, USA

Fall semester research intern, supervised by Dr. Kishore Papineni, working on:

- Generative inner-outer bracket models for word alignment [7];
- Detailed experiments comparing different word alignment approaches.

Summer 2001 (Apr.–Jul.) **Research Intern** — Microsoft Research Asia, Beijing, China

Summer research intern, supervised by Dr. Eric Chang, working on:

- Discriminative training for large scale continuous speech recognition;
- Discriminative training algorithms for vowel recognition;
- Experiments with speech recognition toolkits of HTK and MS-WHISPER.

1998–2001 **Research Assistant** — Chinese Academy of Sciences, China

Graduate research assistant at National Laboratory of Pattern Recognition (NLPR). Supervised by Prof. Taiyi Huang and Prof. Bo Xu, Major works include [17, 18, 19, 20]:

- Mandarin continuous digits recognition system for Nokia;
- Acoustic Model Adaptation: Detailed comparisons of several MLLR based algorithms; MRF based MAP, SMAP, and Quasi-Bayes learning algorithms implementation;
- Chinese Trigram Language Modeling; Chinese pinyin to character conversion.

Thesis for M.S. – A Continuous Chinese Digit Speech Recognition System: Acoustic Modeling, Speaker Adaptation, and Decoding.

1993–1998 **Research Assistant** Univ. of Science & Technology of China

Undergraduate Research at *Information Processing Center*, Advisor: Prof. ZhengKai Liu.

Thesis for B.S. – A Wavelet Transformation based Compression Algorithm for Seismogram.

TEACHING EXPERIENCE

Teaching Assistant:

2005-SPRING CMU undergraduate course 15-381, *Artificial Intelligence*, designs of homework, quiz, exam, grading and office hours

2005-FALL CMU graduate course 11-751, *Speech Recognition and Understanding*, assisted with course design, homework, quiz, exam, project design, grading, and office hours

PROFESSIONAL ACTIVITIES

SMT Related Evaluations:

GALE-PROJ Speech translation from 2005-present

TIDES-PROJ Statistical machine translation from 2002, 2003, 2004, and 2006

OTHER-PROJ Spoken language translation evaluations such as IWSLT-2005, IWSLT-2006, TC-Star, STEEM

Projects Workshops/Meetings:

NIST-SMT WORKSHOPS Workshops of NIST-SMT Evaluation in 2004 and 2006.

GALE PI-MEETINGS GALE project meetings in March 2006 (Boston) and Oct. 2006 (New York)

Membership:

Member of Association for Computational Linguistics (ACL)

PUBLICATIONS

Papers:

[1] **Bing Zhao**, Nguyen Bach, Ian Lane, and Stephan Vogel, “A Log-linear Block Transliteration Model based on Bi-Stream HMMs”, to appear in HLT/NAACL-2007.

[2] **Bing Zhao** and Eric P. Xing, "BiTam: Bilingual Topic AdMixture Models for Word Alignment", in the proceedings of *Joint Conference of Computational Linguists and Meeting of Association for Computational Linguists (ACL/Coling 2006),* July, 2006.

[3] Muntsin Kolss, **Bing Zhao**, Stephan Vogel, Ashish Venugopal, and Ying Zhang, “The ISL Statistical Machine Translation System for the TC-STAR Spring 2006 Evaluation”, TC-Star Workshop on Speech-to-Speech Translation, TC-STAR-WS 2006, Barcelona, Spain, 2006.

[4] Matthias Eck, Ian Lane, Nguyen Bach, Sanjika Hewavitharana, Muntsin Kolss, **Bing Zhao**, Almut Silja Hildebrand, Stephan Vogel and Alex Waibel, “The UKA/CMU Statistical Machine Translation System for IWSLT 2006”, in the proceedings of *International Workshop on Spoken Language Translation (IWSLT 2006)*

[5] Sanjika Hewavitharana, **Bing Zhao**, Almut Silja Hildebrand, Matthias Eck, Chiori Hori, Stephan Vogel and Alex Waibel, “ The CMU Statistical Machine Translation System for IWSLT2005”, in the proceedings of *International Workshop on Spoken Language Translation (IWSLT 2005)*, Sept. 2005.

[6] **Bing Zhao** and Alex Waibel, “Learning a Log-Linear Model with Bilingual Phrase-Pair Features for Statistical Machine Translation”, in the proceedings of *Fourth SigHan workshop on Chinese Language Processing (SigHan 2005),* October, 2005.

[7] **Bing Zhao**, Niyu Ge, and Kishore Papineni, “Inner-Outer Bracket Models for Word Alignment using Hidden Blocks”, in the proceedings of *Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Oct. 2005.

[8] **Bing Zhao** and Stephan Vogel, “A Generalized Alignment-Free Phrase Extraction”, in the proceeding of *ACL 2005 Workshop on Building and using Parallel Texts: Data Driven Machine Translation and Beyond (ACL WPT-05)*, June 2005.

[9] **Bing Zhao**, Eric P. Xing, and Alex Waibel, “Bilingual Word Spectral Clustering for Statistical Machine Translation”, in the proceeding of *ACL 2005 Workshop on Building and using Parallel Texts: Data Driven Machine Translation and Beyond (ACL WPT-05),* June 2005.

[10] **Bing Zhao**, Stephan Vogel, and Alex Waibel, “Phrase Pair Rescoring with Term Weightings for Statistical Machine Translation”, in the proceeding of *Conference on Empirical Methods in Natural Language Processing (EMNLP 2004),* July 2004.

[11] **Bing Zhao**, Matthias Eck, and Stephan Vogel, “Language Model Adaptation for Statistical Machine Translation with Structured Query Models”, in the proceeding of *The 20th International Conference on Computational Linguistics (Coling 2004)*, Aug. 2004.

[12] **Bing Zhao**, Klaus Zechner, Stephan Vogel, and Alex Waibel, “Efficient Optimization for Bilingual Sentence Alignment based on Linear Regression”, in the proceeding of *HLT/NAACL 2003 Workshop on Building and using Parallel Texts: Data Driven Machine Translation and Beyond (HLT/NAACL WPT-03),* May, 2003.

[13] Stephan Vogel, Ying Zhang, Alicia Tribble, Fei Huang, Ashish Venugopal, **Bing Zhao**, and Alex Waibel. “The CMU Statistical Translation System.” in Proceedings of the MT Summit IX. New Orleans, LA. September 2003.

[14] **Bing Zhao** and Stephan Vogel, “Word Alignment Based on Bilingual Bracketing”, in the proceeding of *HLT/NAACL 2003 Workshop on Building and using Parallel Texts: Data Driven Machine Translation and Beyond (HLT/NAACL WPT-03)*, May, 2003.

[15] Ying Zhang, **Bing Zhao**, Jie Yang, and Alex Waibel, “Automatic SIGN Translation”, in the proceeding of *International Conference on Spoken Language Processing (ICSLP2002)*, Aug. 2002.

[16] **Bing Zhao** and Stephan Vogel, “Full-text Story Alignment Models for Chinese-English Bilingual News Corpora”, in the proceeding of *International Conference on Spoken Language Processing (ICSLP2002)*, 2002.

[17] **Bing Zhao** and Stephan Vogel, “ Adaptive Parallel Sentences Mining From Web Bilingual News Collection”, in the proceeding of *the 2002 IEEE International Conference on Data Mining (ICDM 2002)*, December 2002.

[18] Yun Zhou, Chengqing Zong, and **Bing Zhao**, “The corpus oriented analysis of Chinese spoken dialog understanding”, in the proceeding of *International Symposium of Chinese Spoken Language Processing (ISCSLP 2000)*, July, 2000.

[19] **Bing Zhao** and Bo Xu, “MLLR Speaker Adaptation using Acoustic Correlation Information”, in the proceeding of *The National Conference on Man-Machine Speech Communication (NCMMSC 2000),* 2000.

[20] Sheng Gao, Bo Xu, Hong Zhang, **Bing Zhao**, Chengrong Li and Taiyi Huang, “Updated Progress of SINOHEAR: Advanced Mandarin LVCSR System at NLPR”, in the proceeding of *International Conference on Spoken Language Processing (ICSLP 2000), 2000.*

[21] **Bing Zhao** and Bo Xu, “Incorporating HMM State Sequence Confusion for Rapid MLLR Adaptation to New Speakers”, in the proceeding of *International Conference on Spoken Language Processing (ICSLP 2000),* 2000.

Working Paper:

[22] **Bing Zhao** and Eric P. Xing, “A Hidden Markov Bilingual Topical AdMixture Model for Statistical Word Alignment”.

Doctoral Dissertation (Expected in 2007):

[23] **Bing Zhao**, “Statistical Alignment Models for Translational Equivalences”, *Doctoral Thesis*, Expected in Dec. 2006~2007, Language Technologies Institute, School of Computer Science, Carnegie Mellon University.

SELECTED HONORS AND AWARDS

2001 – present	Computer science graduate fellowship, Carnegie Mellon University
1999	Tung’s Oriental Scholarship, Chinese Academy of Sciences
1998	Waive of Mandatoy Graduate Entrance Exam to National Lab of Pattern Recognition (NLPR)
1998	Excellent Bachelor Thesis Award, University of Science and Technology of China (USTC)
1993-1997	Excellent Student, Yu-Cai, DingXin, P&G scholarships at USTC

COMPUTER SKILLS

Programming languages: C++/C, Perl, Java, Matlab, Pascal, Python, TCL/TK, Lisp, Latex, Fortran.
Operating Systems: Linux and Windows.

RESEARCH – STATISTICAL MACHINE TRANSLATION AT CMU

2002 –present **Statistical Translation Models for Word and Phrase Alignment** CMU

In my PhD thesis work, I focus on bilingual topic AdMixture (BiTAM) translation models leveraging bilingual document-level context. The parallel sentence-pairs within a document-pair are assumed to constitute a mixture of hidden topics; each observed word-pair follows a topic-specific translation lexicon. With such topical information inferred from document level context, the translation models are expected to be sharper and the word alignment process less ambiguous. Traditional IBM translation models are word-mixture models, which simply ignore the parallel document boundaries in their generative processes.

I proposed a novel formalism for BiTAM translation models, and the traditional IBM models can be easily integrated within this formalism to formulate new models. Three such BiTAMs with embedded IBM Model-1 alignments were tested extensively; an extended hidden Markov bilingual topic AdMixture model was developed and evaluated on both word alignment accuracy and translation quality, showing improved modeling power.

Currently, I am applying the topic-specific translation lexicons learnt from the proposed BiTAM models: 1) leveraging the topical hidden information during decoding; 2) efficient variational inference algorithms to scale to large size of training data; 3) extensions to other alignment models such as BiTAM IBM-3, in which a fertility table is estimated for better word alignment.

2001 –2003 **Language Models for Statistical Machine Translation** CMU

Part of my work at CMU involves building language models (LM) for statistical machine translation. This includes extending the implementation of a suffix array LM, LM interface within a translation decoder, and detailed experiments of comparing cluster-LM, POS-LM, and the handling of UNK word for statistical machine translation.

To improve LM's effectiveness, I proposed an effective sentence-level LM adaptation with structured query models. Sentence-specific LMs were interpolated with a background LM to re-decode each test sentence. Experiments showed improved perplexity, and translation qualities in terms of BLEU and NIST scores.

2001 – present **Hands-on Experience in Statistical Machine Translation** CMU

I have about 5-year experiences covering main aspects of statistical machine translation. I designed and implemented cross-lingual bag-of-word models to align parallel documents from comparable data, and a dynamic programming algorithm to extract parallel sentence-pairs from the aligned document pairs. I applied these techniques to align the 10-year XinHua news comparable corpora and generated the collection released under LDC2002E18. I implemented a SMT beam search decoder. I also designed models for transliterations of Arabic unknown words. I directly participated in a number of international machine translation evaluations including GALE, NIST, IWSLT and TC-STAR.

COURSES AT CMU

Speech Recognition (11751), Machine Learning (15781), Statistical Approach to Learning and Discovery (15802), Language and Statistics (11761), Information Retrieval (11741), Information Extraction (11748), Grammar and Lexicon (11721), Algorithm for NLP (11711), Software Engineering (11791), Graphical, Statistical and Causal Models (10706).

REFERENCES

Available Upon Request