# Xiaoyi Ma, Ph.D.

---


CURRENT POSITION
Software Team Lead
Linguistic Data Consortium
University of Pennsylvania

CONTACT INFO
Email: xma@ldc.upenn.edu


PROFILE

Researcher and manager with more than ten years of experience in the field of human language technology. Has managed projects to support technologies as diverse as machine translation, topic detection, information extraction, and speech recognition, and continues to be interested in information retrieval, search technology and human language understanding in general. Research has led to the development of new data processing tools and technologies that facilitate the recent progress in machine translation research.

EDUCATION

Ph.D., Computer Science, 2008
University of Pennsylvania, Philadelphia, USA
Dissertation: *Improving Named Entity Recognition with Co-training and Unlabeled Bilingual Data*
Advisor: Mark Liberman
Committee: A. Joshi, M. Marcus, L. Ungar, D. Yarowski (external, JHU)

M.S., Computer Science, 1996
Peking University, Beijing, China
Specialization: Speech Recognition, Text-to-speech Systems

B.S., Computer Science, 1993
Beijing University of Aeronautics and Astronautics, Beijing, China

WORK EXPERIENCE

**2008-present Software Team Lead**
**Linguistic Data Consortium, University of Pennsylvania**

- Research focusing on named entity recognition and alignment, bilingual and multilingual text exploration, comparable text analysis, and textual language identification
- Technical support for RATS, BOLT, DEFT
- Manages DEFT Textual Entailment and inference annotation project

**1998-2007 Senior Program Analyst**
**Linguistic Data Consortium, University of Pennsylvania**

**2004-2007**

- Managed Parallel Text creation and Word Alignment of the DARPA GALE program

- Researched and developed tools to collect, create, and process parallel text and bilingual lexicons
- Created guidelines and tools for word alignment of Arabic-English and Chinese-English parallel text
- Coordinated linguistic resource development activities across multiple external organizations
- Responsible for the technical and financial management of the project
- Oversaw strategic development, planning, and human resources
- Managed translation outsourcing in terms of translation guidelines, translation company selection, QC
- Processed and distributed parallel text corpora and bilingual lexicons
- Major contributor to LDC's REFLEX Entity Translation grant proposal

**2002-2004**

- Managed Parallel Text creation of the DARPA TIDES program
- Researched and developed of tools to collect, create, and process parallel text and bilingual lexicons
- Coordinated linguistic resource development activities across multiple external organizations
- Managed technical and financial aspects of the project
- Managed translation outsourcing in terms of translation guidelines, translation company selection, QC
- Processed and distributed parallel text corpora and bilingual lexicons

**1998-2002**

- Researched and developed speech/text corpora and tools for a diverse set of projects
- Developed BITS program to collect parallel text over the Internet with minimum human intervention
- Created Annotation Graph Tool Kit, an open source tool kit for developing linguistic annotation tools
- Developed forced alignment tools to create faster and more cost effective speech corpora
- Created, collected, processed, and distributed parallel text corpora and bilingual lexicons
- Supported LDC's ongoing projects

**1996–1997 Software Engineer**
**Apple Computers Research Center, China**

- Member of Apple's Chinese and Cantonese speech recognition systems development team

**1993–1996 Research Assistant**

**Founder R&D Center, Beijing, China**

- Developed a domain specific Chinese speech recognition system
- Developed a Chinese text to speech system

PUBLICATIONS

Xiaoyi Ma, *LDC Forced Aligner*, LREC 2012: The eighth international conference on Language Resources and Evaluation, Istanbul, Turkey, May 2012

Xiaoyi Ma *Toward a Named Entity Aligned Bilingual Corpus,* LREC 2010 Workshop on Methods for the Automatic Acquisition of Language Resources and Their Evaluation Methods, Valletta, Malta, May 2010

Stephen Grimes, Xuansong Li, Ann Bies, Seth Kulick, Xiaoyi Ma, and Stephanie Strassel, *Creating Arabic-English Parallel Word-Aligned Treebank Corpora*, LREC 2010: Workshop on Language Resources and Human Language Technologies for Semitic Languages, Valletta, Malta, May 2010

Kazuaki Maeda, Xiaoyi Ma, and Stephanie Strassel, *Creating Sentence-Aligned Parallel Text Corpora from a Large Archive of Potential Parallel Text using BITS and Champollion,* LREC 2008, Marrakech, Morocco, May 28-30, 2008

Xiaoyi Ma, *Champollion: A Robust Parallel Text Sentence Aligner,* LREC 2006: Fifth International Conference on Language Resources and Evaluation, Genoa, Italy, May 2006

Xiaoyi Ma and Christopher Cieri, *Corpus Support for Machine Translation at LDC,* LREC 2006: Fifth International Conference on Language Resources and Evaluation, Genoa, Italy, May 2006

Stephanie Strassel, Christopher Cieri, Andy Cole, Denise DiPersio, Mark Liberman, Xiaoyi Ma, Mohamed Maamouri, and Kazuaki Maeda, *Integrated Linguistic Resources for Language Exploitation Technologies,* LREC 2006: Fifth International Conference on Language Resources and Evaluation, Genoa, Italy, May 2006

Steven Bird, Kazuaki Maeda, Xiaoyi Ma, Haejoong Lee, Beth Randall, and Salim Zayat, *TableTrans, MultiTrans, InterTrans and TreeTrans: Diverse Tools Built on the Annotation Graph Toolkit*, Proceedings of the Third International Conference on Language Resources and Evaluation, Las Palmas, Spain, May 2002

Xiaoyi Ma, Haejoong Lee, Steven Bird and Kazuaki Maeda, *Models and Tools for Collaborative Annotation*, The Third International Conference on Language Resources and Evaluation, Las Palmas, Spain, May 2002

Kazuaki Maeda, Steven Bird, Xiaoyi Ma, and Haejoong Lee, *Creating Annotation Tools with the Annotation Graph Toolkit,* Proceedings of the

Third International Conference on Language Resources and Evaluation, Las Palmas, Spain, May 2002

Kazuaki Maeda, Steven Bird, Xiaoyi Ma and Haejoong Lee, *The Annotation Graph Toolkit: Software Components for Building Linguistic Annotation Tools,* The Human Language Technologies Conference, San Diego, CA, March 2001

Xiaoyi Ma and Mark Liberman, *BITS: A Method for Bilingual Text Search over the Web,* Machine Translation Summit VII, Singapore, March 1999

Xiaoyi Ma, *Parallel Text Collections at the Linguistic Data Consortium,* Machine Translation Summit VII, Singapore, March 1999

PROGRAMMING ENVIRONMENT: C/C++, Java, Perl, SQL, PHP, Apache, *nix

LANGUAGES: Chinese: native
English: fluent