CSCI 599, Spring 2011

Applications of Natural Language Processing:

Machine Translation

Meeting time: TTh 11:00-12:20, VKC 211
Office hours: immediately following each lecture

Instructors

- David Chiang (chiang@isi.edu)
- Liang Huang (lhuang@isi.edu)
- Kevin Knight (knight@isi.edu)

Prerequisites: CSCI 562 or permission of instructor. Students should have familiarity with statistical natural language processing and be comfortable with medium-sized programming projects.

Goals: This is an introduction to the field of machine translation (systems that translate speech or text from one human language to another), with a focus on statistical approaches. Three major paradigms will be covered: word-based translation, phrase-based translation, and syntax-based translation. Students will gain hands-on experience with building translation systems and working with real-world data, and they will learn how to formulate and investigate research questions in machine translation.

Textbook: Philipp Koehn, *Statistical Machine Translation* [Publisher] [Amazon]

Home page: http://nlg.isi.edu/teaching/cs599mt

Requirements

- 4 homework assignments (15% each). Credit for homework assignments will mainly be assigned based on completion of the assigned work, but also in some cases on creativity or ambitiousness of the approach implemented, or its performance on test data relative to other students.
  Topics (subject to change):
  1. Implement a simple word alignment model (IBM Model 1, 2, or HMM). Experiment with improvements to the model.
  2. Implement a phrase extractor and decode using the Moses decoder. Experiment with new features.
  3. Implement a monotone phrase-based decoder. Experiment with different reordering models or contextual features.
  4. Implement a synchronous CFG extractor and decode with cdec or Joshua. Experiment with modifications to extractor or with new features.
- Final project (40%). Individual students or pairs of students will

propose a topic to the instructors for approval, or the instructors will assign a topic. The project will require the students to define a problem clearly (with well defined inputs/outputs and evaluation) and explore it with sufficient depth and creativity.

Resources

- Europarl data for use in homework assignments (revised 2011-01-20): [tgz, 3.8M]

Course overview (subject to change)

Date	Topic	Instructor	Assignments
Jan 11	Overview of machine translation. The statistical approach to MT. [PDF, 1.5M]	Chiang	Required: • Koehn, ch. 1 and 2 • Knight, "Automating knowledge acquisition for machine translation," *AI Magazine* 18(4), 1997. [PDF]
	Part One: Word-based alignment and translation		
Jan 13	IBM Models 1–5.	Knight	Required: • Koehn, ch. 4 • Knight, "A statistical MT tutorial workbook," 1999. [PDF] [RTF] Background: • Koehn, ch. 3 • CSCI 562 notes on EM Supplemental: • Brown et al, "The mathematics of statistical machine translation: parameter estimation," *Computational Linguistics* 19(2). [PDF] • Knight, "Decoding complexity in word-replacement translation models," *Computational Linguistics* 25(4) [PDF]
Jan 18	IBM Models 1–5.	Knight	Required: • Vogel, "HMM-Based Word Alignment in Statistical Translation," Proc. COLING, 1996. [PDF]
Jan 20	IBM Models 1–5.	Knight	

Jan 25	*n*-gram language models. Absolute discounting and Kneser-Ney smoothing.	Chiang	Required: • Koehn, ch. 7 Supplemental: • Chen and Goodman, "An empirical study of smoothing techniques for language modeling," Technical Report 10-98, Harvard University. [PDF]
Jan 27 *Add/drop period ends*	*n*-gram language models continued. Very large language models.	Chiang	*Assignment 1 due.*
Feb 1	MT evaluation. BLEU.	Chiang	Koehn, ch. 8
	Part Two: Phrase-based translation and discriminative training		
Feb 3	Phrase-based MT. Why do we need phrases. Relationship to EBMT. Phrase extraction. Estimating phrase translation probabilities and the problem of overfitting.	Chiang	Koehn, ch. 5 Marcu and Wong, "A phrase-based, joint probability model for statistical machine translation." In *Proc. EMNLP*, 2002. [PDF]
Feb 8	From the noisy channel to linear models. Phrase features.	Chiang	
Feb 10	Phrase reordering models.	Chiang	
Feb 15	Phrase-based decoding.	Huang	Koehn, ch. 6
Feb 17	Phrase-based decoding cont. *k*-best lists.	Huang	*Assignment 2 due.* Huang and Chiang, "Better *k*-best parsing." In *Proc. IWPT*, 2005. [PDF] Koehn, "Pharaoh: a beam search decoder for phrase-based statistical machine

			translation models." In *Proc. AMTA*, 2004. [PDF]
Feb 22	Maximum entropy. Minimum error-rate training.	Chiang	Koehn, ch. 9
Feb 24	Perceptron, max-margin methods.	Chiang	
Mar 1	System combination.	Chiang	
	Interlude: Subword translation		
Mar 3	Transliteration. Integrating traditional translation rules.	Knight	Koehn, ch. 10
Mar 8	Integrating morphology into translation.	Knight	
Mar 10	Decoding with lattices for morphology and word segmentation.	Knight	*Assignment 3 due.*
Mar 15	*Spring break*		
Mar 17	*Spring break*		
	Part Three: Syntax-based translation		
Mar 22	Hierarchical and syntax-based MT. Why do we need syntax. Synchronous context-free grammars and TSGs.	Chiang	Koehn, ch. 11 Chiang, "An introduction to synchronous grammars."
Mar 24	Extracting synchronous CFGs and TSGs from parallel data. Estimating rule probabilities and the problem of overfitting.	Chiang	
	Extracting synchronous		

Mar 29	TSGs from tree-tree data and the problem of nonisomorphism.	Chiang	
Mar 31	CKY decoding.	Huang	Chiang, "Hierarchical phrase-based translation."
Apr 5	CKY with an *n*-gram language model.	Huang	*Assignment 4 due.*
Apr 7	More CKY decoding: Binarization. *k*-best lists. Decoding with lattices.	Huang	Huang et al., "Binarization for Synchronous Context-Free Grammars" Huang and Chiang, "Better k-best Parsing"
Apr 12	Source-side tree decoding. Target-side left-to-right decoding.	Huang	Huang et al., "Statistical Syntax-Directed Translation" Huang and Mi, "Efficient Incremental Decoding for Tree-to-String Translation"
Apr 14	Syntax-based language models.	Knight	
Apr 19	Beyond synchronous CFGs and TSGs.	Knight	Knight, "Capturing Practical Natural Language Transformations"
Apr 21	Towards semantics-based translation.	Knight	
Apr 26	Final project presentations		
Apr 28	Final project presentations		

Course policies
Students are expected to submit only their own work for homework assignments. They may discuss the assignments with one another but may not collaborate with or copy from one another. University policies on academic integrity will be closely observed.

All assignments and the project will be due at the beginning of class on the due date. Late assignments will be accepted with a 7% penalty for each day after the due date, up to a week after the due date. No exceptions can be made except for a grave reason.

Statement for Students with Disabilities
Any student requesting academic accommodations based on a disability is required to register with Disability Services and Programs (DSP) each semester. A letter of verification for approved accommodations can be obtained from DSP. Please be sure the letter is delivered to me (or to

TA) as early in the semester as possible. DSP is located in STU 301 and is open 8:30 a.m.–5:00 p.m., Monday through Friday. The phone number for DSP is (213) 740-0776.

Statement on Academic Integrity
USC seeks to maintain an optimal learning environment. General principles of academic honesty include the concept of respect for the intellectual property of others, the expectation that individual work will be submitted unless otherwise allowed by an instructor, and the obligations both to protect one's own academic work from misuse by others as well as to avoid using another's work as one's own. All students are expected to understand and abide by these principles. *Scampus,* the Student Guidebook, contains the Student Conduct Code in Section 11.00, while the recommended sanctions are located in Appendix A: http://www.usc.edu/dept/publications/SCAMPUS/gov/. Students will be referred to the Office of Student Judicial Affairs and Community Standards for further review, should there be any suspicion of academic dishonesty. The Review process can be found at: http://www.usc.edu/student-affairs/SJACS/.