CSCI 599, Spring 2014
Machine Translation

Meeting time: TTh 11:00-12:20

Instructors:

David Chiang (chiang@isi.edu) and Kevin Knight (knight@isi.edu)
office hours immediately following each lecture

Recommended Preparation: CSCI 562/662 or permission of instructor. Masters students must receive D-clearance from 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*

Requirements:

- 5 homework assignments (12% 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.
    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 using the Moses decoder. Experiment with modifications to extractor or with new features.
    5. Implement a synchronous CFG decoder (without language model). Experiment with extensions.
- 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.

Course overview

Date	Topic	Instructor	Assignments
Jan 14	Overview of machine translation. The statistical approach to MT.	Chiang/Knight	Koehn, ch. 1 and 2 Knight, "Automating Knowledge Acquisition for Machine Translation", *AI Magazine* 18(4), 1997.

	Part One: Word-based alignment and translation		
Jan 16	IBM word alignment models.	Knight	Koehn, ch. 4 Background: Koehn, ch. 3; CSCI 562/662 notes on EM Supplemental: Brown et al, "The Mathematics of Statistical Machine Translation: Parameter Estimation" Knight, "Decoding complexity in word-replacement translation models", *Computational Linguistics* 25(4) Vogel, "HMM-Based Word Alignment in Statistical Translation"
Jan 21	IBM word alignment models, continued.	Knight	
Jan 23	IBM word alignment models, continued.	Knight	
Jan 28	*n*-gram language models. Absolute discounting and Kneser-Ney smoothing.	Knight	Koehn, ch. 7 Supplemental: Chen and Goodman, "An Empirical Study of Smoothing Techniques for Language Modeling"
Jan 30 *Add/drop period ends*	*n*-gram language models continued. Very large language models.	Knight	*Assignment 1 due.*
Feb 4	MT evaluation. BLEU.	Chiang	Koehn, ch. 8
	Part Two: Phrase-based translation and discriminative training		
Feb 6	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"
Feb 11	From the noisy channel to linear models. Phrase	Knight	

	features.		
Feb 13	Phrase reordering models.	Knight	
Feb 18	Phrase-based decoding.	Chiang	Koehn, ch. 6
Feb 20	Phrase-based decoding cont. *k*-best lists.	Chiang	*Assignment 2 due.*
Feb 25	Maximum entropy. Minimum error-rate training.	Chiang	Koehn, ch. 9
Feb 27	Perceptron, max-margin methods.	Chiang	
Mar 4	System combination.	Knight	
	Interlude: Subword translation		
Mar 6	Transliteration. Integrating traditional translation rules.	Knight	Koehn, ch. 10
Mar 11	Integrating morphology into translation.	Knight	
Mar 13	Decoding with lattices for morphology and word segmentation.	Knight	*Assignment 3 due.*
Mar 18	*Spring break*		
Mar 20	*Spring break*		
	Part Three: Syntax-based translation		
Mar 25	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 27	Extracting synchronous CFGs and TSGs from parallel data. Estimating rule probabilities and the problem of overfitting.	Chiang	
Apr 1	Extracting synchronous TSGs from tree-tree data and the problem of nonisomorphism.	Chiang	
Apr 4	CKY decoding.	Chiang	Chiang, "Hierarchical

			phrase-based translation."
Apr 8	CKY with an *n*-gram language model.	Chiang	*Assignment 4 due.*
Apr 10	More CKY decoding: Binarization. *k*-best lists. Decoding with lattices.	Chiang	
Apr 15	Source-side tree decoding. Target-side left-to-right decoding.	Chiang	
Apr 17	Syntax-based language models.	Knight	
	Postlude		
Apr 22	Beyond synchronous CFGs and TSGs.	Knight	Knight, "Capturing Practical Natural Language Transformations"
Apr 24	Towards semantics-based translation.	Knight	
Apr 29	Final project presentations		
May 1	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/.

Emergency Preparedness/Course Continuity in a Crisis

In case of a declared emergency if travel to campus is not feasible, USC executive leadership will announce an electronic way for instructors to teach students in their residence halls or homes using a combination of Blackboard, teleconferencing, and other technologies.

Please activate your course in Blackboard with access to the course syllabus. Whether or not you use Blackboard regularly, these preparations will be crucial in an emergency. USC's Blackboard learning management system and support information is available at blackboard.usc.edu.