

Computer Science 2501/485, Fall 2015

Computational Linguistics

Instructor

Frank Rudzicz. Office TRI 12-175 (Toronto Rehab, 550 University Avenue, East Wing, 12th floor); phone 416-597-3422 x7971; e-mail frank@cs.toronto.edu (include “2501” or “485” in the subject line to prevent spam filtering); office hours Wednesday 10h–11h or by appointment.

Prof. Rudzicz works on speech technology for individuals with speech disorders. He is the President of the international joint ACL/ISCA special interest group on Speech and Language Processing for Assistive Technology and Young Investigator of the Alzheimer’s Society. His work involves machine-learning, human-computer interaction, speech-language pathology, rehabilitation engineering, signal processing, and linguistics. Contributions include the the first speech recognition system for people with speech disorders based on their physical speech articulation and subsequent software that modifies hard-to-understand speech signals to be more understandable. He led development of the speech interaction component for HitchBOT (www.hitchbot.me), a robot that hitchhiked across Canada, Germany, and (part of) the United States. He also teaches CSC2511 / 401, “Natural Language Computing”, in the Spring.

Teaching assistant

Krish Perumal (t4peruma@cdf) is a master’s student in Computational Linguistics, supervised by Prof. Graeme Hirst. He is currently working on semi-supervised machine learning techniques for extracting questions and answers from discussion forums. Previously, he has worked on automatically extracting argumentation from student essays and classifying edits on Wikipedia.

Class meetings

The class will meet on Wednesdays, 13h–15h, and Thursdays, 16h–17h in LM157. Class meetings may be lectures, tutorials, or combinations of both, as necessary.

Syllabus and approximate schedule

Week	Topic	Advance reading*
16–23 Sept	Intro to computational linguistics	J&M: 1; BK&L: 1, 2.3, 4 <i>as necessary</i>
17 Sept	Intro to NLTK	<i>BK&L: 1, 2.3, 4 as necessary</i>
23–30 Sept	Grammars and parsing	RP ; J&M: 5.0–1, 12.0–12.3.3, 12.3.7, 13.1–2; BK&L: 8.0–8.4
30 Sept–1 Oct	Chart parsing	J&M: 13.3–4; A: 3.4, 3.6; BK&L: 8.4 and online extras section 8.2 on chart parsing
1–7 Oct	Parsing with features	RP ; J&M: 12.3.4–6, 15.0–3; A: 4.1–5; BK&L: 9
8–14 Oct	Ambiguity resolution	
15–21 Oct	Statistical attachment disambiguation	RP
22–28 Oct	Lexical semantics	J&M: 19.1–4, 20.8
29 Oct–4 Nov	Word sense disambiguation	RP ; J&M: 20.1–5
5–11 Nov	Statistical parsing	J&M: 5.2–5.5.2, 5.6, 12.4, 14.0–1, 14.3–7
12 Nov	Anaphora resolution	RP ; 21.0, 21.2–8
18–19 Nov	Semantic representations	J&M: 17.0–17.4.1, 17.5; BK&L: 10.0–4
25 Nov	Catch-up class if needed	
26 Nov–2 Dec	No class	

*J&M = Jurafsky and Martin; BK&L = Bird, Klein, and Loper; A = Allen; **RP** = research paper made available online before class; *italics indicates optional additional reading*.

Textbooks

Strongly recommended: Jurafsky, Daniel, and Martin, James H. *Speech and Language Processing, 2nd edition*, Pearson Prentice-Hall, 2009. Available in paper and e-book rental versions (for the latter, go to CourseSmart.com and search for *Jurafsky*.) (See errata list at www.cs.colorado.edu/~martin/SLP/Errata/SLP2-PH-Errata.html)

Strongly recommended: Bird, Steven; Klein, Ewan; and Loper, Edward. *Natural Language Processing with Python*, O'Reilly, 2009. Free (in HTML) with online extras at www.nltk.org/book

Optional: Allen, James. *Natural Language Understanding*, 2nd edition. Benjamin/Cummings, 1995.

Possibly helpful: Mertz, David. *Text Processing in Python*. Addison-Wesley, 2003. Free ascii version at Gnosis.cx/TPiP

Evaluation

Marks in the course are based on five assignments (75% of the final mark), and five short write-ups on assigned research papers ($5 \times 5\% = 25\%$). (There is no final examination.)

Assignments and due dates

Assignment	Weight	Due time and date	Where to hand in
Write-up 1	5%	13h10, Wed 23 Sept	CDF
Asst 1	12%	13h10, Mon 5 Oct	CDF
Write-up 2	5%	13h10, Wed 7 Oct	CDF
Asst 2	18%	13h10, Tues 20 Oct	CDF
Write-up 3	5%	13h10, Wed 21 Oct	CDF
Asst 3	12%	13h10, Mon 2 Nov	CDF
Write-up 4	5%	13h10, Wed 4 Nov	CDF
Asst 4	18%	13h10, Mon 16 Nov	CDF
Write-up 5	5%	13h10, Wed 18 Nov	CDF
Asst 5	15%	13h10, Thurs 3 Dec	CDF

Note: All assignments are submitted electronically using the command `submit` on CDF. For research-paper write-ups, you should keep a copy for yourself (even if electronic) for use during the discussion.

Late submissions of assignments will be penalized 10% of the initially available overall marks per day, and will be accepted no more than **four** days late. The CDF submission command will be locked out soon after the due time, so late submissions must be emailed to the instructor and TA. The lateness penalty may be waived at the discretion of the instructor. Reading write-ups will not normally be accepted after the classroom discussion of the paper; exceptions will be made in reasonable cases.

Course home page and bulletin board

The course home page is www.cs.utoronto.ca/~frank/csc2501. Online materials will be available on this page. You are responsible for checking it for class announcements as well.

The course bulletin board is <https://csc.cdf.toronto.edu/csc2501f> or <https://csc.cdf.toronto.edu/csc485f> (both addresses redirect to the same page). You may use the course bulletin board for discussions of the course and (in general terms) the assignments. The instructor and TAs will check it regularly, but we can't promise to answer all questions, nor to respond rapidly. The board may also be used for course announcements.

Computer accounts

Accounts on CDF will be allocated for this course, and some assignments will require submission of code to be run by the grader on CDF. Grant-funded research machines (such as the *comps** and *apps** machines, and research workstations) should not be used for coursework. You may use your own computer at your own risk — that is, it is your own responsibility to install the requisite software and to deal with any hardware or software problems; such occurrences will not be accepted as an excuse for a late assignment. A program that doesn't work for the grader on CDF, even if it runs on your machine, will get $\sin(\pi)$ (i.e., zero) marks.

Academic offences

All assignments must be done individually; the work that you submit must be your own. You may discuss your work with other people only in general terms; i.e., you should not share words, code, or specific ideas related to evaluation material before it is due. Be your own person! Using someone else's words, code, or ideas in anything you submit is an academic offence.

You may, and in fact *should*, discuss your work with others in general terms. But you should avoid looking at one another's work or talking about it in detail. If you have questions about the definition and harm of plagiarism and academic offences, please see www.cs.toronto.edu/~fpitt/documents/plagiarism.html and www.cs.toronto.edu/~clarke/acoffences/. If you have specific questions about what constitutes plagiarism or academic offences in this course, *please see the instructor for clarification.*