CSC2611 (W2020): Computational Models of Semantic Change

Date/Time: Tuesday, 10am-12pm

Location: MY440

Instructor: Yang Xu
Contact: yangxu@cs.toronto.edu
Office Hours: By appointment

This syllabus may be adjusted as the course progresses.

Course Description: Words are fundamental components of human language, but their meanings tend to change over time, e.g., *face* ('body part →'facial expression), *gay* ('happy→'homosexual), *mouse* ('rodent →'device). Changes like these present challenges for computers to learn accurate representations of word meanings—a task that is crucial for natural language systems. This course explores data-driven computational approaches to word meaning representation and semantic change. Topics include latent models of word meaning (e.g., LSA, word2vec), corpus-based detection of semantic change, probabilistic diachronic models of word meaning, and cognitive mechanisms of word sense extension (e.g., chaining, metaphor). The course involves both seminar-style presentations and a strong hands-on component that focuses on diachronic text analysis.

Note: This graduate course presumes extensive knowledge of Python programming and big data analytics. Undergraduates who are interested in enrolling should obtain special permissions from the instructor. Preferred preparatory courses include CSC108, CSC148, COG260, COG403, CSC401/2511, and other courses in computational linguistics or natural language processing.

Objectives: This course is aimed at the following three objectives.

1. Develop a broad foundation for the interdisciplinary study on semantic change.
2. Develop technical skills in the computational analysis of longitudinal textual data.
3. Develop essential communicative skills in scientific presentation and writing.

Recommended background readings:

- Traugott, E.C., & Dasher, R.B. *Regularity in semantic change*. CUP. 2001.
- Sweetser, E. *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure*. CUP. 1991.
- Hopper, P.J., & Traugott, E.C. *Grammaticalization*. CUP. 2003.
- Lakoff, G. *Women, fire, and dangerous things: What categories reveal about the mind*. UCP. 1987.

© Yang Xu

Deliverables and Assessments:

Paper presentation	25%
Lab assignment (with code repository)	20%
Project proposal	10%
Project milestones	5%
Project final report	20%
Project final presentation	10%
Project code repository	10%

Letter Grade Scale:

90 - 100%	A+	77 - 79%	B+
85 - 89%	A	73 - 76%	B
80 - 84%	A-	70 - 72%	B-
		0 - 69%	Fail

Course Policies:

- **General**
  - Students are required to present and lead discussion on papers from the numerically indexed reading materials in "Schedule".
  - Students with scheduled presentations are required to send the PDF slides to the instructor two days before the presentations.
  - Late submissions will receive a 1 point deduction per delayed hour until no point can be further deducted.
- **Attendance**
  - Attendance is expected in general and required on days of presentation.
  - Students are responsible for all missed assignments due to absence, unless they notify the instructor at least two days prior to the due date.
- **Project**
  - Students are expected to work independently on projects.
  - Students may obtain the instructor's permission to work on their own research projects, provided that the projects are relevant to the course.
  - Students may proceed with their projects only if the initial proposals have been approved by the instructor. Otherwise they may do so until the revised proposals have been approved.

© Yang Xu

Schedule (see course webpage for readings, presentations, and projects):

Date	Content
~~Jan 10~~	**The problem of semantic change**
~~Jan 17~~	**Distributed representations of word meaning** 1. Landauer, T.K. and Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2), 211-240. 2. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *NIPS* 2013. • Basic exercise
~~Jan 24~~	**Detection of semantic change** 3. Gulordava, K. and Baroni, M. (2011). A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. *GEMS* 2011. 4. Kulkarni, V., Al-Rfou, R., Perozzi, B., and Skiena, S. (2015). Statistically significant detection of linguistic change. *WWW* 2015. 5. Recchia, G., Jones, E., Nulty, P., Regan, J., and de Bolla, P. (2016). Tracing shifting conceptual vocabularies through time. *European Knowledge Acquisition Workshop* 2016. • Lab assignment
~~Jan 31~~	**Statistical laws of semantic change** 6. Xu, Y. and Kemp, C. (2015). A computational evaluation of two laws of semantic change. *CogSci* 2015. 7. Hamilton, W.L., Leskovec, J., and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. *ACL* 2016. • Project announcement Reference survey for the project: Tahmasebi, N., Borin, L., and Jatowt, A. (2018). Survey of computational approaches to lexical semantic change. *Preprint at ArXiv* 2018.

© Yang Xu

~~Feb 7~~	**Methodologies and issues in semantic change** 8. Sagi, E., Kaufmann, S., and Clark, B. (2009). Semantic density analysis: Comparing word meaning across time and phonetic space. *Workshop on Geometrical Models of Natural Language Semantics* 2009. 9. Dubossarsky, H., Weinshall, D., and Grossman, E. (2017). Outta control: Laws of semantic change and inherent biases in word representation models. *EMNLP* 2017. 10. Dubossarsky, H., Hengchen, S., Tahmasebi, N., and Schlechtweg, D. (2019). Time-Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change. *EMNLP* 2019. • Lab assignment due
~~Feb 14~~	**Probabilistic models of semantic change** 11. Frermann, L. and Lapata, M. (2016). A Bayesian model of diachronic meaning change. *ACL* 2016. 12. Bamler, R. and Mandt, S. (2018). Dynamic word embeddings. *ICML* 2017. • Project proposal due
~~Feb 28~~	**Word sense induction** 13. Lau, J.H., Cook, P., McCarthy, D., Newman, D., and Baldwin, T. (2012). Word sense induction for novel sense detection. *EACL* 2012. 14. Hu, R., Li, S., and Liang, S. (2019). Diachronic sense modeling with deep contextualized word embeddings: An ecological view. *ACL* 2019.
~~Mar 7~~	**Word sense extension** 15. Xu, Y., Malt, B.C., and Srinivasan, M. (2017). Evolution of word meanings through metaphorical mapping: Systematicity over the past millennium. *Cognitive Psychology*, 96, 41-53. 16. Ramiro, C., Srinivasan, M., Malt, B.C. and Xu, Y. (2018). Algorithms in the historical emergence of word senses. *PNAS*, 115(10), 2323-2328.
~~Mar 14~~	**Semantic bleaching** 17. Luo, Y., Jurafsky, D., and Levin, B. (2019). From insanely jealous to insanely delicious: Computational models for the semantic bleaching of English intensifiers. *ACL Workshop*. • Project milestone

© Yang Xu

~~Mar 21~~	**Relations to cross-linguistic semantics** 18. Youn, H., Sutton, L., Smith, E., ..., and Bhattacharya, T. (2016). On the universal structure of human lexical semantics. *PNAS*, 113(7), 1766-1771. 19. Ammar, W., Mulcaire, G., Tsvetkov, Y., Lample, G., Dyer, C., and Smith, N. A. (2016). Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*
~~Mar 28~~	**Relations to social science** 20. Eisenstein, J., O'Connor, B., Smith, N.A., and Xing, E. (2014). Diffusion of lexical change in social media. *PLOS ONE*, 9(11), e113114. 21. Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *PNAS*, 115(16), E3635-E3644.
~~Apr 4~~	• Project final presentation • Project final report due Friday of the same week

© Yang Xu

Project Report Guidelines

- Report should be 7-8 pages following the LaTeX template of ACL proceedings.
- Report should be structured as follows:

  Abstract - *Introduction* - *Related work* - *Methods* - *Data* - *Results* - *Discussion* - *References*.

- *Methods* should provide GitHub (`github.com`) or OSF (`osf.io`) link to code/data.
- Report not conforming to the above standards will not receive any credit.
- Reporting style should support replication of the analyses and results described.
- Report and appendix should be submitted as a single PDF, with name(s) on page 1.

© Yang Xu

Resources:

- Python:

  Jupyter: https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/

  Natural Language Processing with Python: http://www.nltk.org/book/

  Natural Language Toolkit: http://www.nltk.org/

  Bare essentials: http://www.cs.toronto.edu/~yangxu/PythonBookletV4.pdf

- GitHub:

  Creating a repo: https://help.github.com/articles/create-a-repo/

  Common commands: https://gist.github.com/jedmao/5053440

- Word embeddings:

  Word2vec: https://code.google.com/archive/p/word2vec/

  GLOVE: https://nlp.stanford.edu/projects/glove/

  Lda2vec: https://github.com/cemoody/lda2vec

  tSNE: https://github.com/paulorauber/thesne

  HistWords: https://nlp.stanford.edu/projects/histwords/

- Longitudinal text corpora:

  Project Gutenberg: https://www.gutenberg.org/

  Google N-grams: http://storage.googleapis.com/books/ngrams/books/datasetsv2.html

  Syntactic N-grams: http://commondatastorage.googleapis.com/books/syntactic-ngrams/index.html

  Helsinki Corpus of English: http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/

  Early English Books Online: https://corpus.byu.edu/eebo/

  CHILDES: https://childes.talkbank.org/

- Lexical resources:

  WordNet: https://wordnet.princeton.edu/

  MetaNet: https://metanet.icsi.berkeley.edu/metanet/

© Yang Xu

Metaphor Map of English: http://mappingmetaphor.arts.gla.ac.uk/

Historical Thesaurus of English: http://historicalthesaurus.arts.gla.ac.uk/

Dictionary of Old English: https://www.doe.utoronto.ca/pages/index.html

- Benchmark data:

  WordSimilarity-353: http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/

  SimLex-999: https://www.cl.cam.ac.uk/~fh295/simlex.html

  SemEval-2017: http://alt.qcri.org/semeval2017/index.php?id=tasks

  Stanford Question Answering: https://rajpurkar.github.io/SQuAD-explorer/

- Human behavioural data:

  University of South Florida Free Association Norms: http://w3.usf.edu/FreeAssociation/

  Human Brain Cloud: http://www.humanbraincloud.com/

  Word concreteness ratings: http://crr.ugent.be/archives/1330

  Word affectiveness ratings: http://crr.ugent.be/archives/1003

  Word age-of-acquisition norms: http://crr.ugent.be/archives/806

© Yang Xu