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:

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:

<table>
<thead>
<tr>
<th>Range</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>90 - 100%</td>
<td>A+</td>
</tr>
<tr>
<td>87 - 89%</td>
<td>B+</td>
</tr>
<tr>
<td>85 - 89%</td>
<td>A</td>
</tr>
<tr>
<td>80 - 84%</td>
<td>A-</td>
</tr>
<tr>
<td>77 - 79%</td>
<td>B</td>
</tr>
<tr>
<td>70 - 72%</td>
<td>B-</td>
</tr>
<tr>
<td>0 - 69%</td>
<td>Fail</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Date</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 10</td>
<td><strong>The problem of semantic change</strong></td>
</tr>
<tr>
<td>Jan 17</td>
<td><strong>Distributed representations of word meaning</strong></td>
</tr>
<tr>
<td></td>
<td>• Basic exercise</td>
</tr>
<tr>
<td>Jan 24</td>
<td><strong>Detection of semantic change</strong></td>
</tr>
<tr>
<td></td>
<td>• Lab assignment</td>
</tr>
<tr>
<td>Jan 31</td>
<td><strong>Statistical laws of semantic change</strong></td>
</tr>
<tr>
<td></td>
<td>• Project announcement</td>
</tr>
<tr>
<td>Date</td>
<td>Topic</td>
</tr>
<tr>
<td>-------</td>
<td>--------------------------------------------</td>
</tr>
</tbody>
</table>
* Lab assignment due |
* Project proposal due |
* Project milestone |
<table>
<thead>
<tr>
<th>Date</th>
<th>Relations to cross-linguistic semantics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mar 28</td>
<td>Relations to social science</td>
</tr>
<tr>
<td>Apr 4</td>
<td>• Project final presentation</td>
</tr>
<tr>
<td></td>
<td>• Project final report due Friday of the same week</td>
</tr>
</tbody>
</table>
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.
Resources:

- **Python:**

- **GitHub:**
  - Creating a repo: [https://help.github.com/articles/create-a-repo/](https://help.github.com/articles/create-a-repo/)
  - Common commands: [https://gist.github.com/jedmao/5053440](https://gist.github.com/jedmao/5053440)

- **Word embeddings:**
  - Word2vec: [https://code.google.com/archive/p/word2vec/](https://code.google.com/archive/p/word2vec/)
  - Lda2vec: [https://github.com/cemoody/lda2vec](https://github.com/cemoody/lda2vec)
  - tSNE: [https://github.com/paulorauber/thesne](https://github.com/paulorauber/thesne)
  - HistWords: [https://nlp.stanford.edu/projects/histwords/](https://nlp.stanford.edu/projects/histwords/)

- **Longitudinal text corpora:**
  - Project Gutenberg: [https://www.gutenberg.org/](https://www.gutenberg.org/)
  - Google N-grams: [http://storage.googleapis.com/books/ngrams/books/datasetsv2.html](http://storage.googleapis.com/books/ngrams/books/datasetsv2.html)
  - Early English Books Online: [https://corpus.byu.edu/eebo/](https://corpus.byu.edu/eebo/)
  - CHILDES: [https://childes.talkbank.org/](https://childes.talkbank.org/)

- **Lexical resources:**
  - WordNet: [https://wordnet.princeton.edu/](https://wordnet.princeton.edu/)
  - MetaNet: [https://metanet.icsi.berkeley.edu/metanet/](https://metanet.icsi.berkeley.edu/metanet/)
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:

  SimLex-999: https://www.cl.cam.ac.uk/~fh295/simlex.html
  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 affectiveness ratings: http://crr.ugent.be/archives/1003