CS591 W1: Towards Universal Natural Language Understanding

Term: Fall 2019
Instructor: Derry Wijaya (wijaya@bu.edu)
Lecture: Tue, Thur 9.30 - 10.45 am
Office hours: Wed, 1.10 - 2 pm, MCS 294A
Credits: 4

Course description:

This is a graduate course on statistical natural language processing (NLP) that is research-oriented and geared towards extending natural language processing approaches to more languages in the world even those without large annotated data. The course will start off with instructor-led presentations that introduce students to NLP. The rest of the course will focus on discussing the state-of-the-art algorithms for multilingual NLP or algorithms that are relevant to learning with few annotated data by discussing research papers. Participants take turns as discussion leaders.

Grading:

In team of twos:

Assignment 1: Survey paper (20%): each team picks a task they will like to work and 4-5 research papers relevant to the task. In the survey paper, they will need to (1) explain the task/problem they are going to address and (2) do a survey of the approaches for the task/problem.

Paper presentation (20%): each team will choose a lecture they want to lead, and 2 papers (out of the 4-5 papers they surveyed) that they will present and lead discussion on (each member of the team will present 1 paper)

Participation in class discussions (30%): before each student-led presentation, each non-presenting team will email the instructor their questions for the papers that are to be presented

Assignment 2: Research paper (30%): Based on their findings from the survey paper, each team will need to (1) propose an approach to extend a past multilingual NLP research on the their task of interest, (2) explain the limitations of the previous approaches and which limitation(s) they are addressing with their proposed method, (3) run/implement at least one baseline approach. Bonus points (+10%) if they actually implement their proposed approach. The paper should have a description of your method and why you think it has a good chance of working based on analysis of related works and (optionally) preliminary experimental results

Assignments and Presentation:
The goal of the assignments and presentation is to gain experience doing NLP research

Timeline:
Check point 1: Due September 29: Team picks tasks, 4-5 relevant papers, and presentation papers and d:
Check point 2: Due October 20: Assignment 1 (Survey Paper)
Check point 3: **Due December 11**: Assignment 2 (Research Paper)

**Assignment and final report details:**
Get in group of twos:
Assignment and project reports should be 4-6 pages long (not including references), written by you alone.
**Formatting:** Please use the ACL style files. Reports longer than 6 pages of the ACL format will not be cons

**Suggested tasks:**
as they involve more than one language, however please confirm with the instructor before you start
assignment 1
Machine Translation
Dictionary Induction
Word Sense Disambiguation
Representation Learning
Semantic Similarity
Lexical Substitution
Textual Entailment
Text Simplification
Text Summarization
Text Categorization
Sentiment/Opinion Mining
Question Answering
Information Extraction
Information Retrieval
Part of Speech tagging
Dependency parsing
Named Entity Recognition
Named Entity Linking
Coreference resolution
Event extraction
Vision-Language tasks: Image Captioning, Visual QA
Dialogue systems
etc. etc.

**Suggested datasets:**
WMT — Machine Translation Workshops Tasks: http://www.statmt.org/wmt19/
Suggested conferences and sites to find relevant papers:

CoNLL 2018: https://www.conll.org/accepted-2018
EMNLP 2018: https://emnlp2018.org/schedule/
NAACL 2019: https://naacl2019.org/schedule/
ACL 2019: http://www.acl2019.org/EN/program.xhtml
WMT 2019: http://www.statmt.org/wmt19/papers.html
*SEM 2019: https://starsem.org/2019/
EMNLP 2019: https://www.emnlp-ijcnlp2019.org/program/accepted/
ICLR 2019: https://openreview.net/group?id=ICLR.cc/2019/Conference
ICML 2019: http://proceedings.mlr.press/v97/
NIPS 2019: TBA

Academic Misconduct:
I will assume that you understand BU’s Academic Conduct Code:
http://www.bu.edu/academics/policies/academic-conduct-code
<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>J&amp;M Chapter</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>week-1 (Sep 3, 5)</td>
<td>Introduction, Text Processing, Language Models</td>
<td>1</td>
<td>Derry</td>
</tr>
<tr>
<td>week-2 (Sep 10, 12)</td>
<td>Model formulation, vector semantics, multilingual representation</td>
<td>2, 3</td>
<td></td>
</tr>
<tr>
<td>week-3 (Sep 17, 19)</td>
<td>Neural model for NLP</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>week-4 (Sep 24, 26)</td>
<td>Unsupervised and semi-supervised learning, domain adaptation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>week-5 (Oct 1, 3)</td>
<td>Machine translation</td>
<td></td>
<td>Derry</td>
</tr>
<tr>
<td>week-6 (Oct 8, 10)</td>
<td>Information extraction, multimodal representation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>week-7 (Oct 17)</td>
<td>student-led presentation 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>week-8 (Oct 22, 24)</td>
<td>student-led presentation 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>week-9 (Oct 29, 31)</td>
<td>student-led presentation 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>week-10 (Nov 5, 7)</td>
<td>no classes (conference week)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>week-11 (Nov 12, 14)</td>
<td>student-led presentation 6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>week-12 (Nov 19, 21)</td>
<td>student-led presentation 7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>week-13 (Nov 26)</td>
<td>student-led presentation 8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>week-14 (Dec 3, 5)</td>
<td>student-led presentation 10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>week-15 (Dec 10)</td>
<td>student-led presentation 13</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note: Tentative, may subject to substantial change!