COSI 137: Information Extraction

Spring 2018

Instructor: Nianwen Xue

- [http://www.cs.brandeis.edu/~xuen](http://www.cs.brandeis.edu/~xuen)
- Email: xuen@brandeis.edu
- Office: 136 Volen
- Office hours: Fri, 2pm-4pm

Teaching Assistant: Yuchen Zhang

- Email: yuchenz@brandeis.edu
- Office: 110 Volen
- Office hours: TBA

Time: 11:00am – 12:20pm, Tue/Fri

Location: TBA

Class website: [http://latte.brandeis.edu](http://latte.brandeis.edu)

Course Objective:

The objective of this course is to provide an in-depth exploration of an Information Extraction system. Other related topics such as text mining, sentiment and opinion analysis will also be introduced. Upon completion of the course, the student will have designed and implemented a working system, drawing from linguistic knowledge, computational techniques and statistical modeling. The student will also be required to make a technical presentation and write a technical paper on his/her system, acquiring technical communication skills in the process.

Course Description:

Information Extraction (IE) is the process of extracting semantically meaningful text units from unstructured textual data sources and putting them in a structured format that can be easily accessed and consumed by a computer program. Information Extraction takes on new significance with the rapid expansion of the Internet, which is an inexhaustible source of unstructured linguistic data that contains valuable information. An Information Extraction system requires complex interactions between different natural language components and it typically includes the extraction of person names, organizations, locations, events and time expressions from natural language text, the detection of all text units that refer to specific semantic classes and the linking of these text units that refer to the same object as well as the determination of semantic relations between pairs of entities. These natural language components in turn require underlying natural language technologies such as part-of-speech tagging, phase chunking and syntactic parsing.
Information Extraction often involves the classification of textual units of the same sort based on some semantic criteria and it covers a wide range of tasks, including:

*Name extraction:* identifying the names in a text and classifying them as people, organizations, locations, temporal expressions, etc.

*Entity extraction:* identifying all text units (e.g., phrases) that refer to objects of specific semantic classes, and linking textual units that refer the same object

*Relation extraction:* identifying pairs of entities in a specific semantic relation

*Event extraction:* identifying instances of events of a particular type, and the arguments of each event, the temporal locations of events and the temporal ordering between events.

*Knowledge base population (KBP):* the linking of entities and their attributes as well as events and their arguments identified in a natural language text to a knowledge base (e.g., Wikipedia),

Information research started with newswire, but has since spread rapidly to other domains such as the biomedical field, where there is a growing body of literature that is increasingly difficult for researchers to find relevant information from. Automatic information extraction using Natural Language Processing techniques is a promising approach to tackle this problem.

**Tentative schedule:**

- **Weeks 1 - 5:** Extraction of Named entities, nominals, coreference resolution, relation extraction
- **Weeks 6 – 7:** Extraction of events, time expressions, and temporal location of events, temporal ordering between events, sentiment and opinion analysis
- **Weeks 8 - 9:** Knowledge base population and slot filling
- **Weeks 10 – 13:** Domain-specific IE, survey of related NLP applications (e.g., QA)

**Readings:**

1. NLTK book, Chapters 6 and 7
2. Jurafsky and Martin, chapter 22
3. Selected conference papers and journal articles.

**Grading:**

- Participation: 5%
- Quizzes: 20%
- Small projects: 45%
  - Single tasks: NE-tagging, nominal tagging, coref resolution, relation detection, event detection, event ordering, event-time relation, etc.
- Final project and report: 30%
o Component integration: NE tagging + nominal tagging + coref
o NE tagging + nominal tagging + relation,
  o Event detection + event ordering
  o Knowledge base population

**Academic Integrity:**

You should finish homework assignments, exams, and project reports on your own unless a project is explicitly stated as a collaborative project. Late projects are subject to grade deduction.