CMP711 Natural Language Processing
Project Topics

- You may choose ***your project proposal*** from the following list or you may suggest any other project in NLP field. Each student will select a separate project.

- You should pick your project topic as soon as possible. You should write one page document for your project proposal, and submit ***your project proposal***. You should send a pdf file for your proposal. Your project proposal should include the followings:
    - Project title
    - Project idea
    - The computational work that will be done.
    - The relevant papers with your project proposal. You should read at least one of them before your project proposal.

- You should find at least two-three relevant major papers in your project topic and read them.

- At the middle of the semester, you will submit your ***midway project report***. This means that you should finish some of your project work before the midway point. The midway project report should be in the format of the ***final project report*** and it should be 3-5 pages. Your midway project report should include the related work section and describe the machine learning techniques that you tried upto the midway point and their results. The midway project report is basically a short version of your final project report.

- At the end of the semester, you will submit your ***final project report*** . You should send a pdf file of your final report
    - Prepare your final project report in the format of a conference article using IEEE double column format (6-8 pages).
    - In your final project report, you should use your own words. *Do not cut and paste from the papers that you read.*
    - Your final project report should contain at least the following sections in addition to a *title* and an *abstract*:

      Introduction – Describing the problem that you are attacking

      Related Work – Describe the related works here together with their relations with your work.

      Sections describing your computational work in detail – Describe the details of your computational work in these sections. If your computational work needs evaluation, do not forget to include evaluation sections.

      Conclusion – Give your concluding remarks and possible future works in this section.

      References – Give the references that are cited in your paper.

- With your final project report, you should send an electronic copy of each of the ***major papers that you read in your survey***.

- You should send your ***executable and source files of your project*** at the end of the semester. Make sure that I can execute your project on my PC.

- You should make a ***demo of your project*** to me at the end of semester.

Some Possible Project Topics:

- You may select a project from the following list or you can suggest a project related with NLP.

Creation of Language Models for Turkish

- Collecting a huge Turkish corpus and creating different language models. Some of these language models should depend on Turkish morphology.
- Using created language models in some applications to show their effectiveness.

Developing a Chatbot for a selected domain:

- Many companies are now utilizing conversational bots, called Chatbots to interact with their customers and resolve their issues. You can select a specific domain (such as bank, phone company, municipial government, …) and develop a Chatbot for this domain.

Statistical Machine Translation System (between English and Turkish)

- Creating a statictical machine translation system from bilingual corpora containing translation examples.

Named Entity Recognition

- Named-entity recognition is a subtask of information extraction that seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories such as person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc.

Bilingual Terminology Extraction from Bilingual Corpus

- Creating a terminology extraction system from bilingual corpora.

Creation of A Translation Memory System

- A translation memory system stores a set of translation examples and finds candidate translation examples for a given a source language sentence.

A System to Measure Text Similarity

- This system shouldbe able to find similarities of texts. It can be used for plagiarism dection.

A Morphological Disambiguator for Turkish

- A word can have different part of speech tags (such as noun, verb, …), but its usage in a sentence will be only one of them. For example, English word "fly" can be verb (uçmak) or noun (sinek). In the sentence "A **fly** can **fly**", the first "fly" is a noun and the second "fly" is a verb. A part of speech tagger tries to determine the intended part of speech tag for each word in a sentence.
- Your part of speech tagger should invoke Turkish morphological analyzer (which is available) to find possible part of speech tags of each word, and should try to find the correct part of speech tag of each
- This can be an improvement to our rule-based morphological disambiguator or a new statistical morphological disambiguator.

Text Categorization

- Each written document can be categorized according to its content. For example, the category of a newspaper article can be econmy, sport, etc. A text categorization system determines the category of a given document.

Author Identification

Determining the author of a given text.

A NP-chunker for Turkish

- It should find NPs (noun phrases) a given Turkish sentence.
- For example, NPs in the following sentence are underlined.
  <u>Kırmızı başlıklı kız</u> <u>Ankara'dan</u> <u>İstanbul'a</u> <u>uçakla</u> gitti.
- The system does not need the parse the whole sentence. It should find only the noun phrases in the sentence.

Extracting Domain Terms from Domain Articles

- From legal text documents extracting terms used in legal documents.
- Extracting medical terms from medical texts.

Finding semantic similarities between words for Turkish (or English)

- The system should categories the Turkish words (nouns and verbs) according to their semantic similarities using a corpus.

a) Using Latent Semantic Analysis (LSA)
b) Using other methods different than LSA

Extraction of protein interaction from biomedical texts

- Extraction of the relations between the protein names in the biomedical texts.

Keyphrase Extraction for Turkish

- Generation of keypharases of given texts.

Keyphrase Extraction for English

- Generation of keypharases of given texts.

Text Summarization for Turkish

- Generation of summaries of given texts by selecting the important sentences of the given texts.

a) Using Latent Semantic Analysis (LSA)
b) Using other methods different than LSA

Text Summarization for English

- Generation of summaries of given texts by selecting the important sentences of the given texts.

a) Using Latent Semantic Analysis (LSA)
b) Using other methods different than LSA

Opinion Mining (Sentiment Analysis)

- Deciding the polarity (positive or negative) of views in customer review texts or texts in social media environments. This can be done for Turkish or English texts

Some Possible Project Topics from CMU Machine Learning Course:

Using Vector Space Models for Natural Language Processing

Using NP-context co-occurrence matrix, finding synonyms, finding members of categories (i.e., "is this noun phrase an athlete?"), or clustering noun phrases to automatically induce categories.

- http://www.cs.cmu.edu/~tom/10709_fall09/RTWdata.html
- http://qwone.com/~jason/20Newsgroups/

Peter D. Turney and Patrick Pantel (2010). From Frequency to Meaning: Vector Space Models of Semantics. Journal of Artificial Intelligence Research 37, pp. 141-188.

WebKB

This data set contains webpages from 4 universities, labeled with whether they are professor, student, project, or other pages.
Project idea: Learning classifiers to predict the type of webpage from the text.

- http://www.cs.cmu.edu/~webkb/

Email Annotation

The goal is to identify which parts of the email refer to a person name. This task is an example of the general problem area of Information Extraction. Model the task as a Sequential Labeling problem, where each email is a sequence of tokens, and each token can have either a label of "person-name" or "not-a-person-name".

- http://www.cs.cmu.edu/~einat/datasets.html
- Paper: http://www.cs.cmu.edu/~einat/email.pdf

Enron E-mail Dataset

The Enron E-mail data set contains about 500,000 e-mails from about 150 users.

- http://www.cs.cmu.edu/~enron/

Project idea: Can you classify the text of an e-mail message to decide who sent it?