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 as soon as possible. You should write one page document for your project proposal, and submit your project proposal before **22 March 2016** (***hard copy***).
  - Your project should also include a computational work.
  - In your project proposal, you should talk about your computational work.
  - You may talk with me about what are my expectations for the possible projects described in this document.
  - You should find at least two-three major papers in your project topic and read them.

- At the end of semester, you will submit ***your final project report (its soft copy by email and its hard copy***).
  - Prepare your final project report in the format of a conference article using IEEE double column format (6-10 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, and describe the 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 before you give final project report.

Possible project topics:

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.

Statistical Machine Translation System (between English and Turkish)

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

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.
  Kırmızı başlıklı kız Ankara'dan İstanbul'a uçakla 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

- 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