Course
Language Technologies

Module "Knowledge Technologies"
Jožef Stefan International Postgraduate School
2017

URL *http://nl.ijs.si/et/teach/mps17-hlt/*

Suggested topics for seminar work

2017-04-05

The course assessment is through seminar work on a topic connected with LT, where ½ of the grade is given on the basis of the quality of work, and ½ on the quality of the report, comprising its structure, motivation, review of related work including literature, presentation of the data-set and experiment, conclusions and language. For more substantial projects, students also have the option of performing the work in pairs (they will both get the same grade). Where appropriate, the results should be published in the CLARIN.SI repository.

Some possible topics are given below, but students have the option of choosing the topic themselves, in agreement with the lecturer.

Each student should have consultations with the lecturer regarding their project – send email to arrange for the time of the consultation.

The seminar should ideally be finished by the official exam date in late May / early June. The seminar work should be sent to the lecturer at least one week before the exam date.

Suggested topics

Most of the topics take some (Slovene) language resources as the starting point, and the task is to develop (either with heuristics or with some machine learning approach) a model for a particular aspect of Slovene. Most of the resources mentioned below are available from the CLARIN.SI repository, cf. http://www.clarin.si/.

Language detection

This topic focuses on high-quality language assignment to texts of the Janes corpus. Currently it is done with the Python module lang.py and some heuristics, but could be significantly improved, cf.http://www.informatica.si/index.php/informatica/article/view/746/0

Exploration of user networks in Slovene Tweets, Forums, etc.

This topic uses the Janes corpus of Slovene user-generated content. Here the texts have rich metadata (time of posting, username, gender, sentiment, standardness etc.) which makes it possible to compare different groups of users in terms of the language they use, by clustering, keyword extraction, topic maps etc. Possible topics include:

- Cluster users according to whom they reply to, so identifying various communities, and compare the vocabulary of different groups
- Group tweet threads and compare them by length
- Find multiword keywords ("terms") in the avtomobilizem.com forum

Link: http://nl.ijs.si/janes/

ZRC dictionaries of Slovene

The CLARIN.SI repository will soon host some Slovene dictionaries in XML. It would be interesting to connect the information contained in them with texts (corpora), such as Janes, or other dictionaries.

Term variants

The corpus of Slovene academic writing, KAS, contains PhDs, MScs and diplomas. Concentrate on a few particular terms and try to find them, and their variants, in KAS.

Methods for Hate Speech Detection

Explore comments on news articles in Janes and determine how soon in the thread they turn to "hate speech" and what are the most common topics – which words and phrases are used the most? Discuss related work on hate speech detection and propose an automatic method and annotation campaign.

Slovene Syllables

Words can be split into syllables, i.e. the smallest individually pronounceable parts of a word. For Slovene there currently does not yet exists a public list of syllables. The task is to produce one on the basis of the Sloleks lexicon. A possible way of making the list would be to use the OpenOffice or TeX patterns for word hyphenation, as words are split according to syllable boundaries. The task includes and explanation of what Slovene syllables are, the list of Slovene syllables with frequency of occurrence, and an evaluation of the results.

Developing a Slovene-English machine translation system

While statistical machine translation has quite complex underlying mathematics, it has now become relatively simple to install and use SMT thanks to the Moses toolkit. The project will take an already prepared English-Slovene parallel corpus and train Moses on this data and make an evaluation of the results. Knowledge of Linux is needed.

Adding processing for Slovene to NLTK

Many resources for Slovene are now available under permissive licences, so it is possible to add processing for Slovene to various open source libraries, such as NLTK, http://www.nltk.org/ .

Tagging Slovene

The task is to train and evaluation the RFTagger and / or HunPos on Slovene, using the ssj500k manually annotated corpus and also the automatically annotated GigaFida corpus. You should convert the JOS resources into the format needed by the tagger, train it in various settings and evaluate the results.

Prerequisites: some knowledge of programming.

- http://www.ims.uni-stuttgart.de/projekte/corplex/RFTagger/
- http://mokk.bme.hu/resources/hunpos/
- http://hdl.handle.net/11356/1052
- http://nl.ijs.si/jos/

Writing HLT Wikipedia articles in Slovene

This might be appropriate for non-programmers and native Slovene speakers. The article(s) should also give pointers to specifically Slovene resources. Care should be taken with terminology and, possibly, related articles. One possibility would be to write an article on SSJ project.