Course
Advanced Language Technologies

Module "Knowledge Technologies"
Jožef Stefan International Postgraduate School
Winter 2011 / Spring 2012

URL http://nl.ijs.si/et/teach/mps11-hlt/

Possible topics for seminar work

2012-03-27

The course assessment is through seminar work on a topic connected with HLT, where ½ of the grade is on the basis of the quality of work performed, and ½ on the quality of the report (structure, literature review, how well the experiment is presented).

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

The topics should be chosen today by December 5th, at the latest. Students also have the option of performing the work in pairs (they will both get the same grade).

At the next lecture (March 28th), the students should present preliminary work on their topic (5-10 minutes each). Each student can have 1 hr of consultations with the lecturer either before or after – send email to arrange for the time of the consultation.

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

1. Rule induction for mapping historical words into modern ones

Shortly a corpus of annotated historical Slovene texts will be released on htrtp://nl.ijs.si/imp/ where each word in the corpus is annotated by its modern-day equivalent, e.g. "lubeſn / ljubezen". Make a program that learns rules to transform old words into modern ones, e.g. $l \rightarrow lj$, $ſn \rightarrow zen$, given a lexicon of modern words. The seminar work includes a survey of current approaches to this problem, c.f. Proceedings of LaTeCH (Language technology for cultural heritage) workshops.

2. WordNet and sense similarity

The task is to adapt Ted Pedersen's WordNet::Similarity package for Princeton WN to work with Slovene WN. The task includes a small evaluation of the results.

WordNet::Similarity takes two words and computes their semantic relatedness given wordnet; it can use various similarity measures, not all of then possible for Slovene.

Prerequisites: knowledge of Perl / Linux; knowledge of Slovene for evaluation.

http://nl.ijs.si/slownet/
http://www.d.umn.edu/~tpederse/similarity.html

3. Crowdsourcing for Slovene WordNet – test case

Make a small case study for correcting sloWNet through crowdsourcing. For example, generate a task that shows a Slovene literal and the synset it belongs to so that the user can confirm / reject it. The task also includes a small scale evaluation of the collected results (ask friends to take part)

Prerequisites: Access to a web server and knowledge of web service programing.

http://nl.ijs.si/slownet/

4. Named entity recognition for Slovene

Use the entities marked up in the SSJ learners corpus to develop a system for named entity recognition for Slovene.

5. Tagging Slovene with RFTagger

The task is to train and evaluation the RFTagger on Slovene, using the JOS corpora. The tagger has already been trained on Slovene data, but not very well. You should convert the JOS resources into the format needed by the tagger, train it and evaluate the results.

Prerequisites: some knowledge of programming.

http://www.ims.uni-stuttgart.de/projekte/corplex/RFTagger/
http://nl.ijs.si/jos/

Chosen topics

~~1. Evaluation of sloWTool~~ (Aleš Jurca)

The task is to evaluation the WordNet browser and editor sloWTool in terms of usability and easy of searching and editing, as well as to perform an evaluation of the current Slovene wordnet. The seminar should contain an introduction to wordnets, Slovene wordnet and editors / browsers for wordnets, and an evaluation of sloWTool, with emphasis on possible improvements.

Prerequisites: knowledge of Slovene.

http://nl.ijs.si/slownet/, http://nl.ijs.si/slowtool/search

~~2. Evaluation of MT for Slovene~~ (Blaž Mahnič)

There are only a few publicly available MT systems form Slovene: Presis from Amebis and Google Translate. Perform a contrastive evaluation of the quality of these two MT systems. The task involves a review of MT evaluation methods.

Prerequisites: knowledge of Slovene / English.

3. ~~Named entity recognition for occupations~~ (Jasna Škrbec)

Using the Enrycher program determine Named Entities in English newspaper articles, and then develop a classifier that will determine the field of occupation for persons.

4. ~~True Casing~~ (Tomaž Kompara)

In the JOS corpora (and others) the lemma form of the word is always given in lower case, even for names or abbreviations such as “Janez” or “NASA”. This is unfriendly to the users, and can degrade the precision of programs using these lemmas.

The task is to develop a program that would give the correct case to each lemma, using the information available in the JOS corpus: word-form, lemma, MSD, and possibly the lexicon extracted from a complete text. The task includes a good evaluation of the results.

Reference: http://acl.ldc.upenn.edu/acl2003/main/pdfs/Lita.pdf

5. ~~Widgets for Web services~~ (Anže Vavpetič, Nejc Trdin)

The task is to implement widgets for web-service workflows for the following tasks:

- reading and display of text
- To(Tr)TaLe anayzer
- frequency lexicon generator
- term extractor
- glossary extractor.

6. ~~Izdelava gradnikov za spletne servise~~ (Nej)

V okviru seminarske naloge bo implementiral gradnike (widgete) za števec frekvenc besed, term extractor za slovenščino ter glossary extractor.

7. ~~Survey of crowdsourcing for language resource annotation~~ (Maja Škrjanc)

Crowdsourcing – i.e. using “the general public” to perform task in annotation of data has become very recently very popular. Make a study of crowdsourcing for annotating language resources by going through recent proceeding of ACL, EACL, COLING, LREC. Suggest way in which this approach could be used for annotating Slovene WordNet.

8. ~~Survey of methods for event detection from text~~ (Alexandra Moraru)

Event detection deals with detecting “events of interest” which are usually anomalous events that rarely occur. Such methods are useful in many real-world tasks in surveillance, scientific discovery and data cleaning. The seminar should give a review of the state-of-the-art in event detection, focusing on textual data.

9. ~~Mapping LFG f-structures to CycL~~ (Janez Starc)

Task is to define, apply and evaluate rules for mapping the output of the PARC's XLE parser to the language of Cyc, named CycL. The XLE (http://www2.parc.com/isl/groups/nltt/xle/) consists of cutting-edge algorithms for parsing and generating Lexical Functional Grammars (LFGs).One of the distinct dimensions of the LFG is the representation of grammatical functions (f-structure or feature structure). On the other hand Research Cyc provides rules for semantic translation. These rules help map logic which is still in fairly linguistic form to a more semantic form written in CycL. The missing link in this pipeline are rules which can together with the mentioned rules from Cyc transform F-structures to CycL. All language processing will be done on English language. Rules will be defined and tested on controlled set of articles from Financial domain.

10. ~~Conversion of PDF documents into plain text for Web corpus building~~ (Jovan Tanevski, Nikola Simidjievski)

Many large corpora for individual languages are today produced by automatically crawling the Web. Methods for extracting text from HTML pages are well developed, however, lots of text is also present on the Web in the form of PDF documents, which are also a rich source for corpus building. Programs exists that convert PDF to plain text, but the quality of the resulting text is often poor, because the text can have problems in character encoding (čšž), “fi” and similar ligatures can be incorrectly encoded, or the layout of the text precludes good conversion.

The task is to devise a filter that would categorise (text extracted from) PDF documents into “usable” and “unusable”, depending on the quality of Slovene text inside them. This can be done with (a combination of) the character profile of the text, number of hyphens, n-gram models, percentage of OOV words, etc. The task also involves applying some heuristics to fix the most obvious errors in the text (such as end-of-line hyphenation).

Prerequisites: knowledge of programming.

11. ~~ChatBot using CyC~~ (Luka Bradeško)

The task is to develop a chatbot using AIML, Cyc EBMT and CURE.