Abstracts of Student Presentations

Student Presentations I, Monday, 07/12, 15-16 h, IKP Lecture Hall

Vaclav Nemcik
MU Brno, Czech Republic

Anaphora Resolution for Czech

One of the challenges in natural language understanding is to determine what entities are referred to in the discourse and how they relate to each other. This is a very complex task. But, as the first step, it is useful to determine co-reference classes over the set of referring expressions.
I will present a preliminary version of a system that performs automatic co-reference resolution on the syntactic basis. The system allows the realization of various AR algorithms in a modular way. It can be straightforwardly
used, in principle, with any natural language.

Dennis Mehay
KU Leven, Belgium

Automatic Classification of Semantic Roles

Halliday's Systemic-Functional Theory posits a set of semantic categories through which humans view and describe the world. We have adopted Halliday's categories (and the semantic roles implied therein) as a domain-independent (i.e., *generic*) framework for representing semantic roles and events in natural language discourse. We present a number of experiments in classifying pre-segmented generic semantic roles expressed by a small set of English verbs using five o®-the-shelf, open source classifiers from various
machine learning approaches — viz., support vector machines, maximum entropy modeling (MEM), *k* nearest neighbor classification (with $k = 1$), C4.5 decision trees and na¨ýve Bayes classifiers. Recently, great strides have been made in semantic role detection and classification in free text, but these results
have relied on large semantically annotated corpora, in-house built classifiers and sophisticated linguistic analysis. In our experiments we demonstrate that, by extracting only superficial features such as part of speech (POS) tags and lexical information, a reasonably accurate classification of semantic roles can be achieved through fairly simple means. To this end, we extracted a small subset (¼ 1450 sentences) of the Reuters Corpus (1996-08-20). We lemmatized and POS tagged this subcorpus and then annotated it with Halliday-style semantic roles. Feature vectors were extracted and used to train and test each classifier using 10-fold cross-validation. We achieved verb frequency-weighted average accuracies of up to 68% (MEM), as compared to a baseline (most common role assignment) accuracy of 37%. Our results suggest that reasonably good results can be obtained with freely available software and shallow linguistic analysis and are encouraging for the prospect of rapid deployment of low-cost semantic analysis systems.

Student Presentations II A: Tuesday, 07/13, 17-18 h, IKP Lecture Hall

Tomas Capek
MU Brno, Czech Republic

SAFT - A Software Module for Semantic Tagging Using WordNet

In this presentation we introduce a software module SAFT that enables a user to access Czech WordNet when processing a free Czech text. It can insert into morphologically analyzed text almost any information provided by WordNet, thus creating semantically tagged text (but not disambiguated). SAFT is coupled with the morphological analyzer for Czech called AJKA and also with the module for processing MWE.

A user can choose what data he wants to obtain on the output via numerous options. The tool provides a generic and universal environment for information processing from WordNet databases which can be useful in various applications as e.g. text categorization, partial semantic disambiguation or information extraction. SAFT can be also used within parsers to yield the necessary semantic information for selection the best parse tree as a desired result of parsing.

Marek Grac
MU Brno, Czech Republic

A Morphological Analyser for Slovak

During automatic text processing, the fundamental task is mastering the morphology of the language in question. Our morphological analyser is a
software tool used for classifying words into part-of-speech classes and assigning them grammatical categories. Until recently there was no decent analyser for Slovak available. That is the reason for importing available Slovak language data to the Czech morphological analyser AJKA. The main goal of this presentation is to cover the process of data conversion between two different formats and to describe the methods of solution for this problem. The fruit of this project is a functional morphological analyser able to recognize almost 40.000 word stems.

Student Presentations II – B: Tuesday, 07/13, 18-19 h, Lecture Room in IKP Pavilion

Yi Pan
KU Leuven, Belgium
Tools for a Sentence Compressor and an English Grammar for ShaRPa

Two parts of work during the internship are described. The first part mainly introduces several tools developed for the ATraNoS project, and also mentions a different way, the new ASD concept, of evaluating the existing compressor. These tools are used together with the compressor as a complete application for the ATraNoS project which generates the subtitle for the hearing-impaired people. The second part is mainly about the English chunking grammars I developed, which gets a high accuracy for the identification of intermediate level NPs and PPs based on the BNC C5 tagset. The chunker also gave the information of the concrete steps of the chunking procedure. The grammars are developed for usage within ShaRPa for English (V. Vandeghinste, 2003).

Nico Juhász
Universität Stuttgart, Germany

Text-to-Speech Syntheses - A Version for German TTS-Syntheses

The presentation will give an overview of different modules and necessary components as well as architectures of a TTS system. Furthermore, as an explicit example the Philipps TTS-System for German will be introduced.

Student Presentations III: Thursday, 07/15, 13-14 h, IKP Lecture Hall

Simon Meers
KU Leuven, Belgium

A Dialogue System for a Railway Application for Dutch I

My presentation deals with the work I have done during an internship at the company of Natlanco, Gent, within the framework of the Master's in Artificial Intelligence program at the KULeuven. The project dealt with the design of a dialogue system for a railway application. The dialogue system should be able to help customers who want information about train hour tables, prices, connections etc… The project was done in cooperation with another student, Mss Kamakshi Rajagopal, who is also a student of the MAI program.

The aim of the project, to create an intelligent written dialogue system for the Belgian railways and, by extension, a basic language model for the Dutch language has unfortunately not been entirely reached, as some parts of the dialogue system have not been designed and/or implemented due to time constraints. My presentation describes the different modules that are part of the dialogue system. It first gives an overview of these sections, i.e. corpus collection, syntax, morphology, lexicons, semantics and dialogue. Each section is described according to the initial goals that were defined and the results as they are. The theory behind the design and the program that was used are given a great deal of attention. The largest part deals with the development of the syntactic grammar. Additional attention also goes to the morphology, the syntactic and morphological lexicons and the correlation that exists between these elements.

The semantics and the dialogue models are discussed as well. Since these parts were not worked on profoundly during the internship due to time constraints, they are only discussed here in a minimal way.

Kamakshi Rajagopal
KU Leuven, Belgium

A Dialogue System for a Railway Application for Dutch II

The proposed presentation is based on a report of an internship at the company of Natlanco, Belgium, done within the framework of the Master in Artificial Intelligence program at the KULeuven, Leuven, Belgium. The aim of the project was to design a dialogue system for a railway kiosk application. This was done in cooperation with another student, Mr. Simon Meers, of the same program.

The internship report describes the different modules that are part of the dialogue system. These are corpus collection, morphology, syntax, lexicons, semantics and dialogue. For each module, it gives the initially defined aims and the eventually reached results. Much attention is given to the theoretical basis of each section and the practical issues that arose during its design. The actual design of the dialogue system was divided between the two students. This report contains an in-depth discussion of the morphological grammar and the lexicons. The decisions made regarding the structure of the grammar and the direct correlation that exists between the grammar and the lexicons are dealt with in much detail. Constant coordination was required with Mr. Meers, whose work included the design of the syntactic grammar. Finally, some time is devoted to the semantics and the dialogue models as well. Since these parts were not worked on profoundly during the internship due to time constraints, they are only discussed in a minimal way.

Poster Presentations: Wednesday, 07/14, 13-15 h, IKP Entrance Hall

Ying Sun
KU Leuven, Belgium

Recognizing Directory Information

This report will give a description of the work done during my internship at the ESAT-PSI Speech Group under supervision of professor Hugo Van hamme and Kris Demuynck. The content can be divided into three parts as follows:
The first part will give an overview of the two techniques – the classical all-in-one approach and the layered approach used in recognizing directory information. Furthermore, a graph search algorithm used for searching the best path in the phoneme lattice which is produced in the first phase of the layered approach is described.
In the second part, attention will be paid to the experimental resources. Here I will describe the methods used in handling the directory information given by the CGN corpus and the data provided by Phonetic Topegraphics / Tele Atlas.
In the third part, the experimental results will be shown in the third part, with the focus on a detailed error analysis.

Georgi Petrov
KU Leuven, Belgium

Training an ID3 tagger for a Swedish text-to-speech synthesis system

There are three important questions when developing the tagger for a (Swedish) text-to-speech system, which I am going to answer:
1) Is the tagger needed: how many words are written in the same way and pronounced differently in the language; will a tagger be able to disambiguate them?
2) How can the tag set be reduced to below 64 tags without losing ambiguities?
3) What are the best parameters for the tagger and how can a specially derived ambiguous lexicon be implemented for better performance on the ambiguous words?

Yanfen Hao
KU Leuven, Belgium

Switchboard Language Model Improvement with Conversational Data from Gigaword

http://www.esat.kuleuven.ac.be/~spch/MAIN.html

This paper aims to report my project at the ESAT-PSI speech group under supervision of Dr. Dong Hoon Van Uytsel and Dr. Jacques Duchateau. The goal of the project is to extract the conversational data from a newswire corpus (Gigaword) and to improve the conversational speech (Switchboard) language model.

The project consists of two stages. At the beginning of the first part, an overview of different approaches for text classification was given. Then a unigram classifier using cross-entropy for text categorization was built. What we achieved at the end of the first phrase was a conversational unigram classifier that showed high accuracy in classifying newswire text and transcriptions of conversations over the telephone.

In the second stage, the classifier was used to select additional conversational data from the newswire corpus in order to augment limited spontaneous speech data. The experiments were conducted to see how efficiently these additional data can make contribution in improving the spontaneous speech language model

Umer Imtiaz
KU Leuven, Belgium

Experimental results on a different feature extraction method in speech recognition

Speech recognition accuracy degrades when speech is corrupted by noise. This fact is more prominent on recognition accuracy when system has been trained on clean data. Noise in speech is difficult to handle especially when its prior knowledge is difficult to obtain or unknown. That is why noise compensation algorithms do not improve much recognition accuracy and even failed to do so. In this internship we have exploited an approach based on inherent properties of the speech signal. The speech signal is decomposed into harmonic and noise like components. These components are processed independently and then recombined. Recognition experiments are performed using these features on the Aurora 2 framework.

Karina Gössl
Universität Bonn, Germany

Is sarcasm universal or language specific? A preliminary acoustic study.

A sarcastic utterance is characterised by conveying the negation of its semantic meaning. One of the ways of signalling sarcasm in German appears to be a drastic lengthening. In this study, the universality of this acoustic signalling of sarcasm is tested on language material in various languages and language families.

Xiuli Jing
KU Leuven, Belgium

Evaluation of Headline Generation

My presentation is focused on the evaluation and refinement of headline generation.

Headline generation is a useful application in Natural Language Processing. Headlines are usually built by condensing a sentence from a document (in news corpora, usually the first sentence). A vital task in headline generation is to critically evaluate the approach and refine it. To build evaluation metrics, we used the manual headlines from the Document Understanding Conference (DUC) 2003[1]. Through a great deal of statistics and evaluation work of these corpora and also based on a study of headlines in the literature, we drafted the criteria for good headlines. Based on syntactical, semantic and length-oriented criteria we manually judged the headline generation algorithm, Hedge Trimmer[2]. Hedge Trimmer uses a compression technology which relies on the syntactic analysis of the sentence output by a statistical parser[3] and a minimum of linguistic knowledge.

We refined the Hedge Trimmer algorithm by creating new syntactical rules and produced new automatic headlines. We evaluated the new output on a small corpus from DUC 2004. The evaluation results of the new output showed that the modified Hedge Trimmer algorithm could generate very good headlines.

Cheng Dezhi
KU Leuven, Belgium

Music Perception of Cochlear Implant Recipients

I have been working at the Medical Electronics Lab of Antwerp University for three months, and have enjoyed working in such a vivid group. The internship started from 8th March and lasted till 28th May. During the 12 weeks, I have got a profound understand of the cochlear implant technology, strategies and design. Also I have learned some digital signal processing knowledge and designed common used filters. I got some ideas about DSP applications used in the speech and language technology. I understood the work flow of the cochlear implant CIS strategy and simulated the sounds after the processing that recipients receive. After that, I collected and studied the relative papers and wrote a 30-page literature review about the music perception of cochlear implant recipients. Afterwards, I got familiar with the music acoustic features and tested the music processed by cochlear implant. I analyzed the reasons result in cochlear implant recipients' poor music perception effects. Many experiments and programs were designed in Matlab and Visual Studio.net environment.

[1] The provided English newspaper texts: corpus of the DUC (Document Understanding Conference)
[2] Hedge Trimmer: A parse-and-trim approach to headline generation (B. Dorr et al, 2003)
[3] Statistical Parser: A maximum-entropy-inspired parser (E. Charniak, 2000)

Yi Zhou
KU Leuven, Belgium

An Implementation of Fixed-point Digital Filter/Filterbank For Cochlear Implants

Two platform independent fixed-point C code libraries, FIR filter and IIR filter bank were made in the internship. They are parts of the sound processing chain used for cochlear implants. Both floating-point and fixed-point versions of the libraries were developed on a PC. A set of tests were made to decide the word length and Q value for the internal fixed-point data form in calculation. Different signals (Noise, Sine wave and Speech) were used in the tests. Based on the analysis of the results of the tests, the 16 bits Q15 fixed-point data form was chosen for FIR filter and 32 bits Q20 for IIR filter bank. The fixed-point version of the programs was then implemented on a TI DSP processor (TMS320VC5509). A practical problem, 32 bits fixed-point multiplication, was solved here by using a fractional multiplication algorithm. The fixed-point programs yielded the same results on both PC and DSP processor.

Angeliki Papagiannopoulou
KU Leuven, Belgium

Optimizing and Localizing a rule – based approach to subtitle generation

The objective of this report is to give a general overview of the different steps that have been followed in optimizing and localizing a rule–based approach to subtitle generation. This work has been done in CNTS, in the University of Antwerp and it is part of the European project MUSA: Multilingual Subtitling of Multimedia Content. The aim of this project is to develop a multimodal multilingual system that converts audio streams into text transcriptions and then to generate automatically subtitles for television programs, for hearing impaired people. These subtitles will be translated in other languages. To be more specific, the source and target language is English and the generated English subtitles will be automatically translated in French and Greek. The subtitling part of the project, which will involve this report, aims to compress the sentences originally uttered by the speaker, so as to fit in the space available for subtitles. In order to proceed to subtitling generation there have been followed two aspects within the cadres of the MUSA project: the first one is based on hand-crafted deletion rules and the second one concerns learning sentence reduction from a parallel corpus.

Wen Zhao
KU Leuven, Belgium

Building a Chinese Language Model

With the growing interest in Chinese language processing, more and more NLP tools have been developed to segment, tag, or parse the Chinese sentences. However, due to the special features in Chinese, compared to English, the research in Chinese processing is far from mature. People cannot make a consensus in some syntax problems. In my internship at Natlanco, I built a Chinese language model, mainly syntactic and lexical aspects, with LingBench IDE. In the process of developing the model, I tried to sum up the guidelines of POS-tagging and parsing Chinese. Along with the optimization of the grammar, 300 sentences varying in length and complexity have been parsed and most sentences can get correct analysis without ambiguity in syntactic level. An annotated lexicon of 10000 Chinese word forms covers most of the basic words in daily life, all the words are annotated by hand; the results are accurate and consistent.

Shu-Ju Lee
Universität Bonn, Germany

Co-ordinating conjunctions as an indicator of text types in German?

Conjunction looks at inter-connections between processes: adding, comparing, sequencing, or explaining them. These are logical meanings that link figures in sequences. These meanings cannot be detected on the basis of a traditional word class specification. A large corpus consisting of different text types is examined and enriched with attributes according to the different logical relationships introduced by the conjunctions.