Language &
Communication
Technologies

Annual Meeting 2014

Rovereto

Students' Session

Abstract Booklet

Exploiting text corpora for visual recognition

Le Dieu Thu

Combining text and image processing has recently received an increased interest in both natural language processing and computer vision communities. Concept detection in images such as object detection and human action recognition are challenging tasks that often require expensive training data. We present a framework that aids concept detection in images by using general knowledge exploited from large text collections. This framework allows detecting high level concepts in images without expensive training images through automatic knowledge extraction from universal large text corpora.

Coloring objects: towards adjective nouns composition in images

Dat Tien Nguyen

In this project, we study how image representations can help avoiding the extraction of big amount of observed phrases from text to learn the DS meaning representations of functional words. To achieve such goal, we propose to exploit existing translations from a Distributional Visual Model (DVM) into a DSM. We take such vector representations in place of the text observed ones and run some of the standard CDSM evaluation task. In order to evaluate our system, we consider data containing phrases that occur rarely in a text. In particular, following (E. Bruni, G. Boleda, M. Baroni and N. Tran. 2012.) who shows that colors are captured better by DVM than DSM, we will look at adjective noun phrases involving colors.

Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world

Angeliki Lazaridou

Following up on recent work on establishing a mapping between vector-based semantic embeddings of words and the visual representations of the corresponding objects from natural images, we first present a simple approach to cross-modal vector-based semantics for the task of zero-shot learning, in which an image of a previously unseen object is mapped to a linguistic representation denoting its word. We then introduce fast mapping, a challenging and more cognitively plausible variant of the zero-shot task, in which the learner is exposed to new objects and the corresponding words in very limited linguistic contexts. By combining prior linguistic and visual knowledge acquired about words and their objects, as well as exploiting the limited new evidence available, the learner must learn to associate new objects with words. Our results on this task pave the way to realistic simulations of how children or robots could use existing knowledge to bootstrap grounded semantic knowledge about new concepts.

Algebraic Effects and Handlers in Natural Language Interpretation

Jiří Maršík

Phenomena on the syntax-semantics interface of natural languages have been observed to have links with programming language semantics, namely computational effcts and evaluation order. We explore this connection to be able to profit from recent development in the study of effects. We propose adopting algebraic effects and handlers as tools for facilitating a uniform and integrated treatment of different non-compositional phenomena on the syntax-semantics interface.

A Stylometric View of Individual Style Development

Carmen Klaussner

In authorship analysis, it is a natural idealization to treat different works of an author as synchronous events even though this is tantamount to the impossibility that they were all written at the same instant. Therefore, this takes into account neither the individual changes that an author's style might undergo over time, nor the general underlying language change influencing all contemporaneous writers.

This study addresses stylistic change over time of two authors of the late 19th to early 20th century, namely Henry James and Mark Twain. Using time series analysis, we observe features that undergo a change in terms of frequency of usage for an individual author and try to find co-dependent relationships between those features, in the sense of one growing stronger while the other is weakening over time. Having identified early and late profiles as well as related features, we interpret profiles in a psychological context relating to the author's persuasion at that time.

Het Woordspelletje: a serious game for building a Dutch word association network

Tuur Leeuwenberg
Sara van de Moosdijk

Het Woordspelletje aims to let users contribute to a Dutch word association network through game play, thereby creating a new linguistic resource for performing research on the Dutch language. By building the network through the use of a game, this project aims to capture the use of everyday Dutch and give every Dutch-speaker a chance to contribute to research. The resulting association network is freely available for any and all research purposes.

A Game with a Purpose for Semantic Role Annotation

Lena Rampula

Semantic role assignment can be a very challenging task for a computer, but much simpler task for a human. Given an adequate explanation, a non-expert player can easily choose the right answer. Meanwhile agreement between the players ensures that the role is indeed assigned correctly. My focus is on the automatic generation of questions from the Groningen Meaning Bank. The texts in the corpus are annotated using DRS. I use this information to depict a sentence for the question and extract the arguments of the verb. To simplify the question, the highlighted argument was limited to the head noun of the noun phrase. VerbNet was used for the generation of multiple-choice answers. For each verb there is a different set of possible roles, thus a different set of possible answers. In each question, the answers include only the relevant roles for the highlighted verb.

Nominal Coercion in Space: Investigating Mass and Count Nouns Using Distributional Semantics

Manuela Hürlimann

It is generally assumed that when a noun shifts from mass to count ("three *beers* please"), or vice versa ("there is *apple* in the salad"), its meaning changes. We try to shed light on these shifts using Distributional Semantics. Following Katz and Zamparelli, we look at the singular-plural similarities of both mass and count nouns. In line with their results, we find a greater distance between singular and plural for mass nouns than for count nouns and interpret this as an indication that mass nouns need to undergo coercion in order to be pluralised. We attempt to find the direction and destination of a shift in the semantic space, and are working on identifying groups of nouns that shift in a similar way using clustering. This should allow us to bridge Theoretical Linguistics and Distributional Semantics by verifying whether the claims made in the Theoretical Linguistics literature about "shift classes" are reflected in cooccurrence statistics extracted from corpora. In doing so, we seek to discover the fingerprint of nominal coercion in the semantic space.

The Predictive Powers of Properties: Visual Concept Knowledge from Linguistically Derived Attributes

Kim Heiligenstein

Corpus-based semantic models such as Strudel (Baroni et al 2010) allow us to reach conceptual knowledge linguistically through induction using distributional semantics to derive property-based concept representation. Because humans base their meaning representation and acquire semantic knowledge from perceptual information as well, the combination of language and vision to reach a more cognitive view of meaning seems to be the natural extension. We suggest a model that uses semantically derived concept attributes from a corpus of naturally-occurring text applied to an image-based model to arrive at meaning representations endowed with stronger cognitive qualities. The discussion consists of a qualitative data analysis which encourages the idea that the concept attributes extracted from text are in fact plausible candidates to achieve conceptual knowledge grounded in visual perception.

Automatic rule acquisition for transliteration to East Asian languages

Matthew Smith

Chinese, Japanese, and Korean possess scripts and phonological systems that are radically different from those of European languages. The goal of this thesis is to: a) automatically develop rules for transliteration from various European languages to these East Asian languages insofar as this is possible, b) predict cases where the rules fail to provide correct transliterations, and c) show that Chinese, due to its logographic script and its tonality, is much more difficult to transliterate to than the other two languages. The results could be used both to build a machine transliteration system as well as to teach second-language learners how to transliterate proper nouns.

Combining Statistical and Symbolic Methods for Parsing

Anastasia Shimorina
Anastasiia Tsukanova

Using symbolic methods is advantageous as in contrast to statistical methods because they allow to handle much more complex linguistic phenomena. But there are two quite well-known drawbacks of symbolic parsing: it is not robust and it is very costly. The task is to try different ways of combining some statistical tools with symbolic methods. We have chosen the Leopar parser for French by the Sémagramme team, the Sequoia corpus, the MaltParser and the Talismane parser.

Robustness: The easiest way is to parse a sentence with Leopar hoping to capture complex linguistic phenomena first, and then, if Leopar fails, to use a statistical parser as a backoff tool. A better way is to analyze the given sentence as the whole, and to provide some parts of it that Leopar may be able to parse.

Computational cost: We consider an output of a statistical parser for a sentence where Leopar fails, because of a timeout error, and give a selection of its POS tags to Leopar. However, this approach requires an analysis how to make the tagging compatible.

Composition in Distributional Semantics & Sentence Meaning

Nghia Pham

Distributional semantic methods to approximate word meaning with context vectors have been very successful empirically, and the last years have seen a surge of interest in their compositional extension to phrases and sentences. We present here a new model that, like those of Coecke et al. (2010) and Baroni and Zamparelli (2010), closely mimics the standard Montagovian semantic treatment of composition in distributional terms. However, our approach avoids a number of issues that have prevented the application of the earlier linguistically-motivated models to full-fledged, real-life sentences. We test the model on a variety of empirical tasks, showing that it consistently outperforms a set of competitive rivals.

Mining texts at discourse level

Sara van de Moosdijk

Linguistic discourse refers to the meaning of large chunks of text, from phrases to whole documents. It could be very useful for guiding attempts at text mining, which focus on the goals of document selection, document summarization, or other knowledge extraction goals. Hence the aim of this project is to apply discourse information in textual data to Knowledge Discovery in Databases. As far as we know, this is the first attempt at combining these two very different fields, so the goal is to create a basis for this type of knowledge extraction. We approach the problem by extracting discourse relations using unsupervised methods, and then model the data using pattern structures in Formal Concept Analysis, which are ideal for handling complex data. Our method is applied to a corpus of medical articles compiled from PubMed. This medical data can be further enhanced with concepts from the UMLS MetaThesaurus, which are combined with the UMLS Semantic Network to apply as an ontology in the pattern structures. The results show that despite having a large amount of noise, the method is promising and could be applied to domains other than the medical domain. We explore the pitfalls and suggest ways in which the process could be improved.

MDI Adaptation for the Lazy: Avoiding Normalization in LM Adaptation for Lecture Translation

Nicholas Ruiz

We provide a fast alternative to Minimum Discrimination Information-based language model adaptation for statistical machine translation. We provide an alternative to computing a normalization term that requires computing full model probabilities (including back-off probabilities) for all n-grams. Rather than re-estimating an entire language model, our Lazy MDI approach leverages a smoothed unigram ratio between an adaptation text and the background language model to scale only the n-gram probabilities corresponding to translation options gathered by the SMT decoder. The effects of the unigram ratio are scaled by adding an additional feature weight to the log-linear discriminative model. We present results on the IWSLT 2012 TED talk translation task and show that Lazy MDI provides comparable language model adaptation performance to classic MDI.

LCT master Annual Meeting 2014

Student session

Exploiting text corpora for visual recognition
Le Dieu Thu

Coloring objects: towards adjective nouns composition in images
Dat Tien Nguyen

Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world
Angeliki Lazaridou

Algebraic Effects and Handlers in Natural Language Interpretation
Jiří Maršík

A Stylometric View of Individual Style Development
Carmen Klaussner

Het Woordspelletje: a serious game for building a Dutch word association network
Tuur Leeuwenberg, Sara van de Moosdijk

A Game with a Purpose for Semantic Role Annotation
Lena Rampula

Nominal Coercion in Space: Investigating Mass and Count Nouns UsingDistributional Semantics
Manuela Hürlimann

The Predictive Powers of Properties: Visual Concept Knowledge from Linguistically Derived Attributes
Kim Heiligenstein

Automatic rule acquisition for transliteration to East Asian languages
Matthew Smith

Combining Statistical and Symbolic Methods for Parsing
Anastasia Shimorina, Anastasiia Tsukanova

Composition in Distributional Semantics & Sentence Meaning
Nghia Pham

Mining texts at discourse level
Sara van de Moosdijk

MDI Adaptation for the Lazy: Avoiding Normalization in LM Adaptation for Lecture Translation
Nicholas Ruiz

www.lct-master.org
www.unitn.it
booklet by Antonio F. García Sevilla

Erasmus Mundus