Kleinwalsertal Retreat

6–8 July 2017

SFB 732: *Incremental Specification in Context*

Integrated Research Training Group (MGK)

version of July 5, 2017

Program

Thursday

12:30 – 14:00	lunch	
14:00 – 14:20	Michael Neumann (A8) *Speech Emotion Recognition with Convolutional Neural Networks*	short talk
14:20 – 14:40	Daniel Ortega (A8) *A neural-based dialog act classifier for spoken dialog systems*	short talk
14:40 – 15:00	Moritz Stiefel (A8) *First Steps Towards Dependency Parsing on ASR Output*	short talk
15:00 – 15:10	short break	
15:10 – 15:30	Jeremy Barnes (D12) *Cross-lingual Sentiment Analysis for Under-resourced Languages*	short talk
15:30 – 15:50	Evgeny Kim (CRETA) *Emotions in Fiction*	short talk
15:50 – 16:10	Sarah Schulz (CRETA) *Non-Standard Text Processing at the Example of POS Tagging of Historical Languages*	short talk
16:10 – 16:40	coffee break	
16:40 – 17:00	Florian Groll (B8) *Charting beneficiency*	short talk
17:00 – 17:20	Alassane Kiemtoré (B8) *Clause embedding and logophoric marking*	short talk
17:20 – 17:40	Anne Temme (B6) *Experiencers, attitudes and evaluation*	short talk
17:40 – 17:50	short break	
17:50 – 18:10	Fabian Bross (ILG) *Scope-Taking Strategies in German Sign Language*	short talk
18:10 – 18:30	Ina Rösiger (A6; CRETA) *Bridging resolution - task, data and challenges*	short talk
18:30 – 19:30	dinner	
19:30 – 20:30	Self evaluation	

Friday

08:00 – 09:00	breakfast	
09:00 – 09:20	Anna Hätty (Bosch) *Term extraction for standard and user-generated text: scoring term specificity*	short talk
09:20 – 09:40	Max Kisselew (B9) *Making Sense of Morphological Derivation*	short talk
09:40 – 10:00	Abhijeet Gupta (D10) *Relation prediction through distributed word embeddings*	short talk
10:00 – 10:10	short break	
10:10 – 10:30	Sylvia Springorum (D12) *A cognitive perspective on German particle verbs*	short talk
10:30 – 10:50	Maximilian Köper (D12) *Modelling German Particle Verb Analogies using Vector Space Models*	short talk
10:50 – 11:10	Dominik Schlechtweg (D12) *Distributional Models of Semantic Change*	short talk
11:10 – 11:40	coffee break	
11:40 – 12:00	Xiang Yu (D8) *A General-Purpose Tagger with Convolutional Neural Networks*	short talk
12:00 – 12:20	Agnieszka Faleńska (D8) *Source selection for delexicalized parsing of low-resource languages*	short talk
12:20 – 12:30	short break	
12:30 – 13:15	Sabrina Stehwien (A8) *CNN-based Prosodic Event Recognition for Speech Understanding*	progress report
13:15 – 14:00	Markus Gärtner (INF) *A universal modeling and query toolkit for multi-layer and multi-modal corpora*	progress report
14:00 – 18:30	free time	
18:30 – 19:45	dinner	

Saturday

08:00 – 09:00	breakfast	
09:00 – 09:45	Yvonne Viesel (A6) *Information structure and discourse particles*	dry run
09:45 – 10:30	Kyle Richardson (D2) *Investigations into Semantic Parser Induction*	dry run
10:30 – 10:45	short break	
10:45 – 11:30	Anders Björkelund (D8) *Online Learning of Structured Predictors with Approximate Search and Latent Structure*	dry run
11:30 – 12:15	Collaborations • Yvonne Viesel and Arndt Riester (A6): *Information structure in wh-questions: Evidence from discourse structure* • Ina Rösiger (A6) and Kim-Anh Nguyen, Maximilian Köper, Sabine Schulte im Walde (SemRel): *Using predicted semantic relations for coreference and bridging resolution* • Ina Rösiger (A6) and Sabrina Stehwien (A8): *Using automatic prosodic annotations for coreference resolution* • Diego Frassinelli et al.: *ExPsy Group*	
12:15 – 13:15	lunch break	
13:15 – 14:00	Jason Utt (D10) *Constructing Syntax-Based Distributional Semantic Models for Novel Languages*	dry run
14:00 – 14:45	Patrick Ziering (D11) *Indirectly Supervised Determination and Structural Analysis of Nominal Compounds*	dry run

Abstracts

Short talks

Speech Emotion Recognition with Convolutional Neural Networks

Michael Neumann (A8) [short talk]

Project: SFB 732, A8 "Investigating the Interaction between Speech and Language Processing for Spoken Language Understanding: A Case Study for Sentiment Analysis"
E-mail: michael.neumann@ims.uni-stuttgart.de
Website: http://www.ims.uni-stuttgart.de/institut/mitarbeiter/neumanml/
Institute: Institut für Maschinelle Sprachverarbeitung

Speech emotion recognition is an important and challenging task in the realm of human-computer interaction. Prior work has proposed a variety of models and feature sets for training a system. We have conducted extensive experiments using an attentive convolutional neural network. We compare system performance using different lengths of the input signal, different types of acoustic features and different types of emotion speech (improvised/ scripted). Our experimental results on the Interactive Emotional Motion Capture (IEMOCAP) database reveal that the recognition performance strongly depends on the type of speech data independent of the choice of input features. In this talk I will give an overview on the task of speech emotion recognition and present our convolutional neural network model. Further, I will present first results from multi-lingual experiments incorporating English and French speech data.

A neural-based dialog act classifier for spoken dialog systems

Daniel Ortega (A8) [short talk]

Project: SFB 732, A8 "Investigating the Interaction between Speech and Language Processing for Spoken Language Understanding: A Case Study for Sentiment Analysis"
E-mail: daniel.ortega@ims.uni-stuttgart.de
Website: -
Institute: Institut für Maschinelle Sprachverarbeitung

Our research focuses on neural-based task-oriented spoken dialog systems. As part of the spoken language understanding module, we approach the capability to identify automatically the dialog act (DA) of each utterance in a dialog using not only lexical features obtained from manual or automatic transcriptions, but also the acoustic features contained in the speech signal, in order to deal with ambiguous DAs that cannot be correctly distinguished by only looking at lexical information. Moreover, this approach is intended to reduce the impact of the word error rate caused by automatic speech recognitionin. Our proposed model employs convolutional neural networks, recurrent neural networks and attention mechanisms to process lexical and acoustic features producing combined high-level features for the final classification.

First Steps Towards Dependency Parsing on ASR Output

Moritz Stiefel (A8) [short talk]

Project: SFB 732, A8 "Investigating the Interaction between Speech and Language processing for Spoken Language Understanding"
E-mail: moritz.stiefel@ims.uni-stuttgart.de
Website: http://www.ims.uni-stuttgart.de/institut/mitarbeiter/stiefemz/index.html
Institute: Institut für Maschinelle Sprachverarbeitung

Parsing natural speech is a precursor to its understanding. Creating a representation that allows parsing speech is the goal of this work. We frame our problem as joint modeling of automatic speech recognition (ASR) and part-of-speech (POS) tagging to enrich ASR word lattices. To that end, we manipulate the ASR process from the pronouncing dictionary onwards to use word-POS pairs instead of words. We evaluate ASR, POS tagging and dependency parsing (DP) performance demonstrating a successful lattice-based integration of ASR and POS tagging at cost.

Cross-lingual Sentiment Analysis for Under-resourced Languages

Jeremy Barnes (D12) [short talk]

Project: SFB 732, D12 "Sense Discrimination and Regular Meaning Shifts of German Particle Verbs"
E-mail: jeremy.barnes@upf.edu
Website: https://jbarnesspain.github.io/
Institute: Institut für Maschinelle Sprachverarbeitung

The data created on the Internet often contain information about private states, such as emotions, sentiment, or polarity, which are of interest to researchers in sentiment and emotion analysis. For many languages, however, there are not enough resources to extract this information reliably. In these cases, we can use cross-lingual methods by which we leverage information available in one language to perform a task in an under-resourced language. An attractive alternative to commonly used machine translation is to represent words using a cross-lingual distributional semantic model (CDSM). These methods are particularly appropriate for under-resourced languages, but can also provide a simple solution for any language that does not have sentiment or emotion resources. However, their current performance for cross-lingual sentiment analysis is not optimal. Therefore, it is necessary to improve the current models for these tasks.

Emotions in Fiction

Evgeny Kim (CRETA) [short talk]

Project: CRETA "Centre for Reflected Text Analytics"
E-mail: evgeny.kim@ims.uni-stuttgart.de
Website: http://www.ims.uni-stuttgart.de/institut/mitarbeiter/kimey
Institute: Institut für Maschinelle Sprachverarbeitung

In this talk, I will give a broad overview of my PhD plans and briefly discuss two directions of my research. First, I will present some results on literary genre classification based on the concept of emotional arcs, and will argue that this approach is complementary to the state-of-the-art. Then I will briefly outline an ongoing work on the annotation of directed emotions in the literary fiction and will explain how that fits my PhD goals.

Non-Standard Text Processing at the Example of POS Tagging of Historical Languages

Sarah Schulz (CRETA) [short talk]

Project: CRETA "Center for Reflected Text Analytics"
E-mail: sarah.schulz@ims.uni-stuttgart.de
Website: http://www.ims.uni-stuttgart.de/institut/mitarbeiter/schulzsh/
Institute: Institut für Maschinelle Sprachverarbeitung

Tools that have been developed for standard language show a crucially decreased performance on non-standard data. The best POS taggers that are available for English reach an accuracy of up to 97.6% on the Wall Street Journal (Choi 2016) and the best German models perform around 97% for newspaper text (Brants 2000, Schmid 1994), the performance for texts from the web drops to 90–93% (Giesbrecht and Evert 2009) and more significantly decreases for Middle High German to 45%. Texts serving as basis of Digital Humanities research often deviate from standard text. To provide an example of how to work with such non-standard text, I focus on POS tagging of historical text. Historical stages of languages do not just differ from the modern stage of the language but commonly also show a large diversity within what is considered to be one stage of a language due to missing regulation of spelling and grammar and a not yet unified lexis. To be able to compare techniques for different kinds of text, I evaluate the task of POS tagging throughout different data sets with a varying degree of divergence from the standard form and differences in terms of availability of preprocessing tools. I examine the influence of the quality and quantity of training data on tagging performance as well as the suitability of different techniques for low-resource non-standard data processing.

Charting beneficiency

Florian Groll (B8) [short talk]

Project: SFB 732, B8 "Alternations and binding"
E-mail: florian_groll@yahoo.de
Website: http://www.uni-stuttgart.de/ilg/institut/mitarbeiter/groll/index.html
Institute: Institut für Linguistik/Germanistik

Although a lot of typological work on beneficiency has been carried out in the past ten years no proper semantic analysis and subclassification of this cross-linguistic phenomenon has been developed up to the present day. Thus, the primary goal of our presentation will be to show the correspondence between event mereology and different subtypes of beneficiency. We will come up with a major dividing line between perceptive benefaction and causative benefaction and we will review different beneficiency-related phenomena like, for example, the contiguity constraint in Dutch or engager-benefaction in Korean.

Clause embedding and logophoric marking

Alassane Kiemtoré (B8) [short talk]

Project: SFB 732, B8 "Alternations and binding"
E-mail: akiemtor04@yahoo.fr
Website: –
Institute: Institut für Linguistik/Germanistik

Cross-linguistically logophoric marking can be seen as a strategy of coreference marking involving report or attribution of speech, attitudes and other mental states. One crucial syntactic property of logophoric markers (pronouns, clitics and affixes) is that they occur in clause subordinate to another one (the matrix clause); hence a common assumption in the literature is that logophoric marking is sensitive to semantics of some embedding predicates, so called *logophoric verbs* (Culy 1994, Adesola 2005, 2006, Koopman and Sportiche 1989, Speas 2004, Sundaresan 2012 among others). In the talk I wish to show that this picture is too simple and cannot account for all aspects of logophoric marking. I further propose that (at least) for some constructions the logophoric marker seems to be licensed by some factors inherent to the embedded clause (the logophoric domain) itself.

Experiencers, attitudes and evaluation

Anne Temme (B6) [short talk]

Project: SFB 732, B6 "Underspecification in Voice systems and the syntax-morphology interface"
E-mail: anne.temme@posteo.de
Website: www.annetemme.com
Institute: Institut für Linguistik/Germanistik

Experiencer object verbs license sentential complements (i), a property they share with attitudinal (ii) and dispositional/evaluative predicates (iii/iv):

i. That Peter is a spy surprised/annoyed everyone.
ii. Everyone around him knows/regrets that Peter is a spy.
iii. That Peter is a spy endangered everyone around him.
iv. That Peter is a spy is not good for everyone around him.

I my talk, I compare the verb classes based on this common feature in order to get closer to the evaluative nature of psych verbs.

Scope-Taking Strategies in German Sign Language

Fabian Bross (ILG) [short talk]

Project: –
E-mail: fabian.bross@ling.uni-stuttgart.de
Website: http://www.uni-stuttgart.de/ilg/institut/mitarbeiter/bross/
Institute: Institut für Linguistik/Germanistik

This talk investigates the manual and non-manual encoding of high and low scope-taking operators in German Sign Language (DGS) and adopts a cartographic approach to clausal syntax (Cinque 1999, 2006). The central hypotheses are that the scope-taking indicated by concatenation for lower operator switches to nonmanual suprasegmental facial operator expressions for higher operators, thereby establishing the vertical axis as a scopally relevant dimension: high operators are encoded high, and low operators are encoded low. The findings have repercussions for the comparison of DGS and spoken German on the one hand, and for a general theory of scope-taking in sign languages, on the other. It is assumed that all sign languages follow this simple rule: The higher an operator is, the higher its expression on the body.

Bridging resolution – task, data and challenges

Ina Rösiger (A6; CRETA) [short talk]

Project: SFB 732, A6 "Encoding of Information Structure in German and French";
CRETA "Center for Reflected Text Analytics"
E-mail: Ina.Roesiger@ims.uni-stuttgart.de
Website: http://www.ims.uni-stuttgart.de/institut/mitarbeiter/roesigia/
Institute: Institut für Maschinelle Sprachverarbeitung

Bridging resolution is the task of linking certain anaphoric noun phrases (bridging anaphors) and their antecedents, where both do not refer to the same referent, but are related in a way that is not explicitly stated. The resolution of bridging links is important because it can help in tasks which use the concept of textual coherence, e.g. Barzilay and Lapata (2008)'s entity grid. They might also be of use in higher-level text understanding tasks such as textual entailment or summarisation. Given that there are no clear syntactic or other surface markers to indicate the existence of a bridging anaphor, bridging resolution is considered challenging. In this talk, I will introduce the task and give a short overview on previous work before I focus on our current annotation and computational modelling efforts and the challenges that we are faced with.

Term extraction for standard and user-generated text: scoring term specificity

Anna Hätty (Bosch) [short talk]

Project: Bosch, group: "User Monitoring and Modeling", project: "Ubiquitous Personal Assistance"
E-mail: Anna.Haetty@de.bosch.com
Website: –
Institute: –

Automatic Term Extraction (ATE) or Automatic Term Recognition (ATR) is a subtask of Information Extraction which deals with identifying terms in domain-specific corpora. The focus of this work will be both on finding a domain-independent approach to define and to identify terms, and to characterize terms for specificity. The thesis will consist of three parts: Creating a gold standard annotated according to domain-independent term annotation guidelines, identifying terms automatically and scoring them for specificity. For the first task, I will introduce (technical) terms and domains, and the hardnesses to define them and their relation to each other. For the second part, a first approach will be presented where several conventionally used statistical feature classes are combined to identify terms. Strengths and weaknesses of this approach will be discussed, as well as advantages compared to standard approaches for term extraction. In the end, I will give an outlook on how to address specificity and how this is applicable to determine user expertise.

Making Sense of Morphological Derivation

Max Kisselew (B9) [short talk]

Project: SFB 732, B9 "Distributional Characterization of Derivation"
E-mail: kisselmx@ims.uni-stuttgart.de
Website: http://www.ims.uni-stuttgart.de/institut/mitarbeiter/kisselmx/
Institute: Institut für Maschinelle Sprachverarbeitung

The goal of my PhD project is to investigate how Distributional Semantics can contribute to our understanding of morphological derivation. Over the last years I conducted several studies which aimed at investigating how derived word forms in German can be predicted from their base forms using methods from Compositional Distributional Semantics. But these approaches were afflicted with the fact that a distributional vector usually conflates different senses of a word. However, as Plank (2010) points out, for some cases of derivation the derived word often maintains only a specific sense of its base. Taking this into account I employ recently introduced models for constructing disambiguated word vectors to investigate to what extent they can improve the performance of Compositional Distributional Semantics Models used to generate vectors for derived words in English. In my talk I will present the results of this study.

Relation prediction through distributed word embeddings

Abhijeet Gupta (D10) [short talk]

Project: SFB 732, D10 "Incrementality in Compositional Distributional Semantics"
E-mail: abhijeet.gupta@gmail.com
Website: http://www.abhijeetgupta.info
Institute: Institut für Maschinelle Sprachverarbeitung

Word embeddings represent the meaning of a word by capturing the word's meaning through (observing) its contexts. Such embeddings have been extensively and successfully applied in machine learning models to address various phenomena and tasks in computational linguistics and NLP. However, there is a general assumption that embeddings encode relatively coarse-grained concept-level information (dogs and cats are more similar than dogs and beetles). The topic of my PhD thesis is on analyzing to what extent more fine-grained semantic properties and relations are also represented in embeddings. The focus is on Named Entities which are generally associated with specific properties and relations (Italy has a population of about 60 million, Italy borders Croatia). In an earlier work, we extracted numeric relations from word embeddings to a reasonable degree of accuracy. In the current work (and presentation), we focus on predicting categorical relations and analyse various aspects which make these relations easy or difficult to predict.

A cognitive perspective on German particle verbs

Sylvia Springorum (D12) [short talk]

Project: SFB 732, D12 "Sense Discrimination and Regular Meaning Shifts of German Particle Verbs"
E-mail: Sylvia.Springorum@ims.uni-stuttgart.de
Website: http://www.ims.uni-stuttgart.de/institut/mitarbeiter/riestesa/
Institute: Institut für Maschinelle Sprachverarbeitung

I will present an overview of various experiments on particle verbs (PV). At first, we showed that prepositional particles can be associated with arrow pictographs, which we understand as visual representations of directional concepts. This conceptual directionality also influences the compositionality of a PV, if the base verb (BV) is also connected to a conceptual direction. A directional mismatch results in longer processing time. Furthermore, I will introduce our data on PV neologism with analyses, focusing on concrete and abstract concepts, which also determines whether a PV is understood as literal or non-literal. The domains of the verb and its context are thereby central. Our data collection with annotation of common source domains for BVs and common target domains for PVs will give some deeper insights into the role of domains, with respect to PV readings and their meaning-shifts.

Modelling German Particle Verb Analogies using Vector Space Models

Maximilian Köper (D12) [short talk]

Project: SFB 732, D12 "Sense Discrimination and Regular Meaning Shifts of German Particle Verbs"
E-mail: Maximilian.Koeper@ims.uni-stuttgart.de
Website: http://www.ims.uni-stuttgart.de/institut/mitarbeiter/koepermn/index.en.html
Institute: Institut für Maschinelle Sprachverarbeitung

Vector-space representations provide geometric tools for reasoning about the similarity of a set of objects and their relationships. On the other hand German particle verbs (PVs) such as anlachen (laugh at) are compositions of a base verb (BV) such as lachen (smile/laugh) and a verb particle such as an. PVs are highly productive and the particles are notoriously ambiguous. In this talk I will present vector space models of analogy (beyond man:king::women:queen !) and their potential application to find regular meaning shifts between base verb – particle verb combinations.

Distributional Models of Semantic Change

Dominik Schlechtweg (D12) [short talk]

Project: SFB 732, D12 "Sense Discrimination and Regular Meaning Shifts of German Particle Verbs"
E-mail: Dominik.Schlechtweg@gmx.de
Website: http://www.ims.uni-stuttgart.de/institut/mitarbeiter/schlecdk/
Institute: Institut für Maschinelle Sprachverarbeitung

The thesis shall examine the distributionally measurable properties of semantic change in German. It shall not focus on the reasons but rather on the effects of semantic change and exploit those that are measurable in order to make it accessible to automatic detection. Therefore, it shall be determined in the first part which effects of semantic change are automatically measurable in German text corpora. Such are, for instance: frequency, semantic generality, polysemy, grammatical status, subjectivity, evaluativeness or semantic similarity. Different evaluation resources related to semantic change phenomena displaying these effects shall be gathered, implemented or created. In a second part different models shall be implemented and tested for their capacity to quantify the effects and to predict specific phenomena. In a last part the models shall be combined in order to investigate how well they can distinguish different types of semantic change such as desemanticization (as part of grammaticalization) or metaphoric change.

A General-Purpose Tagger with Convolutional Neural Networks

Xiang Yu (D8) [short talk]

Project: SFB 732, D8 "Data-driven Dependency Analysis – Context Factors in Parsing and Anaphora Resolution"
E-mail: Xiang.Yu@ims.uni-stuttgart.de
Website: http://www.ims.uni-stuttgart.de/institut/mitarbeiter/xiangyu/index.en.html
Institute: Institut für Maschinelle Sprachverarbeitung

We present a general-purpose tagger based on convolutional neural networks (CNN), used for both composing word vectors and encoding context information. The CNN tagger is robust across different tagging tasks: without task-specific tuning of hyper-parameters, it achieves state-of-the-art results in part-of-speech tagging, morphological tagging and supertagging. The CNN tagger is also robust against the out-of-vocabulary problem; it performs well on artificially unnormalized texts.

Source selection for delexicalized parsing of low-resource languages

Agnieszka Faleńska (D8) [short talk]

Project: SFB 732, D8 "Data-driven Dependency Analysis – Context Factors in Parsing and Anaphora Resolution"
E-mail: Agnieszka.Falenska@ims.uni-stuttgart.de
Website: http://www.ims.uni-stuttgart.de/institut/mitarbeiter/falensaa/index.en.html
Institute: Institut für Maschinelle Sprachverarbeitung

A straightforward approach to parsing low-resource languages is to use delexicalized parser transfer. The idea is to train a parser on a source treebank using only non-lexical features and apply it on sentences from the target language. Selecting the best source treebank for parsing a given target language is a challenging task. I will show two different methods: (1) using language specific features from Wals and (2) using treebank characteristics. I will analyze how both these methods behave in two tasks: the best source selection and weighting sources for re-parsing. The parsing experiments will be held on Universal Dependencies treebanks with very small amounts of training data i.a. surprise languages from the CONLL 2017 Shared Task.

Progress report talks

CNN-based Prosodic Event Recognition for Speech Understanding

Sabrina Stehwien (A8) [progress report]

Project: SFB 732, A8 "Investigating the Interaction between Speech and Language Processing for Spoken Language Understanding: A Case Study for Sentiment Analysis"
E-mail: sabrina.stehwien@ims.uni-stuttgart.de
Website: http://www.ims.uni-stuttgart.de/institut/mitarbeiter/stehwisa/index.en.html
Institute: Institut für Maschinelle Sprachverarbeitung

The importance of including prosodic knowledge in the automatic processing and understanding of speech has already been well established. For example, prosody can help disambiguate sentence meaning and guide the listener through the information structure of speech. My PhD project consists of two main goals:

1) the development and detailed analysis of efficient prosodic event recognizion methods using Convolutional Neural Networks (CNNs) that can readily be applied to new datasets in both English and German.
2) the investigation of the use of including automatically obtained prosodic information in various speech and language processing tasks, especially on automatically recognized text.

A universal modeling and query toolkit for multi-layer and multi-modal corpora

Markus Gärtner (INF) [progress report]

Project: SFB 732, INF "Informationsinfrastruktur"
E-mail: Markus.Gaertner@ims.uni-stuttgart.de
Website: http://www.ims.uni-stuttgart.de/institut/mitarbeiter/gaertnms/
Institute: Institut für Maschinelle Sprachverarbeitung

In my PhD project I work towards a corpus query toolkit that is able to operate across modalities and which can access most if not all possible linguistic annotation layers. As such the goals revolve around a) modeling access to a very diverse universe of potential input data, and b) designing a query system expressive enough to exploit this richness. Especially the interesting contact points of different modalities pose challenges for existing state-of-the-art modeling and query systems and limit their usability to subfields, preventing easy exploitation of synergies between multiple layers, theories or modalities. Since this talk is part of my progress report, I will not focus on one specific topic, but rather give an overview of the work done so far, open and resolved issues, as well as summarize next steps.

Dry run talks

Information structure and discourse particles

Yvonne Viesel (A6) [dry run]

Project: SFB 732, A6 "Encoding of Information Structure in German and French"
E-mail: yvonne.viesel@uni-konstanz.de
Website: http://ling.uni-konstanz.de/pages/home/viesel/index.html
Institute: Institut für Maschinelle Sprachverarbeitung

An extensive set of natural data on German discourse particles (DiPs), mainly 'ja' (lit. 'yes', roughly 'as we know'), shows that previous accounts do not capture the DiP's distribution, especially in non-root environments. I argue that the pragmatic function of 'ja' is not the same as its meaning, from which the function is derived as a non-cancelable conversational implicature. Similarly, DiPs like 'ja' interact with focus syntactically for pragmatic reasons: since 'ja' relates 'old information' to ongoing discourse, this information must enter the discourse tree in a coherent way and hence be information-structurally marked for addressing a specific (solved) issue. Eventually, DiPs like 'ja' are not confined to non-restrictive subsentential domains and can thus not be used to test for assertive types of (subordinate) clauses, but they always involve foci that need not coincide with the sentence foci and are therefore an indicator of layers of information structure in complex sentences.

Investigations into Semantic Parser Induction

Kyle Richardson (D2) [dry run]

Project: SFB 732, D2 "Combining Contextual Information Sources for Disambiguation in Parsing and Choice in Generation"
E-mail: kyle@ims.uni-stuttgart.de
Website: http://www.ims.uni-stuttgart.de/institut/mitarbeiter/kyle/
Institute: Institut für Maschinelle Sprachverarbeitung

I will give an overview of my thesis project, tentatively titled "Investigations into Data-driven Semantic Parser Induction". The main topic, semantic parser induction, relates to the problem of learning to map input text to full meaning representations from parallel datasets. Such resulting "semantic parsers" are often a core component in various downstream natural language understanding applications, including automated question-answering and generation systems. In my thesis we look at learning within several novel domains and datasets (e.g., software documentation) and under various types of data supervision (e.g., learning from logical forms, entailment, denotation).

Online Learning of Structured Predictors with Approximate Search and Latent Structure

Anders Björkelund (D8) [dry run]

Project: SFB 732, D8 "Data-driven Dependency Parsing – Context Factors in Dependency Classification"
E-mail: anders@ims.uni-stuttgart.de
Website: http://www.ims.uni-stuttgart.de/~anders
Institute: Institut für Maschinelle Sprachverarbeitung

Many NLP problems exhibit an inherent structure. The size of the output space for a given input instance is typically exponential in the input size, implying that pure enumeration of potential output structures is intractable. This search problem is typically solved either by (1) exact search algorithms (e.g., dynamic programming) that sacrifice expressivity in terms of features or (2) approximate search algorithms that allow for richer feature sets while sacrificing exactness. We study the second type and devise novel learning algorithms and evaluate the approach across several NLP tasks, sometimes allowing the target structure to be latent.

Constructing Syntax-Based Distributional Semantic Models for Novel Languages

Jason Utt (D10) [dry run]

Project: SFB 732, D10 "Incrementality in Compositional Distributional Semantics"
E-mail: jason.utt@ims.uni-stuttgart.de
Website: http://www.ims.uni-stuttgart.de/institut/mitarbeiter/uttjn/
Institute: Institut für Maschinelle Sprachverarbeitung

Distributional models have proven useful as a reliable source of unsupervised representations of word meaning having found use across numerous applications and fields. In the last decade, such models have been refined to include syntactic information such as grammatical relations between words. This leads to richer representations allowing for more fine-grained distinctions and analogies to be drawn, bringing distributional semantics closer to the goal of semantics in general: to model human lexical knowledge. Beyond simple text corpora, however, SDSMs require accurate syntactic analyses of the source corpus sentence. In addition, as the observed contexts contain more information, more data is required to make the counts for these syntactic cooccurrences reliable, calling for larger corpora. This reveals a problem faced in many NLP domains: the resource gradient. While reliable parsers are available today – and the further development of these is an active field of research – most languages currently are still lacking in this respect. This inequality of resource distribution affects the construction of SDSMs for potential target languages in two ways: languages can be lacking in both regards or they might have a good situation with respect to corpora but still have under-developed parsers. I will consider different approaches and discuss some important results of inducing SDSMs across different languages and levels of resource availability.

Indirectly Supervised Determination and Structural Analysis of Nominal Compounds

Patrick Ziering (D11) [dry run]

Project: SFB 732, D11 "A cross-lingual approach to the analysis of compound nouns"
E-mail: Patrick.Ziering@gmx.de
Website: http://www.ziering.de
Institute: Institut für Maschinelle Sprachverarbeitung

This presentation provides an overview of my PhD work, which focuses on the determination and analysis of nominal compounds. The compound formation process is highly productive and corpus studies (e.g., Baroni et al., 2002) observed that half of all word types in German texts are compounds. In this presentation, I will focus on the three main contributions of my PhD thesis. Firstly, despite their abundance, the definition (and even the existence) of compounds is controversially discussed in linguistics literature (Lieber and Stekauer, 2009). Focusing on the predominant category of nominal compounds, I inspected the relevance of various established linguistic criteria for compoundhood and provide new insights on the phenomenon. Secondly, for the analysis of compounds, I aimed to avoid methods that rely on direct supervision in terms of hand-labeled training data or lexical resources, and instead focused on indirect supervision by means of naturally-occurring cross-lingual evidence, the main reason being to produce language-independent, resource-lean applications. A cross-lingual corpus study on English compounds revealed the large space of surface variations across languages, which allows for cross-lingual supervision for compound analysis. I will present two analysis tasks that enjoy cross-lingual supervision: identification and parsing. Thirdly, I addressed the task of compound splitting. Here, I exploited a form of indirect supervision that relies on monolingual morphological regularities, thereby eschewing a limitation of cross-lingual supervision, i.e., the dependence on parallel data. In summary, the resulting methods achieve competitive results that are state-of-the-art within the scope of indirectly supervised methods. Moreover, the nature of the approaches, which are for the most part motivated by linguistic theorys, shed light on the complex phenomenon of compoundhood in cross-lingual as well as monolingual settings.