Language & Communication Technologies

Annual Meeting

2015

Groningen

Students' Session

Abstract Booklet

A real-time multi-modal visualisation of articulatory data

Kristy James

Speech researchers use a variety of techniques to gain an insight into vocal tract movements including ultrasound, electropalatography, magnetic resonance imaging (MRI), and X-ray microbeams, as well as optical tracking measures. However, many of these are unsuitable for repeated measurements from the same user due to radiation exposure, expense, or are unable to capture movement inside the mouth.

Recent advances in point tracking measurement techniques such as Electromagnetic Articulography (EMA) allow for the capture of tongue movement data with ease. In newer systems (NDI Wave, Carstens AG501) such data can be captured inexpensively and more comfortably for the user (who sits within a magnetic field) from sensors glued to the articulators. An additional benefit of this technique is that it can be combined with other imaging techniques such as ultrasound and optical tracking without affecting the measurement quality.

With increased processing speeds, in addition to the visualisation of articulator movement graphically in 2D (for example, VisArtico), real-time manipulation of these data points is possible, even allowing 3D avatars to be manipulated in real-time. Moreover, a growing body of work suggests that real-time visualisation feedback can facilitate the learning of the pronunciation of a foreign language or to help patients with hearing impairment.

This project presents a tool that aims to improve upon current visualisation software by utilising open source tools so that researchers can implement their own tongue-models and avatars. In addition, it aims to make use of multi-modal data inputs, allowing for the visualisation of the audio signal, ultrasound recordings and a video feed so that users can make use of the tools they feel most comfortable with, and experiment with various ways of displaying the goal movements.

Combining character N-gram and layout dissimilarity for authorship verification

Esther van den Berg

This paper describes a method of authorship verification developed for the PAN 2015 shared task. The shared task challenged participants to determine for pairs of documents in 4 different languages (English, Dutch, Greek and Spanish) whether they share the same author. We approached this problem by experimenting with a range of features, including character N-gram dissimilarity, entropy, POS-tag and dependency dissimilarity, basic features like sentence length and layout dissimilarity measures. The latter included determining visual aspects of documents like the distribution of sentences and punctuation marks across lines to detect differences across styles of poetry. We finally selected those features that were found most informative by feature ablation tests for a final feature set to be used in classification with a Support Vector Machine. Character N-grams and layout dissimilarity were revealed to be the most useful, which is in line with previous findings pointing to the limited added value of more complicated features. Our best combination of features performed well above the baseline for all four languages.

Combining metaphor comprehension models with knowledge mismatch

Betül Aksu

The fascinating process of how people comprehend figurative language is of extensive interest to researchers in the fields of Psycholinguistics and Computational Linguistics. Due to its observable yet vague deviation from literal meaning, figurative language findings has been suggesting quite complex and contradictory models to explain the general mechanisms behind figurative language processing. With regard to an important focus on metaphor in the recent issues of language science and language comprehension studies, I am conducting my thesis within the scope of metaphor studies, particularly on direct/indirect access of metaphor comprehension with a new perspective on knowledge mismatch. Typically, models of metaphor comprehension try to explain the conceptual mapping between the base and the target terms regarding the effect of context while ignoring the role of perspective taking. However, it is not clear whether taking other's perspective influences metaphor comprehension in any way. In my thesis project, I aim to answer whether the processing of a metaphor is different when the addressee is naïve/knowledgeable to the context of the metaphor use, or it is independent of other's knowledge. My stimuli consist of metaphorical expressions such as "The intern was a shark" with figuratively biased and neutral contexts. Direct access models predict that the processing of a shark would differ under those contexts while career of metaphor/gradient saliency models predict no difference between a shark in supportive and neutral contexts. If readers are sensitive to fictional character's knowledge state, processing a shark should be more difficult when the addressee is naïve in supportive context and, perhaps, there should be same processing difficulty when the addressee is naïve and knowledgeable in neutral contexts. The results of this psycholinguistic experiment will be analyzed within the scope of current models of metaphor comprehension, and are expected to bring insight to computational approach to metaphor studies since it is envisaged that insights from human language comprehension can be used to build systems that automatically recognize metaphors.

Mapping semantic representation to pixel space

Dat Tien Nguyen, Angeliki Lazaridou, Marco Baroni, Raffaella Bernardi

Semantic representation plays an important role in many NLP applications. Recently, high-quality semantic representations of linguistic/visual units can be effectively constructed thanks to advances in modern neural network methods. However, the understanding of these representations remains limited, so they need to be assessed in an intuitive way. The goal of this project is to build a complete pipeline, in which word representations are transformed into visual vectors using cross modal mapping and these vectors are projected to pixel space using inversion algorithms. This suggests that there might be a groundbreaking way to inspect and evaluate the semantics encoded in word representations by generating pictures that represent it.

An online development platform for computational grammars

Antonio F. García Sevilla

Modern science has seen the rise in prominence of group research projects and other many-person endeavours, in what has been called "Big Science". Computational linguistics is no exception to that, and especially the development of large linguistic resources is a task best suited for collaborative approaches.

The Internet has been a powerful medium for this exercise. E-mail, web pages, FTP servers, etc. were the tools that allowed collaboration over the web 1.0. Development still happened within the researcher's local environment, but synchronization was able to be done online. Data could be shared, opinions exchanged. With the advent of the web 2.0, even more powerful tools are available for collaboration. Servers are fast and affordable (the "cloud"), modern browsers are versatile. Web applications can nowadays perform most of the tasks required by their users, with no disadvantage compared to their desktop counterparts.

The first aim of this project is to take advantage of this new model of computation, and build an online platform for the linguistic development of computational grammars. The main goal of the platform will be to enable on-site collaboration, by allowing researchers to both share knowledge and data directly online, and to develop resources and computational models directly within the platform.

A second aim is to integrate an HPSG toolkit within the system. This toolkit will both demonstrate the potential of the platform for grammar development, and the practical improvements made possible by the technology to the underlying theory. A small HPSG grammar of Spanish will be written using this toolkit, and a data-centric application will also be integrated. For example, this might be an application where linguist experts can rate parse trees, and the information obtained can be compiled as weights for grammar rules in order to implement a probabilistic grammar.

Towards a Part-of-speech Tagger for Middle Low German

Mariya Koleva, Melissa Farasyn, Anne Breitbarth, Veronique Hoste

This work presents our continuing effort to create a POS-tagged and syntactically parsed corpus of Middle Low German (MLG). MLG is a group of dialects used between 1250 and 1600 in Northern Germany. As is the case with most historical language varieties, their written form is characterized by significant linguistic variation in orthography and grammar. This poses a major challenge to standard NLP tasks such as POS-tagging, as it exacerbates the inherent issue of data sparsity. Furthermore, even texts within the same dialect show differences, depending on the specific locality from which they originate, effectively creating different domains. As a starting point for tackling this issue, we manually tag a portion of legal texts, selected for our corpus, and we train models for POS-tagging for one of the chosen MLG varieties (Westphalian) –one for each chosen locality, and combinations thereof. The models are trained with a CRF learner, using a set of standard and corpus-specific custom features. Furthermore, preliminary experiments with feature selection with a genetic algorithm are presented.

Assessing the impact of manual corrections in the Groningen Meaning Bank

Benno Weck

Developing large-scale annotated corpora requires time and money, which are often limited. Nonetheless, many NLP tasks are in constant need for gold standards (e.g. training data for machine learning). The Groningen Meaning Bank (GMB) project (Basile et al., 2012) develops a corpus of English texts with rich syntactic and semantic annotations. Annotations in GMB are generated semi-automatically and stem from two sources:

1. Initial annotations from a set of NLP tools which use only their respective standard models.
2. Corrections/refinements by human annotators.

Annotators are either experts, who apply corrections in a wiki-like fashion, or non-experts, who help annotation by playing a game with a purpose called Wordrobe (Venhuizen et al., 2013). These corrections are called Bits of Wisdom (BOWs). At the moment there are more than 150,000 BOWs which improve tokenization, lexical and semantic annotation. Our main question is: How can the BOWs be used to effectively retrain the tools to eventually improve annotations on the entire GMB corpus? We hypothesize that (near) gold standard annotations can be acquired by applying an iterative bootstrapping approach to the annotations.

This will be investigated by addressing a number of sub-questions:

- How many mistakes does one BOW fix?
- What factors are important (type of mistake, source of correction, type of annotation corrected, additional (external) training data)?
- How can we ensure that correcting errors does not introduce new ones?
- How can we minimize supervision of the retraining process?

Our main focus is on the lexical annotations of the corpus. We start with POS annotations because they do not require any other layer of annotation and are thus not affected by mistakes made on other layers. Once a set of techniques shows promising results on the POS tagging we can evaluate that technique on named entity tagging.

Synonym Extraction from Unannotated Text

Artuur Leeuwenberg

Synonyms are used in many NLP tasks (e.g. machine translation or text mining). Often, WordNet is used as the resource to get these synonyms (synsets). Since the coverage of WordNet is strongly language dependent we try to extract synonyms from unannotated text, using distributional word vectors, such as word2vec (Mikolov, 2013). Since synonyms are distributionally quite similar to hypernyms, hyponyms and antonyms we try to find characteristic properties of synonyms that, ideally, can be extracted from unannotated text as well.

Query expansion for medical information retrieval

Feraena Bibyna

Unified Medical Language System (UMLS) is a set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems. Among the tools that contained in UMLS are metathesaurus and semantic network. Terms and phrases is assigned a concept id, which are a group of similar terms/phrases in the metathesaurus coming from several sources. This concept structures represents a synonymy relationship between terms/phrases. These can be ambigious, i.e. have more than one possible concept ids. Concepts are also connected by other relationships defined in each source vocabularies, mainly (but not limited to) is-a relationship. Every concept in the metathesaurus is assigned to at least one semantic type in the semantic network. There are previous works that found that using thesaurus or ontology for query expansion can improve precision of the retrieval. Lay people (i.e. non-medical experts) often used common terms in their queries when looking for health-related information. Medical documents, however, are written using medical terms that are unknown to non-medical-experts. CLEF eHealth Task 2 aims to evaluate the effectiveness of information retrieval systems when searching for health content on the web given this mismatch in terms between queries and documents. In this year task, provided queries aimed to mimic search behaviour of lay people that are confronted with a sign, symptom or condition and attempt to find out more about the condition they may have. These queries are often circumlocutory in nature. The aim of this thesis is to evaluate the effect of different use of semantic relations between terms gained from UMLS for query expansion to the precision of medical information retrieval in this task. For our purpose, we use Terrier as our retrieval tool with a few modification. We use queries from previous year's task to tune our retrieval model.

Converting AWA corpora into relANNIS with Pepper

Talvany Carlotto

AWA is a scheme for representing linguistic annotations at many levels, ranging from simple tokenization to syntactic and semantic analysis. The main goal of AWA is to facilitate communication between linguistic processors in a variety of NLP applications, and thus it can deal with issues such as linguistic ambiguity phenomena in a proper way. AWA is particularly well suited to deal with morphologically rich languages such as Basque, and there exists Basque corpora of considerable size (million of words) which are annotated following the AWA paradigm. Nowadays, the interface to access AWA is considerably limited, and a better way to access the corpora in AWA in required. ANNIS is a different annotation scheme that comes with a powerful web interface, with a good representation of features such as complex regular expressions, overlapping and span. However, relANNIS, the format used by the ANNIS interface, is not suitable to represent complex phenomena present in Basque in the same way that AWA is. The ANNIS project recommends the use of the Pepper framework to convert documents of different formats into relANNIS. As a matter of fact, Pepper allows conversions of data between various formats, and that is carried out by the conversion from the source format into a format called Salt. All the document formats available in Pepper can be converted into Salt documents, as well as Salt documents can be converted into documents of any of the available formats. Pepper provides tools for the development of new conversion modules for new formats not covered by the standard library. The project presented here consists in the development of a new module in Pepper to allow conversions from AWA documents into relANNIS, with the ultimate goal of using the ANNIS interface to interact with AWA documents. In the current phase of the project, it is already possible to convert AWA documents and visualise them in the ANNIS interface.

Generating stories from different timelines: a statistical approach

Anastasia Shimorina

Traditionally, Natural Language Generation (NLG) systems generate texts from non-linguistic data. A standard NLG system consists of three components: Document Planning, Microplanning, and Surface Realisation. In most conventional NLG systems those stages are executed in a sequence, so this could lead to errors propagating along the pipeline. Since decisions taken upstream inevitably affect those taken downstream in a pipeline architecture, it is quite possible for early decisions to result in problems later on. By taking a statistical approach, one tends to blur the boundaries between the modules of a traditional NLG architecture. The objective of the master thesis is to find statistically significant language patterns and make use of them in generating new texts. This corpus-driven approach can be a part of a template-based NLG system which uses a statistical learning process.

The work aims at exploring the temporal relations in a narrative. Time expressions and grammatical categories of verb, namely tense and aspect, are under scrutiny. To investigate the narrative temporal structure, a specially designed corpus of stories is used. It is made of 161 stories written by native English speakers; for each story, the main events and their chronological and discourse-level orders are known. This corpus allows us to identify common temporal patterns for specific orderings of events at the discourse level. The most probable sequences of tenses are found for each event ordering. The Conditional Random Fields method is applied to predict the output tense sequence. The best tense sequence for a particular event order is considered as a model for the NLG system. This system takes a story from the corpus and an intended event order as input and re-generates the story by putting the events in the prescribed order. Generated stories are evaluated by human subjects.

A Coarticulation Model for Articulatory Speech Synthesis

Anastasiia Tsukanova

The state-of-the-art techniques for speech synthesis rely on concatenation of acoustic units taken from a vast pre-recorded speech database. This approach yields synthesis of good quality, but is a purely technical solution which brings no or very little information about the acoustics of speech or about how the articulators (mandible, tongue, lips, velum...) are controlled.

In contrast, the biomechanical approach generates the speech signal from the vocal tract shape and its modelled acoustic phenomena. The vocal tract deformation control comprises slow anticipation of the main constriction and fast and imperatively accurate aiming for consonants.

The system predicts the sequence of articulatory gestures from a sequence of phonemes to be articulated and a model of the coarticulation effects in it. We use Magnetic Resonance Imaging (MRI) captures of the vocal tract shape when producing phonemes in various contexts, thus following an approach by Birkholz (2013). The evaluation of the model is done by comparing the synthesis results with the available dynamic data and by running acoustic simulation on the system-generated vocal tract shape evolution so as to verify the quality of the synthesized speech signal. The speech signals are to be compared perceptively and computationally, by the formant frequencies produced.

The work is in progress, but so far it seems feasible. Some problems with the captured MRI images have arisen. When, apparently, the speaker produced a wrong gesture for a phoneme, or a technical problem occurred, we correct the shape manually or disregard the image completely. It has also turned out to be necessary to experiment with modelling the articulators which bear an acoustically important function in producing speech.

Learning to Rank CVs for Jobs

Min Fang

We investigate the challenging problem of matching semi-structured CVs to job ads where each ad is represented by an automatically generated search query that is issued to a search engine. Our main goal is to improve the initial ranking of candidates as returned by the search engine within a learning to rank framework. For this purpose we experiment with a real-world dataset obtained from a recruiting agency in the Netherlands, which contains a large number of job offers from various companies and corresponding applicants' CVs as well as hiring decisions made by recruiters. This initially unstructured dataset is first passed through proprietary machine learning software which parses the unstructured CVs into semi-structured documents to make them searchable and transforms the job ads into search queries based on heuristic rules and gazetteer-matching. Our work follows these preprocessing steps and focuses on extracting features from the parsed CVs and the queries that can be used as signals in learning to rank algorithms. While most work in the learning to rank literature investigates different ranking algorithms, little to no research has been done on features, especially when the domain characteristics diverge from web search as in our case. We are interested in investigating whether and how it is possible to leverage domain knowledge and insights from traditional information retrieval research, especially from work on language modelling for IR and retrieval models for semi-structured data, to construct features that can be used in a learning setting in meaningful ways. We will evaluate our work using the set of hiring decisions as well a smaller set of human judgements (fewer queries but on average larger pool of documents).

From books to characters: named entity recognition and relation extraction in novels

Aitor M. Egurzegi, Ionut T. Sorodoc, Quan P. Ngoc

This project is an ongoing project developed by the first year HLST students in the University of Malta. The task of relation extraction has been studied or subject of much research during the last decade. However, comparatively little effort has been put into the case of relation extraction within novels of fiction. The development of this aspect of relation extraction could make a valuable contribution toward the development of a reading aid that could help summarize the main elements within such work. Our project takes the first step towards the development of such an aid which is based on the identification of main characters in such works and the relationships between them.

The main issues we are currently facing are the following: the detection of named entities within a book, the classification of those entities and the identification of co-reference chains. Secondly, the selection of promising pairs of entities, between which a relationship is likely to exist. Thirdly, the deployment of computational techniques which are capable of identifying that relation. It is important that, given the limited time constraints that we have in order to carry out this project, we are favoring precision over recall.

The final consideration concerns the problem of presenting the results to the reader in such a way that it will help them improve and facilitate the understanding of the novel. They will be carried out by designing a suitable interface that will present the results to the reader.

Cross-Lingual Pronoun Prediction

Ngoc Quan Pham, Ionut-Teodor Sorodoc, Lonneke van der Plas

In this poster we would like to present the system the UoM-LCT team built for the DiscoMT shared task on cross-lingual pronoun prediction. This task is concerned with finding the right French translations (mostly pronouns) for the English third-person subject pronouns 'it' and 'they'. The task is set up in such a way that the system needs to choose between 9 classes of French pronouns: ce, elle, elles, il, ils, ça, cela, on, and OTHER in bitexts in which the pronoun aligned to the English pronouns 'it' and 'they' are substituted by placeholders. It is a hard task, because the French translations we find for these English pronouns vary a lot and several factors are at play: number and gender of the antecedents of the French pronoun, abstractness vs concreteness of the antecedents, level of definiteness of antecedents, grammatical relation of the pronoun etc. These types of information are shown to be captured by word embedding in previous research.

Therefore, the core of our system is inspired by the Probablistic Neural Network language model (Bengio et al, 2003), in which the predicted word is learnt from the distributed representation of the previous words. Similarly, we set up a network which predicts the desired French pronouns based on the English and French words projected to continuous spaces. The training progress is alleviated by training the word embedding models and the neural network models separately. This setup outperforms the informed baseline predictions based on the language model KenLM (48.3% F-score on the development data), by reaching 60.0%. In addition, we incorporate morphological information for French words provided by an automatic analyser, hence increasing the final F-score to 65.5%.

Semantic relation extraction from unstructured data in the business domain

Ilana Rampula

Text Analytics in the business domain is a growing field of research. A significant part of databases of large companies is in the form of raw unstructured text (e.g. clients' emails) often linked to some structured data (e.g. clients' profiles). Nevertheless, this information is rarely utilized. In order to take advantage of the textual information, it has to be first converted to a structured form. One way to accomplish this is by extracting named entities in the text and relations between them. The extracted relations could be then used in predictive models and play a role in the process of decision-making.

The goal of the thesis is to apply methods for semantic relation extraction to enrich available structured data with relations extracted from unstructured textual data. The work includes comparison of several methods for relation extraction, such as semi-supervised Snowball and Distant Supervision and their modification to fit the use-case (e.g. by utilization the existing structured data). The work applied on a set of Czech texts from a business domain provided by an industrial partner and evaluated on manually annotated sample of the data.

Does What improve on Where? Using Lexical Knowledge to Predict Spatial Relations in Images

Manuela Hürlimann

Both Natural Language Processing and Computer Vision have advanced considerably over the past decades. This has paved the way for research linking the two fields, such as the automatic generation of labels for images, or the translation of text into visual scenes. One interesting task with respect to model-theoretic semantics and visual input is to automatically derive a semantic model from an image. Potential applications include image search, generation of image labels, and Question Answering about images.

In order to achieve this goal one needs to identify both the objects present in the image and the relations which hold between them. Our focus is on predicting spatial relations between objects ("part of", "touches", "supports"). Specifically, we propose to answer the following research questions:

1. What spatial relations are suitable for automatic prediction, and what formal properties should these relations have?
2. To what extent is simple spatial information (e.g. coordinates of objects) useful for predicting spatial relations between objects in images?
3. In addition to simple spatial information, do we need linguistic/world knowledge?
4. If so, what kind of world knowledge (e.g. knowledge about properties of objects) is useful?

In the absence of broad-coverage object recognition systems, we manually annotated objects in a hundred public domain images. The resulting resource will also be discussed. Using a supervised approach, we explore the impact of linguistic and world knowledge on prediction performance, i.e., we compare two types of classifiers:

i. One which only receives the coordinates of the bounding boxes as input
ii. One which is additionally provided with world knowledge (e.g. about object properties)

We expect that classifiers of the second type (ii) perform better than those of the first type (i), i.e., that world knowledge and lexical information are crucial to the task of adequately predicting spatial relations in images.

Preliminary results show that incorporating meronymy information from WordNet greatly improves performance for the part of relation: using only simple spatial information, recall for this relation is between 0 and 15%, while adding meronymy information boosts this number to between 70 and 95%. This suggests that lexical information can indeed enhance prediction of spatial relations in images.

QGP-Sarsa: a combination of Q-learning and GP-Sarsa

Thanh Le, Filip Jurčíček

Gaussian processes (GP) have shown its power in many learning tasks ranging from supervised to reinforcement learning. Optimising policy online for adaptable systems has been significantly improved by GP-Sarsa algorithm. However, the policy learning process in GP-Sarsa is a random process which strictly depends on randomly selected actions. Another inevitable characteristic of learning online is that the system poorly behaves at the early of learning. In this paper we propose combining Q-learning and GP-Sarsa to overcome these problems. We evaluate the proposed method (QGP-Sarsa) by comparing to GP-Sarsa on toy problem simulating a simple spoken dialogue system. The results suggest that the QGP-Sarsa algorithm is more stable during learning and converges faster.

Speaker recognition using neural networks

Bich Ngoc Do

Recently, mainly due to the use of computational power of GPU (Graphical Processing Units – CUDA), it was shown that artificial neural networks provides one of the best results in many machine learning tasks. This topic aims to explore the use of neural networks in speaker identification. We have chosen recurrent neural network as our model because of its ability to infer time series and to capture non linear relations within hidden layers. The evaluation will be performed on the standard 2004 NIST Speaker Recognition Evaluation tasks.

An Approach to Context-Relevant Off-Activity Talk Suggestion in a Conversational Quiz-Game Setting

Omid Moradiannasab

Conversational agents (e.g. virtual characters, chat-bots) are software programs that interact with users employing natural language processing capabilities. Traditional conversational systems are designed for specific purposes. Dialog in such systems is limited to task-bound talks. These talks have specific purpose and follow particular structures. Relational agents on the other hand, are defined in literature as "computational artifacts designed to establish and maintain long-term social-emotional relationships with their users" (Bickmore and Picard, 2005). In order to achieve such relationships, certain amount of user trust and engagement is required. Various conversational strategies are employed in Relational agents that comprise models of social dialogues with the aim to raise user trust and solidarity. As an instance of a conversational strategy employment, small talk (also social talk)

is discussed in literature. It is introduced as a kind of talk which executes such conversational strategies and, while interleaved between task-bound talks, indirectly builds trust through the natural progression of a conversation. (Bickmore and Cassell, 2001) define small talk as "any talk in which interpersonal goals are emphasized and task goals are either non-existent or de-emphasized". Small talk is useful to develop the conversation and to avoid pauses. It can be used to ease the situation and to make a user feel more comfortable in a conversation with an agent (Cassell and Bickmore, 2000).

In this work, we focus on off-activity talk (also called non-activity talk) which was first defined by (Kruijff-Korbayova et al., 2014) in resemblance to small talks. Both small talk and OAT enrich the task-oriented dialogue via opening the structure of the conversation. However, it can be differentiated from small talk by the topic and the purpose of the talk. OAT in this definition has a specific purpose (e.g. knowledge exchange) and is about a specific topic while a small talk is an independent talk without any functional topics.

Current project is an exploration toward verifying the hypothesis that providing talks by the agent out of the main activity (i.e. off-activity talk), leads to building up engagement of the users. Within this wider vision, the concrete objective is to propose a tool which automatically suggests off-activity talks in form of some sentences relevant to the dialogue context, to be used in conversational quiz-game settings.

This idea of context determination and off-activity talk suggestion for dialogue contribution, in a broad sense, can be used in any task oriented dialogue settings. However, due to its predominantly verbal character and naturally constrained interaction structure, a conversational quiz-game setting is chosen as a good test bed for current project.

LCT master Annual Meeting 2015
Student session