Language & Communication Technologies

Annual Meeting 2016

San Sebastián

Abstract Booklet

Student's Session

Table of Contents

A Tool for automatic analysis of linguistic clues in dialogue transcripts

Olga Leticevscaia, Ruixue Liu and Yiqing Liang

Our project is involved in the research project named SLAM, which aims at systematizing the study of pathological dialogues: dialogue with patients of schizophrenia. The main topic of it is the study of dialog interaction, which includes several levels of linguistic annotation such as disfluence tagging and part of speech (POS) tagging. The current research object is to develop a tool that integrates several other tools for annotation, to automatically identify pathological linguistic clues in transcriptions. Two stages are involved in the development of the tool. The first stage is to re-implement the SLAMtk tools in a generic manner. Developed with the Python programming language, the tool works on different corpora with specific characteristics. In order to build an XML structure to represent the generic version of different corpora, we perform a series of pre-treatments to normalize them. With the XML structure, all the necessary information needed for annotation and data extraction are stored, including annotation from Distagger for disfluencies (euh, self-correction, repetition of words, etc.) and morpho-syntactic segmentation (part-of-speech) from Melt. The second stage refers to the analysis of linguistic features and identification of clues. Several treatments are performed to create representations of these interactions and mathematics tests are applied to identify specific clues.

Compiling a Germanic cognate dictionary by automated cognate recognition

Martin Kroon

This research focuses on the automatic compilation of a Germanic cognate database. This is done by aligning the parallel Europarl corpus (Koehn, 2005) on word level using the fast-align tool from the cdec framework (Dyer, Lopez, Ganitkevich et al., 2010), and then by checking if the aligned words are in fact cognates. Before alignment, all words in the corpus are lemmatized and compounds split to improve results. Checking for cognacy is done using orthographic, phonological, and syntactic information, and will be based on (multi-level) analogical rules found using (an adaptation of) the Minimal Generalization Learner (Albright and Hayes, 2006). All words found to be cognates will be saved to the database and then be used to extract more analogy rules and find more cognates. This research also investigates the value of cognate prediction, where it is predicted what cognates would look like using the analogy rules found earlier. Attempts will then be made to find this predicted form in the corpus. The resulting database will contain English, Danish, Dutch, German and Swedish cognates (which are the five Germanic languages in the Europarl corpus), and will be comprehensive and searchable.

Cooking recipes

Laura-Ana-Maria Bostan

Recipes are an interesting specific genre of how-to instructions as they let us discover insights into culinary and cultural preferences. Most of the work studying recipes relies on simple ingredient bag-of-word representations, which most of the times fail to capture much of the recipes' internal structure. Little research has been done with a focus on semantic representations of cooking recipes with an eye toward more complex and deep understanding. I want to use natural language processing techniques that have been recently used in computational semantics to interpret cooking instructions. My goal is to explore the potential for computational creativity in gastronomy by developing a data-driven recipe narrative generation system. Briefly, the aim of the proposed system is to fill the generated recipe with ingredients, which have semantic information associated with them. With the ingredients in place, I can then generate interacting events for each ingredient, conditioning on the ingredient semantics. For example, I can try generating ingredients corresponding to potential type (obtained by clusters discovered in the data) like: "spice", "baking powder", and "fat", and then generate events appropriate for each type. The main idea is that we will try to discover these ingredients clusters in a data-driven fashion: instead of hard-coding what kinds of ingredients there are, I will attempt to discover ingredient semantics from corpora of existing recipes. This will reduce the amount of expert knowledge required to build an ingredient-driven recipe narrative system. Instead of artificially imposing ingredient classifications, I will try to see how well our semantics can predict existing classes. Once I have our ingredients, I will need events for them. I propose using a language model using distributionally learnt representations of events to generate events for our recipes. I predict that the guiding framework of the ingredient semantics will constrain the event generator to produce a recipe that is both fluent and coherent. This project requires a suitable corpus. The ideal corpus would be a collection of cleaned recipes with a few central strong recipes. To evaluate the quality of the automatically generated recipes, I plan to conduct a blind taste test like, using 10 generated recipes against 10 human designed recipes, I could ask some testers to rank the dishes and also to tell me which one they think is generated by computer and which one is human designed. Apart from subjective evaluating the system I plan to use some information theory measures to check the quality of the language model, e.g.: perplexity.

Copying syntactic patterns on Reddit

Talvany Carlotto

When people interact with each other they adapt their verbal and non-verbal behavior to that of their partner. A particularly interesting example of such adaptations is

syntactic copying. Copying other people's syntactic rules is not thought to be under conscious control, and hence analyzing this phenomenon helps us understand different aspects of human language processing, such as how do we adapt to each other and how production is related to comprehension. Additionally, understanding how we use social copying, both in oral and in virtual settings, is of crucial importance to build computational systems aiming at simulating human conversation. Previous experiments on syntactic copying have not analyzed conversations on internet forums, which are a popular form of modern communication and thus might give us valuable input regarding human interaction. This project studies the level of syntactic copy in conversations between two people in the social network Reddit. The huge amount of available reddit content allows us to find very specific forms of computer interaction, so we can obtain conversations that are very close to natural human dialogue. The level of syntactic copying in this project is measured using Healey et al's (2014) method of comparing subtrees of syntactic structures. The method was applied to the conversations extracted from reddit and to conversations artificially created with random turns from the extracted conversations. Results from both databases are then compared to analyze whether the level of syntactic copying found in the conversations is greater than what would be expected by chance. At the current stage of the project, these comparisons were not yet made, so we cannot yet draw any conclusions.

Determining Confidence Measures on Fundamental Frequency Estimations

Boyuan Deng

Fundamental frequency estimation acts as the basis of many applications related to speech or music, such as speaker recognition, development of speech synthesis voices and music information retrieval. Various algorithms working in time or frequency domain have been developed in the past and work quite well on good-quality signals. They become more error prone on lower-quality signals but few of them indicate how reliable the estimation results are, which is a crucial metric if the overall system makes decisions based on probabilities or multiple estimation algorithms are combined using ensemble methods. In our work, we tackle this problem by developing and evaluating confidence measures based on machine learning approaches for several existing fundamental frequency estimation algorithms. Such confidence measure can be seen as the probability of the estimation result to be correct. Sophisticated neural network architectures, including long short-term memory networks (LSTM), are chosen for the construction of these measures. And numerous features computed on the signals and generated during estimations are used as input. Besides confidence measures, our work also illustrates some

empirical analyses on the robustness of existing fundamental frequency estimation algorithms. The audio corpus is artificially distorted in a quantitatively controlled way, by modifying signal levels or adding certain types of noises at given signal-to-noise levels, etc. Changes in error rates as well as the different error types (deviation or resulted from voiced/unvoiced decisions) are visualized and the underlying causes are analyzed in detail. The conclusions drawn may help guide the choice of fundamental frequency estimation algorithms depending on situations and needs.

Evaluating Automatic Keyword Extraction Systems for Internet Reviews

Alice Leung

Keyword extraction and topic indexing are long-explored tasks in NLP circles. The task consists of extracting a set of terms or phrases that sum up the content of a piece of text. Professional indexers or the text's authors themselves are usually the individuals responsible for the manual task of keyword extraction; however, many studies in the literature are exploring the idea of automatic keyword extraction, because the process of manual keyword extraction is time-consuming and arduous, and the volume of content to be handled is only getting larger. While this task is often used to help with information retrieval, text mining, and summarization in text corpora, it has taken a slightly different direction when it comes to textual content on the internet. A form of keyword extraction called tagging has been used often for internet content, like blogs, review sites like TripAdvisor, and news articles. Unlike traditional keyword extraction, tagging of internet content serves purposes such as marking content based on users' interests, content discovery or recommendation of new content, and search engine optimization. Furthermore, the tagging process is not done by professional indexers, but by the users and consumers of the content. RealSelf provides a platform for consumers to write reviews about plastic surgeons and cosmetic treatments, as well as a forum for new consumers to gain knowledge about various cosmetic procedures. There is a large volume of textual content on RealSelf, the majority of which is community-generated, and is tagged manually by moderators of RealSelf. However, with the various automatic keyword extraction systems available now, we explore the feasibility and accuracy of automating the tagging of content. Three different automatic keyword extraction systems are compared: (1) IBM Watson's Alchemy Language, a commercial API whose algorithms are not publicly known; (2) RAKE, a simple yet powerful unsupervised keyword extraction method based on stopword removal and term frequency; (3) and Maui, a supervised keyword extraction method using novel features and a Naïve Bayes algorithm. Preliminary results show that Maui performs the best when evaluated with ten-

fold cross-validation, while the other two systems achieve low precision but relatively high recall.

Language Model Contextualization for Automatic Speech Recognition by Dynamic Adjustment

Anna Currey

Out-of-vocabulary words can pose a particular problem for automatic speech recognition of broadcast news. The language models of ASR systems are typically trained on static corpora, whereas new words are continually introduced in the media. Additionally, such OOVs are often content-rich proper nouns that are vital to understanding the topic. In this work, we explore methods for dynamically updating a language model by adding topic-relevant OOVs to the model using direct estimation. We concentrate on adaptation of the bigram LM used in the first pass of our ASR system. We propose two strategies for direct estimation of n-gram parameters. The first relies on finding IV words similar to the OOVs; OOV behavior is modeled after the behavior of these similar IVs. We use word embeddings to define similarity and examine various word vector training algorithms, finding skip-gram with a context size of two words to be ideal for our application. Our second strategy leverages a small contemporary corpus to estimate OOV unigram probabilities and to find bigrams containing the OOVs. We find that it is best to limit the number of bigrams added to the language model; this allows us to increase the coverage of the model without creating a significant amount of noise. We use two experimental setups to evaluate our proposed methods. In experiment 1, we create only one adapted LM for each algorithm; these LMs are then incorporated into our existing ASR system and recognition error rates are calculated. In experiment 2, we create a separate LM for each article and for each algorithm; the LMs are evaluated using perplexity. In both experiments, our adapted models improve over the baseline; improvements are greatest in experiment 2. The corpus-based adaptation method outperforms the method based on word similarity in all cases.

Language Models with Convolutional Neural Networks

Ngoc Quan Pham

This work investigates a convolutional neural network architecture used for language modeling, which shows the likelihood that a word string makes sense by predicting words given the context. The original feed-forward language model concept is enhanced with a temporal convolutional layer, which acts as feature detectors for word prediction. The convolutional kernels are shown to be able to represent distinctive patterns in the data with respect to the position in the context. We conducted experiments on various corpora with various vocabulary sizes and observe that, not only does the proposed model improve the original baseline by a

considerate amount without adding many parameters, it also performs on par with the state-of-the-art models such as Long-Short Term Memory (LSTM) networks, or Memory Networks. We also analyzed the behaviour of the network with respect to the attention to each position in the context.

Learning to quantify from images

Ionut-Teodor Sorodoc

This work presents an investigation of an important class of function words : natural language quantifiers(e. g. some, all, no, few, most). We consider that we can learn the meaning of quantifiers by analyzing their usage in visual context. We observed that in grounded contexts, children learn to make quantification estimates before being able to count, using their Approximate Number Sense. Following this observation, we evaluate several neural network models, with and without counting abilities, on to the task of assigning a non-cardinal(e. g. some, all) to a referent in a grounded scenario. We create an experimental setup in which, given a scenario with a set of images which contain a variety of objects(dogs, cats, cakes, etc.) with different properties(furry, black, white) and a query like "Dogs black", the model learns to apply the correct quantifier to the situation. The main model implemented for this task is a memory network model which we show that it can learn the scope of quantifier words such as all/most/some/few/no, given some visually grounded training data. We show that, as observed in children, our best model doesn't need to be able to count in order to quantify.

Modelling Noise in a Neural Network approach to Named Entity Recognition

Esther van den Berg

The performance of even very good classifiers is harmed when the labels used for training are unreliable. In fact, noisy labels have been shown to be more harmful than noisy features. Previous research has found one particular solution, namely to estimate a noise distribution using the EM algorithm and to use estimations for true labels to improve performance of a deep learning approach, which is then called a NLNN (Noisy Label Neural Network). The focus of much machine learning research is currently on non-supervised learning. Non-supervised methods have the advantage of being useable for tasks in low-resource language for which human annotators are scarce, but they also tend to generate noisier data. Modelling a noise distribution for automatically generated data and training an NLNN to better classify new data would be a considerable step forward for non-supervised machine learning. However, it is not known whether the proposed EM-based method works for more complex linguistic tasks. The aim of my thesis is therefor to apply the proposed method first to a moderately complex task, text classification of the classic Reuters-21578 collection, and then to a more complex task, Named Entity Recognition (NER) for

Hausa and Turkish. The first step has been to replicate a NLNN on handwriting recognition data. During the LCT conference I intend to report the results of an application to text classification in English, and give an update on the plans for application to NER.

Predicting Lexical Differences with Distributional Space Models

Alicia Krebs

The distributional hypothesis states that words that occur in similar contexts tend to carry similar meanings. Following that hypothesis, distributional semantics have been concerned with building models that can accurately represent word meaning through the collection of occurrence counts. In a distributional space model, each word is represented by a distributional vector which consists of the occurrence counts of every context of that word. These models have been used to solve linguistic tasks such as word similarity or lexical ambiguity, and to pass linguistic tests like the TOEFL synonymy and analogy tests. The cosine of two vectors is a standard measure used to assess the semantic similarity of two words. But the nature of semantic differences has yet to be explored in distributional models. Given two words, what is the difference between their meanings? Modeling such differences can not only help capture individual aspects of word meaning, but also help us evaluate the extent to which distributional models can encode such meanings. My work consists of building a system that can use word embeddings to predict differences between two related words. For example, given the vectors for [apple] and [banana], the system should predict [round] and [red]. Practical applications of this system include conversational agents (where choosing lexical items with contextually relevant features can help create more human-like dialogues) and machine translation (where taking into account semantic differences between translation variants can improve the quality of the output).

Splitting word compounds

Jonathan Oberländer

Languages such as German, Swedish, or Hungarian can have arbitrarily long compounds which pose problems for Machine Translation, as well as many other NLP tasks (e.g. Information Retrieval, Distributional Semantics, Speech Recognition). In many cases it can be helpful to split compounds into smaller parts ("decompounding") before further processing. Decompounding systems have been built for specific languages (mostly for German), examples of existing systems are JWord Splitter, Banana Splitter, or the ASV Toolbox. In this work, we will build a mostly decompounding tool, based on both previous work on the subject and new approaches, and compare its performance with that of existing systems. The goal is to make it both language-agnostic and un- or semisupervised, so it can be adapted to many languages. One of the novel features

we will try to integrate is a distributional model based on the assumptions that compounds are semantically similar to their parts, and the parts are often somewhat similar to eachother. As a dataset, both for training and testing purposes, we will use a medical corpus (both free text and dictionaries) that is available in several languages. Besides being a big available resource, the medical domain is also specifically suited for the task because it contains many compound words.

The role of pitch contour in the categorization of emotional speech

Aitor Egurtzegi

In the absence of emotion-conveying semantic information, low-level perceptual features associated with emotional prosody are assumed to be critical for categorizing the emotion conveyed by a speaker (Murray & Arnott 1993). A number of such emotional-prosodic features have been identified, including pitch, intensity, voice quality, fundamental frequency (Huttar 1968; Schrerer & Scherer 1981; Murray & Arnott 1993). There is currently no consensus as to which of these many proposed features are both necessary and sufficient for comprehenders to be able to identify the emotion conveyed in spoken language. However, several studies have argued that pitch contour is the most important feature (Vroomen, Collier & Mozziconacci, 1993; Schröder, 1999; Mozziconacci 2002; Rodero, 2011). A large body of research argues that "core" human emotions can be categorized along two orthogonal dimensions, valence (i.e., negative to positive) and arousal (i.e., low to high), which can be depicted on a 2D coordinate map (Cowie et al., 2000; Plutchik 1980, 1997, 2001). We hypothesize that pitch contour can be decomposed into two analogous orthogonal dimensions: (1) pitch range (i.e., the difference between maximum and minimum pitch within an utterance) may correspond to emotional arousal, and (2) the frequency of large pitch components (i.e., the number of large peaks and valleys within the pitch trace of an utterance) may correspond to emotional valence. We will refer to this proposed mapping between emotion dimensions and pitch contour dimensions as the pitch contour hypothesis. The proposed research will test the pitch contour hypothesis in three perception experiments. Separate experiments will investigate the influence of pitch range, component frequency, and their interaction. In each experiment, participants will listen to recordings of semantically-neutral sentences (e.g., Everyday I eat an omelet for breakfast) in five emotional-prosody conditions (neutral, angry, happy, sad, serene), plus some artificially created conditions. These conditions will be constructed by digitally manipulating one or both of the pitch contour dimensions in the neutral prosody condition to be equivalent to those in each of the four remaining emotional-prosody conditions. If the proposed mapping between emotion dimensions and pitch contour dimensions

is accurate, then the artificial conditions should be perceived as natural emotions.

Universal Dependencies for Russian: The distinction of ccomp and xcomp

Elena Badmaeva

Parsing has been considered to be one of the most challenging issues in the processing of natural languages. Fortunately, due to the rapid development of the various treebanks in the last years, it has achieved significant accomplishment. However, different languages alter in their annotation scheme because each of them rely on the different grammatical traditions. These discrepancies cause severe obstacles when comparing the outcome of parsing across languages (Nivre et al., 2007). One of the latest attempts to unify the cross-linguistic framework under the single annotation scheme was the Universal Dependencies project (McDonald et al., 2013). This initiative represents syntax with the basis on word-to-word dependencies, i.e. Dependency Grammar developed by Lucièn Tesnière (Tesnière, 1959). Since this theory does not require a strict word order, it has been traditionally used for Slavic languages which have high degree of word order variation. While converting the different language treebanks to the same set of dependencies, the process has been posing some problems in several constructions. Russian language joined the project in the last release issued on May 15, 2016 in two variants. The first one is the converted to the UD format manually annotated treebank SynTagRus and another one is an automatically labeled treebank carried out by the Gold Parser Standard. One of the difficulties appeared while conversion is the distinction between a clausal complement ccomp and an open clausal complement xcomp. A clausal complement is a dependent clause which serves as core argument and functions like an object of the verb or adjective. But for those cases when the subject of the clausal complement is under control, the relation is xcomp. This subtle characteristic is computationally not straightforward. Our main focus is on the investigation of the methods which can make this distinction feasible and accurate.