PROGRAM BOOKLET

STUDENT POSTER SESSION

JUNE 20-21, 2019

Table of contents

Student Abstracts

Akshay Aggarwal (BASQUE COUNTRY)

With the exploding size and number of treebanks being created/generated, it is not possible for human annotators to manually verify and/or annotate the data. In the UD Project, covering 83 languages in 146 dependency treebanks, it's even more essential to verify the data and correct them, preferably automatically. This is important because inconsistencies affect the quality of parsers/taggers trained on the data. The current thesis attempts to study some of the most commonly pointed out inconsistencies in the UD data, and correct them. The majority of the work was done on UDv2.3, with a limited portion on UDv2.4.

[scroll back to beginning]

Thalita Anthonio (BASQUE COUNTRY)

Hyperpartisan news is a type of news characterized by extremely one-sided content from a left-wing or right-wing political perspective. This thesis is concerned with automatically detecting such hyperpartisan news through supervised text classification. We work with data from the recent shared task on hyperpartisan news detection (SemEval 2019 task 4). We experiment with two classification techniques: support vector machine classifiers (SVM) and neural networks. In the former, we build document representations using bag-of-words/characters, bag-of-clusters and word embeddings (trained with Word2vec, GloVe or FastText). We try to improve these classifiers by adding local features, such as POS n-grams, stylistic features and the sentiment of a text. We use recurrent neural networks in our second approach, in which we work with contextual character-based embeddings (Flair). We are able to receive accuracy scores that are close to the winning system of the shared task, but fail to obtain similar results with recurrent neural networks.

Ronald Cardenas Acosta (MALTA)

The development of basic language tools for extremely low-resource languages calls for approaches not dependant on annotated corpora and modest on their raw-text requirements. Within the approaches to morphological analysis, neural transducers (Aharoni and Goldberg, 2017; Makarov and Clematide, 2018) have shown promis-ing results for the tasks of lemmatization and reinflection. In these architectures,transduction consumes one character at a time and applies operations inspired by the edit-distance algorithm. In this work, we propose an architecture that transduces the whole word form in every step, and introduce a merging strategy inspired by Byte–Pair–Encoding that reduces the space of valid operations by merging frequent adjacent operations. In addition, we propose a pipelined approach for the tasks of lemmatization and tagging that leverages annotations from several high-resourced languages at the same time. We find that the inferred operations are interpretable, resemble word for-mation processes, and capture important associations with fine-grained morpho-syntax labels.

[scroll back to beginning]

Natallia Chaiko (BASQUE COUNTRY)

Emotion Recognition from Speech using ML algorithms

Automatic emotion detection is currently among the most popular research directions in the field of computer science. A lot of applications have been developed to be able to recognize human emotions. These applications can serve as separate systems, for example used in call centres for distinguishing emotional calls from customers, or as parts of larger systems such as personal assistants, information providers etc.

Our goal is to compare existing algorithms recognizing emotions and propose a new one or improve some aspect of an existing one. All the existing solutions were developed using different algorithms and tested on different data, so it is impossible to compare them as they are. In our project we use the same data for all the algorithms to find the best-performing one, and at the same time by solving all the issues while installing and running the solutions, we provide step by step instructions for those who would like to use an emotion detector in their projects.

María Andrea Cruz Blandón (BASQUE COUNTRY)

The Navarro-Lapurdian dialect is a Basque dialect spoken in the French side of the Basque country. This dialect differs from the standard Basque in terms of its phonology, as well as at the grammatical and lexical levels. Additionally, passages in this dialect are code-switched texts with French. TTS systems for this dialect need to handle both Navarro-Lapurdian and French phonemes repertoire. Inaccurate processing of the French words can result in using the Basque phonology to transcribe them or even in a wrong verbalisation. Previous TTS system has shown that failing to identify and correctly preprocess the French words cause a drop in the quality of the system.

In this work, we propose a multilingual approach for the linguistic module of the system to improve the phonetic transcription of French words. We included a language identification (LID) task at the first stage of the process and a multilingual Grapheme-to-Phoneme (G2P) model at the last stage. A Max-Entropy classifier and a Conditional Random Field (CRF) classifier are used to identify the language at the word-level. Besides, the Transformer architecture, a deep neural network, is used to train the multilingual G2P model. CRF outperforms the Max-Entropy classifier achieving a 0.828 F1-measure for the French words in the LID task, showing an improvement of 0.126 over the Max-Entropy classifier. The best G2P model trained on monolingual sentences and tested on the code-switched corpus achieves a PER of 17.68% and a WER of 31.22%.

[scroll back to beginning]

Sandipana Dowerah (NANCY)

Grapheme-to-phoneme (G2P) conversion is the task of converting a given sequence of letters termed as grapheme, into a sequence of pronunciation symbols termed as phonemes. Grapheme-to-phoneme converter is a vital part in Text-to-Speech system and Automatic Speech Recognition system for providing accurate pronunciations for words that are not covered by the lexicon. There have been various approaches applied for the task of grapheme-to-phoneme conversion over the years.

The goal of the proposed study is to elaborate a reliable and robust approach for predicting the pronunciation of words for speech synthesis purpose. The proposed approach is relying on a set of various grapheme-to-phoneme models for producing more reliable results comparing to an individual model. The first approach is ensemble of two models and is followed by ensemble of more than two models. For the task, we have considered Sequitur and Phonetisaurus as the baseline models, they both rely on statistical modelling of spelling and pronunciation subsequences. For an unbiased result, the evaluation is carried out with our write-up code. Then, we have adapted the Sequence-to-Sequence architecture from OpenNMT for our task by carrying out various modifications. We train the Sequence-to-Sequence G2P model with various model structures such as RNN-LSTM, Bi-LSTM, Attention and non-Attention with varied layers (1-1, 4-4) for ensemble and for investigating their performance.

[scroll back to beginning]

Our approach of ensembling model has outperformed the baseline models. We have tried various combinations of two of the above models. So far, Sequitur combined with the Sequence-to-Sequence model gives the best result among them. It achieves 1.11% of phone-error-rate and 4.44% of word-error-rate on the pronunciations that are identically predicted by the two approaches. The evaluation was conducted on a subset (test set) of the CMU French pronunciation dictionary. We have also carried out experiments on the CMUdict for English; the corresponding performance achieved is 2.98% of phone-error-rate and 13.07% of word-error-rate on the English test set.

We are currently analysing in more details the performance achieved by each individual model, and by combining models, in order to improve the global performance. We will also explore the application of other types of neural network approaches for grapheme-to-phoneme conversion.

Meisyarah Dwiastuti (PRAGUE)

In this work, we conduct a study on Neural Machine Translation (NMT) for English-Indonesian (EN-ID) and Indonesian-English (ID-EN). We focus on spoken language domains, namely colloquial and speech languages. We build NMT systems using Transformer model for both translation directions and implement domain-adaptation in which we train our pre-trained NMT systems on speech language (in-domain) data. Moreover, we conduct an evaluation on how the domain adaptation method in our EN-ID system can result more formal translation outputs.

Nina Hosseinikivanani (TRENTO)

Both behavioral and neuroimaging studies frequently investigate the mental processing of different categories. What do we talk about when we talk about "though, hate, joy" in our daily life? Answering this question is quite complex and needs to consider a lot of themes about the difference in the representation of words in the brain. In cognitive neuroscience, the neural representation of semantic knowledge is current interest in recent years. Previous studies of brain imaging have provided relatively little supporting evidence for the representation of abstract concepts in the brain. Based on authors knowledge, this study is one of the first studies that investigates the representation of different kind of abstract words in the brain. All in all, this study investigates the representation of meaning, in other words, the similarity in the neural activity produced by concepts reflects the semantic similarity between those concepts in brain regions that encode meaning. The original assumption of this study is that the more similar words should have more similar neural representations in brain regions encoding the meaning of those words.

References: *Fiebach, C. J. and Friederici, A. D. (2003). Processing concrete words: fMRI evidence against a specific right-hemisphere involvement. Neuropsychologia, 42(1):62–70. | Kiefer, M. (2012). Conceptual representations in mind and brain. Cortex, 48(7):805–825. | Mikolov, T. et al. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781. | Wang, J. et al. (2010). Neural representation of abstract and concrete concepts. Human Brain Mapping, 31(10):1459–1468. | Wang, X., et al. (2018). Organizational principles of abstract words in the human brain. Cerebral Cortex, 28(12):4305–4318.*

Dan Kondratyuk (SAARLAND)

Recent research has shown promise in multilingual modeling, demonstrating how a single model is capable of learning tasks across several languages. However, typical recurrent neural models fail to scale beyond a small number of related languages and can be quite detrimental if multiple distant languages are grouped together for training. We introduce a simple method that does not have this scaling problem, producing a single multi-task model that predicts universal part-of-speech, morphological features, lemmas, and dependency trees simultaneously for 124 Universal Dependencies treebanks across 75 languages. By leveraging the multilingual BERT model pretrained on 104 languages, we apply several modifications and fine-tune it on the concatenation of all available Universal Dependencies training data. The resulting model, we call UDify, can closely match or exceed state-of-the-art UPOS, UFeats, Lemmas, (and especially) UAS, and LAS scores, without requiring any recurrent or language-specific components. We evaluate UDify for multilingual learning, showing that low-resource languages benefit the most from cross-linguistic annotations. We also evaluate UDify for zero-shot learning, with results suggesting that multilingual training provides strong UD predictions even for languages that neither UDify nor BERT have ever been trained on.

Elizaveta Kuzmenko (TRENTO)

When we talk about objects that exist in the real world, how differently are they represented in our speech compared to their real properties? Which entities have similar representations in our speech but are completely different when it comes to life? In this project we want to investigate how much the world depicted in the distributional space differs from the real world. It was previously shown that the representations for colors of some entities in our speech are modified according to Gricean maxims (Rawee 2018). Now we want to investigate these differences not only for the color subspace but on a large scale.

In order to perform such a comparison, we build a model representing the real world from the Visual Genome, a large database of images marked with objects, attributes, and relations. The text space is built from the English Wikipedia, which represents encyclopaedic language, and BNC, which contains a variety of styles. Comparison methods include nearest neighbor comparison for different words, vocabulary comparison for models and learning a mapping function between spaces. We also check the stability of each space across original corpora perturbations.

References: *Rawee J. (2018) The Color Subspace in Distributional Semantics: Between Utterance Conservation and World Transformation*

Gosse Minnema (TRENTO)

Events are an important part of natural language meaning but are difficult to model in distributional semantics. My thesis investigates *named events* (e.g., ‘Hurricane Sandy’, ‘Battle of Waterloo’) and proposes distributional representations for these events. My work so far focused mostly on representations derived from encyclopedic definitions of these events. We experiment both with traditional additive methods for vector composition and with contextual sentence embeddings obtained using BERT, a state-of-the-art language representation model. Inspired by previous studies on extracting structured information from distributional vectors, we evaluate our representations by using them to predict referential attributes of the events that they encode. For this purpose, we construct a new dataset, derived from Wikipedia, containing event descriptions and attribute-value pairs for various types of historical events. Our results show that simple classification models can achieve good prediction performance on many of the attributes, even with limited training data. Moreover, we found that BERT embeddings only marginally outperform traditional distributional representations, and that this crucially depends on the method used for extracting and pooling embeddings.

Aria Nourbakhsh (GRONINGEN)

Gender identification is the task of detecting the gender of the author of a textual data. Training and testing within genre using lexical representation of the documents such as combination of word bi- and trigrams with character 3-6 grams, yields a good result. But when we want to train on one genre and test it on another, we lose accuracy. Our goal in this task is to see if we can capture the difference in writing between men and women (if there is any differences) regardless of its source. For this purpose, we are working with PAN dataset extracted from twitter and blog posts for English and Spanish.

Our approach consists of two parts. One is to use non-lexicalized ways such as different types of word embeddings, such that it can generalize to both Twitter and blog spaces. For the second approach, with the presence of the second tags in our dataset (age and native language of the authors) we are going to see if these two variables can help gender identification in a multi-task learning setting. Feature and result analysis are a part of our evaluation method.

Maria Obedkova (BASQUE COUNTRY)

In ASR systems, dictionaries are usually used to describe pronunciations of words in a language. These dictionaries are typically hand-crafted by linguists. One of the most significant drawbacks of dictionaries created this way is that linguistically motivated pronunciations are not necessarily the optimal ones for ASR. Furthermore, such dictionaries are not usually available for low-resource languages. The goal of this research is to explore approaches of data-driven pronunciation generation for ASR. We investigate several approaches of lexicon generation, implement our data-driven solution based on the pronunciation clustering and compare the current hand-crafted dictionaries with our approach. The application for the proposed lexicon generation approach is not limited to the creation of a lexicon from scratch. This approach can also be used as an alternative to G2P handling of OOV words.

Ataur Rahman (SAARLAND)

Determining Action Tendencies in the Domain of Hate Speech Towards Migrants

The growing inrush of immigrants and refugees in the EU has been pursued by a striking increase in online hate speech towards migrants in most of the European countries. Those hateful comments often contain action tendencies such as threat or call to physical violence, sometimes even killing intentions. The main motivation of this research thus came from the automatic detection of these hateful action tendencies specific to immigrants.

Simon Preissner (TRENTO)

Over the course of the last years, continuous advances in distributional semantics have produced elaborate language models that represent conceptual knowledge with increasingly high accuracy. New architectures continue to push the limits of this field, but their increasing complexity comes at the costs of inflexibility towards new amounts of text and, most notably, interpretability. These two issues are tackled by the Fruit Fly Algorithm (FFA), an adaption of a lightweight locality-sensitive hashing algorithm. It produces semantic spaces from continuously growing resources (i.e., it implements incrementality) and remains transparent and fully interpretable at every stage of the process. The poster shows the algorithm's architecture as well as its application in an incremental setup, its performance in comparison to other distributional semantics approaches, and the interpretation of the obtained semantic space.

Micha de Rijk (PRAGUE)

People are very good at associating words with one another: holiday, sun, beach. These associations come to mind easily. There exist some crowd-sourced lexicons and a decent body of psycholinguistics research on word association, but data driven models of word association haven't received as much attention. Building models of word association using existing NLP methods provides valuable information on how good semantic resources like word embeddings are at capturing these relations. Additionally, we can establish a baseline for how well the task can be performed by computers. We introduce several models for finding related words using proximity in word embeddings and pointwise mutual information of sentence-level and dependency-based bigrams. To measure the performance of these methods, we let people judge the relatedness of words produced by each model in the context of a word association game. The game is a singleplayer version of a Czech game called Codenames (Krycí jména), where you have to guess 9 words out of a set of 25 words using a limited number of hints. This allows us to score and compare the models based on the number of correct and incorrect words selected by the player. We achieve scores well above random chance, but many improvements can still be made.

[scroll back to beginning]

Alumni abstracts

Angelo Basile

Inspired by Labov's seminal work on stylistic variation as a function of social stratification, we develop and compare neural models that predict a person's presumed socio-economic status, obtained through distant supervision, from their writing style on social media. The focus of our work is on identifying the most important stylistic parameters to predict socioeconomic group. In particular, we show the effectiveness of morpho-syntactic features as predictors of style, in contrast to lexical features, which are good predictors of topic.