THE DEPARTMENTS OF INFORMATICS & PPLS PRESENT

THE 75th LANGUAGE LUNCH @ EDINBURGH

16/09/2022
1800-2000 (UTC +1)

For more information, visit:
https://blogs.ed.ac.uk/languagelunch/

SCHEDULE

1800-2000 (UTC +1) I
16/09/2022
LIVE ON ZOOM

1800 I SESSION #1

1810 I POROMA MOSTAFIZ

Morphological variation in Bangla: Exploring patterns of gender-based variation in child directed speech

1825 I SHUN SHAO

Gold Doesn't Always Glitter: Spectral Removal of Linear and Nonlinear Guarded Attribute Information

1840 I XIAOWEN ZHAO

The Representation of RASIM from Ukraine in News Coverage: A Corpus-Based Critical Discourse Analysis

1855 I MEET THE SPEAKERS #1

1905 I SESSION #2

1910 I SIGNE KAZAKA

The Function of the Marker dah in Bahasa Melayu Pakning

1925 I NICK FERGUSON

Integrating Paraphrasing into the FRANK Query Answering System

1940 I SASHA CORDIER

Iconicity in word learning: Implications for generalization within the meaning space

1955 I MEET THE SPEAKERS #2

MORPHOLOGICAL VARIATION IN BANGLA: EXPLORING PATTERNS OF GENDER-BASED VARIATION IN CHILD DIRECTED SPEECH

This study investigates morphological variation in child-directed speech (CDS) of native Bangla-speaking families from Bangladesh. The paper analyses style-shift between standard and non-standard Bangla with a focus on present-tense verb forms: e.g. korchi do.1SG.PRES (standard) vs kortesi do.1SG.PRES (non-standard). From informal observation, we noticed a tendency for parents to shift to standard verb forms when addressing young children in their care, particularly with girls. Our study therefore aimed (1) to gather quantitative data on the variable use of standard verb forms in Bangla CDS; and (2) to investigate both quantitatively and qualitatively language attitudes of potential relevance for patterns of style shift.

We base our primary analysis of verb forms on a corpus of YouTube videos, for which a total of five families with children (age ranging from a few months to twelve years old) were chosen. 'Parentese' (i.e. speech directed at children) and inter-adult speech were both analysed with special emphasis on the social context (pragmatic frames) of each utterance. Additionally, focus-group discussion and an online survey were used to gather data on adult speakers' attitudes to standard vs non-standard Bangla. The aim of this was to gain a better understanding of the motivation behind interlocuters' choosing to style shift or not. In accordance with our predictions, style shifting was found to occur variably in CDS depending on the age and gender of the addressee as well as the gender of the addresser. The results also suggest that language attitudes for standard vs non-standard varieties of the language play a significant role in style choices adopted by the parents. Thus, a complex interplay between language attitudes, socio-pragmatic context and sociolinguistic variables underlie style-shifting tendencies in Bangla CDS.

Keywords: Style-shift; Language regard; Child-directed speech; Mixed methodology; Gender.

POROMA MOSTAFIZ | PPLS

A CERTIFIED BOOKWORM, POROMA STARTED HER JOURNEY AS AN UNDERGRADUATE STUDENT AT NORTH SOUTH UNIVERSITY IN BANGLADESH WITH THE HOPES OF PURSUING A LITERATURE DEGREE. SHE HAD TO TAKE ONE COURSE IN LINGUISTICS TO FULFIL CREDIT REQUIREMENTS AND THAT COURSE CHANGED HER DIRECTION. FORSAKING LITERATURE, SHE IS NOW ON HER WAY TO UNCOVER MORE DEPTHS OF LINGUISTICS AS SHE PURSUES AN MSC IN LINGUISTICS AT EDINBURGH.

GOLD DOESN'T ALWAYS GLITTER: SPECTRAL REMOVAL OF LINEAR AND NONLINEAR GUARDED ATTRIBUTE INFORMATION

Language models are successful in modelling various applications. However, the decision-making by language models is often affected by undesirable biases encoded in the real-world training data. The current research on the debiasing method shows strengths in mitigating gender biases but hardly removes other demographic biases and is criticised for hurting the language model's performance. We describe a simple and effective method (Spectral Attribute removaL; SAL) to remove private or guarded information from neural representations that can be applied to any kind of bias. We use matrix decomposition to project the input representations into directions with reduced covariance with the guarded information rather than maximal covariance. We begin with linear information removal and proceed to generalize our algorithm to the case of nonlinear information removal using kernels. SAL outperforms the recent method of Ravfogel et al. (2020) aimed at solving the same problem, and is able to remove guarded information much faster while retaining better performance for the main task with only 3 lines of code. We demonstrate that our algorithm retains better main task performance after removing the guarded information compared to previous work. We also demonstrate that we need a relatively small amount of guarded attribute data to do so, which lowers the exposure to sensitive data and is more suitable for low-resource scenarios.

Keywords: Guarded Attribute Removal, Deep Learning, Debiasing, Fair Classification.

SHUN SHAO | INFORMATICS

SHUN SHAO IS A RESEARCH MASTER'S STUDENT AT THE INSTITUTE FOR LANGUAGE, COGNITION AND COMPUTATION (ILCC), UNIVERSITY OF EDINBURGH. HIS RESEARCH FOCUSES ON PROTECTED ATTRIBUTE REMOVAL, & ERASURE OF UNALIGNED ATTRIBUTES FROM NEURAL REPRESENTATION. HIS AIM IS TO ALLEVIATE UNFAIRNESS & MAINTAIN PRIVACY IN NATURAL LANGUAGE PROCESSING. PREVIOUSLY, HE FINISHED HIS M ENG AT UNIVERSITY COLLEGE LONDON.

THE REPRESENTATION OF RASIM FROM UKRAINE IN NEWS COVERAGE: A CORPUS-BASED CRITICAL DISCOURSE ANALYSIS

This paper examined the representation of Ukrainian refugees, asylum seekers, immigrants and migrants (RASIM) in New York Times, The Guardian, and RT between 24th February 2022 and 31st May 2022, the 2022 Russia-Ukraine war. A combination of corpus-based approach and critical discourse analysis (CDA) was adopted. Specifically, LancsBox was selected as the corpus analyzing software to produce the content collocates relating to RASIM, which were grouped into categories based on the examination of concordance lines and fine-tuned according to topoi/topics. The discourse-historical approach (DHA) was chosen as the analytical framework for CDA, focusing on the discursive strategies employed in the RASIM-related news discourse. The results revealed both negative (e.g., water metaphors, the illegality of refugees) and positive (e.g., help from host countries) presentation of RASIM and collocates (e.g., trauma, desperate) evoking compassion for refugees. Overall, The Guardian and RT indicated a more positive attitude towards RASIM than NYT. NYT and The Guardian represented the stance of NATO, expressing anti-Russia ideology through constructing the Ukrainian RASIM as victims of Putin's offensive invasion. Such ideology, however, was not evidenced in RT, where Russia was presented as the victims of the western oppression in some cases.

Keywords: Corpus-based critical discourse analysis; Refugees; Asylum; News; Ukraine.

XIAOWEN ZHAO | PPLS

AFTER FINISHING HER UNDERGRAD AT THE UNIVERSITY OF NOTTINGHAM, XIAOWEN PURSUED THE MSC APPLIED LINGUISTICS PROGRAM AT EDINBURGH. HER RESEARCH INTERESTS ARE CORPUS-BASED CRITICAL DISCOURSE ANALYSIS, QUANTITATIVE + QUALITATIVE APPROACHES, MULTIMODALITY. SHE IS ALSO INTERESTED IN NEWS DISCOURSE AND MEDIA COMMUNICATION. HER FUTURE PHD RESEARCH FOCUSES ON THE PRESENTATION OF COVID-19 ON CHINESE MAINSTREAM MEDIA.

THE FUNCTION OF THE MARKER DAH IN BAHASA MELAYU PAKNING

This study explores whether marker dah in Bahasa Melayu Pakning is similar in meaning to English perfect, perfective aspect, and adverbial already. The main focus of the study is the semantic factors and syntactic distribution of the marker, which are based on direct elicitation data obtained while working with a language consultant who is a native speaker of the language. Markers similar to dah are found in East and Mainland Southeast Asian and Austronesian languages, and their semantics have been widely discussed in a variety of studies.

The two main suggestions on how to analyze these markers are to treat them as adverbial already, or to introduce a new grammatical category which combines both properties of perfect aspect and adverbial already - iamitive. Another suggestion is that the semantics of these markers is determined by their syntactic distribution. The results of the study suggest that dah combines semantic characteristics similar to perfect and perfective aspect, and adverbial already. Preverbal dah may be analyzed as perfect or perfectly, yet it does not exclude analysis of adverbial already. Postverbal and sentence/clause final dah may be analyzed as already, however they exhibit an intriguing pattern which suggests that they might have different semantics based on their syntactic distribution.

Keywords: Grammatical category; Aspect; Adverbial already; Bahasa Melayu Pakning.

SIGNE KAZAKA | PPLS

SIGNE GRADUATED WITH A DEGREE IN ENGLISH AND SWEDISH PHILOLOGY FROM THE UNIVERSITY OF LATVIA. DURING THAT TIME, SHE SPENT A YEAR STUDYING IN SWEDEN. AFTER THAT SHE WORKED AS A SWEDISH TUTOR AND LATER BECAME LECTURER IN SWEDISH AT THE LATVIAN ACADEMY OF CULTURE. SHE HAS JUST FINISHED HER MSC LINGUISTICS PROGRAMME AT EDINBURGH AND IS HOPING TO CONTINUE IN A PHD PROGRAMME WHERE SHE WOULD LIKE TO PURSUE LINGUISTIC FIELDWORK.

INTEGRATING PARAPHRASING INTO THE FRANK QUERY ANSWERING SYSTEM

We present a study into the ability of paraphrase generation to increase the variety of natural language queries that the FRANK Query Answering system can answer. FRANK decomposes natural language queries into subgoals, which are executed over knowledge bases such as Wikidata. Results from these knowledge base queries are then combined into an answer according to the intent of the user's query. Rather than simple factoid retrieval, FRANK specialises in synthesising new knowledge using statistical reasoning. For example, take the query 'What will the population of France be in 2029?'. As this data does not exist, FRANK will look up past population data and perform regression over it in order to predict an answer. However, FRANK's parser is template-based, and limited in the number of question forms that it can be asked. In order to increase its coverage, we tested if paraphrase generation methods could reformulate unparsable queries into parsable ones. We chose an English-French backtranslation model to generate paraphrases, which we test using a small challenge dataset. We concluded that this method is not useful for improving the variety of natural language queries that FRANK can answer. Based on our observations, we recommend future work in the following directions: (1) allowing the ability to specify a form to paraphrase an input into; (2) constrained paraphrasing by masking specified terms to avoid loss of information about query intent; and (3) the need for an automatic evaluation metric which captures semantic similarity, promotes syntactic variation, and rewards preservation of query intent.

Keywords: Question Answering, Paraphrasing, Backtranslation.

NICK FERGUSON | INFORMATICS

NICK IS A SECOND-YEAR PHD STUDENT IN THE CDT IN NATURAL LANGUAGE PROCESSING. AFTER A BSC IN PHYSICS AND AN MSC IN ARTIFICIAL INTELLIGENCE AT CARDIFF UNIVERSITY, NICK NOW WORKS ON A QUESTION ANSWERING SYSTEM AS A VEHICLE FOR DELIVERING A COMPOSITIONAL AND EXPLAINABLE APPROACH TO AI. NICK'S RESEARCH FOCUSES ON THE GENERATION OF NATURAL LANGUAGE EXPLANATIONS WHICH DESCRIBE THE REASONING PROCESS BEHIND THIS SYSTEM.

ICONICITY IN WORD LEARNING: IMPLICATIONS FOR GENERALIZATION WITHIN THE MEANING SPACE

Iconicity has been documented to influence lexical learning (Ortega, 2017), but this research focuses on form production or mapping to meanings. This study consequently investigates how iconicity influences the meaning assumptions made in word learning and the resulting generalizations accepted by learners through an artificial sign language learning experiment.

Participants are trained on a small set of either iconic or arbitrary gesture-image pairs, and then tested on a broader range of candidate meanings. No significant difference was recorded in the average number of meanings selected based on iconicity, but there was a significant interaction between iconicity and meaning type. In the iconic condition, participants were significantly more likely to select meanings that were dissimilar apart from one shared iconic feature. In the arbitrary condition, participants were significantly more likely to select meanings that shared all features apart from one (which in the iconic condition would have been the iconic feature).

These results may demonstrate the broader ability of iconicity to unite disparate meanings, but further research is required that offers a gradient of candidate meanings within each relationship to the trained stimuli.

Keywords: Iconicity; Word learning; Meaning extension; Abstraction.

SASHA CORDIER | PPLS

SASHA HOLDS A DOUBLE UNDERGRADUATE DEGREE IN LINGUISTICS AND SPANISH. RECENTLY, SHE COMPLETED THE MSC EVOLUTION OF LANGUAGE AND COGNITION AT EDINBURGH. HER DISSERTATION OFFERS PRELIMINARY EXPERIMENTAL RESULTS REGARDING THE NATURE OF ICONIC MEANING ASSOCIATIONS, EXPLORING THE POSSIBILITY FOR GENERALIZATION, & FURTHER DISTINGUISHING DIFFERENT ROLES OF ICONICITY DURING LANGUAGE ACQUISITION.

ORGANIZING COMMITTEE

ANNA LAOIDE-KEMP | PPLS
GEORGIA CARTER | INFORMATICS
SYDELLE DE SOUZA | INFORMATICS
ABIGAIL LA LIBERTE | PPLS

ACKNOWLEDGEMENTS

THIS EVENT WAS MADE POSSIBLE BY THE GENEROSITY OF THE CDT IN NLP AND THE SCHOOLS OF PPLS AND INFORMATICS

THE UNIVERSITY *of* EDINBURGH
UKRI Centre for Doctoral Training
in Natural Language Processing

THE UNIVERSITY *of* EDINBURGH
informatics

THE UNIVERSITY *of* EDINBURGH
School of Philosophy, Psychology
& Language Sciences