Université de Lorraine - UFR Math-Info
Master SCA – M1
2016-2017

Fiche de projet tutoré / Project form

Dialogue processing on Twitter

Encadrement / Supervisors

- Main supervisor: Christophe Cerisara (cerisara@loria.fr)

Description / Description

The language used in Twitter is (1) dynamic, evolving quickly ("topic drift"); (2) noisy: standard NLP tools may fail on it; (3) contextually very rich: lots of redundancy, contains time, user and localisation information, hyperlinks, #tags, @mentions, retweets...

This project will focus on studying Twitter dialogues in French, which occur when users recursively reply-to a previous tweet, hence forming a tree-structured dialogue graph. Based on millions of French dialogue tweets that are already available in the Synalp team, the objective is to study these dialogues, to adapt/simplify the annotation scheme derived from (Scheffler, 2016) for French, and to annotate manually a gold corpus of about 1000 tweets with dialogue acts, where a dialogue act characterizes a phrase as, e.g., an open or closed question, an aswer, an informative statement, etc. I recommend to simplify the set of dialog acts to about 10 easy-to-annotate standard dialogue acts. A part of the 1000 tweets shall be annotated by 2 or 3 persons to compute the inter-annotator agreement ratio.

Then, you will investigate lexical features, such as punctuation marks, to automatically annotate these tweets with dialogue acts. You may finally use this automatic annotation to filter-out uncertain tweets, train a deep learning model on the most reliable tweets, and evaluate it on the annotated gold corpus.

Informations diverses : matériel nécessaire, contexte de réalisation /
Various information: material, context of realization

The corpus of dialogue tweets in French is available in the Synalp team of the LORIA laboratory. For the last part of the project, training deep learning models, the Keras toolkit will be used because it makes programming deep learning models very easy in python. For running the experiments, GPU computers are available at LORIA.

Livrables et échéancier / Deliverable and schedule

M0+1 : annotation guide validated on 100 examples to be annotated
M0+2 : annotated gold corpus of 1000 examples
M0+3 : validation of feature-based classifiers on the gold corpus

M0+4 : validation of deep learning Keras models on the gold corpus

Bibliographie / References

Tatjana Scheffler and Elina Zarisheva. Dialog Act Recognition for Twitter Conversations. In: *Proceedings of the Workshop on Normalisation and Analysis of Social Media Texts (NormSoMe)*, pages 31-38, collocated with LREC, Portorož, Slovenia. 2016.

Elina Zarisheva and Tatjana Scheffler. Dialog Act Annotation for Twitter Conversations. In: *Proceedings of the SIGDIAL 2015 Conference*, pages 114–123, Prague, Czech Republic, 2-4 September 2015.