

Fiche de projet tutoré / Project form

What's the answer: dialogue annotation

Encadrement / Supervisors

- encadrant principal / main supervisor

Maxime Amblard Sémagramme-Loria maxime.amblard@univ-lorraine.fr

- autres encadrants / other supervisors

Maria Boritchev Sémagramme-Loria maria.boritchev@loria.fr

Description / Description

Studying dialogical interactions is a major issue in natural language processing, since dialogues compose the basis of human communication. Addressing this problem is a complex task linking approaches from fields such as semantics, pragmatics, and more generally logic and cognition.

This project is part of a more general study of dialogical interactions conducted by the SÉMAGRAMME team (Loria, Inria Nancy Grand Est). It will focus on a systematic study of dialogues from the TCOF (Traitement des Corpus Oraux en Français) corpus in terms of questions and their answers. This corpus is composed of both oral dialogues and their transcriptions. The main task is to propose an annotation scheme of dialogical interaction.

A first step will be to create a questions annotation scheme derived from [3] and [6] and to annotate manually a gold corpus of several (exact amount to be determined) dialogues. Then, in order to automatically produce the annotation, it will be interesting to investigate lexical features such as punctuation marks to be able to automatically spot questions, as well as semantic features such as word usage in order to determine the span of an answer (as questions can be follow-ups of previous ones).

The annotated corpus can also be used to train a machine learning algorithm or a deep neural network one. Then we can compare the two approaches.

Informations diverses : matériel nécessaire, contexte de réalisation /

Various information: material, context of realization

The corpus is freely available online on the Ortholang platform.

The steps below can be followed:

- get familiar with corpus annotation
- choose interesting phenomena with respect to the question-answer pairs
- annotate relevant fragments of the corpus

- compare the results with those of other annotation formalisms (DRT, MRS)
- propose and implement a question-and-answer searching algorithm
- implement a machine learning solution
- run experiments and compare results

Livrables et échéancier / Deliverable and schedule

- November-December: bibliography
- February: annotation scheme definition, manual annotation of the corpus
- March: implementation testing
- April-May: analysis of the results, report writing and defense

Bibliographie / References

- 1- Virginie André, Christophe Benzitoun, Emmanuelle Canut, Bertrand Gaiffe, et al.. Traitement informatique de données orales : quels outils pour quelles analyses ? Recherche Qualitative, Hors série (n° 9), 2010
- 2- Maria Boritchev, Approaching dialogue modeling in a dynamic framework, juin 2017, Université de Lorraine
- 3- Alex Boulton, « Alors la question c'est... ? Questions pragmatiques et annotation pédagogique des corpus », Cahiers de l'Association for French Language Studies, Vol. 13.2., pp. 31-60, 2007
- 4- Emmanuelle Canut, Virginie André, Bertrand Gaiffe. Mise à disposition de corpus oraux interactifs : le projet TCOF (Traitement des Corpus Oraux en Français). Pratiques : théorie, pratique, pédagogie, 2010, Interactions et Corpus Oraux, pp.147-148
- 5- Ivano Ciardelli, Jeroen Groenendijk, Floris Roelofsen, Inquisitive semantics , 2017
- 6- Jonathan Ginzburg and Ivan Sag. Interrogative investigations. Stanford: CSLI publications, 2000.
- 7- Jan Gorisch, Emina Kurtic, Ella Page, Bill Wells, Guy Brown, et al.. Prosodic Matching and Turn Competition in Multi-Party Conversations. *Proceedings of Prosodic Constructions in Dialog*, 2015, Antwerp, Belgium. 2015. 〈hal-01231883〉
- 8- Ide, Nancy, and James Pustejovsky, eds. *Handbook of Linguistic Annotation*. Springer, 2017.
- 9- Marie-Paule Péry-Woodley, Nicholas Asher, Patrice Enjalbert, Farah Benamara, Myriam Bras, et al.. ANNODIS : une approche outillée de l'annotation de structures discursives. *TALN 2009 (Conférence sur le Traitement Automatique des Langues Naturelles)*, Jun 2009, Senlis, France. paper_TALN_52, 2009. 〈hal-00410590〉