Université de Lorraine - UFR Math-Info
Master SCA - M1
2015–2016

Master Sciences de la Cognition et Applications

Titre: Outil d'analyse d'indices linguistiques dans les transcriptions de dialogue
Title: A Tool for automatic analyse of linguistic clues in dialogue transcripts
Public : M1 Projet Tutoré

1 Encadrement / Supervisors

Encadrant : Maxime Amblard

équipe : Sémagramme-Loria
contact : maxime.amblard@univ-lorraine.fr

2 Description / Description

Ce projet s'inscrit dans le projet de de recherche SLAM [1]. Il vise à systématiser l'étude des conversations pathologiques dans le cadre d'une approche interdisciplinaire alliant psychologie, linguistique informatique et philosophie. Il se concentre notamment sur des conversations impliquant des personnes souffrant de troubles psychiatriques (schizophrènes).

Le sujet principal du projet est l'étude de l'interaction dialogique, en incluant plusieurs niveaux de description linguistique. Après plusieurs études sur la pertinence de la modélisation de ces dialogues avec des outils formels et le développement de plusieurs tests, nous souhaitons développer un outil qui intègre d'autres outils du TAL (au niveau de l'état de l'art) pour identifier automatiquement des indices de dysfonctionnement dans des transcription d'entretiens.

L'objet est donc de reprendre l'outils SLAMTK en cours de développement pour le rendre générique. L'outil est développé en python. Il fonctionne pour différents corpus qui possèdent des caractéristiques spécifiques. Il s'agit alors de comprendre les actions sur chacun des corpus et de définir le fonctionnement générique attendu afin de construire une nouvelle version des corpus fondée sur une structure XML. Puis, une fois cette cartographie réalisée, il sera nécessaire d'étendre les traitements pour enrichir l'information et l'identification d'indices.

This project is involved in the SLAM[1] research project. The aims is to systematize the study of pathological dialogues. It is part of an interdisciplinary approach combining psychology, computational linguistics and philosophy. It will focus particularly on dialogue involving people with psychiatric disorders (schizophrenia).

The main topic of the current project is the study of dialogic interaction, including several levels of linguistic annotations. After several studies on the relevance of modeling these dialogues with formal tools and the implementation of several tests, we want to develop a tool that integrates several other tools (at the state of the art) to automatically identify pathological linguistic clues in transcriptions.

The first part is to re-implement the SLAMTK tools, which is still under development, in a generic matter. The tool is develope with the planguage python. It works on different corpus which have specific characteristics. Students must understand the process on each of them and define the expected structure in order to build a generic version of the corpus (with an XML structure). Then, once this mapping carried out, it will be necessary to extend the treatment to enrich the

1. http://semagramme.loria.fr/doku.php?id=projects:slam

information and identification of clues.

Actuellement, l'outil prend en entrée un corpus transcrit. Il réalise une suite de pré-traitements pour normaliser les structures. Ensuite, l'outil DISTAGGER est utilisé pour identifier les disfluences ('heu...', les répétitions de mots, etc.), puis l'outil MELT pour réaliser la segmentation morpho-syntaxique (*part-of-speech*). Ensuite, plusieurs traitements sont réalisés pour créer des représentations de ces interactions et des tests mathématiques sont appliqués pour identifier des indices particuliers.

Le projet suivra plusieurs évolutions :

1. Prise en main et adaptation de l'outil SLAMTK : définition de l'actuelle cartographie d'action
2. Implémentation d'une version générique de SLAMTK (qui produisent à la fois la nouvelle ressource annotée, ainsi que les analyses et les représentations)

Dans la suite de cette première étape, il sera nécessaire d'identifier comment étendre la chaîne de traitement en identifiant des outils au niveau de l'état de l'art (*speech to text*, analyse syntaxique, etc.) et/ou en implémentant de nouvelles analyses automatiques.

— Identifier des outils au niveau de l'état de l'art pour augmenter la couverture de la chaîne de traitements
— Intégrer ces outils à l'outil SLAMTK
— Implémenter de nouvelles analyses automatiques (au niveau phonétique et/ou syntaxique)
— Ouvrir l'outil à d'autres langues

Currently, the tool takes as input a corpus transcribed. It performs a series of pre-treatments to normalize it. Distagger tool is then used to identify disfluencies ('euh ...', repetition of words, etc.) and also the MELT tool for achieving the morphosyntactic segmentation (part-of-speech). Several treatments are performed to create representations of these interactions and mathematics tests are applied to identify specific clues.

The project will follow several developments :

1. Get started and adaptation with the SLAMTK tool : defining the current map of treatments
2. Implement a generic version of SLAMTK (which produce both the new annotated resource and analyzes)

After this first stage, it will be necessary to identify how to extend the processing chain by identifying tools at the state of the art (speech to text, parsing, etc.) and / or by implementing new automatic analysis.

— identify tools at the state of the art to increase the coverage of the chain
— Integrating these tools to the tool SLAMTK
— Implemented new automatic analysis (at phonetic and / or syntactic level)
— An interesting challenge is to open the tool to onther languages than french

3 Informations diverses / Various information

Ce projet s'inscrit dans le cadre du projet SLAM de la MSH-Lorraine, il fait suite à plusieurs autres stages précédents. Les implémentations seront réalisées en python.

This project falls under the SLAM project of the MSH-Lorraine, it follows several other previous courses. All the code will be done in python. It is not necessary to speak french for this project.

4 Livrable et échéancier / Deliverable and schedule

— cartographie de l'outil SLAMTK
— implémentation d'une version générique de SLAMTK
— identification de nouveaux indices linguistiques pertinents

— identification et utilisation de nouveaux outils pour augmenter la couverture de la chaîne de traitement

— Mapping the SLAMTK tool
— Implementing a generic version of SLAMTK
— Identifying new relevant linguistic cues
— Identify and use of new tools to increase coverage of the processing chain

Références

[1] Maxime Amblard, Karën Fort, Caroline Demily, Nicolas Franck, and Michel Musiol. Analyse lexicale outillée de la parole transcrite de patients schizophrènes. *Traitement Automatique des Langues*, 55(3) :25, August 2015.

[2] Nicholas Asher and Alex Lascarides. *Logics of Conversation*. Studies in Natural Language Processing. Cambridge University Press, 2003.

[3] Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. Transcriber : a free tool for segmenting, labeling and transcribing speech. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 1373–1376, Granada, Spain, May 1998.

[4] Olivier Baude, Claire Blanche-Benveniste, Marie-France Calas, Paul Cappeau, Pascal Cordereix, Laurence Goury, Michel Jacobson, Isabelle De Lamberterie, Christiane Marchello-Nizia, and Lorenza Mondada. *Corpus oraux, guide des bonnes pratiques 2006*. CNRS Editions, Presses Universitaires Orléans, 2006.

[5] Christophe Benzitoun, Karën Fort, and Benoît Sagot. TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe. In *Traitement Automatique des Langues Naturelles (TALN)*, pages 99–112, Grenoble, France, 2012.

[6] Karën Fort. *Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus*. PhD thesis, Université Paris XIII, LIPN, INIST-CNRS, December 2012.

[7] Hans Kamp and Uwe Reyle. *From Discourse to Logic*. Kluwer Academic Publishers, 1993.

[8] Michel Musiol, Maxime Amblard, and Manuel Rebuschi. L'hypothèse des troubles du langage et de la pensée au risque de la formalisation sémantique du discours. . In *Analyse des discours Hors-Normes : approches, concepts et méthodes.*, Sherbrooke, Canada, June 2015.

[9] Association American Psychiatric. *Diagnostic and Statistical Manual of Mental Disorders DSM-IV-TR*. 30. American Psychiatric Association, Washington DC, fourth (text revision) edition, 1994.