What's The Answer: Dialogue Annotation

1 Introduction

Our project focused on identifying and classifying question-answer pairs in spoken language. We had three main goals: *(1)* designing an **annotation schema** for classifying questions and answers, *(2)* writing an **annotation guide** and **manually annotate** dialogues in different languages, and *(3)* exploring **machine learning approaches** to automate our work. Our work took place in the context of the SLAM (*Schizophrénie et Langage: Analyse et Modélisation*) and aims to contribute to the development of computational models of both impaired and unimpaired discourse.

2 Annotation schema

Our annotation schema comprises five types of questions classified based on their form (syntax) and function (semantics and pragmatics) and seven types of answers based on the possible answers a particular type of question can have. **Question types**: Yes/No (YN), Wh-question (WH), Disjunctive question (DQ), Phaic question (PQ) and Completion suggestion (CS). **Answer types**: Positive answer (PA), Negative answer (NA), Feature answer (FA), Phatic answer (PHA), Uncertainty answer (UA), Unrelated topic (UT) and Deny the assumption (DA). The annotation schema also includes additional information such as: *Feature* in the case of WH or DQ questions, for example for the question *When do you leave?* The feature will be Temporal (TMP).

3 Annotations

To test our annotation schema, we wrote an annotation guide that describes the tagging process. We then used this annotation guide to tag dialogues from three different corpora: the **Saarbrücken Corpus of Spoken English (SCoSE)**, the **CallFriend (Spanish)** corpus, and the **Corpus of Spoken Dutch (CGN)**. One of the dialogues in SCoSE serves the 'golden standard' for our project; a significant part of it was tagged by all of the group members, and based on this we computed agreement scores. Agreement was moderate for question types ($\kappa = 0.63$), but less so for answer types ($\kappa = 0.49$). Since our annotation task is quite complicated and requires making often subtle distinctions, this was only to be expected. As a final step, we analyzed our disagreements and improved our annotation guide based on these.

4 Machine learning

We experimented with two approaches for automatic question type classification: a **classical statistical approach** (decision trees) and several **neural network approaches.** The decision tree was based on feature structures that were extracted from the data using simple heuristics (e.g. 'contains a wh-word', 'contains a disjunction'). For the neural approaches, we tried out three different architectures and data representations: a multi-layer perceptron based on feature structures, a linear classifier that uses 'bags of words' (BOW), and an RNN classifier that represents questions as sequences of words. The **decision tree** produced the best results (accuracy 73%, $F1 = 0.58$ on unseen data); the neural networks generally performed well on seen data but were unable to generalize. We tried to improve the results of the RNN classifier by using pre-trained word embeddings, but this did not work. All models suffered from a **lack of sufficient training data** (only ~200 data points as input); until more human annotations are produced, it is too early to draw any definitive conclusions.