Polish human-human spoken dialog transcriptions – Experience from LUNA project

Krzysztof Marasek (Polish-Japanese Institute of Information Technology) and Ryszard Gubrynowicz

LUNA (6th framework project IST 033549) concentrates efforts to create a robust and effective spoken language understanding (SLU) module, which can be used to improve the speech-enabled telecom services in multilingual context. Polish was chosen as one of the languages for which the methods and solutions will be prepared and on which the elaborated methods will be tested. In order to do that a database of 500 spoken human-to-human dialogs has to be collected, transcribed and annotated. As a result of common consortium efforts a multilevel annotation scheme has been prepared. The recorded speech signal is annotated on several tiers:

- speech signal segmentation into turns and transcription of speech and acoustic events including special marking of spellings, foreign words, acronyms, external and speaker noises as well as mispronunciations, filler sounds , hesitations, etc.,
- morphosyntactic annotation,
- multilevel semantic annotation,
- dialog act annotations.

Those annotations will be used to train models of the SLU on the levels of language and semantic modeling as well as for testing and preparing of multilingual portability of the SLU module.

The paper presents the current status of the first level annotations, with focus on transcription of acoustic events (including special marking of those), selection of recordings and spontaneous speech effects. Polish spoken dialogs are collected at the call center of Warsaw Transport Authority (ZTM Warszawa) and are split into 4 main topics:

(1) dialogs referring to timetable of transportation lines,

(2) dialogs on path routes,

(3) dialogs referred to stops (closest stop to a given point) and

(4) dialogs concerning line routes.

These groups are additionally divided into subgroups male/female users and good/poor recording quality. The last distinction is done referring to the subjective evaluation of the S/N level, and caller's speech quality.

Substantial problems in speech transcription are observed:

- noisy calls: most of them are done in an adverse acoustic conditions (many calls are carried when driving or waiting on a noisy street, thus a lot of external noises, wind noises, etc. disturb the speech),
- low quality of speech transmission: many calls are over GSM when moving or using low quality microphones, what again influences the overall quality of recording, and causes overall low level of speech signal,
- long pauses and long speeches, quite often not at all relevant to the main topic of the dialog,
- strong emotions (quite often negative) of the speaker influencing her/his articulation.

Most of the speakers cannot speak fluently, even if they have time to prepare the talk. Breaks, hesitations, verbal deletions, non-grammatical expressions, problems with proper sentence formulation

and first of all, careless articulations are often observed. Despite of this, the speech transcription is progressing well.

We hope this database (first of such kind for Polish) will be an important resource for development of Polish spoken dialogs systems and overall studies on Polish speech.