Towards an Open-Source Dutch Speech Recognition System for the Healthcare Domain

C. Tejedor-García (cristian.tejedorgarcia@ru.nl)*, B. van der Molen†, H. van den Heuvel*, A. van Hessen‡, T. Pieters†
*CLST (RU, Nijmegen), † Freudenthal Institute (UU, Utrecht), ‡ HMI (UT, Twente), the Netherlands

Context

➤ Annually +15,000 hospital admissions in the Netherlands
  ➤ Avoidable misuse of medicines

➤ Main reasons:
  ➤ Functional illiteracy
  ➤ Forgetfulness
  ➤ Misuse of prescribed usage

➤ Consequences:
  ➤ Diverse + inappropriate forms of use
  ➤ Low levels of adherence
  ➤ Waste of scarce financial resources

➤ How to solve it?
  ➤ Better understanding of the explicit and implicit attribution of meaning to medicines as part of the information processing
  ➤ Effective + efficient transcriptions of doctor-patient interviews: ASR technology

➤ Context with considerable privacy-sensitive constraints

Our proposal: HoMed Project

➤ Proposes a new research infrastructure and method:
  ➤ Automatic transcription of sensitive audio-visual (AV) recordings

 ➤ General Data Protection Regulations (GDPR)

➤ Current largest open-source generic ASR for Dutch (Kaldi_NL):
  ➤ Vocabulary: not healthcare jargon: needs adaptation:
    1. Semantic adaptation (LM): this paper
       Material: Medicijnjournaal + lists of medical terms
    2. Acoustic adaptation (AM):
       Material: Nivel AV recordings + previous material

➤ INPUT: healthcare-related material
  ➤ Transcription files + healthcare word lists
  ➤ AV-recording files (highly sensitive)

➤ OUTPUT: ASR models + methodology to other domains:
  1. CLARIAH’s Infrastructure (Media Suite)
  2. Stichting Open Spraaktechnologie
  3. Nivel: standalone version

ASR in (Dutch) Healthcare

➤ ASR Advantages:
  ➤ Increases medical staff’s productivity
  ➤ Facilitates the completeness of medical documentation
  ➤ Inspires patient engagement

➤ CGN (Corpus Gesproken Nederlands):
  ➤ Generic/daily conversations
  ➤ WER: ∼7-8%

➤ Commercial ASR systems:
  ➤ Jargon + Data privacy + Costs

➤ Related current research projects (Google ASR) in the NL
  ➤ Care2Report, CAIRE-lab

ASR Systems Development & Evaluation

1. CGN: LM: General conversations, AM: Adult speech, Lexicon: 255000 tokens

WER & Error Analysis (Categories)

<table>
<thead>
<tr>
<th>Type of error</th>
<th>Kaldi_NL</th>
<th>HoMedV2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Spelling variant</td>
<td>457</td>
<td>798</td>
</tr>
<tr>
<td>2. Compound word</td>
<td>156</td>
<td>19</td>
</tr>
<tr>
<td>3. Morphological variant</td>
<td>31</td>
<td>11</td>
</tr>
<tr>
<td>4. Error within lexicon</td>
<td>598</td>
<td>872</td>
</tr>
<tr>
<td>5. OOV</td>
<td>286</td>
<td>78</td>
</tr>
</tbody>
</table>

Table: Comparison of the ASR systems performance

Funding & Useful Links

➤ Platform Digitale Infrastructuur Social Science and Humanities
  PDI-SSH 2020: https://pdi-ssh.nl

➤ Project webpage: https://homed.ruhosting.nl

Future Work

➤ Netherlands Institute for Health Services Research (Nivel)