Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

cjvt Center za jezikovne vire in tehnologije

Online Notes

A Real-Time Speech Recognition and Machine Translation System for Slovene University Lectures

Marko Bajec, Iztok Lebar Bajec, Tjaša Šoltes, Jernej Cvek
Jaka Čibej, Kaja Gantar, Sara Sever, Simon Krek
Laboratory for Data Technologies | Centre for Language Resources and Technologies
Faculty of Computer and Information Science
University of Ljubljana

Digital Inclusion in the Information Society (DIGIN 2023), Ljubljana, 11 October 2023

Presentation Outline

- Motivation for the Project
- Project Overview
- The Online Notes system
- Automatic Speech Recognition
- Machine Translation
- Workflow
- Main Interface
- Pilot Lectures and Feedback
- Conclusion and Future Work

Motivation

- development strategy adopted by the University of Ljubljana (2022–2027)
- a strong focus on internationalization, equal opportunity, solidarity, and inclusiveness
  - non-Slovene-speaking students
  - students with disabilities
- internationalization
- systematic support for the accessibility of studies to individuals with special needs

Motivation

- survey conducted by the University of Ljubljana in February 2021
- 38% of students with disabilities are dyslexic, 2% are hearing-impaired, 4% are sight-impaired, and 13% are cognitively impaired

- Online Notes – one of the projects financed by the University to tackle both the issues of internationalization and inclusiveness
  - Official title of the project: Upgrading a Machine Translation System for Learning Communities (and Special Needs Students)

About the Project

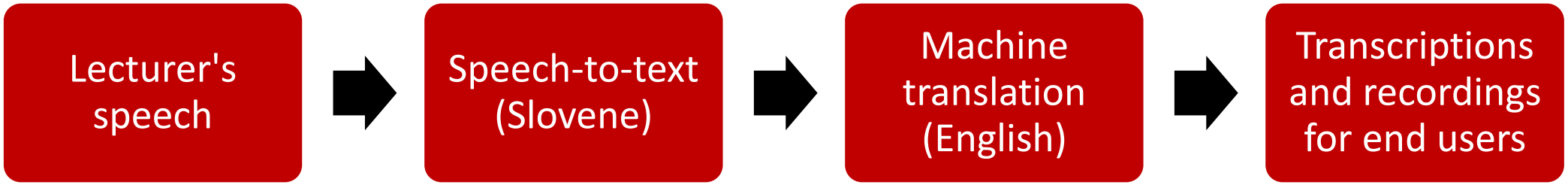
- Faculty of Computer and Information Science (University of Ljubljana)
  - Laboratory for Data Technologies
  - Centre for Language Resources and Technologies
- Project timeline: 2020–2024
- main goal: to develop a system for the real-time speech recognition and automatic translation of Slovene lectures
  - non-Slovene-speaking students can attend lectures held in Slovene
  - transcriptions and recordings of lectures are provided for students with disabilities

Project Activities

- (a) the development of speech recognition models for the domain of Slovene university-level lectures
- (b) the development of the system backend and frontend
- (c) the organization of pilot lectures at the University of Ljubljana
- (d) collection of user feedback

The Online Notes System

- The Online Notes system roughly consists of the following "stages":
  - automatic speech recognition
  - machine translation
  - interface



Speech Recognition Models

- two domain-specific speech recognition models for Slovene were developed (based on the Kaldi toolkit):
  - social science lectures (ON-DR)
  - technical science lectures (ON-NT)
- collection of audio- or video-recordings of lectures from previous years
  - automatically converted to text, then manually corrected
- corrected transcriptions were used to extract vocabulary, which was added to the appropriate ASR-model
- segmenter and automatic punctuator

Machine Translation

- Each segment is translated separately to minimize the delay in the display of translations.
- two machine translation systems are supported:
  - RSDO (Slovene-English) → https://www.slovenscina.eu/prevajalnik
  - Google Translate (Slovene-English), allows users to choose their preferred language

Main Components

- **administrator site** (to schedule lectures in the system)
- **streaming site** (to record lectures)
- **lecturer site** (to edit transcriptions and translations)
- **student site** (to follow transcriptions and translations)

Workflow

- (1) lectures are scheduled in the ON system by an administrator
- (2) a lecturer runs the ON speech-recognition and machine-translation system during their scheduled lecture through the streaming site
- (3) students can access the real-time transcriptions and translations during the lecture through the student site
- (4) once the lecture is complete, the lecturer can edit the transcription and translation (if necessary) through the lecturer site
- (5) the recording, (edited) transcription and (edited) translation are archived in the system and can be accessed again at a later date through the student or lecturer site

Main Interface

Clean | Edited | Low Confidence

A, veš potem pokazal Oke.	Ah, you know then showed Oke.
Vidiš zanimivo naloge beljakovin v organizmu, živali so zelo mnogovrstne!	You see the interesting functions of proteins in the body, animals are very diverse!
Vav so sestavni deli orodja: telesnih celic, vseh tkivo in organ.	Vav are components of tools: body cells, all tissues and organs.
Opravljajo zaščitne funkcije imajo kontrolna. Funkcijo. So: učinkovine.	They perform protective functions and have control functions. Function. They are: active ingredients.
Prenašajo rudninske snovi in kisik po krvnem obtoku, omogočajo prenos metabolitov v celice, sodeluje v obrambnem sistemu, aminokisline se lahko izkoristijo kot vir energije.	They transport mineral substances and oxygen through the bloodstream, enable the transfer of metabolites into cells, participate in the defense system, amino acids can be used as a source of energy.
Določene vrste beljakovin so za organizem škodljive.	Certain types of proteins are harmful to the body.
Naš tukaj imeti zdaj na ene stvari ne zapiše.	Our here to have now on one thing does not write down.
Ja, saj zato gledam, a veš recimo, tam sem rekla, pa sem vam bodo naredili določene, pa jih ni bilo, a ne?	Yes, because that's why I'm watching, but you know, for example, I said there, and they will make certain ones for you, but there weren't any, right?
Če je treba naučit o pomanjkanju beljakovin ali posameznih esencialnih aminokislin v obroku, se pri živalih pojavijo motnje v rasti in prireji.	If it is necessary to learn about the lack of protein or individual essential amino acids in the meal, disturbances in growth and breeding occur in the animals.

Pilot Lectures

- Since 2021 – a total of 29 pilot lectures (with two consisting of multiple lectures) have been held at 8 different faculties of the University of Ljubljana so far
  - social sciences (e.g. the Faculty of Arts, the Faculty of Social Sciences)
  - STEM subjects (e.g. the Faculty of Mathematics and Physics, the Faculty of Computer and Information Science).
- Mostly in 2022 and 2023 (previously, it was much more difficult because of the pandemic)

Feedback So Far

- collection of feedback currently in progress
- positive feedback from lecturers
- positive feedback from Erasmus students
- general feedback from students with disabilities (e.g. on interface design – GWAC guidelines)
  - We are still trying to organize pilot lectures for students with disabilities specifically.

Future Work and Tool Limitations

- improvement of ASR-models (currently, the word-error rate is between 12.6% and 32%)
  - lecturer’s speech rate, clarity, and degree of standardness
- speech segmentation
  - speech segments do not necessarily coincide with semantically coherent units
  - if segments are too long, there is a delay in translation
- successful use of ON depends on lecturer‘s speech as well!
  - avoiding segmented speech with numerous false starts
  - filler words such as *ne* in Slovene (which can also mean *no* or *not*, which causes problems in machine translation)

Conclusion

- By the end of the project, the plan is to develop the Online Notes system so that it can be successfully implemented at the University in order to:
- (1) help increase the accessibility of lectures for people with disabilities;
- (2) contribute to the internationalization of the University of Ljubljana by allowing non-Slovene-speaking students to attend Slovene lectures;
- (3) allow teachers to archive their lectures and edit their transcriptions/translations;
- (4) contribute to the wider language infrastructure for Slovene.

Univerza *v Ljubljani*
Fakulteta *za računalništvo*
in informatiko

cjvt Center za jezikovne vire in tehnologije

Thank you for your attention.