Euronews:
a Multilingual ASR Benchmark

EU-BRIDGE Partners

Fondazione Bruno Kessler (FBK)

Contacts

Fondazione Bruno Kessler (FBK)
Via Sommarive, 18
I-38123 Trento
http://hlt.fbk.eu
Roberto Gretter
gretter@fbk.eu

Description and Exploitable Knowledge

Recording data from TV and monitoring the Web to look for new audio resources is a fundamental activity in Automatic Speech Recognition (ASR). The TV channel Euronews, which broadcasts news in several languages, is an attractive source of comparable data, which were used to design a multilingual speech corpus for ASR purposes, made of recordings from TV and downloads from the Web.

The corpus includes data in 10 languages: Arabic, English, French, German, Italian, Polish, Portuguese, Russian, Spanish and Turkish; it was designed both to train Acoustic Models (AMs) and to evaluate ASR as well as Language Identification (LID) performance. For each language, the corpus is composed of about 100 hours of speech for training (60 for Polish) and about 4 hours, manually transcribed, for testing. Training data include the audio, some reference text coming from the Euronews portal which sometimes is a partial orthographic transcription, the ASR output and their alignment. Thanks to the light supervision technique, about 60 hours per language can be considered correctly transcribed.

These data are used inside the EU-BRIDGE consortium as a multilingual benchmark to evaluate ASR progress in the 10 languages, using similar amount of training data and comparable evaluation data. More details about this corpus can be found in the paper "Euronews: a multilingual speech corpus for ASR" by Roberto Gretter, in Proceedings of LREC, Reykjavik, Iceland, 2014.

Nationalfeiertag: Frankreich gedenkt der Opfer des Ersten Weltkriegs

NACHRICHTEN | The Corner

Deutschland zum vierten Mal Fußball-Weltmeister

14/07 07:29 CET

Euronews Portal

Application Sectors

ASR and LID training and evaluation

Terms of Availability

In 2013 the EU-BRIDGE consortium signed an agreement with Euronews, which gives to EU-BRIDGE partners the right to use Euronews material for research purposes, and to exchange it within the project.

Concerning the availability of these data for the whole research community, in 2014 Euronews agreed to make it available for research purposes. At present, part of the data is available as AM training data for the ASR multilingual evaluation benchmark for IWSLT 2014. We plan to make available more data for the next IWSLT evaluations.

language	#videos	speech duration	#ref words	#rec words	#common words	aligned speech
Arabic	4406	107:22:58	650,146	756,100	379,000	49:54:03
English	4512	112:18:29	973,210	1,032,727	699,850	63:02:34
French	4434	108:56:37	954,242	1,123,709	796,997	62:13:46
German	4438	108:33:23	809,289	896,387	653,372	61:46:27
Italian	4464	110:35:51	900,291	1,012,521	765,559	61:31:36
Polish	2626	58:32:03	350,729	454,977	278,854	27:42:35
Portuguese	4431	108:03:27	841,148	966,586	699,681	59:29:38
Russian	4418	107:42:24	714,363	828,060	611,347	60:49:15
Spanish	4465	109:16:50	939,408	1,053,255	797,698	63:29:23
Turkish	4387	106:30:31	683,041	764,329	556,760	60:52:55

Project Coordinator

Karlsruhe Institute of Technology
Adenauerring 2
D-76131 Karlsruhe
http://isl.anthropomatik.kit.edu
Prof. Alex Waibel
alexander.waibel@kit.edu

EU★BRIDGE

The work leading to these results has received funding from the European Union under grant agreement n°287658

SEVENTH FRAMEWORK PROGRAMME

www.eu-bridge.eu

EU-BRIDGE - the Project

EU-BRIDGE is a European Integrated Project that aims at developing automatic transcription and translation technology that will permit the development of innovative multimedia captioning and translation services of audio-visual documents between European and non-European languages.