Euronews: a multilingual speech corpus for ASR

Roberto Gretter
FBK
Via Sommarive, 18 - I-38123 POVO (TN)
gretter@fbk.eu

Abstract

In this paper we present a multilingual speech corpus, designed for Automatic Speech Recognition (ASR) purposes. Data come from the portal Euronews and were acquired both from the Web and from TV. The corpus includes data in 10 languages (Arabic, English, French, German, Italian, Polish, Portuguese, Russian, Spanish and Turkish) and was designed both to train AMs and to evaluate ASR 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, the ASR output and their alignment. We plan to make public at least part of the benchmark in view of a multilingual ASR benchmark for IWSLT 2014.

Keywords: multilingual, speech corpus, light supervision

1. Introduction

In the last few years, Web crawling and TV recording have become main data sources for research and progress in Automatic Speech Recognition (ASR).

Monitoring the Web to look for new resources is a fundamental activity, in particular for what concerns manually transcribed audio/video data and parallel corpora. In the past we started to collect data, on a daily basis, from the portal Euronews, which broadcast news in several languages.

Audio data can be used to train Acoustic Models (AMs) for a target language in a complete unsupervised way, as described in (Falavigna and Gretter, 2011; Bisazza and Gretter, 2012), where AMs are bootstrapped from those of a well trained source language (Italian in our case), without using any transcribed data in the target language.

When the video data come along with some text, which could be a partial transcription of it, other approaches are possible which exploits the so called "lightly supervised" training (Lamel et al., 2002) which uses transcriptions close, but not exact, to what is spoken. In the past, several multilingual resources were created and made available to the research community, such as Europarl (Koehn, 2005) for MT translation and GlobalPhone (Schultz, 2002) for multilingual ASR.

In this paper we describe a multilingual corpus, designed both to train AMs and test ASR systems in 10 languages: Arabic, English, French, German, Italian, Polish, Portuguese, Russian, Spanish and Turkish. These languages cover most of the languages that are addressed in the European project EU-BRIDGE.[1]

2. Euronews: sources and data description

Euronews provides multilingual data through two main channels: a satellite TV and a Web portal. We acquire data from both of them.

2.1. Data from TV

International news are acquired from the satellite TV channel in different languages, potentially attractive as a source of comparable speech data. At present news are transmitted in 13 different languages (new languages are added over time): Arabic, English, French, German, Greek, Hungarian, Italian, Persian, (in the past Polish), Portuguese, Russian, Spanish, Turkish and Ukrainian, each one over a different audio channel in the digital TV stream. As the video content is the same for all of the audio channels, the news are time aligned.

From ASR perspective, data cannot be considered really clean, in the sense that several phenomena take place: often, in case of interviews, some seconds of speech in the original language are played before the translation starts; commercials are often in English; there is the presence of music; sometimes a particular piece of news has not been translated yet in all of the languages, so that some channels may contain the original audio (in another language).

For Machine Translation (MT) purposes, it has to be said that news in different languages are **not** exact translations one of each other. Normally, first the journalists of the various countries meet and discuss about the news to be recorded, and after that each one goes in his/her office and writes the text in his/her own language. This makes the news comparable but not parallel.

Every day we record one hour of audio stream: we first extract the list of audio channels and then, for each language, its audio track which is stored at 16 kHz sampling rate. Given the particular structure of the stream, it is quite easy to segment the different audio tracks into news, because the boundaries among news are characterized by background noise over all of the channels. So, from a practical point of view, an effective and simple approach is to detect pauses on each separate channel, and consider as probable news boundaries the pauses that are common to all channels. In this paper, TV data will be used only as part of the test set.

2.2. Data from the Web

Every day, Euronews stores in its portal[2] videos and text describing the main news of the day. These latter are some of the news that are broadcasted via TV. The text associated with a news is sometimes a summary, sometimes a

[1] https://www.eu-bridge.eu/

[2] http://fr.euronews.com/

complete transcription, sometimes a partial transcription of the content of the news. By exploiting some information, it is possible to link a video to the same news in several languages.

It is worth to note that Euronews does not keep a complete archive of the video news: due to storage limitations, only the text version of the news is retained. Concerning the videos, they are removed after some weeks, so only the most recent ones can be found on the Web pages. This means that to get an archive of videos it is important to download them frequently.

Approximately, Euronews publish on its portal about 30 videos per day for most of the languages. The average duration of a video is about 90 seconds, so in one month we are able to collect about 900 videos for a total duration of about 22 hours per language. However one of the EU-BRIDGE languages, Polish, is under-represented: it is currently not broadcasted via TV and the number of videos produced is less than 6 per day. This means that Polish data are different from the other languages.

To crawl news content, in the past we developed a program, called HLT WebManager (Girardi, 2011), (free download from http://wit3.fbk.eu/ from the section "Tools"), that is able to download audio/video data, web pages, and to extract relevant text from them. It also provides a mechanism to find the parallel documents for each site.

3. Multilingual data for AM training

In order to prepare material for AM training, it is necessary to collect a set of audio recordings together with their orthographic transcription. Our target was to collect about 100 hours of raw speech - including silence, music, etc. - for each language.

In order to obtain a reliable transcription of each news in an automatic way, we did the following (Lamel et al., 2002): perform ASR recognition on a news, align the ASR output with the text associated with the news, and retain only the speech segments where there is a perfect match. The amount of retained data mainly depends on two factors: the effectiveness of the ASR for the given language and the reliability of the transcription, and in our case ranges from about 45% to about 60% of the material.

Data preparation for AM training for the 10 languages is described below. The data were published from January 1st, 2013 to May 31st, 2013 (5 months) and they roughly correspond to 100 hours of speech for each language. For Polish we used videos published from January 1st, 2012 to March 31st, 2013 (15 months) roughly corresponding to 60 hours of speech.

Concerning the text data, we only downloaded the texts associated with the videos. At this stage it was fundamental to extract from the Web pages only the relevant text, i.e. title and body of the news, discarding all the other stuff. Depending on the language, in this way we were able to collect from 350 to 970 Kwords, that was enough to build a Language Model (LM) which resulted quite small but focused on the material, and at the end appropriate for our purposes.

For each of the languages we applied the following processing steps:

#	time	ref word	ASR output
D	0.00	LKW	-
D	0.00	will	-
D	0.00	auf	-
S	0.00	wenden	@bg
S	0.04	zwei	ich
S	0.61	Tote	@bg
C	1.60	lettischer	lettischer
C	2.10	LKW	LKW
C	2.54	Fahrer	Fahrer
C	15.88	blieb	blieb
C	16.12	-	@bg
C	16.14	unverletzt	unverletzt
I	56.38	-	Hund
I	56.55	-	gebissen
# u: 113 e: 13 s: 6 i: 1 d: 6 c: 101			
# ua: 88.50% pc: 89.38% uer: 11.50%			

Table 1: Example of alignment between reference text and ASR output. D, I, S, C, indicate Deletion, Insertion, Substitution, Correct.

1. download videos and corresponding reference text from the website http://fr.euronews.com/. The reference text could be either a relatively precise manual orthographic transcription, or just a summary, or a long text related to the news but not corresponding to the audio.

2. find the cross-lingual links (we are able to detect the same news in different languages). This is not directly related to AMs training purposes, but could be useful for future cross-lingual issues.

3. extract the audio stream from the video, and store it in sphere format.

4. normalize the reference text for every news: this includes capitalization, punctuation removal, processing of number and is a language-dependent processing. We thank people from the Polish-Japanese Institute of Information Technology for their help with Polish.

5. join the texts from all the news and build a relatively small N-gram LM, to be used by the ASR. Using text related to the news, this will minimize the out-of-vocabulary problems, particularly for what concerns proper names.

6. perform ASR on all the audio data and align, news by news, the reference text with the ASR output.

7. compute some statistics for each file (duration/words, wer, #words ref, #words rec, etc.).

The processing of each video resulted into the following files:

- **.sph file**: audio in sphere format
- **.info file**: original text associated with the news (normally title + content, 2 lines)

language	recording	#videos	speech	#ref	#rec	#common	aligned speech
	period		duration	words	words	words	second (first) run
Arabic	01/01/13 - 31/05/13	4406	107:22:58	650,146	756,100	379,000	49:54:03 (42:39:46)
English	01/01/13 - 31/05/13	4512	112:18:29	973,210	1,032,727	699,850	63:02:34 (52:44:08)
French	01/01/13 - 31/05/13	4434	108:56:37	954,242	1,123,709	796,997	62:13:46 (57:55:07)
German	01/01/13 - 31/05/13	4438	108:33:23	809,289	896,387	653,372	61:46:27 (53:04:06)
Italian	01/01/13 - 31/05/13	4464	110:35:51	900,291	1,012,521	765,559	61:31:36 (54:01:00)
Polish	01/01/12 - 31/03/13	2626	58:32:03	350,729	454,977	278,854	27:42:35 (24:55:03)
Portuguese	01/01/13 - 31/05/13	4431	108:03:27	841,148	966,586	699,681	59:29:38 (43:24:16)
Russian	01/01/13 - 31/05/13	4418	107:42:24	714,363	828,060	611,347	60:49:15 (58:08:39)
Spanish	01/01/13 - 31/05/13	4465	109:16:50	939,408	1,053,255	797,698	63:29:23 (59:31:38)
Turkish	01/01/13 - 31/05/13	4387	106:30:31	683,041	764,329	556,760	60:52:55 (55:00:43)

Table 2: Amount of Web data collected for AM training for the 10 EU-BRIDGE languages. The last two columns report respectively the number of common words resulting from the alignment among reference text and ASR output, and the resulting duration of speech (hh:mm:ss) after the second (first) run.

- **.ref file**: reference text (UTF-8 text obtained by processing the .info file in order to remove punctuation, handle acronyms (uppercase: LKW, US), expand numbers (e.g. 777 → sieben hundert sieben und siebzig), etc.
- **.ctm file**: ASR output, UTF-8.
- **.lgn file**: alignment between .ref and .ctm files. Apart comments (lines beginning with ”#”), each line contains 4 tokens:
  - D I S C (Deletion, Insertion, Substitution, Correct);
  - word start time (from the .ctm);
  - reference word (or ”-”), from .ref file;
  - ASR output word (or ”-”), from .ctm file. @bg means silence.
  Some row samples are in Table 1.

In addition, there is one summary file containing some statistics about each video, including the Word Error Rate (WER) which is just an indicator of the match among reference text and the ASR output: sometimes it is very high just because the reference text is extremely short, maybe just the title of the news. What is really important is the number of common words and the amount of reliable speech, computed excluding empty segments (music, noise, etc.). All the statistics are reported in Table 2.

We repeated the whole process twice; in both cases the LMs were trained only on the texts downloaded with the videos. The first run was performed using AMs trained at FBK in the past, exploiting composite corpora and performing quite differently. In some cases (Portuguese and Polish) we used multilingual AMs, trained on speech coming from 6 other languages, which are part of a language identification system (Giuliani and Gretter, 2013). The aligned speech resulting from the first run was then used to train a new AM for each language, which was finally used to perform the second run. As expected, in all cases the amount of reliable speech increased.

4. Multilingual data for ASR evaluation

To evaluate ASR performance it was decided to prepare and manually transcribe about 4 hours of speech for each language. Ongoing transcription is currently carried out by various EU-BRIDGE partners and, after some unexpected delay, is now scheduled to be completed by the end of April, 2014. Transcription guidelines consider usual phenomena like false starts, hesitations, speaker change, etc. Particular care was posed to the identification of segments of foreign speech (e.g. Arabic speech in the Italian data), quite common in Euronews data.

Temporal markers should clearly identify segments in other languages. Most of the transcriptions were done by EU-BRIDGE partners working in the field of subtitling, which tend to use their software for captioning TV programs during their daily work. As a consequence, time markers were often approximate, and a checking phase was needed to ensure proper time marker placement. This last step was done using the popular software Transcriber[3].

We collected more than 2 hours from the Web and nearly 2 hours from the TV, see Table 3 for details. Concerning the Web data, we acquired videos from the first week of July 2013, then a subset of these was selected, according to the following criteria:

- for each language, 120 files were selected, roughly corresponding to 155 minutes of audio
- precedence was given to news common to all languages, which resulted to be true for all languages except Polish. As a result, the same news compose the test sets for 9 languages.

Concerning TV data, we carried out recordings starting from the first week of October 2013. Then we applied a simple news segmenter, based on the observation that news boundaries in Euronews are marked by silence over all the channels. Automatic news boundaries were manually checked and corrected when needed. We retained approximately 15 minutes per day, divided into news.

Again, for Polish we had to change something. Since the TV is no longer broadcasting the Polish channel (it was active in previous years), for this language we had to rely only on Web data. We kept data from April 1st, 2013 to

[3] http://trans.sourceforge.net/

language	recording period	Web videos	speech duration	recording period	TV videos	speech duration	total duration
Arabic	01/07/13 - 07/07/13	120	02:35:20	01/10/13 - 09/10/13	99	01:41:33	04:16:53
English	01/07/13 - 07/07/13	120	02:35:20	01/10/13 - 09/10/13	99	01:41:33	04:16:53
French	01/07/13 - 07/07/13	120	02:34:29	01/10/13 - 09/10/13	99	01:41:33	04:16:02
German	01/07/13 - 07/07/13	120	02:35:20	01/10/13 - 09/10/13	99	01:41:33	04:16:53
Italian	01/07/13 - 07/07/13	120	02:35:20	01/10/13 - 09/10/13	99	01:41:33	04:16:53
Polish	01/04/13 - 09/08/13	212	03:03:37	01/10/13 - 31/10/13	93(*)	01:28:50	04:32:27
Portuguese	01/07/13 - 07/07/13	120	02:35:20	01/10/13 - 09/10/13	99	01:41:33	04:16:53
Russian	01/07/13 - 07/07/13	120	02:35:20	01/10/13 - 09/10/13	99	01:41:33	04:16:53
Spanish	01/07/13 - 07/07/13	120	02:34:29	01/10/13 - 09/10/13	99	01:41:33	04:16:02
Turkish	01/07/13 - 07/07/13	120	02:35:20	01/10/13 - 09/10/13	99	01:41:33	04:16:53

Table 3: Amount of Web and TV data collected for development and evaluation for the 10 EU-BRIDGE languages. Polish TV (*) comes from YouTube Euronews channel. Last column report the whole duration of the speech data (hh:mm:ss).

about mid August 2013. This correspond to about 3 hours of speech, so we did a second acquisition in October 2013. Meanwhile, Euronews is migrating to YouTube, so this last dataset comes from this source.

In view of future evaluations, a cut-off date is set to June 30th, 2013 for all the languages but Polish, for which it is March, 31st, 2013. A summary of the amount of collected data is reported in Table 4.

	dates	size
Web training	01/01/2013 - 31/05/2013	∼ 109 h
(Polish only)	01/01/2012 - 31/03/2013	∼ 59 h
cut-off date	30/06/2013	
(Polish only)	31/03/2013	
Web test	01/07/2013 - 07/07/2013	∼ 02:35:00
(Polish only)	01/04/2013 - 09/08/2013	03:03:37
TV test	01/10/2013 - 09/10/2013	01:41:33
(Polish only)	01/10/2013 - 31/10/2013	01:28:50

Table 4: Summary of information about the Euronews corpus.

5. Data availability

The EU-BRIDGE consortium contacted Euronews and an agreement was signed, which gives to all 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, we are currently discussing an agreement with Euronews and we plan to organize a multilingual evaluation benchmark for IWSLT 2014. An overview of the IWSLT 2012 Evaluation Campaign can be found in (Federico, et al., 2012).

6. Conclusion

In this paper we presented a multilingual benchmark, designed for ASR purposes. For each of the 10 languages, the benchmark is composed of about 100 hours of speech for training (60 for Polish) and about 4 hours for testing purposes. Training data include the audio, some reference text, the ASR output and their alignment, that results in about 50 to 63 hours of speech (28 for Polish) with a reliable transcription, depending on the language. Test data come both from the Web and from TV, and are being manually transcribed. We plan to make public at least part of the benchmark in view of a multilingual ASR benchmark for IWSLT 2014.

7. Acknowledgements

This work has been partially funded by the European project EU-BRIDGE, under the contract FP7-287658. The author wish to thank the EU-BRIDGE partners who contributed to transcriptions; Danijel Korzinek and Krzysztof Marasek from the Polish-Japanese Institute of Information Technology for their support for Polish text processing; Christian Girardi from FBK for its valuable help in downloading and cleaning text and video from the Web. A special thank to Walid Chamak from Euronews who gave us permission to use the corpus inside EU-BRIDGE.

8. References

Bisazza, A. and Gretter, R. (2012). Building a turkish asr system with minimal resources. In *Proceedings of First Workshop on Language Resources and Technologies for Turkic Languages*, Istanbul, Turkey, May.

Falavigna, D. and Gretter, R. (2011). Cheap bootstrap of multi-lingual hidden markov models. In *Proceedings of INTERSPEECH*, Firenze, Italy, August.

Federico, M. and Cettolo, M. and Bentivogli, L. and Paul, M. and Stüker, S. (2012). Overview of the iwslt 2012 evaluation campaign. In *Proceedings of IWSLT*, Hong Kong, December.

Girardi, C. (2011). The hlt web manager. Technical Report 23969, FBK, Trento, Italy.

Giuliani, D. and Gretter, R. (2013). Esperimenti di identificazione della lingua parlata in ambito giornalistico. In *AISV*, January.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Lamel, L., Gauvain, J., and Adda, G. (2002). Lightly supervised and unsupervised acoustic model training. *Computer Speech and Language*, 16(1):115–129.

Schultz, T. (2002). Globalphone: a multilingual speech and text database developed at karlsruhe university. In *INTERSPEECH*.