TED-LIUM: an Automatic Speech Recognition dedicated corpus

Anthony Rousseau, Paul Deléglise, Yannick Estève

Laboratoire Informatique de l'Université du Maine (LIUM)
University of Le Mans, France
firstname.lastname@lium.univ-lemans.fr

Abstract

This paper presents the corpus developed by the LIUM for Automatic Speech Recognition (ASR), based on the TED Talks. This corpus was built during the IWSLT 2011 Evaluation Campaign, and is composed of 118 hours of speech with its accompanying automatically aligned transcripts. We describe the content of the corpus, how the data was collected and processed, how it will be publicly available and how we built an ASR system using this data leading to a WER score of 17.4%. The official results we obtained at the IWSLT 2011 evaluation campaign are also discussed.

Keywords: Speech recognition, Data collection, Unsupervised learning

1. Introduction

This corpus has been created within the context of our participation to the IWSLT 2011 evaluation campaign. The Talk Task consisted in decoding and translating speeches from the TED (Technology, Entertainment, Design) conferences, from English to French. Thus, all the data we used was extracted from the freely available video talks on the TED website[1] despite the fact that training data for the task was not constrained.

The remainder of this paper is structured as follows: first, in section 2., we present our TED-LIUM corpus and the method we used to built it. In section 3., we expose the architecture of our ASR system. Then, in section 4., we detail our participation to the IWSLT campaign. Lastly, this paper concludes on our corpus availability to the community.

2. The TED-LIUM ASR corpus

For our ASR system training, we aimed at using in-domain audio and transcripts, *i.e.* from the TED talks. In order to collect the desired data, we developed a specific tool intended to extract videos and closed captions from the TED website. This led us to dispose of 818 audio files (talks), along with their corresponding closed captions, for a total of 216 hours of audio (192 hours of real speech) distributed among 698 unique speakers. Among these speakers, we identified 129 hours of male speech and 63 hours of female speech.

Unfortunately, the TED closed captions are not verbatim transcripts of what the speakers pronounce in their talks. For instance, they lack speech disfluencies like repetitions or hesitations and some expressions or contractions are either missing or transcribed differently. Moreover, since these "transcripts" are closed captions, their segmentation is adapted to on-screen reading, not to speech recognition, and the timings we could extract aren't precise, thus they can't be directly used to train an ASR system.

[1] http://www.ted.com

2.1. Building and refining the corpus

In order to turn the collected data in a real ASR training corpus, the first step we needed to achieve was to generate proper alignments between the speech and the closed captions.

We started by generating an automatic segmentation of the audio data using our in-house speaker segmentation and clustering tool (*LIUM_SpkDiarization*), presented in (Meignier and Merlin, 2010). First, we initiated the process by decoding all the available audio data using the default acoustic models provided in the CMU Sphinx 3 package and a 4-gram language model trained with the SRILM toolkit (Stolcke, 2002) on all the text contained in the closed captions. Then, using the NIST Scoring Toolkit *sclite* tool compiled with the *diff* algorithm option enabled, we were able to map the unaligned text to our outputs, thus creating rough reference STM files for our audio data, based on the decoding output CTM timings. Doing so, by scoring these decoder outputs against the newly-created reference files, we were able to get a general idea of the quality of our alignments, even if the WER score is not a good metric to measure this. This helped us remove the worst-aligned talks, and left us with 794 talks representing 135 hours of speech: 91 hours of male and 44 hours of female.

This first iteration (the bootstrap) led us to train new acoustic models based on these 135 hours of speech. Then, by performing a forced alignment and decoding all of our speech data again, we were able to generate a more accurate set of reference STM files. In this second iteration, besides the very strict forced alignment between the decoder output and the speech data (in case of alignment issue, the automatic segment is discarded), we only kept the segments where the decoding output and the unaligned text from the closed captions agreed on the first and last word of the considered segment. Using this method, we

were able to keep 779 talks, for an amount of speech of 152 hours, 106 hours of male and 46 hours of female. In spite of the more aggressive filtering performed here, the total amount of speech is superior to the one from the first iteration. This can be explained by the difference in quality and coverage between the default acoustic model used in the first iteration decoding and the one we produced based on our TED collected and selected data from the previous iteration.

Starting from this data, and for a third time, we trained new acoustic models and decoded all of our speech, keeping the forced alignment policy. We then selected only automatic segments which were consistent enough, *i.e.* the segments which were perfectly aligned (word by word) with the original text from the closed captions. This way, we were able to circumvent the fact that this text was approximative.

In the end, our TED corpus is composed of a total of 774 talks, representing 118 hours of speech: 82 hours of male and 36 hours of female. For all iterations, we managed the disfluencies in the following way: the repetitions are transcribed, the hesitations are mapped to a specific filler word and the false starts are not taken into account. Moreover, the filler words are not taken into account in the alignment evaluation process. The table 1 resumes the statistics of our corpus for each iteration, the WER score is computed on the development corpus described in section 2.2.

Acou. model + selection	#Talks	Speech hours	Gender		WER
			M	F	
Original data	818	192	129	63	N/A
Default model + sclite-diff	794	135	91	44	22.6
1st model + 1st&last word	779	152	106	46	20.2
2nd model + all words	774	118	82	36	18.4

Table 1: TED audio training corpus statistics by iteration.

On top of that, in order to enhance our corpus with some diversity and stability, parts of the 1997 English Broadcast News Speech corpus (HUB4) were added in each iteration, for a total of 65 hours of speech (41 hours of man speech and 23 hours of woman speech).

2.2. Development corpus for ASR

When training an ASR system, it is mandatory to dispose of a development corpus well related to the training data, with precise if not exact transcriptions. This helps achieving fine tuning of all weights used by the system. In order to get such data, we took the talks used within the IWSLT 2010 dev and test corpora, and transcribed them manually to get references as precise as possible (since the TED closed captions are not verbatim transcriptions).

In terms of size, this development corpus, composed of 19 talks, represents a total of 4 hours and 13 minutes of speech. Among these, male speech counts for 3 hours and 14 minutes, while female speech represents 59 minutes.

2.3. Availability of the corpus

We plan to release at our local website and on other sites like Voxforge[2] a package containing the corpus as soon as possible. This package will be constituted of every talk we kept, along with their corresponding aligned transcripts and the pronunciation dictionary, which contains about 150k words. In order to cope with legal aspects, the release will be made under the exact same license as the TED material, which is a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported license (CC BY-NC-ND 3.0) (Commons, 2012).

2.4. Characteristics of the final corpus

The table 2 and table 3 respectively summarize the characteristics of the textual and audio data from the final corpus destined to be released.

Characteristic	Train	Dev
Number of talks	774	19
Number of segments	56,8 k	2 k
Number of words	2,56 M	47 k

Table 2: TED-LIUM corpus text characteristics.

Characteristic	Train	Dev
Total duration	118h 4m 48s	4h 12m 55s
- Male	81h 53m 7s	3h 13m 57s
- Female	36h 11m 41s	58m 58s
Mean duration	9m 9s	13m 18s
Number of unique speakers	666	19

Table 3: TED-LIUM corpus audio characteristics.

The figure 1 shows an example of a reference STM file. This is a classical STM format, one line per segment, where the first field is the show name, the second one the channel, third one is the speaker ID, fourth and fifth are start and end times, sixth is band and genre identification, the rest of each line being the actual text of the segment. The numbers between brackets correspond to the pronouncing variants present in our dictionary. The text between braces matches with the different fillers present in our phonemes list. All text is lowercased, and there is no punctuation at all.

[2] www.voxforge.org

```
CraigVenter_2005G 1 S11 31.82 43.00 <F0_M> what(2) i'm(2) going to tell you about in my eighteen minutes is
{FILL3} how we're <sil> about to switch from reading(2) the genetic code {FILL1} to(2) {FILL3} the first
stages of beginning <sil> to write <sil> the code ourselves <sil>

CraigVenter_2005G 1 S11 43.65 57.16 <F0_M> {FILL2} it's {FILL2} only {FILL2} ten {FILL2} years ago {FILL3}
this <sil> month when(2) {FILL2} we published the(2) first sequence of a(2) free living organism that {FILL3}
of {FILL3} haemophilus {FILL2} influenzae {FILL3} that(2) {FILL4} took {FILL3} a(2) genome project from
{FILL4} thirteen years {FILL1} down to four months {FILL1} <sil>

CraigVenter_2005G 1 S11 57.97 68.76 <F0_M> we can(2) now do that same genome project in the order <sil> of
<sil> two to(2) {COUGH} eight hours {FILL4} so in the last(2) decade a large number of genomes have been
added {FILL1} most human {FILL2} pathogens {FILL4}

CraigVenter_2005G 1 S11 69.78 76.96 <F0_M> a couple of plants {FILL3} several <sil> insects {FILL4} and
several mammals including {FILL3} the(2) human genome <sil>

CraigVenter_2005G 1 S11 88.15 96.35 <F0_M> it's on the(2) order of several hundred(3) {FILL4} we {FILL3} just
got a(2) grant from the gordon and betty moore foundation to sequence one(2) hundred and {FILL1} thirty
genomes this year <sil>

CraigVenter_2005G 1 S11 96.76 104.10 <F0_M> as(2) a side {FILL5} project {FILL1} from {FILL2} environmental
organisms <sil> so the rate of reading(2) the genetic code has changed <sil> <sil>

CraigVenter_2005G 1 S11 104.66 112.38 <F0_M> but as we look {FILL5} what's out there {FILL5} we've barely
scratched the surface {FILL1} on what {FILL3} is available {FILL4} on this planet <sil>

CraigVenter_2005G 1 S11 136.66 149.13 <F0_M> and(2) on(2) the(2) order of <sil> ten million {FILL3} viruses
<sil> less than(2) five thousand microbial species have been characterized as(2) of two years ago <sil> and
so we decided to do something about it and(2) we started the sorcerer {FILL1} ii {FILL2} expedition {FILL4}

CraigVenter_2005G 1 S11 149.49 156.44 <F0_M> where {FILL5} we <sil> were(2) <sil> as with(2) great
oceanographic expeditions trying to sample {FILL4} the(2) ocean every(2) two hundred miles <sil> <sil>
```

Figure 1: Example of a STM reference file.

3. Architecture of the LIUM's ASR system

3.1. Vocabulary and language modeling

In order to select the optimal vocabulary for our system, we trained unigram language models on each monolingual corpus proposed for the IWSLT 2011 task, plus TED and HUB4. Using these models, we interpolated them to get a global unigram model whose interpolation coefficients were optimized to minimize the perplexity on the IWSLT development data described previously. That global model was then sorted according to the word probabilities in reverse order, which allowed us to select the most likely words appearing in the corpora, a method described in (Allauzen and Gauvain, 2004).

For our system, we selected the 150k most likely words, and we added all of the TED and HUB4 words that were not already retained within the 150k words to ensure that our system training would be consistent. This left us with a vocabulary size of 157,6k words. Pronunciations for this vocabulary were taken from the CMU dictionary (*CMUdict v 0.7a*. Missing pronunciations were generated using the Festival Speech Synthesis System (CSTR, 2012).

To train our language models (LM), we used the SRILM toolkit. The selected vocabulary is exactly the same as the one described above to keep the system's consistency. We trained several 4-gram LMs, one for each monolingual corpus, which were then interpolated to create the final LM, a 4-gram back-off model with modified Kneser-Ney discounting. The interpolation weights are computed with an EM procedure, using the textual data from our development corpus mentioned above. Given the vocabulary limited size of our system, we didn't apply any cut-offs on the final language model.

3.2. Description

Our in-house ASR system is a five-pass system based on the open-source CMU Sphinx framework (version 3 and 4), quite similar to the LIUM'08 french ASR system described in (Deléglise et al., 2009). The acoustic models were trained in the same manner, to the exception that we added a multi-layer perceptron (MLP) using the Bottle-Neck feature extraction as described in (Grézl and Fousek, 2008).

The input speech representation of our MLP is a concatenation of nine frames of thirty-nine MFCC coefficients (twelve MFCC features, energies, Δ and Δ^2 derivatives). The topology of the MLP is the following: the first hidden layer is composed of 4000 neurons, the second one, used as the decoding output, of 40 neurons and the third one, used for training, of 123 neurons (41 phonemes, 3 states per phoneme). For the decoding, we first perform a Principal Components Analysis (PCA) transformation on the 40 parameters. Then two streams are decoded: the

first one is composed of the 40 parameters from the PCA transformation while the second one is made of 39 standard PLP features. The streams likelihoods are weighted in order to obtain a resulting likelihood dynamic similar to one single PLP stream. Training of the MLP features is performed using the ICSI QuickNet libraries (ICSI, 2012).

Here is a summary of the five passes performed by the system for decoding:

- # 1 The first pass uses generic acoustic models and a 3-gram language model.
- # 2 The best hypotheses generated by pass # 1 are used to compute a CMLLR transformation for each speaker. The decoding # 2, using SAT and Minimum Phone Error (MPE) acoustic models with CMLLR transformations, generates word-graphs.
- # 3 During the third pass, the computed MLP features are used to rescore the word-graphs obtained during the second pass.
- # 4 The fourth pass consists in recomputing the linguistic scores from the updated word-graphs of the third pass with a 4-gram language model.
- # 5 Finally, the last pass generates a confusion network from the word-graphs and applies the consensus method to extract the final one-best hypothesis.

For a better consistency, the system was learned with lowercased text and no punctuation at all.

4. The IWSLT 2011 evaluation campaign

The International Workshop on Spoken Language Translation (IWSLT) is an annually scientific workshop, associated with an open evaluation campaign on spoken language translation, where both scientific papers and system descriptions are presented. The 8th International Workshop on Spoken Language Translation took place in San Francisco, USA on December 08 and 09, 2011 (Federico et al., 2011).

In this context, we participated in several tasks, including the ASR Talk one. This task was to recognize the recordings made available by TED on their website, which is a repository of public speeches held in English, covering a variety of topics. While the speech in TED lectures is rather well articulated and recorded in high quality, issues arise from the large domain due to the many varying topics that can be the subject of a TED talk, and from the fact that talks in English are also often given by non-native speakers.

The official results from the IWSLT 2011 organizers showed that our system performed well, be it on development and test data from IWSLT 2010, where it was the second best system, or on test data from this year's campaign, where it was the third best system.

Regarding the size of other systems training data, we know that the best performing one, from MIT, used approximatively the same collection technique than us, except that they filtered utterances with a Word Error Rate superior to 20%. This yielded 164 hours of audio data from TED talks (Aminzadeh et al., 2011). The second best performing system, used about 450 hours of audio from various sources (EPPS, HUB4, Quaero...) but no TED audio data. Official results are presented in the table 4.

Data set	LIUM system (WER)	Best system (WER)	2nd system (WER)
Dev 2010	19.2%	17.8%	21.2%
Test 2010	18.2%	15.8%	19.7%
Test 2011	17.4%	15.3%	17.1%

Table 4: Official results for LIUM ASR system, in WER.

5. Conclusion

In this paper, we presented a new corpus dedicated to Automatic Speech Recognition named TED-LIUM. This corpus has been built in an unsupervised way, based on iterations refining the alignment between audio data and raw text from closed captions. It represents a total of 118 hours of speech, with corresponding transcripts, for a global Word Error Rate of 18.4 percent. A manually transcribed development corpus accompanies the training corpus, for a total of 4 hours of speech. This corpus will be released for public availability in the near future, as we believe that it could be useful to the community.

In a second part, we also described the Automatic Speech Recognition system built upon this corpus, a five-pass in-house system based on CMU Sphinx, with the addition of a multi-layer perceptron. We then detailed his system's participation in the IWSLT 2011 evaluation campaign, where it ranked third (17.4 percent WER).

6. References

A. Allauzen and J.-L. Gauvain. 2004. Construction automatique du vocabulaire d'un système de transcription. In *Journées d'Étude sur la Parole*.

A. Ryan Aminzadeh, Tim Anderson, Ray Slyh, Brian Ore, Eric Hansen, Wade Shen, Jennifer Drexler, and Terry Gleason. 2011. The MIT-LL/AFRL IWSLT-2011 MT System. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, San Francisco (CA), December.

Creative Commons. 2012. Cc by-nc-nd 3.0. `http://creativecommons.org/licenses/by-nc-nd/3.0`.

CSTR. 2012. The festival speech synthesis system. `http://www.cstr.ed.ac.uk/projects/festival`.

P. Deléglise, Y. Estève, S. Meignier, and T. Merlin. 2009. Improvements to the LIUM french ASR system based

on CMU Sphinx: what helps to significantly reduce the word error rate? In *International Conference on Spoken Language Processing (Interspeech 2009)*, Brighton (United Kingdom), 6-10 september.

M. Federico, L. Bentivogli, M. Paul, and S. Stueker. 2011. Overview of the IWSLT 2011 evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, San Francisco (CA), December.

F. Grézl and P. Fousek. 2008. Optimizing bottle-neck features for LVCSR. In *2008 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008)*, pages 4729–4732. IEEE Signal Processing Society.

ICSI. 2012. Quicknet. `http://www.icsi.berkeley.edu/Speech/qn.html`.

S. Meignier and T. Merlin. 2010. LIUM SpkDiarization: an open source toolkit for diarization. In *CMU SPUD Workshop*, Dallas (Texas, USA), mars.

A. Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *International Conference on Spoken Language Processing (Interspeech 2002)*, pages 257–286, November.