REAL-TIME SPEECH RECOGNITION CAPTIONING OF EVENTS AND MEETINGS

*Gilles Boulianne, Maryse Boisvert, Fre´de´ric Osterrath*

# Centre de recherche informatique de Montre´al 550 Sherbrooke St. West, suite 100

Montre´al, Que´bec, Canada

### ABSTRACT

Real-time speech recognition captioning has not progressed much, beyond television broadcast, to other tasks like meet- ings in the workplace. A number of obstacles prevent this transition, such as proper means to receive and display cap- tions, or on-site shadow speakers costs. More problematic is the insufﬁcient performance of speech recognition for less formal and one-time events. We describe how we developed a mobile platform for remote captioning during trials in several conferences and meetings. We also show that sentence se- lection based on relative entropy allows training of adequate language models with small amounts of in-domain data, mak- ing real-time captioning of an event possible with only a few hours of preparation.

***Index Terms*—** Speech recognition, handicapped aids, unsupervised learning.

### INTRODUCTION

The high cost and chronic shortage of stenographers and sign language interpreters has spurred the recent development of real-time speech recognition captioning for TV broadcasts, and several systems have been successfully deployed in a number of languages [1][2], including our French system [3]. To accommodate the wide range of channel quality and speaking styles, and be able to summarize when desirable, these systems use shadow speakers (also called re-speakers or voice writers) who act as intermediates between the raw audio signal and the speech recognition system. Because shadow speakers can be trained in weeks rather than years, these systems help reach the 100% closed-captioning target

set in many countries by television regulatory authorities.

However, these systems cannot be yet deployed in the workplace, for daily meetings or even large conferences, and there the deaf and hard-of-hearing community contin- ues to face a shortage of stenographers and sign language interpreters.

Meetings and conferences present several new challenges when compared to broadcast programs. First, there is a lack of

This work was done as part of the E-Inclusion research network funded in part by Heritage Canada New Media Funds.

proper equipment, because none of the usual broadcast equip- ment for transmitting audio and video feeds is available, and there is no TV set to receive, decode and display the captions. More importantly, speech recognition itself becomes more difﬁcult in meeting situations: the mismatch in dic- tionaries, language models, and speaking style produces error rates 5–10 times higher than for read broadcast speech [4]. This mismatch between the actual task, and the data that was available to train the system, arises because there is very little in-domain data available in advance for meetings or one-time events. In contrast, broadcast programs are better matched with large written text archives, and transcribing already aired

programs provides another well-matched source of text.

This work describes our progress in solving these chal- lenges, moving in steps from TV broadcasts to large events and then meetings. In the next section, we will shortly de- scribe our current system for broadcast captioning; following sections will present our new solutions for mobile equipment and the language modeling problem. In section 4 we will re- port our results in actual live caption production of several meetings, conferences and events.

### LIVE CAPTIONING OF TV BROADCASTS

Our real-time system for broadcast television was described in more details in [3], and is currently used in production by several broadcasters, as well as provided as a remote service to several others. Remote captioning is an important element in reducing cost and personnel shortage, because shadow speakers do not have to be present at the captioned event.

### Remote captioning for TV broadcast

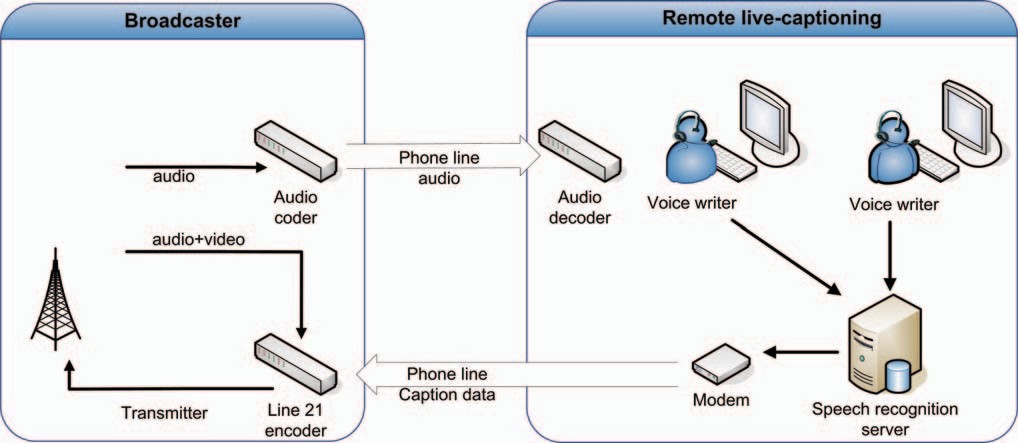
In the typical broadcast setup illustrated in ﬁgure 1, audio from the broadcaster is transmitted through a specialized low- delay coder over a phone line. Video provides context to shadow speakers and can be taken from a cable or satellite feed (delays are acceptable). The audio is decoded at the re- mote captioning site, repeated by the shadow speakers, and converted into captions by the speech recognition server. Cap- tions are streamed back to the broadcast encoder by modem over a second phone line. This encoder inserts the captions

978-1-4244-3472-5/08/$25.00 ©2008 IEEE

197

## SLT 2008

into line 21 of the broadcast video signal prior to transmis- sion. Television receivers are equipped with a mandatory cap- tion decoder which extracts captions from line 21 and displays them in real-time on the TV screen.



**Fig. 1**. Typical setup for broadcast captioning.

### Language model and vocabulary selection

Background texts are taken from a collection of French Cana- dian newspapers and broadcaster news archives which pro- vides a total of a 175 million words of text.

In-domain texts are obtained from the broadcaster (when available) or more typically collected from manual correction of captions produced from actual programs. A development subset and a test subset are witheld from this data. Interpola- tion weights are chosen to minimize perplexity on the devel- opment subset [5] for a mix of unigram models from all back- ground and in-domain sources. These weights are then used to select the vocabulary following [2], then trigram models are interpolated to optimize the perplexity for this vocabulary on the development set. The ﬁnal language model is entropy- pruned [6] to ﬁt memory and real-time constraints, and con- verted into a ﬁnite-state transducer with roughly 2 - 3 million transitions.

Table 1 shows the amount of in-domain texts used for our current production models, their perplexity on the test set, and word error rates obtained for actual live caption productions. These techniques allow us to create successful language mod- els using moderate amounts of text data; a new domain can be trained in a few weeks by transcribing previously recorded programs.

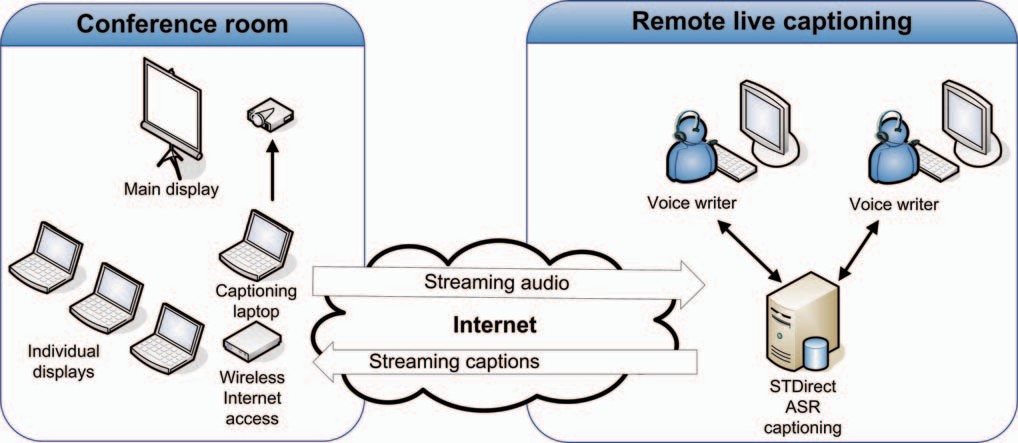
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Domain** | **In-domain** | **Perplexity** | **Test** | **WER** |
| Parliament debates | 73.4 M | 81 | 76 h | 6.9% |
| Hockey play | 3.5 M | 59 | 63 h | 5.1% |
| Hockey analysis | 3.5 M | 70 | 17 h | 6.1% |
| Political reviews | 1.0 M | 74 | 24 h | 7.9% |
| Interviews | 268 K | 86 | 7 h | 8.7% |

**Table 1**. N. of in-domain words, perplexity, test size and word error rates for current broadcast tasks.

### LIVE CAPTIONING OF MEETINGS AND EVENTS

Meetings and conferences take place in unpredictable envi- ronments, where means of capturing the audio and transmit- ting to the remote captioning site, as well as receiving the captions and display them, may not be available. We started by reproducing the TV broadcast setup in our ﬁrst live events but setting all the various pieces of equipment involved re- quired specialized technicians on each side and the risks of malfunctioning during an actual production were high. We needed a single piece of equipment, easy to move and setup, which would not rely on any installed phone line or Internet connection.

Figure 2 illustrates the solution we arrived at after several experiments described in section 4. All communications go through a portable wireless long-range modem. Audio and video are captured on-site with a laptop and sent with Skype to the remote captioning site. The captions are pushed by a Web server with dynamic HTML and can be displayed in real- time on any Web browser on individual laptops, in a format mimicking a TV decoder. If needed the browser screen can be projected from the capture laptop to a large screen. This ﬁnal solution provides a good quality display and receives positive comments from the users.



**Fig. 2**. Final mobile equipement and caption delivery setup.

### Language model and vocabulary selection

Meetings and one-time events are difﬁcult speech recognition problems, even when the acoustic problems are solved using close-talking microphone and shadow speakers, because of the difﬁculty in obtaining training data matched for vocabu- lary, language model, and speaking style [4].

In these situations, texts which are deemed pertinent are manually collected from the Web or other sources, and these constitute the in-domain texts. The amount so collected will be in the 30,000 to 50,000 word range, and will contain around 3000-4000 unique words. In preliminary experiments, we tried our interpolation-based methods on these tasks, and we obtained perplexities well over 100, which is quite high compared with the perplexities for our broadcast data, which range from 60 to 90 (see table 1). These perplexities lead to word errror rates of 15% and more.

* + 1. *Entropy-based sentence selection*

Given a small amount of in-domain text, several methods have been proposed to extract new relevant texts from the Web or other sources to train a language model. These methods select sentences or documents that best match in-domain texts based on measures such as perplexity, information retrieval TF-IDF or BLEU score [7][8]. These methods tend to select sentences which are already well represented in the language model. To address this shortcoming, [9] proposed an approach for building topic-dependent language models by selecting a set of sentences which has a distribution as similar as possible to the in-domain data distribution, instead of being concentrated at the center of it.

The sequential and greedy algorithm starts with the lan- guage model *P* built from in-domain data, and an initial se- lection of sentences. The entropy of the current selected sen- tence set, relative to the in-domain model *P* is:

( 1) = ( ) ln *P* (*i*)

Σ*H j − − P i*

*W* (*i*)*/N*

*i*

where *W* (*i*) is the set of counts for words *i* in the selected set, and *N* = *i W* (*i*) is the total number of words.

Σ

The corpus is scanned sentence by sentence. Let *mij* be

Σ

the count of word *i* in the *jth* sentence *sj*, and *nj* = *i mij* the number of words in *sj*. Sentence *sj* is added to the current selected set if it reduces the entropy relative to *P* :

( ) = ( ) ln *P* (*i*)

Σ*H j − P i*

(*W* (*i*)+ *mij*)*/*(*N* + *nj*)

*i*

Direct computation of this expression would have a very high computational cost. The algorithm described in [9] pro- vides an efﬁcient way to compute the relative entropy change for every sentence in the corpus.

Applying this to our problem, we took the background texts themselves as the corpus from which sentences are se- lected. Using the algorithm and in-domain texts, we created a set of selected sentences amounting to about 10% of the back- ground texts. The sentence set was then used as an additionnal source in the usual interpolation method.

Table 2 shows the results on two domains with less than 50 K words of in-domain data. The baseline test set perplex- ity (column 3) was obtained by interpolating models from the background sources and the in-domain text. Column 4 shows the test set perplexity obtained by adding a source created by relative entropy selection of sentences from the same background texts as the baseline. In both cases, we observed reductions in perplexity (from 113 to 107 and 116 to 95). Columns 5 and 6 show a 10% reduction of out-of- vocabulary rate between the baseline, obtained by applying our vocabulary selection technique to the background and in-domain sources only, or after adding the relative entropy

selected texts. In contrast, applying the relative entropy se- lection to domains in Table 1 improved only the one with the smallest amount of in-domain data, *Interviews*, which saw its perplexity reduced from 86 to 81.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Domain** | **In-domain**  **words** | **Perplexity** | | **% oov** | |
| **base** | **entr** | **base** | **entr** |
| BdC  IRD3 | 48 K  31 K | 113  116 | 107  95 | 0.74  2.0 | 0.66  1.8 |

**Table 2**. Perplexity and out-of-vocabulary rates for baseline and entropy-based text selection.

### EXPERIMENTS

Here we give some details about the improvements on equip- ment and techniques in a string of experiments designed to progress from easier to more difﬁcult tasks.

*TRANSED and PresCup -* The ﬁrst two events in Table 3 were the opening and closing ceremonies of media-covered events. TRANSED was the 11*th* International Conference on Mobility and Transport for Elderly and Disabled Persons held at the Congress Center of Montreal. PresCup was the Presi- dent’s Cup golf tournament held at the Royal Montreal Golf Club by the PGA. Both were large public events and fully sup- plied with audio-visual equipment, including projected video close-ups of the talkers on large screens. For these two events we were able to use a typical broadcast setup: audio was transmitted over a phone line using a codec at each end, dig- ital video was sent over a leased T1 line with a low delay coder, and captions were returned by modem and phone line to an on-site caption encoder for video. Properly setting all the equipment was a complex process which involved spe- cialized teams on both local and remote sites. PresCup was held in an outdoor location, and installation complexity was compounded with random electricity cut-offs and problems with temporary phone lines.

*BdC -* The next event in Table 3 was a large tradeshow also held at the Congress Center of Montreal. This time the audio/video installation was simpler, both being sent with Skype through a provided Internet connection. A video cap- tion encoder installed on-site was fed the video used for projecting close-ups of the talker on large screens. There were phone line disconnects and problems with audio levels coming from the simultaneous interpreters console, impair- ing the respeakers’ ability to acurately understand spoken content.

*IRD meetings -* IRD1, IRD2 and IRD3 in Table 3 were the annual general assembly, a children work group meeting, and a user committee meeting. There were deaf and hard of hear- ing participants in each. Participants answered a short survey about understandability, delivery rate, delay and readability of the captions.

For IRD1, a java applet running on a laptop requested cap- tions from a Web server, and they were projected to a screen in front of the audience. The Internet trafﬁc on the provided connection resulted in audio problems and an uneven display rate that made captions difﬁcult to read. For IRD2, a bet- ter display was found by having a simple terminal applica- tion connected to the speech recognition server through a pri- vate virtual connection. For IRD3, all data was transmitted through a long-range wireless Internet modem. Captions were displayed in a browser on individual laptops using server push and dynamic HTML. Laptops were positioned in front of each deaf or hard-of-hearing participant, so as to be in line-of-sight with sign language interpreters. This is the ﬁnal conﬁguration illustrated in Figure 2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Event** | **Type** | **Part.** | **WER** | **Rating** |
| TRANSED | Conference plenary | 600 | 6.0% | - |
| PresCup | Media event | TV | 7.2% | - |
| BdC | Conference plenary | 1200 | 9.6% | - |
| IRD1 | Annual meeting | 48 | 10.6% | 53% |
| IRD2 | Committee meeting | 22 | 9.6% | 58% |
| IRD3 | Committee meeting | 13 | 10.7% | 94% |

**Table 3**. Live-captioned events, type, number of participants, word error rate and overall user rating.

### DISCUSSION

For TRANSED and PresCup, we had access to moderate amounts of in-domain texts, and we used the normal in- terpolation method to create language models, but for the remaining tasks, we used relative entropy. The word error rates (column 4 of Table 3) for the ﬁrst two events were rel- atively good, as should be expected in these rather formal settings. For BdC, the task was to caption ofﬁcial opening ceremonies and invited talkers David Suzuki and Thierry Vandal. For these speeches we had only a one paragraph summary; furthermore, the talkers were using a more infor- mal style and tended to improvise on other subjects during their talks. This is reﬂected in the higher error rate (9.6%). The worse error rates are observed in the less formal events IRD1, IRD2 and IRD3. Overall, our word error rates still grow worse as less in-domain data is available and the event becomes less formal, but the degradation is graceful.

User rating of caption readability greatly improved from IRD1 to IRD2, from 64% to 77%, and this caused the im- provement in overall user rating seen in Table 3. The ﬁnal setup IRD3 scored very high on all accounts, including un- derstandability, even though the measured word error rate was similar at around 10% for all 3 events. The setup allowed par- ticipants to follow both captions and sign language interpre- tation, and they observed that these were two complementary sources of information.

### CONCLUSION

In order to make practical the transition from real-time cap- tioning of broadcast television to meetings and events, we de- veloped a mobile captioning platform to simplify installation at remote sites and reduce the probability of malfunctions. The method used for displaying the captions had a large im- pact on user acceptance of the technology.

We also showed that for these new tasks, selecting sen- tences from background texts based on a relative entropy cri- terion allowed us to create well-performing language models even with small amounts of in-domain data. The required amount of in-domain texts can be easily collected manually, making real-time captioning of an event possible with only a few hours of preparation.

### REFERENCES

1. A. Lambourne, J. Hewitt, C. Lyon, and S. Warren, “Speech-based real-time subtitling services,” *Interna- tional Journal of Speech Technology*, vol. 4, no. 7, pp. 269–279, 2004.
2. A. Matsui, H. Segi, A. Kobayashi, T. Imai, and A. Ando, “Speech recognition of broadcast sports news,” Tech. Rep. No. 472, NHK Laboratories, 2001.
3. G. Boulianne, J.-F. Beaumont, M. Boisvert, J. Brousseau,

P. Cardinal, C. Chapdelaine, M. Comeau, P. Ouellet, and

F. Osterrath, “Computer-assisted closed-captioning of live TV broadcast in French,” in *Proc. Interspeech 2006*, Sept 2006, pp. 273–276.

1. A. Waibel, H. Yu, M. Westphal, H. Soltau, T. Schultz,

T. Schaaf, Y. Pan, F. Metze, and M. Bett, “Advances in meeting recognition,” in *Proc. HLT 2001*, 2001, pp. 11– 13.

1. A. Stolcke, “SRILM - an extensible language modeling toolkit,” in *Proc. ICSLP 2002*, Sept 2002, pp. 901–904.
2. A. Stolcke, “Entropy-based pruning of backoff language models,” in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 270–274.
3. L. Chen, J.-L. Gauvain, L. Lamel, and G. Adda, “Unsu- pervised language model adaptation for broadcast news,” in *Proc. ICASSP 2003*, Apr 2003, vol. I, pp. 220–223.
4. R. Sarikaya, A. Gravano, and Y. Gao, “Rapid language model development using external resources for new spo- ken dialog domains,” in *Proc. ICASSP 2005*, March 2005, vol. Volume 1, pp. 573–576.
5. A. Sethy, G. G. Panayiotis, and S. Narayanan, “Text data acquisition for domain-speciﬁc language models,” in *Proc. EMNLP 2006*, 2006, pp. 382–389.