MODEL ADAPTATION FOR SENTENCE SEGMENTATION FROM SPEECH

Sébastien Cuendet[1,2] *Dilek Hakkani-Tür*[2] *Gokhan Tur*[3]

Ecole Polytechnique Fédérale de Lausanne[1]
International Computer Science Institute[2]
{cuendet,dilek}@icsi.berkeley.edu

SRI International[3]
Speech Technology and Research Lab
gokhan@speech.sri.com

ABSTRACT

This paper analyzes various methods to adapt sentence segmentation models trained on conversational telephone speech (CTS) to meeting style conversations. The sentence segmentation model trained using a large amount of CTS data is used to improve the performance when various amounts of meeting data are available. We test the sentence segmentation performance on both reference and speech-to-text (STT) conditions on the ICSI MRDA Meeting Corpus using the Switchboard CTS Corpus as the out-of-domain data. Results show that the sentence segmentation performance is significantly improved by the adapted classification model compared to the one obtained by using in-domain data only, independently of the amount of in-domain data used: 17.5% and 8.4% relative error reductions with only 1,000 and 3,000 in-domain sentences, respectively, and 3.7% relative error reduction with all in-domain data of 80,000 words.

1. INTRODUCTION

Sentence segmentation from speech is part of a process that aims at enriching the unstructured stream of words outputted by standard speech recognizers. Its role is to find the sentence units in this stream of words. It is of particular importance for speech related applications, as most of the further processing steps, such as parsing, machine translation, information extraction, assume the presence of sentence boundaries [1, 2].

Sentence segmentation can be seen as a binary classification problem, in which every word boundary has to be labeled as a sentence boundary or as a non-sentence boundary. In the usual learning task, when provided with data for domains, such as conversational telephone speech, broadcast news, or meetings, one has to manually label a consequent amount of them to perform automatic learning. This is an extremely time-consuming and thus very costly task. On the other hand, lots of data in various domains have been labeled throughout the years [3, 4, among others]. *Adaptation* is a general concept which can be used to reduce the human labeling effort by using the already available labeled data (*out-of-domain*) to build or improve a classification model for the new data (*in-domain*). To build a more accurate model, one could decide to label some data of the in-domain, what is referred to as supervised adaptation. The case where no labeling is provided for in-domain data is called unsupervised adaptation.

In this work, we mainly focus on supervised adaptation of phone conversations to meetings. Although these two types of speech could look similar because they are both conversational speech, as opposed to other genre, such as broadcast news, they have significant differences (two speakers vs. multi-speaker environment, visual contact with the interlocutor, etc.). These environmental differences are of particular importance for speech irregularities such as disruptions, backchannels and floor grabbers/holders, which are frequent in meeting conversation style. These similarities and differences make phone conversations and meetings ideal candidates for the adaptation task. We perform adaptation of CTS data on different amounts of meeting data and show that the sentence segmentation performance is significantly improved, especially when little or no meeting data are available.

In the following section, we present related work on sentence segmentation and adaptation. In Section 3, we present our adaptation methods. After the presentation of our results in Section 4, we conclude by discussing the current and future issues.

2. RELATED WORK

2.1. Sentence Segmentation from Speech

Different approaches and classifiers have been studied for the sentence segmentation problem. [5] and [6] use a method that combines hidden Markov models (HMM) with N-gram language models containing words and sentence boundary associated with them, i.e. tags [7]. This method is extended with confusion networks in [8]. [9] provides an overview of different classification algorithms (boosting, hidden-event language model, maximum entropy and decision trees) applied to this task for multilingual Broadcast news. Besides the type of classifier, the features have widely been studied. [6, 10] showed how the sentence segmentation task can benefit from prosodic features. Investigations on prosodic and lexical features in the context of phone conversation and broadcast news speech are presented in [10]. More recently, syntactic features were studied as part of a reranking technique in [11].

2.2. Adaptation

In a typical classification problem, given a set of training data $D = \{(x_n, l_n) \in \mathcal{X} \times \mathcal{L} : 1 \leq n \leq N\}$[1], the goal is to find a function $f : \mathcal{X} \rightarrow \mathcal{L}$, where $\mathcal{X}$ is the feature space, $\mathcal{L}$ is the finite set of possible labels, N is the number of training examples x_n and their associated label l_n. The underlying assumption is that a distribution $p(x_i, l_i)$ exists for each $(x_i, l_i) \in \mathcal{X} \times \mathcal{L}$, but is unknown. In the adaptation problem, we assume *two* data sets, the *in-domain* (or task specific domain) $D^{(i)}$ and the *out-of-domain* $D^{(o)}$ data sets, with $|D^{(i)}| << |D^{(o)}|$. The goal is to find a function $f(x)$ that can predict the classification label l for each example in $D^{(i)}$ by using $D^{(i)}$ *and* $D^{(o)}$. This makes it clear that we assume the distributions $p^{(o)}(x_i, l_i)$ and $p^{(i)}(x_i, l_i)$ of the out-of-domain and the in-domain

[1] We use the same notation as in [12].

respectively not to be independent, in which case $D^{(o)}$ would be useless for classifying $D^{(i)}$.

As far as we know, model adaptation has never been applied to the problem of sentence segmentation. It has however been shown to be useful in other speech processing tasks, such as language modeling (LM) and probabilistic context free grammars using the *maximum a posteriori* adaptation (MAP) method [13, 14]. LM adaptation using linear interpolation and training data filtering is presented in [15]. Adaptation combined with active learning for spoken language understanding is presented in [16]. While these techniques consider one distribution for the in-domain and one for the out-of-domain, a recent work introduces the idea of learning one general distribution, and then using this in conjunction with the in-domain and out-of-domain data [12]. This approach has however not yet been applied to speech related tasks.

3. APPROACH

The adaptation methods that we present in this section are independent of the classifier, except for adaptation with boosting. We chose to use the AdaBoost.MH algorithm[2], which has been shown to be among the best classifiers for the sentence segmentation task [9]. Boosting is an iterative procedure that builds a new weak learner h_t at each iteration. Every example of the training data set is assigned a weight. These weights are initialized uniformly and updated on each iteration so that the algorithm focuses on the examples that were wrongly classified on the previous iteration. At the end of the learning process, the weak learners used on each iteration t are linearly combined to form the classification function:

$$f(x, l) = \sum_{t=1}^{T} \alpha_t h_t(x, l)$$

with α_t the weight of the weak learner h_t and T the number of iterations of the algorithm. More details on Boosting can be found in [17].

In this work, a sample is represented by 9 features which contain lexical information (combination of word unigrams, bigrams and trigrams) and the pause duration between two words.

3.1. Adaptation Methods

The goal of this work is to use the existing labeled data or models to improve the classification performance in a new domain. The combination of the two sets of data can be implemented at different levels, such as the data level (e.g. concatenation), the feature or classifier level (e.g. boosting adaptation) and the classifier output level (e.g. linear interpolation). Note that these different implementations of adaptation *are* equivalent under certain conditions. For example while one can interpolate the outputs obtained by two models, the same effect can be represented in a single interpolated model, as typically done in language models [18]. Similarly, data concatenation can be seen as an unweighted linear interpolation in certain cases.

- **Data concatenation**: the simplest way of combination is to train the classifier on the concatenation of out-of-domain and in-domain data.
- **Logistic interpolation**: each sample of the held-out set is evaluated by the classifier $C^{(i)}$ trained on the in-domain data and the classifier $C^{(o)}$ trained on the out-of-domain data. This evaluation yields probabilities $P^{(i)}(\text{“s”}|x)$ and $P^{(o)}(\text{“s”}|x)$ that the event associated with the sample x is a sentence boundary according to the classifier $C^{(i)}$ (resp. $C^{(o)}$). The final decision is made from the combination of these two probabilities using the logistic function:

$$P(\text{“s”}|\, x) = \frac{1}{1 + e^{(-b_1 - b_2 P^{(i)}(\text{“s”}|\, x) - b_3 P^{(o)}(\text{“s”}|\, x))}}$$

  where b_1, b_2, b_3 are parameters optimized on a held-out set with logistic regression.
- **Using out-of-domain model confidences as an extra feature**: $C^{(o)}$ is run first; the probability it outputs is then used as an extra feature while training a model with the in-domain data. The final decision is made by $C^{(i)}$ trained on this enriched set of features.
- **Boosting adaptation**: Using the same method as in [16] a model is first build with the out-of-domain data and then using boosting adapted to the small amount of in-domain labeled data. This is the same as minimizing a weighted sum of the logistic loss function and the binary relative entropy of the prior probabilities of both models. The weights are optimized using a held-out set.

[2]In this paper, we abusively use the term “Boosting” to designate the AdaBoost.MH algorithm.

	MRDA (bed)	**SWBD**
Training set size (words)	83,959	379,498
Test set size	31,310	-
Held-out set size	29,285	-
Vocabulary size	4,467	13,109
Average utterance length	6.54	7.57

Table 1. Data characteristics for the reference conditions. Sizes and sets are given in number of words.

4. EXPERIMENTS AND RESULTS

We have evaluated the proposed adaptation methods in the case of meetings and phone conversations. Meetings are the target domain and the phone conversations corpus is thus considered as the out-of-domain data. The language of both corpora is English.

4.1. Data Sets and Metrics

The meetings data that we used are from the ICSI meeting corpus (MRDA) [19]. This corpus contains 75 meetings which are grouped in three main types (according to the speakers, the conversations type, etc.). We use the same split of training, test and held-out set as specified in [20]. For reasons of consistency, we limited ourselves to only one type of these meetings (the “bed” type). The phone conversations are the subset of the Switchboard (SWBD) corpus provided by the LDC (RT04). The main characteristics of the data sets are shown in Table 1. In all experiments, we trained the model on the reference transcriptions and tested it on both the reference and the STT transcriptions [21]. The STT transcriptions are automatically obtained from the automatic speech recognizer (ASR) as opposed to the reference transcriptions which have been created by humans on the basis of the audio recording. The STT transcriptions incorporate the errors made by the ASR in the process of recognizing the words (the word error rate on the MRDA corpus is 35.4%) and the classifier

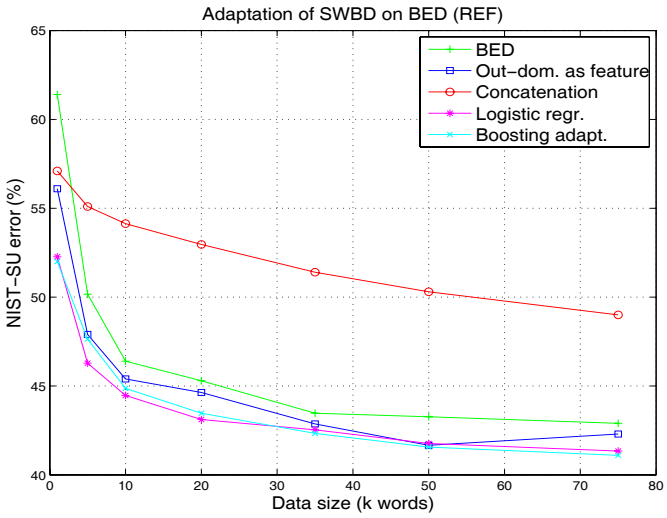

Fig. 1. NIST-SU error for all the methods presented in Section 3; in reference conditions.

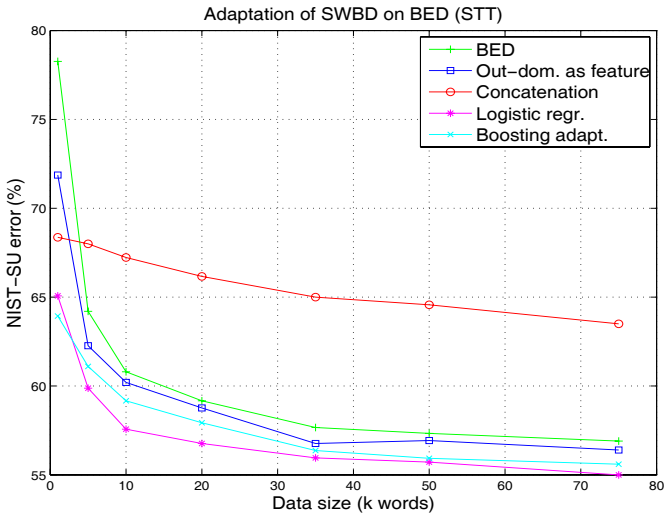

Fig. 2. NIST-SU error for all the methods presented in Section 3; in STT conditions.

performance is thus expected to be worse on them than on the reference transcriptions. The study under the STT conditions is however of big interest in the effort of reducing the human work.

We consider the event associated to an example as a sentence boundary if the posterior probability $P(\text{“s”}|\, x)$ emitted by the classifier for the sample x is bigger than 0.5 (as optimized on the held-out set), and as a non-sentence boundary otherwise.

Metrics. To measure the performance of a classification, we used the F-measure and the NIST-SU error. The F-measure is the harmonic mean of the recall and precision measures of the sentence boundaries hypothesized by the classifier to the ones assigned by human labelers. The NIST-SU error rate is the ratio of the number of wrong hypotheses made by the classifier to the number of reference sentence boundaries. So if no boundaries are marked by sentence segmentation, it is 100%, but it can exceed 100%; the maximum error rate is the ratio of number of words to the number of correct boundaries.

4.2. Results

The learning curves in Figures 1 and 2 show the evolution of the NIST-SU error in reference and STT conditions when the number of training samples of meetings is increased. The F-measure is shown in a similar way in Figures 3 and 4. All results are averaged on 3 experiments with 3 different subsets of the training data as training set.

Boosting adaptation and logistic interpolation are the methods that perform the best when there is very little meeting data available. The “as feature” method is more tightly related to the in-domain model which penalizes it when there is little amount of in-domain data. Logistic interpolation is the method that performs the best independently from the size of the in-domain training data. It reduces the NIST error rate by 17.5% relative for 1k, 8.4% for 3k and 3.7% for 80k, which are all statistically significant improvements according to a Z-test with 95% confidence range.

All figures show that the more meeting data, the smaller the difference between the classifier trained on meetings only and the mixed ones. However, it should be noticed that even with the full meeting training data, all adaptation methods but the data concatenation perform better than the classifier built only on the meeting data.

The performances on the STT conditions show the same pattern as the reference ones, although they are per se lower of 10%-15%. The addition of the ASR error to the classification error can explain this difference. However, in both STT and reference conditions, one would need to label 30k of the meeting data to reach the same performance as the one of out-of-domain data only.

5. DISCUSSION AND CONCLUSIONS

We have presented supervised adaptation methods for sentence segmentation from speech. We have shown that using phone conversations can drastically reduce the error rate on meeting data, especially when these data are scarce. We have also shown that logistic interpolation improves the performance independently of the amount of the meeting data used. One disadvantage of this method is that it requires an extra held-out set to train the regression weights. The results on STT conditions and on the reference are the same qualitatively. Future work includes extending this study to Broadcast news and conversations and finding new ways of interpolation to more effectively take advantage of the out-of-domain knowledge. We also plan to use more prosodic features (we used only the pause duration), because they are intuitively more domain independent than the lexical ones. Unsupervised learning and active learning techniques should also be studied in an effort to reduce the labeling work without decreasing the performance.

Acknowledgments We would like to thank Elizabeth Shriberg, Matthias Zimmerman, Mathew Magimai Doss, and Andreas Stolcke for many helpful discussions. This work was partly supported by the Swiss National Science Foundation through the research network IM2 and Defense Advanced Research Projects Agency (DARPA) GALE (HR0011-06-C-0023) and CALO (NBCHD-030010) fundings at ICSI and SRI, respectively. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

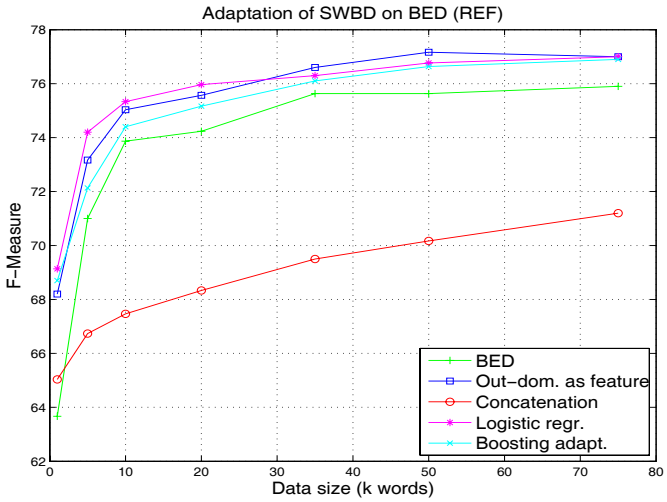

Fig. 3. F-measure for all the methods presented in Section 3; in reference conditions.

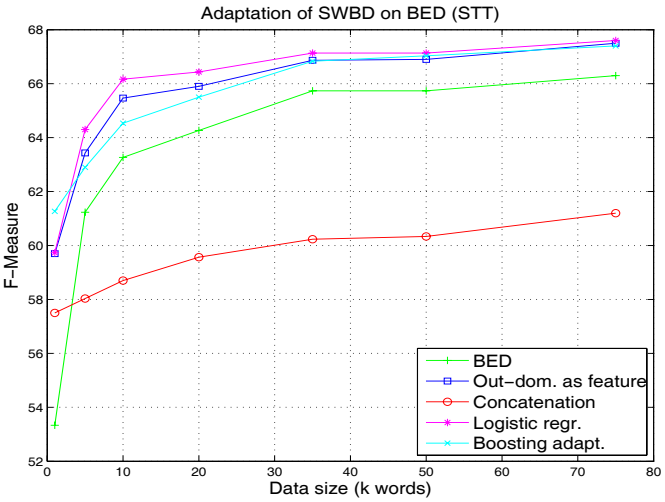

Fig. 4. F-measure for all the methods presented in Section 3; in STT conditions.

6. REFERENCES

[1] Joanna Mrozinski, Edward W. D. Whittaker, Pierre Chatain, and Sadaoki Furui, "Automatic sentence segmentation of speech for automatic summarization," in *Proc. ICASSP*, Philadelphia, PA, 2005.

[2] J. Makhoul, A. Baron, I. Bulyko, L. Nguyen, L. Ramshaw, D. Stallard, R. Schwartz, and B. Xiang, "The effects of speech recognition and punctuation on information extraction performance," in *In Proc. of Interspeech*, Lisbon, 2005.

[3] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proc. ICASSP*, Albuquerque, NM, 1990.

[4] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, , and H. Carvey, "The ICSI meeting recorder dialog act (MRDA) corpus," in *Proc. SigDial Workshop*, Boston, MA, 2004.

[5] Y. Gotoh and S. Renals, "Sentence boundary detection in broadcast speech transcripts," in *Proc. ISCA ITRW Workshop*, Paris, 2000.

[6] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tur, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, 2000.

[7] Stolcke A. and Shriberg E.E., "Automatic linguistic segmentation of conversational speech," in *Proc. ICSLP*, Philadelphia, PA, 1996, vol. 2.

[8] D. Hillard, M. Ostendorf, A. Stolcke, Y. Liu, and E. Shriberg, "Improving automatic sentence boundary detection with confusion networks," in *Proc. HLT-NAACL*, Boston, MA, 2004.

[9] M. Zimmermann, D. Hakkani-Tür, J. Fung, N. Mirghafori, E.Shriberg, and Y. Liu, "The ICSI+ multi-lingual sentence segmentation system," in *Proc. ICSLP*, Pittsburgh, PA, 2006.

[10] Y. Liu, E. Shriberg, A. Stolcke, B. Peskin, J. Ang, D. Hillard, M. Ostendorf, M. Tomalin, P. Woodland, and M. Harper, "Structural metadata research in the EARS program," in *Proc. ICASSP*, 2005.

[11] B. Roark, Y. Liu, M. Harper, R. Stewart, M. Lease, M. Snover, I. Shafran, B. Dorr, J. Hale, A. Krasnyanskaya, and L. Yung, "Reranking for sentence boundary detection in conversational speech," in *Proc. ICASSP*, Toulouse, France, 2006.

[12] H. Daumé III, *Practical Structured Learning Techniques for Natural Language Processing*, Ph.D. thesis, University of Southern California, Los Angeles, CA, 2006.

[13] M. Bacchiani, B. Roark, and M. Saraclar, "Language model adaptation with MAP estimation and the perceptron algorithm," in *Proc. HLT-NAACL*, Boston, MA, 2004.

[14] B. Roark and M. Bacchiani, "Supervised and unsupervised PCFG adaptation to novel domains," in *Proc. HLT-NAACL*, Edmonton, Canada, 2003.

[15] J. Li X. Fang, J. Gao and H. Sheng, "Training data optimization for language model adaptation," in *Proc. EUROSPEECH*, Geneva, Switzerland, 2003.

[16] G. Tur, "Model adaptation for spoken language understanding," in *Proc. ICASSP*, Philadelphia, PA, 2005.

[17] R. E. Schapire, "The boosting approach to machine learning: An overview," in *Proc. MSRI Workshop on Nonlinear Estimation and Classification*, Berkeley, CA, March 2001.

[18] A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit," in *Proceedings of the ICSLP*, Denver, CO, September 2002.

[19] A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, J. Macias-Guarasa, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters, and B. Wrede, "The ICSI meeting project: Resources and research," in *Proc. ICASSP*, Montreal, 2004.

[20] J. Ang, Y. Liu, and E. Shriberg, "Automatic dialog act segmentation and classification in multiparty meetings," in *Proc. ICASSP*, Philadelphia, PA, 2005.

[21] A. Stolcke, X. Anguera, K. Boakye, O. Cetin, F. Grzl, A. Janin, A. Mandal, B. Peskin, C. Wooters, , and J. Zheng, "Further progress in meeting recognition: The ICSI-SRI spring 2005 speech-to-text evaluation system," in *Proc. MLMI*, Edinburgh, Scotland, 2005.