IMPACT OF DYNAMIC MODEL ADAPTATION BEYOND SPEECH RECOGNITION

Fernando Batista[1,2,3], *Rui Amaral*[1,2,4], *Isabel Trancoso*[1,2], *Nuno Mamede*[1,2]

[1]L^2F - Spoken Language Systems Laboratory - INESC ID Lisboa
R. Alves Redol, 9, 1000-029 Lisboa, Portugal
http://www.l2f.inesc-id.pt/
[2]IST – Technical University of Lisbon, Portugal
[3]ISCTE – Instituto de Ciências do Trabalho e da Empresa, Portugal
[4]EST – Escola Superior de Tecnologia de Setúbal

ABSTRACT

The application of speech recognition to live subtitling of Broadcast News has motivated the adaptation of the lexical and language models of the recognizer on a daily basis with text material retrieved from online newspapers. This paper studies the impact of this adaptation on two of the blocks following the speech recognition module: capitalization and topic indexation. We describe and evaluate different adaptation approaches that try to explore the language dynamics.

Index Terms— Speech processing, Speech intelligibility, Natural language interfaces, Language Dynamics, Unsupervised learning

1. INTRODUCTION

The application of automatic speech recognition (ASR) to close-captioning of Broadcast News (BN) has motivated the adaptation of the lexical and language models of the recognizer on a daily basis with text material retrieved from online newspapers [1, 2]. The vocabulary and language model adaptation approaches use 3 corpora as training data: the manual transcriptions of the BN speech training data, a large newspaper text database with 741M words and a relative small adaptation set consisting of the 7 last days of online text. The selection of the 100k dynamic vocabulary is POS-based. Relative to our first ASR version that used a fixed vocabulary of 57k words, the dynamic version achieves a relative reduction of 65% in OOV (out-of-vocabulary) word rate and of 5.7% in WER (word error rate). Roughly half of this improvement is due to the increased size of the vocabulary, as shown by the WER results obtained with a baseline version using a static vocabulary of 100k words.

These improvements have an obvious impact on the quality of the automatically produced subtitles, live since March 2008, on the Portuguese national TV channel (RTP). These subtitles also include online punctuation and capitalization. An offline topic segmentation and indexation module splits the BN show into stories and assigns one or more topics to each story from a closed set of topics. For the time being, a very crude extractive summarization technique also assigns a first-sentence summary to each story.

These post-ASR modules were originally trained with the material available until a certain date, in no way taking advantage of the online newspapers which are daily collected. The goal of this paper is to try to use this data to train better models for the capitalization and the topic indexation modules.

This paper is split into two main sections, the first one dealing with capitalization and the second one with topic indexation. For each of these sections, we shall describe the available corpora and baseline versions, the training of new models and the corresponding results. Although the two sections are quite separate, the last one will try to derive some joint conclusions.

2. CAPITALIZATION

The capitalization task consists of assigning the proper case information to each input word, which may depend on the context. The recognition output benefits from this information in terms of improved readability and for further automatic processing. The capitalization problem has been previously addressed by [3, 4, 5, 6].

Despite the fact that most of the words and constructions of a human language are kept in use for many years or never change, new words are introduced everyday and the usage of others decays with time. [7] conducts a study analyzing the relation between corpora variation over time and the performance of named entity recognition, concluding that as the time gap between corpora increases, the similarity between the corpora and the names shared between those corpora decreases. The language adaptation problem concerning the capitalization has also been addressed by [8, 9], revealing that the capitalization performance is influenced by the training data period.

This section analyses the capitalization performance when performed either with a static capitalization model (CM) or with dynamic capitalization models retrained over time. This work assumes that the capitalization of the first word of each sentence is performed in a separated processing stage (e.g. after punctuation), since its correct graphical form depends on its position in the sentence. Only three ways of writing a word will be considered: lower-case, first-capitalized, and all-upper. Mixed-case words, such as "McGyver", are treated by means of a small lexicon, but not evaluated in the scope of this paper. The evaluation is performed using: Precision, Recall and SER (Slot Error Rate) [10]. Only capitalized words (not lowercase) are considered as slots and used by these metrics. Hence, the SER is calculated by dividing the number of capitalization errors by the number of capitalized words in the reference data.

2.1. Data sources

The capitalization model currently used for BN close-captioning was trained with the content of a newspaper corpus, collected from 1999 to 2004, and containing about 148M words. This CM, denoted as BaseCM, provides the baseline performance for the experiments presented in this paper.

The CM adaptation uses online text, daily collected from the web, and corresponding to last minute news published by the

"Público" newspaper. This corpus, denoted as LMN, is being collected since the beginning of 2005 and currently contains about 30M words. The original text is normalized and all the punctuation marks are removed, making it close to speech transcriptions.

The evaluation is performed over 5 recent BN shows, of about 1 hour each, from the Portuguese public TV channel (RTP). The corpus contains about 40k words and was collected during last June and July, with an 8 day time span between each BN show. The manual orthographic transcription of this corpus provides the reference data, and includes information such as punctuation marks, capital letters and special marks for proper nouns and acronyms. Besides the manual orthographic transcription, we also have available two automatic transcriptions, sharing the same preprocessing segmentation: S_ASR – produced using a static LM and a static 100k word vocabulary; and D_ASR – produced by the current recognition system using a dynamic LM and vocabulary, built specifically for the corresponding day.

Whereas the manual transcriptions already contain a reference capitalization, this is not the case of the automatic transcriptions. The required capitalization reference was produced by means of an alignment between the manual and automatic transcription. The alignment was performed using the NIST SCLite tool[1], followed by an automatic post-processing stage, for correcting possible SCLite errors and aligning compound words which can be written/recognized differently.

2.2. Maximum entropy approach

The capitalization task is performed using a discriminative modeling approach, based on maximum entropy (ME) models, firstly applied to natural language problems by [11]. An ME model estimates the conditional probability of events given the corresponding features. This framework provides a clean way of expressing and combining several knowledge sources and different properties of events, such as word identification and POS (part-of-speech) tagging information. This approach requires all information to be expressed in terms of features, causing the resultant data file to become several times larger than the original one. This constitutes a training problem, making it difficult to train with large corpora. The classification however, is straightforward, making it interesting for on-the-fly usage. The current experiments use only features comprising word identification, sometimes combined as bigrams: w_i (current word); and $\langle w_{i-1}, w_i \rangle$, $\langle w_i, w_{i+1} \rangle$ (bigrams).

The memory required for this approach increases with the size of the corpus (number of observations), preventing or making it difficult to use large corpora for training. For example, training with 2M events requires about 6GB of RAM to process. This problem is solved by splitting the corpus into several subsets, and then iteratively retraining with each one separately. The first subset is used for training the first ME model, which is then used to provide initial weights for the next iteration over the next subset. This process goes on until all subsets are used. Although the final ME model contains information from all corpora subsets, events occurring in the latest training sets gain more importance in the final model. As the training is performed with the new data, the old models are iteratively adjusted to the new data. This approach provides a clean framework for language dynamics adaptation, offering a number of advantages: (1) new events are automatically considered in the new models; and (2) with time, unused events slowly decrease in weight.

[1] available from http://www.nist.gov/speech.

Evaluation set	%Precision	%Recall	%SER
Manual transcription	86.0	87.6	26.6
S_ASR	70.5	78.5	54.0
D_ASR	72.2	80.8	50.1

Table 1. Baseline capitalization results produced using BaseCM, where the SER is shown as an absolute value.

Approach	Model period	%Man	%S_ASR	%D_ASR
baseline	2008-05-20	26.6	54.0	50.1
LMN only	2008-05-20	26.5	**53.3**	**49.5**
adapt-base	2008-05-20	26.0	54.4	50.2
adapt-iter	2008-05-20	**25.0**	**53.8**	**49.6**
adapt-iter	daily model	25.0	53.6	49.8

Table 2. Capitalization SER achieved for all different approaches.

2.3. Baseline results

The baseline results, achieved using the CM currently in use for daily subtitling (BaseCM), are shown in Table 1. The capitalization performance decreases when moving to automatic transcriptions. Even so, a better performance is achieved for the D_ASR transcription, where both the LM and vocabulary are daily computed, and a lower WER is achieved.

2.4. Adaptation results

The adaptation and retraining experiments performed in the scope of this work use LMN corpora subsets of 2M words each, and the previously described retraining method. Each subset is referred by the day corresponding to the latest data in that subset. Accordingly, the CM that results from retraining with a given corpora subset is also referred by the day corresponding to the latest data in that subset.

Three adaptation approaches were tested: (1) using only the LMN corpus for training; (2) adapting the BaseCM to a target period, by retraining with the latest data from that period; and (3) iteratively retraining BaseCM with all the available corpora subsets. While the first approach assumes that using only the most recent data (LMN) is sufficient for training, the other two approaches use this data to retrain the baseline CM, assuming that former data also provide important capitalization information. The second approach assumes that BaseCM already contains most of the capitalization information and a simple retrain with data from a target period is sufficient. The last approach assumes that all corpora periods provide important capitalization information and contribute for a better final model. Table 2 shows the final capitalization results for each approach. Concerning the manual transcription, all the proposed approaches yield better results than the baseline, and the best result is produced using the third approach (lines 4 and 5), which combines the BaseCM with the LMN information. Concerning the automatic transcriptions, the first approach proved to be the best, despite achieving only small improvements over the third approach, specially for the D_ASR transcriptions, currently in use. Results show that the LMN information alone is sufficient to beat the baseline, revealing the importance of training data periods closer to the testing data. The table reveals that results are not further improved by using daily CMs, which corresponds to retraining the 2008-05-20 CM with the latest 2M words former to the testing data (5 daily models were used), suggesting that a periodic retraining is suitable for this task.

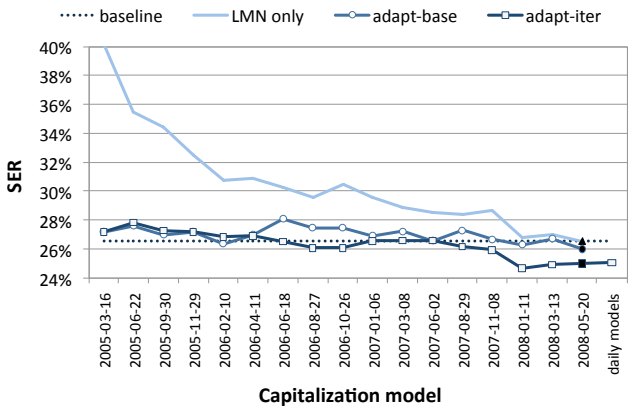

Fig. 1. Manual transcription results, using all approaches.

Topic	Train	Dev	Test	%Acc
National	3558	526	518	83.10
International	1859	227	233	87.13
Economy	946	194	149	90.65
Education	196	17	52	96.22
Environment	235	34	33	94.91
Health	315	90	50	95.69
Justice	496	63	91	94.26
Meteorology	69	7	6	99.41
Politics	1838	285	357	87.24
Security	1037	138	158	87.99
Society	1455	193	260	74.43
Sports	719	118	98	96.86
Total				90.65

Table 3. Number of stories in each topic in the training, development and test sets of the MW corpus, and corresponding accuracy.

Figure 1 illustrates the results achieved for the manual transcription, using different capitalization models and all different approaches. All the approaches depict clear trend lines. However, the capitalization models produced with the third approach are more stable, achieving the best results after a certain period of time.

3. TOPIC INDEXATION

Our recent work in terms of topic indexation involves the classification of stories using a small set of topics adopted by a professional media watch company. Although the stories are automatically segmented, this paper only addresses the classification of manually segmented stories. The fact that this classification is thesaurus-oriented makes it significantly different from the work involved in the TREC SDR Track [12].

Our corpus, denoted as MW (Media Watch), was collected since the beginning of 2007, and was topic-segmented and topic-labelled by the company. The orthographic transcriptions were automatically provided by our ASR system, using the above mentioned fixed 100k vocabulary. The recordings of the RTP daily evening shows were done independently at our lab and at the company, which implied some synchronization problems. The last 9 months of 2007 were used for training topic models (167 shows); January 2008 was used for development (23 shows) and February 2008 (25 shows) was used for testing. Each show has approximately 1h duration, with a single publicity break marked by a jingle detection module.

The number of topics used by the media watch company has evolved during the last year. Whereas during the first months the main distinction was between national and international news, these two broad categories were further subdivided in the following months. In the subdivision, one could distinguish 9 new topics that could be further subdivided, although this hierarchical structure was not consistently used, and was not quite visible. In fact, it was provided to us as an HTML file, in which all topics of each story were written in the same line, separated by a punctuation sign. Table 3 shows the amount of stories for each topic in our training, development and test sets. The topic "meteorology" (or weather forecast) was rarely identified as topic. In fact, the stories on this topic were classified as "national", but included weather forecast as the title of the piece. Because of the importance of this topic for our segmentation module, we extracted the topic information from the title.

For each of the 12 classes, topic and non-topic unigram language models were created using the Good-Touring discount strategy, on the basis of the stories of the MW corpus which were preprocessed in order to remove function words and lemmatize the remaining ones. Topic detection is based on the log likelihood ratio between the topic likelihood $p(W/T_i)$ and the non-topic likelihood $p(W/\overline{T_i})$. The detection of any topic in a story occurs every time the correspondent score is higher than a predefined threshold. The threshold is different for each topic in order to account for the differences in the modeling quality of the topics.

3.1. Baseline results

The last column of Table 3 shows the accuracy values obtained for the MW test corpus. As expected, best results were obtained for the meteorology topic. The bad results obtained for the society topic can be justified by the fact that it is some sort of a miscellaneous topic. The difficulty in assigning the national and international topics may be also partly justified by the fact that stories on how the country is viewed abroad may be manually classified with both topics. This type of indexation will be specially hard to implement automatically.

3.2. Unsupervised topic model adaptation

Our unsupervised topic model adaptation approach is performed daily, on the basis of the 7 last days of newspaper news available from the Internet. The process starts with the automatic indexing of each article in the news. For each of the 12 topics, the topic models based on the ASR transcriptions of the MW corpus and on the recent newspaper texts are linearly interpolated to create a new topic model which will be used in the indexations task of that day. The weights of the linear interpolation of the models are chosen to minimize the perplexity of the interpolated model in an interpolation tuning corpus.

To evaluate the impact of the unsupervised topic models adaptation as well the use of a dynamic vocabulary in the topic indexation task, three experiments were performed using the evaluation set of the MW corpus. In the first scenario, S1, the unsupervised topic models adaptation is performed using a fixed 100k vocabulary and the interpolation tuning corpus is based on newspaper texts from the last 20 days. The newspaper texts of this interpolation tuning corpus are topic indexed to create material for each topic. The second scenario, S2, differs from first since it uses the ASR daily-selected vocabulary. The third and last scenario, S3, differs from the second

Topic	S1 [%Acc]	S2 [%Acc]	S3 [%Acc]
National	83.36	83.25	83.00
International	89.89	89.85	89.76
Economy	91.54	91.53	91.57
Education	98.20	98.20	98.39
Environment	94.23	94.77	95.60
Health	95.47	95.21	96.35
Justice	95.76	95.62	94.76
Meteorology	95.15	95.49	99.23
Politics	88.37	88.25	87.63
Security	90.73	90.88	90.57
Society	74.62	75.43	76.18
Sports	96.66	96.25	96.90
Total	91.17	91.23	91.66

Table 4. Accuracy values for several adaptation scenarios.

one because the interpolation tuning corpus for each topic is created from the transcriptions of the BN shows of the last 20 days. In the first two experiments, S1 and S2, the MW development corpus is used to redefine log likelihood ratio thresholds for each topic, before decoding the broadcast news program. These topic threshold values correspond to the best accuracy results attained in the MW development set. In the S3 experiment, the log likelihood ratio thresholds are fixed. The results concerning the unsupervised topic model adaptation are presented in Table 4.

By comparing the best adaptation scenario (S3) with the fourth column of Table 3, we conclude that the adaptation improves the topic indexation accuracy by 1% absolute. Part of the better results of the S3 scenario are due to an improved meteorology topic model. The S1 and S2 experiments showed a degradation of the meteorology topic model during the adaptation process. Since the unigram statistics of the meteorology model are mainly based on numbers and on a fixed number of cities, some of the newspaper texts concerning lottery or games whose scores are of the same order of magnitude of temperature or precipitation values were decoded with that topic. To overcome this problem in the S3 experiment, some negative examples were extracted from newspaper texts to retrain the meteorology topic. The use of an interpolation tuning corpus based on transcriptions in the S3 experiment allowed us to disregard the recalculation of the log likelihood ratio threshold without decreasing significantly the accuracy result (-0.15%) because this interpolation corpus gives more weight to the topic model based on the ASR transcriptions which is the one most adapted to the evaluation task, since the MW evaluation set is also built from BN shows.

4. CONCLUSIONS AND FUTURE

This paper described the on-going work on adapting two of the blocks that follow the speech recognition module: capitalization and topic indexation. Concerning the capitalization, three different approaches were proposed and evaluated. The most promising approach consists of iteratively retraining a baseline model with the new available data, using fixed corpora subsets. When dealing with the manual transcription the performance grows about 1.6%. Results reveal that producing capitalization models on a daily basis does not lead to a significant improvement. Therefore, the adaptation of capitalization models on a periodic basis is the best choice. Concerning the topic indexation, results achieved in several unsupervised topic models adaptation scenarios were compared. The unsupervised adaptation process takes advantage of the daily newspaper collection. The use of a dynamic vocabulary in the adaptation process has also been explored. The evaluation showed an improvement of 1% using the unsupervised adaptation strategy but the impact of a dynamic vocabulary was not significant. The small improvements gained in terms of capitalization and topic indexation lead us to believe that dynamically updated models may play a small role, but the updating does not need to be done daily, a fact that is also according to our intuition.

The results of the offline processing of each BN show are shown daily on our website[2]. We are currently working on porting our BN processing chain to other varieties of Portuguese (spoken in South America and Africa) and Spanish.

5. ACKNOWLEDGMENTS

The present work is part of Rui Amaral's PhD thesis, initially sponsored by a FCT scholarship. This work was partially funded by the FCT project PoSTPort (PTDC/PLP/72404/2006) and by the European project Vidi-Video. Authors are grateful to Hugo Meinedo for his support during this work.

6. REFERENCES

[1] Ciro Martins, António Teixeira, and João P. Neto, "Dynamic language modeling for a daily broadcast news transcription system," in *Proc. of the ASRU 2007*, 2007.

[2] Ciro Martins, António Teixeira, and João Neto, "Vocabulary selection for a broadcast news transcription system using a morpho-syntatic approach," in *Interspeech 2007*, Sep. 2007.

[3] Lucian Vlad Lita, Abe Ittycheriah, Salim Roukos, and Nanda Kambhatla, "tRuEcasIng," in *Proc. of the 41st annual meeting on ACL*, Morristown, NJ, USA, 2003, pp. 152–159, ACL.

[4] C. Chelba and A. Acero, "Adaptation of maximum entropy capitalizer: Little data can help a lot," *EMNLP '04*, 2004.

[5] Fernando Batista, Nuno Mamede, Diamantino Caseiro, and Isabel Trancoso, "A lightweight on-the-fly capitalization system for automatic speech recognition," in *Proc. RANLP'07*, 2007.

[6] Ji-Hwan Kim and Philip C. Woodland, "Automatic capitalisation generation for speech input," *Computer Speech & Language*, vol. 18, no. 1, pp. 67–90, 2004.

[7] Cristina Mota and Ralph Grishman, "Is this ne tagger getting old?," in *Proc. of the LREC'08*, ELRA, Ed., 2008.

[8] Fernando Batista, Nuno Mamede, and Isabel Trancoso, "The impact of language dynamics on the capitalization of broadcast news," in *Interspeech 2008*, Sep. 2008.

[9] Fernando Batista, Nuno Mamede, and Isabel Trancoso, "Language dynamics and capitalization using maximum entropy," in *Proc. of ACL-08: HLT, Short Papers*. 2008, pp. 1–4, ACL.

[10] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, "Performance measures for information extraction," in *Proc. of the DARPA Broadcast News Workshop*, Herndon, VA, Feb. 1999.

[11] Adam Berger, Stephen Pietra, and Vincent Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[12] J. Garofolo, G. Auzanne, and E. Voorhees, "The TREC spoken document retrieval track: A success story," in *Proc. RIAO'2000*, Paris, France, April 2000.

[2] https://tecnovoz.l2f.inesc-id.pt/demos/asr/legendagem/