Textual Data Selection for Language Modelling in the Scope of Automatic Speech Recognition

Freha Mezzoudj, David Langlois, Denis Jouvet, Abdelkader Benyettou

Textual Data Selection for Language Modelling in the Scope of Automatic Speech Recognition

Freha Mezzoudj*†, David Langlois‡, Denis Jouvet‡, Abdelkader Benyettou*
*Université des Sciences et de la Technologie d'Oran Mohamed Boudiaf, BP 1505, El M'Naouer, 31000, Algérie
†Université Hassiba Benbouali de Chlef, Hai Essalem, 02000, Chlef, Algérie
{freha.mezzoudj, a_benyettou}@yahoo.fr
‡ Speech Group, LORIA
Inria, Villers-lès-Nancy, F-54600, France
Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
CNRS, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
david.langlois@loria.fr, denis.jouvet@inria.fr

***Abstract*—The language model is an important module in many applications that produce natural language text, in particular speech recognition. Training of language models requires large amounts of textual data that matches with the target domain. Selection of target domain (or in-domain) data has been investigated in the past. For example [1] has proposed a criterion based on the difference of cross-entropy between models representing in-domain and non-domain-specific data. However evaluations were conducted using only two sources of data, one corresponding to the in-domain, and another one to generic data from which sentences are selected. In the scope of broadcast news and TV shows transcription systems, language models are built by interpolating several language models estimated from various data sources. This paper investigates the data selection process in this context of building interpolated language models for speech transcription. Results show that, in the selection process, the choice of the language models for representing in-domain and non-domain-specific data is critical. Moreover, it is better to apply the data selection only on some selected data sources. This way, the selection process leads to an improvement of 8.3 in terms of perplexity and 0.2% in terms of word-error rate on the French broadcast transcription task.**

I. Introduction

A Statistical Language Model constitutes one of the key components in several applications that produce natural language texts, such as large vocabulary speech recognition [2], [3], entity disambiguation [4], statistical machine translation [5], information retrieval [6], language text identification [7], handwriting recognition [8] and so on [9].

The goal of Automatic Speech Recognition (ASR) is to accurately and efficiently convert a speech signal into a text message corresponding to the transcription of the spoken words, independently of the device used to record the speech, the speaker, or the environment [10]. The decoder, which is used to identify the pronounced words and sentences, exploits three types of knowledge corresponding to acoustic models which represent the acoustic realisation of sounds, to the lexicon which specifies the possible pronunciations of each word and to the language model (LM) which specifies the possible word sequences.

Our study focuses on French automatic speech transcription systems recently developed around projects and evaluation campaigns of automatic transcription of radio broadcasting programs. The initial ESTER1 campaigns of 2003 and 2005 [11] targeted radio broadcast news, the 2009 edition ESTER2 [12] introduced accented speech and news shows with spontaneous speech. The ETAPE 2011 evaluation [13] focused on TV material with various level of spontaneous speech and multiple speakers speech. The EPAC project of the French National Research agency (ANR) contributed to build the EPAC corpus of conversational speech manually and automatically transcribed [14].

The textual data that matches the best with this task are the textual data corresponding to the manual transcription of radio broadcast shows. This kind of textual data is costly to produce [15]. Hence, the amount of such training data is limited, and this impacts on the performance of a language model trained on this data only.

Besides the training data corresponding to the domain of interest, the development of language models can take benefit of data coming from other sources or other domains. When dealing with heterogeneous corpora, the conventional approach is first, to train an individual language model on each corpus (data source), and then, to combine them so as to maximize the fitting of the resulting (interpolated) language model with some development data representing the target task. Such an approach is used for our baseline LM, which is trained on a large text corpus of about two billion words. The text data comes from heterogeneous sources such as newspapers, news agency reports, web data, and a small quantity of manual transcriptions of radio broadcast programs.

Corpora may be noisy because of the variable quality of the sources. Noise data may lead to under-performing language models. To avoid such phenomenon, it might be useful to select a subset of relevant data for each source corpus.

This paper investigates methods for selecting data from textual corpora in view of improving language modelling for automatic speech transcription of broadcast news and TV shows. The selection methods used rely on computing, for each sentence, a score which represents how close the sentence is to in-domain data compared to non-domain-specific data. Several variants of scoring are proposed and analysed using perplexity measures. Finally, evaluations are conducted with respect to

the automatic speech transcription of radio and TV shows.

The paper is organised as follows. Section 2 exposes the related work on data selection for language modelling and some associated techniques. Section 3 presents the experimental set-up, including the corpora used and the baseline LM. Textual data selection approaches and experiments are presented and discussed in Section 4. Section 5 presents conclusions and research directions for future work.

II. Data selection for language modelling

Classically, a high-performance language model is trained using a small corpus close to the target task (called in-domain data) and a huge data set not close to the task (called general domain data, or non-domain specific data). The in-domain data is carefully prepared by manual transcription. Unfortunately, this high-quality preparation leads to small data. Indeed, high-performance language models must deal with huge data in sake of coverage. To obtain huge data, one uses various sources easily available, often including web data. This leads to huge, but low-quality, general domain data.

This general domain data can contain relevant as well as irrelevant sentences with respect to in-domain data. The use of the irrelevant general domain data is probably more harmful than beneficial. In order to tackle this problem, various approaches were proposed and used in the literature to identify the most relevant portions of the general domain data prior to be used for training target LMs.

Klakow [16] uses a log-likelihood based criterion to select newspapers articles from a training corpus; and proposes two strategies for article removal. In the one-pass strategy, the criterion is computed for each article and then the top scoring articles are selected. The alternative is the iterative strategy, which, for each iteration, calculates the criterion for all articles and remove from the corpus a small fraction of worst scoring articles. This last strategy leads to a LM based on the selected corpus that provides a perplexity 25% lower than the perplexity of the model based on the whole general domain corpus. Shen and Xu [17] use a one-pass paragraph selection based on the perplexity criterion, this leads to improved speech recognition performance. Wang et al. [18] and Gao et al. [19] score each unit (variable number of consecutive sentences) in the general domain corpus by the perplexity according to an in-domain LM, then they retain the units with lowest perplexity. The method of [19] is also applied to sentence selection in the scope of machine translation [20], [21]. Toral [22] uses linguistic information (lemmas and named entities) with a simple perplexity criterion for selecting training data. The resulting models yield lower perplexity than that of the baseline.

These previous works deal with two corpora only, the in-domain data and the general domain (non-domain-specific) data. [16] and [17] select a subset of data from which they directly build a single LM; whereas [18] and [19] build automatically optimal sub-parts of general domain data and then interpolate the corresponding LMs. We face a completely different situation, where four corpora (corresponding to different sources of data) are used, each one having a different "in-domain" degree. This leads to four LMs which are interpolated, and this contrasts with previous works which consider only tow corpora ("in-domain" data and "non-domain-specific" data). Our experimental conditions are therefore very different and more challenging.

Another efficient approach is proposed by Moore and Lewis [1]: two LMs are used for sentence scoring, one is trained on the whole in-domain data and the other one is trained on a random subset of the non-domain-specific data, with a size similar to the in-domain one. Each sentence s from the non-domain-specific data is ranked using the cross-entropy difference $H_{(LM\_in)}(s) - H_{(LM\_out)}(s)$ and the sentences with the lowest scores are selected. This criterion leads to selecting sentences that are similar to the in-domain data and dissimilar to the non-domain-specific data. The method has been adapted to Machine Translation [5], [23], [24]. In Schwenk et al. [24], the perplexity decreases by 20% (when considering LM individually) using about only 20% of available data. But, after LMs interpolation, the improvement vanishes, the perplexity is 86.6 instead of 87.0 when all the data from the general domain corpus are used.

These last works are closer to our experimental conditions because the unit for selection is the sentence, and several language models are interpolated [23], [24]. From the literature, the cross-entropy difference approach [1] leads to the best performance in the scope of corpus selection. Therefore, our work is based on this approach. But we consider the case of several corpora (corresponding to different sources of data), which lead to an interpolation of the individual LMs.

III. Experimental Setup

Experiments are conducted to investigate and analyse the selection of textual data to be used for developing language models dedicated to speech transcription of radio and TV shows. The transcription task considered is the one of the Etape French evaluation campaign. The language models are developed using various text corpora available at the Loria laboratory and corresponding to different sources: manual transcription of radio broadcast shows from 1998 to 2005 (noted Tr); Newspapers data from 1987 to 2007 (noted Np); Web data collected from various web sites (TV, newspapers, magazines, etc.), mainly in 2010 and 2011 (noted Web); and the corpus Gigaword second edition [25] (noted Gw).

In the following, the whole training data set is designed by $Tr + Web + Np + Gw$. The textual training data was normalized, lowercased and tokenized. Table I indicates the sizes of the training data sets (for each source and for the whole corpus). Each line in the table reports the number of sentences, as well as the number of words (resulting from the tokenization process; begin-of-sentence and end-of-sentence tokens are not included in the counts).

TABLE I. Sizes of training corpora, in millions of sentences and millions of words.

Sources	# sentences [M]	# words [M]
Tr (radio broadcast transcriptions)	5	114
Web (web data)	17	334
Np (newspapers)	23	526
Gw (gigaword corpus)	29	783
$Tr + Web + Np + Gw$	74	1 757

The Etape training set corpus ($ETAPE\_Train$), which contains about 300 K running words, is used as a validation corpus for optimising the weights of the linear combination of the individual models estimated separately on each source. This corpus is also called $DevLM$, when used for computing perplexity results. The Etape development set ($ETAPE\_Dev$), which contains about 90 K running words, is used for evaluation purposes. This data set was not used, neither for building the language models, nor for building the acoustic models used for speech transcription evaluations. When reporting perplexity results, this corpus is also called $TestLM$. In all the reported experiments, 3-gram language models are considered; the vocabulary used contains about 100 K words and it was selected as described in [26]. SRILM tools [27], [28] are used for creating and evaluating the language models in the reported experiments. The reported perplexity values do not consider the begin-of-sentence and end-of-sentence tokens (this corresponds to the $ppl1$[1] values provided by the SRILM tools.). Also, thanks to the -unk option, all out-of-vocabulary words are mapped to the <unk> symbol, both for training the language model and for computing the sentences perplexities.

As the training corpora associated to the various sources have very different sizes, the conventional way of training a language model is a 3-step process. In the first step, a separate language model is trained on each source corpus (i.e. one model for Tr, one model for Web, and so on) using the modified Kneser-Ney smoothing method [29]. In the second step, the weights of the linear combination of the individual source models are estimated so as to maximize the likelihood of the LM development data ($DevLM$ corpus). Finally, the individual models are interpolated according to the optimal weight values.

Table II displays informations pertaining to the baseline model built using all the textual training data according to this 3-step procedure. The table shows large differences in the perplexity provided by the individual models (on the LM development data), which translates in very different weights in the interpolated LM. It is striking to observe that the weight associated to the Gigaword LM is very low, which means that the Gigaword data brings very little information in the final LM, although the Gigaword corpus is a very large text corpus.

Hence the goal of the paper is to investigate if a selection of textual data could provide better results.

TABLE II. BASELINE LM, INTERPOLATED FROM THE INDIVIDUAL SOURCE LMS.

Sources	**Individual LMs**	**Interpolated LM**		
	ppl $DevLM$	**weights**	**ppl** $DevLM$	**ppl** $TestLM$
Tr	215.6	0.685	185.7	218.9
Web	264.7	0.246		
Np	364.2	0.062		
Gw	531.4	0.007		

A few other language models were also developed using different collections of training data sources, as reported in Table III. Training a language model using only the Tr corpus leads to a perplexity of 253.0 on the $TestLM$ corpus. The addition of the newspaper and web data (Np and Web corpora) leads to a significant improvement, and yields a perplexity of 218.9 on the $TestLM$ corpus. This perplexity is equal to the one achieved with the baseline model which uses the four data sources. This confirms the fact that, in the interpolated models, the contribution of the model trained on the whole Gigaword corpus seems useless for modelling the ETAPE data.

TABLE III. PERPLEXITIES WITH RESPECT TO TRAINING DATA SOURCES USED FOR TRAINING THE LMS.

LM	**ppl** $DevLM$	**ppl** $TestLM$
LM_Tr	215.6	253.0
$LM\_(Tr + Web + Np)$	185.8	218.9
$LM\_(Tr + Web + Np + Gw)$	185.7	218.9

IV. DATA SELECTION STRATEGY

The experiments reported in this section are conducted to validate the implementation of the selection procedure based on the difference of cross-entropy, as described in [1]. To do so, a context similar to the one used in that paper is defined: only two sources of data are considered. One represents the in-domain data; here we choose the Tr corpus, since the manual transcriptions of broadcast news are the most similar to the ETAPE data according to the perplexity values reported in Table II. The other one represents general domain data, or non-domain-specific data to keep with the terminology defined in [1]; here we choose the Gigaword data for that.

To compare with [1], two selection processes are evaluated. The first one is based on a random selection of the data. A set of LMs are thus trained on subsets corresponding to a random selection of 5%, 10%, etc. of the Gw data. The second one is the data selection method described in [1], based on the difference between the sentence cross-entropy for in-domain LM and for a non-domain-specific LM. For each sentence s of the Gigaword corpus, the cross-entropy difference $dXent(s)$ is computed using two LMs of similar size: one, LM_Tr, is estimated on the transcription data, and the other one, LM_GwTiny, is developed with a randomly selected subset of the Gw corpus of approximately the same size as the Tr corpus (i.e., about 114 M words):

$$dXent(s) = H_{(LM\_Tr)}(s) - H_{(LM\_GwTiny)}(s) \qquad (1)$$

For each threshold applied on the $dXent$ criterion, a LM is trained on the sentences selected with the lowest scores.

Figure 1 displays the perplexity on the $TestLM$ corpus with respect to the percentage of data (percentage of words) selected for training the language models. This figure shows the results obtained with a random selection, and the ones obtained with the strategy presented in this section. Moreover, the perplexity (671.4) obtained by using the whole Gigaword corpus is indicated. Obviously, using smaller training subsets resulting from a random selection on the Gw data degrades the perplexity. On the opposite, using the difference of cross-entropy $dXent$ criterion for selecting data in the Gigaword corpus leads to an improvement of the perplexity (with the best value equal to 454.7). This is due to the fact that this selection criteria is able to select sentences that are close to the in-domain corpus Tr and far from the non-domain-specific data represented here by the LM_GwTiny language model.

[1] www.speech.sri.com/projects/srilm/.../srilm-faq.7.html.

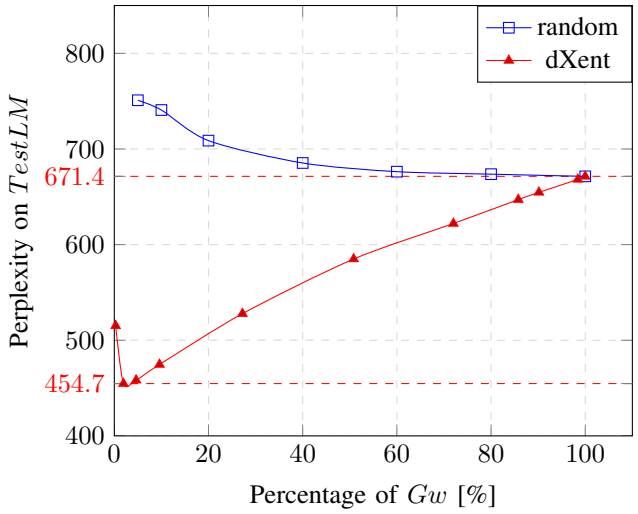

Fig. 1. Perplexity on the $TestLM$ corpus with respect to the percentage of Gigaword data selected for the LM training process; the selection is applied only on the Gigaword data based on a random selection process (random) and on the $dXent$ criterion computed with LM_Tr and LM_GwTiny (dXent).

The behaviour of these results are very similar to the behaviour reported in [1] (Figure 1); best results are obtained with a small set of data selected because they match the best with the target data. This validates the implementation of the selection procedure.

V. Application to multisource-based LM

This section investigates the selection of textual data in the context of multisource-based language models. Using the selection criterion based on the difference of cross-entropy, several choices of models to represent in-domain (or target domain, here ETAPE), and non-domain-specific data are analysed.

Similarly to what was done in previous section, a random selection process is also evaluated. Subsets of respectively 5%, 10%, etc. of randomly selected data are extracted from each source corpus. The perplexity obtained with the interpolated models built from these randomly extracted subsets is reported in Figure 2 (blue curve); and as expected, the perplexity degrades as the amount of randomly selected data gets smaller.

This section is devoted to exploring various choices of language models for representing the in-domain and the non-domain-specific data; and whether the selection process should be applied on all the data sources or only on the Gigaword data.

A. Approach 1

As in Section IV, the in-domain data is represented by the language model LM_Tr trained on the manual transcriptions (Tr), and the non-domain-specific data is represented by the language model LM_GwTiny trained on a random subset of the Gigaword of similar size as the Tr corpus. The $dXent(s)$ criterion is computed for each sentence s of the Tr, Web, Np and Gw corpora (i.e., on each source data) using these two LMs:

$$dXent(s) = H_{(LM\_Tr)}(s) - H_{(LM\_GwTiny)}(s) \quad (2)$$

For each threshold applied on $dXent$, an interpolated LM is trained using the subsets of selected sentences. The perplexity achieved on the $TestLM$ corpus is reported in Figure 2 (red curve). Although the results are better than those achieved with the random selection, there is no improvement in the perplexity, compared to the baseline model. The selection is the same as the one described in Section IV. However when creating an interpolated model from several sources, the combination weights are optimized to fit the LM development data $DevLM$. This optimization step may mask the benefit of data selection on some of the subsets; or the models used to represent the in-domain and the non-domain-specific data might not be good enough.

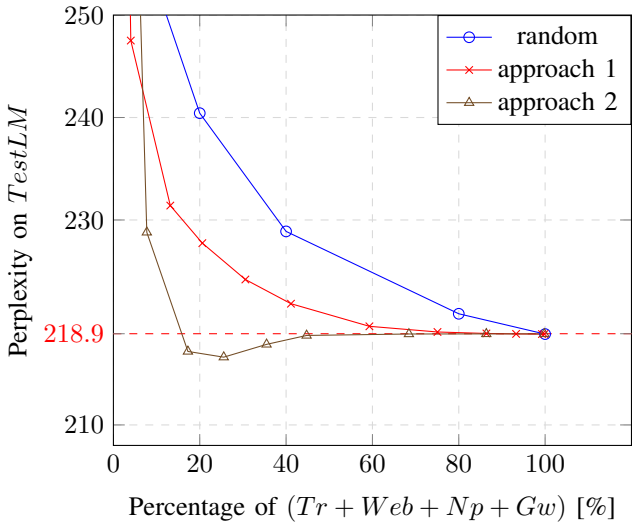

Fig. 2. Perplexity on the $TestLM$ corpus with respect to the percentage of data selected in $(Tr+Web+Np+Gw)$ for the training process; the selection is applied on each source data based: a) on a random process (rand); b) on the $dXent$ criterion computed with LM_Tr and LM_GwTiny (approach 1); and c) on the $dXent$ criterion computed with $LM\_(TrWebNpGw)$ and LM_Gw (approach 2).

B. Approach 2

Here the in-domain data is now represented by the interpolated baseline model $LM\_(Tr+Web+Np+Gw)$ which better represents the ETAPE data than the model LM_Tr trained on only the transcriptions (see for example Table II, where the perplexity on the development data $DevLM$ for the individual model LM_{Tr} (215.6) is higher than that of the interpolated LM (185.7)). For representing the non-domain-specific data, the model LM_{Gw} trained on the whole Gigaword corpus is used. Again the $dXent(s)$ criterion is computed for each sentence s of the Tr, Web, Np and Gw corpora (i.e., each source data) between these two LMs:

$$dXent(s) = H_{(LM\_(TrWebNpGw))}(s) - H_{(LM\_Gw)}(s) \quad (3)$$

As above, for each threshold applied on $dXent$, an interpolated LM is trained using the subsets of selected sentences. The perplexity achieved on the $TestLM$ corpus is reported in Figure 2 (grey curve). Encouraging results are obtained with this approach using only 25% of $(Tr + Web + Np + Gw)$ corpus (more precisely 88% of Tr, 62% of Web, 26% of

Np and 0,2% of Gw). The corresponding model leads to a perplexity 216.6 on the $TestLM$ corpus. Details of this LM are given in Table IV. It can be observed that the weight given to the Gigaword data (or more precisely to the 0.2% of data selected from Gw) is more important in this interpolated model than in the baseline model (0.096 instead of 0.007). Comparing results achieved with approaches 1 and 2, shows that the choice of the models for representing in-domain and non-domain-specific data plays an important role.

TABLE IV. BEST LM, INTERPOLATED FROM THE INDIVIDUAL SOURCE LMS, AFTER DATA SELECTION USING APPROACH 2.

Sources	**Individual LMs**	**Interpolated LM**		
	ppl $DevLM$	**weights**	**ppl** $DevLM$	**ppl** $TestLM$
Tr (88%)	217.6	0.608	185.1	216.6
Web (62%)	262.2	0.234		
Np (26%)	333.0	0.062		
Gw (0.2%)	435.6	0.096		

C. Approach 3

Here the in-domain data is again represented by an interpolated model trained using several data sources (Tr, Web and Np), and the non-domain-specific data is represented by the model trained on the whole Gigaword corpus. However, here, the cross-entropy difference is evaluated only for each sentence s of the Gw corpus using these two models:

$$dXent(s) = H_{(LM\_(TrWebNp))}(s) - H_{(LM\_Gw)}(s) \quad (4)$$

For each threshold applied on the $dXent$ criterion, an interpolated LM is trained using the data selected from the Gigaword corpus, and the whole corpus corresponding to the other sources (Tr, Web and Np). The perplexity achieved on the $TestLM$ corpus is reported in Figure 3, where a logarithmic scale is used on the horizontal axis (percentage of Gigaword data selected). This strategy gives good results; the best LMs are obtained with a small amount of data selected from the Gigaword added to the three other data sources. In the best case, the perplexity decreases to 210.5. The details of the corresponding model are reported in Table V. Again, the weight given to the model estimated from the Gigaword selected data is higher than in the baseline model (0.054 instead of 0.007).

TABLE V. BEST LM, INTERPOLATED FROM THE INDIVIDUAL SOURCE LMS, AFTER DATA SELECTION USING APPROACH 3.

Sources	**Individual LMs**	**Interpolated LM**		
	ppl $DevLM$	**weights**	**ppl** $DevLM$	**ppl** $TestLM$
Tr (100%)	215.6	0.660	179.9	210.6
Web (100%)	264.7	0.240		
Np (100%)	364.2	0.065		
Gw (0.05%)	2822.8	0.054		

VI. TRANSCRIPTION EXPERIMENTS

A selected set of language models resulting from the previous experiments are selected for speech transcription experiments. As indicated in Table VI, this includes the baseline models, as well as the models having the lowest perplexities after selection of data with the approaches 2 and 3. The performance of the speech transcription is evaluated on the $ETAPE\_Dev$ data.

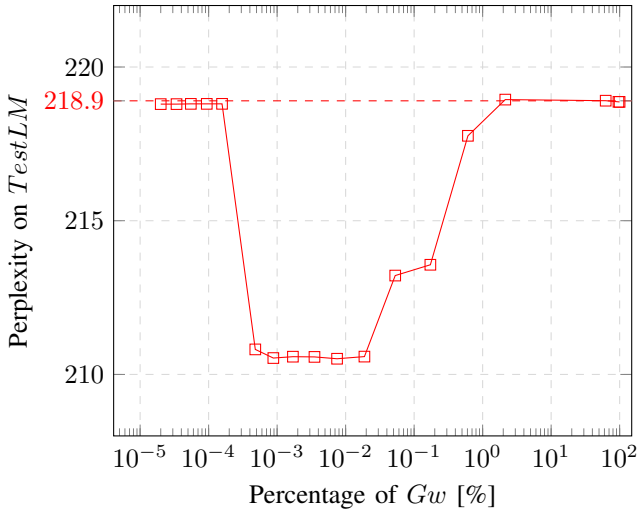

Fig. 3. Perplexity on the $TestLM$ corpus with respect to the percentage of Gigaword data selected for the LM training process; the selection is applied only on the Gigaword data based on the $dXent$ criterion computed with $LM\_(TrWebNp)$ and LM_Gw (approach 3).

TABLE VI. AUTOMATIC SPEECH TRANSCRIPTION RESULTS ON THE ETAPE DEV CORPUS.

LM	**Size** (gz file)	**Etape Dev corpus**	
		ppl	**WER**
$LM\_(Tr + Web + Np + Gw)$	1.2 Go	218.9	27.84%
$LM\_(Tr + Web + Np)$	809.8 Mo	218.9	27.82%
LM_(approach 2, threshold -0.3)	391.3 Mo	217.2	28.07%
LM_(approach 2, threshold -0.2)	501.6 Mo	216.6	27.89%
LM_(approach 3, threshold -0.8)	809.3 Mo	210.5	27.75%
LM_(approach 3, threshold -0.7)	809.3 Mo	210.6	27.72%
LM_(approach 3, threshold -0.6)	809.3 Mo	**210.6**	**27.68** %
LM_(approach 3, threshold -0.1)	809.9 Mo	217.8	27.73%
LM_(approach 3, threshold 0)	881.1 Mo	218.9	27.85%

The best improvement in terms of perplexity for our LMs is about 8.3 (which corresponds to 3.8% relative), whereas the corresponding improvement in terms of WER is only about 0.2%. The interpolated LM created after data selection has a smaller size (reduction by a factor of $2/3$).

VII. CONCLUSION

We presented and analysed multi-source data selection for the training of LMs dedicated to the transcription of broadcast news and TV shows.

The test set perplexity for the LM trained on the $(Tr + Web + Np + Gw)$ corpora, which is our baseline, is 218.9. We noticed that the Tr and Web corpora are the closest to our task, unluckily the huge Gw corpus contains a lot of heterogeneous and irrelevant data. Keeping the three data sources (Tr, Web and Np) and selecting data from the Gw corpus with the cross-entropy difference leads to better results than when selecting data randomly or with cross-entropy difference from the whole corpora (Tr, Web, Np, and Gw). An optimum perplexity of 210.5 is obtained with an LM built from 55.4% of the $(Tr +$

$Web + Np + Gw$) corpora. The best improvement is about 8.3 in terms of perplexity, and results in a reduction of 0.2% absolute in terms of WER.

This work leads to several interesting conclusions. First, the choice of the models that represent in-domain and non-domain-specific data is important. The results are very different from approach 1, approach 2 and approach 3. The results indicate that selection should be applied only on the corpus which is the furthest from the in-domain, that the entire non-domain-specific data should be used to estimate the non-domain-specific LM, and that it is better to avoid overlapping between the non-domain-specific data and the in-domain data. Additional experiments will be conducted to confirm these indications.

The conclusion seems to be that the Gigaword corpus is not very useful for language modelling for our task. We guess this is not true because of the great coverage of this corpus. Therefore, we have now to explore other ways to select data from Gigaword in order to improve the performance. As the vocabulary is also a crucial module for transcription, one way is to take into account the time period of data sub-parts.

REFERENCES

[1] R. C. Moore and W. Lewis, "Intelligent selection of language model training data," in *Proceedings of the ACL 2010 Conference Short Papers*. Association for Computational Linguistics, 2010, pp. 220–224.

[2] L. Lamel, J. Gauvain, V. Le, I. Oparin, and S. Meng, "Improved models for mandarin speech-to-text transcription," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4660–4663.

[3] A. Rousseau, P. Deléglise, and Y. Estève, "Enhancing the ted-lium corpus with selected data for language modeling and more ted talks," in *Proc. of LREC*, 2014, pp. 3935–3939.

[4] B. Dalvi, C. Xiong, and J. Callan, "A language modeling approach to entity recognition and disambiguation for search queries," in *Proceedings of the first international workshop on Entity recognition & disambiguation*. ACM, 2014, pp. 45–54.

[5] P. Koehn and B. Haddow, "Towards effective use of training data in statistical machine translation," in *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2012, pp. 317–321.

[6] P. Goyal, L. Behera, and T. M. McGinnity, "A novel neighborhood based document smoothing model for information retrieval," *Information retrieval*, vol. 16, no. 3, pp. 391–425, 2013.

[7] R. D. Brown, "Finding and identifying text in 900+ languages," *Digital Investigation*, vol. 9, pp. S34–S43, 2012.

[8] M. Hamdani, P. Doetsch, M. Kozielski, A. E.-D. Mousa, and H. Ney, "The rwth large vocabulary arabic handwriting recognition system," in *Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on*. IEEE, 2014, pp. 111–115.

[9] R. Rosenfield, "Two decades of statistical language modeling: Where do we go from here?" 2000.

[10] L. R. Rabiner and B. Juang, "Statistical methods for the recognition and understanding of speech," *Encyclopedia of language and linguistics*, 2004.

[11] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ester phase ii evaluation campaign for the rich transcription of french broadcast news." in *Interspeech*, 2005, pp. 1149–1152.

[12] S. Galliano, G. Gravier, and L. Chaubard, "The ester 2 evaluation campaign for the rich transcription of french radio broadcasts." in *Interspeech*, vol. 9, 2009, pp. 2583–2586.

[13] G. Gravier, G. Adda, N. Paulson, M. Carré, A. Giraudel, and O. Galibert, "The etape corpus for the evaluation of speech-based tv content processing in the french language," in *LREC-Eighth international conference on Language Resources and Evaluation*, 2012, p. na.

[14] Y. Esteve, T. Bazillon, J.-Y. Antoine, F. Béchet, and J. Farinas, "The epac corpus: Manual and automatic annotations of conversational speech in french broadcast news." in *LREC*, 2010.

[15] T. Bazillon, "Transcription et traitement manuel de la parole spontanée pour sa reconnaissance automatique," Ph.D. dissertation, Université du Maine, 2011.

[16] D. Klakow, "Selecting articles from the language model training corpus," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 3. IEEE, 2000, pp. 1695–1698.

[17] X. Shen and B. Xu, "The study of the effect of training set on statistical language modeling." in *INTERSPEECH*, 2001, pp. 721–724.

[18] H.-F. Wang, J. Gao, K.-F. Lee, and M. L. Li, "A unified approach to statistical language modeling for chinese," June 2000. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=68833

[19] J. Gao, J. Goodman, M. Li, and K.-F. Lee, "Toward a unified approach to statistical language modeling for chinese," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 1, no. 1, pp. 3–33, 2002.

[20] K. Yasuda, R. Zhang, H. Yamamoto, and E. Sumita, "Method of selecting training data to build a compact and efficient translation model." in *IJCNLP*, 2008, pp. 655–660.

[21] G. Foster, C. Goutte, and R. Kuhn, "Discriminative instance weighting for domain adaptation in statistical machine translation," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010, pp. 451–459.

[22] A. Toral, "Hybrid selection of language model training data using linguistic information and perplexity," in *Proceedings of the Second Workshop on Hybrid Approaches to Translation, Sofia, Bulgaria*, 2013, pp. 8–12.

[23] A. Axelrod, X. He, and J. Gao, "Domain adaptation via pseudo in-domain data selection," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 355–362.

[24] H. Schwenk, A. Rousseau, and M. Attik, "Large, pruned or continuous space language models on a gpu for statistical machine translation," in *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*. Association for Computational Linguistics, 2012, pp. 11–19.

[25] A. Mendona, G. David, and D. Denise, "French gigaword second edition," *Web Download*, 2009.

[26] D. Jouvet and D. Langlois, "A machine learning based approach for vocabulary selection for speech transcription," in *Text, Speech, and Dialogue*. Springer, 2013, pp. 60–67.

[27] A. Stolcke *et al.*, "Srilm-an extensible language modeling toolkit." in *INTERSPEECH*, 2002.

[28] A. Stolcke, J. Zheng, W. Wang, and V. Abrash, "Srilm at sixteen: Update and outlook," in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, 2011, p. 5.

[29] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," in *Proceedings of the 34th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1996, pp. 310–318.