Textual Data Selection based on Mean Square Difference Probability for Language Modeling

F. Mezzoudj
UHBC Chlef, USTO-MB Oran, Algeria
fmezzoudj@gmail.com

A. Benyettou
USTO-MB Oran, Algeria
abdelkaderbenyettou@univ-usto.dz

Abstract— The language model (LM) is an important module in many applications that produce natural language text such as Automatic Speech Recognition, Machine Translation systems, etc. Generally, the amount of training data which are suitable for training language models dedicated to specific target task is limited. Hence this kind of textual data are too costly to produce, the use of textual data selected from others domains can be useful. This paper proposes to investigate the Mean Square Difference Probability (MSDP) criteria between two models representing respectively in-domain and out-domain-specific data for textual data selection. This technique is analyzed and tested on French broadcast news and TV shows transcription data. Results show that, the selection data based on Mean Square Difference Probability is competitive compared to other criteria of state of the art such as Difference Cross-Entropy (dXent) data selection.

Index Terms—cross-entropy, data selection, mean square difference probability, n-gram language model, perplexity, textual corpus.

I. Introduction

THE Large Vocabulary *Continuous* Speech Recognition (LVCSR) [1] is the process of converting audio waves, or speech acoustic signals, to corresponding written texts with a large considered vocabulary. LVCSR system generally consists of two modules: the acoustic and language models. The method of Weighted Finite State Transducers is used to combine these two models: acoustic models and language models.

The acoustic models are trained on the features extracted from acoustic data using stochastic models such as Hidden Markov Model (HMM) or Deep Neural Network (DNN). However, the language models [2], which provide statistical information in order to avoid ambiguities between the transcribed words, are trained on textual data. Despite the use of neural network language models [3] [4], the statistical n-gram language models still play a central role in modern LVCSR systems. An n-gram is a sequence of *n* words accepted in the considered natural language. An n-gram language model is obtained by a training process using preferably a large textual corpus.

The textual data for training language models dedicated to a specific task are usually limited and costly to produce such as transcribed broadcast data, medical data, touristic data, etc. We can take benefit of textual data coming from other general domains. However, the heterogeneity of these kinds of data can be noisy, which leads us to select from the general data the subset data closest to the specific domain.

Generally, the statistical language modeling has many applications in a large variety of areas, including speech recognition, machine translation, optical character recognition, etc. This paper focuses on the training of statistical n-gram language models using three corpora: an in-domain corpus and two out-domain corpora. In order to improve this training process, we propose to use the Mean Square Difference Probability (MSDP) criterion between two models representing respectively in-domain and out-domain specific data for the textual data selection. So, the used data selection method relies on computing a score which represents how the sentences from the out-domain specific data are close to the in-domain data. Two other variants of scoring based on random selection and difference cross-entropy criterion are used too.

The data selection methods are used on French broadcast news and TV shows data, in order to improve the language models. We are contributing to the efforts of providing an adequate strategy for data selection dedicated to language modeling. All our results of data selection methods are evaluated using perplexity measure, the standard metric of language models.

The paper is organized as follows. Section 2 describes the principle of language models. Section 3 exposes the related work on data selection for language modeling and our proposed strategy in this context. Section 4 presents the experimental set-up, including the corpora used. Textual data selection approaches and experiments are presented and discussed in Section 5. Section 6 presents conclusions and research directions for future works.

II. Language Models

C. N-gram language models

The goal of the language modeling [5] is to provide a task of syntax that defines acceptable spoken input sentences. In order to avoid ambiguities between the words with similar sounds and different meanings, these models estimate the probability of a following word in a text sequence and match high probabilities for the correct sequence of words. For example, it is difficult to distinguish between the words *cent*, *sent* and *scent*, in a sequence speech, without the context information provided by the language models.

The LVCSR system is about to find the correct transcribed sentence s from the given acoustic input. The probability of each sentence from the textual training data $s = w_1 w_2 \cdots w_k$, where $w_{i\in\{1..k\}}$ are the words that compose the sentence s, is given by the equation (1) :

$$P(s) = P(w_1)P(w_2|w_1)\cdots P(w_k|w_1 \cdots w_{k-1}) \qquad (1)$$

According to the markovian assumption, a possible estimation is given by multiplying the probabilities of a predicted word w_i according to only the preceding $n-1$ words (or n-grams): w_{i-n+1}^{i-1}. This n-gram language model estimation of $P(s)$ is given by the equation 2:

$$P(s) \approx P(w_1)P(w_2|w_1)\prod_{i=3}^{k} P(w_i|w_{i-2}, w_{i-1}) \qquad (2)$$

where the first two terms, in the equation 2, are called a unigram and a bigram, respectively and the last term is a 3-gram. A unigram (n=1) represents the probability of each word in the considered text. A bigram (n=2) models the probability of a word given its previous word. A trigram (n=3) takes into account the previous two words, and so on. The n-gram probabilities can be calculated according to the count likelihood estimation of the

words sequences in the textual training data.

D. Perplexity

In order to evaluate a language model, a simple measure called perplexity and noted P is used. It represents the average branching factor i.e., the average number of words that need to be distinguished anywhere in the sequence assuming all words. This value is an indication of how well the language model can predict the sequence.

The perplexity is defined by the following formula $P = 2^{H(p)}$ where $H(p)$ is the cross-entropy of the language model. The cross-entropy is defined using the equation 3 :

$$H(p) = -\frac{1}{|T|}\log P(T) \qquad (3)$$

where T is test textual corpus and $|T|$ is its size.

By comparing perplexities of two language models, the lesser one is for the best model, when computed on an unseen text material during the training step.

E. Smoothing methods

In training language models, we usually face a serious problem. To train a specific domain model, we must deal with the data sparseness problem, because large amount of specific domain data are not available. To overcome this kind of problems, many different methods have been suggested. Smoothing techniques are usually used to better estimate probabilities when there is insufficient data to estimate probabilities accurately. Due to the data sparsity, most of the possible bigrams and the vast majority of trigrams will not occur at all in any text corpora. Hence, smoothing techniques are needed in order to obtain accurate and non-zero probability estimates for all possible n-grams.

Chen [5] and Mezzoudj [6] surveyed different smoothing techniques of the n-gram language models. The basic idea of the smoothing method is to adjust the probability of the seen n-grams downward and allocate this probability to the unseen n-grams during the training step. This allocation is based on the probability distribution of lower order model. Often, Katz-back-off method combined to the strategy of Good-Turing estimation is used. Among these smoothing methods, the Modified Kneser-Ney algorithm, introduced by Chen [5], outperforms the other language models and it can be used in computationally- limited environment. Stupid back-off [7] gives good results too but needs extreme big data for training, which requires important hardware and enough data resources.

Entries from the resulting n-gram files for n-grams language model, look as in the following example:

-5.126541 diplomate -0.1502105

-4.126541 directeur -0.3631682

where the numeric values on the left side represent the estimated log-probabilities for the n-grams (in this case unigrams) "*diplomate"* and *"directeur"*, and the values on the right side represent the logarithm of the Katz back-off coefficients.

VI. Data Selection Strategy

A. Related work

The large amount of any training data can lead to models too large and very general for real applications. Textual data selection is an effective solution to domain adaptation in statistical language modeling. The dominant methods are perplexity-based ones, which tend to select short sentences. Also, other different approaches are proposed in the literature.

Sethy et al. [8] propose a data selection algorithm that selects a sentence from the web set, if adding the sentence to the already selected set reduces the "entropy" with respect to the in-domain data distribution. The algorithm appears efficient in producing a rather small subset of the web data.

An intelligent approach is introduced by Moore and al. [9]: two language models are used for sentence scoring; one is trained on the whole in-domain data and the other one is trained on a random selected subset of the out-domain specific data, with a similar size to the in-domain one. Each sentence s from the out-domain-specific data is ranked using the "cross-entropy difference" $H_{(M\ -in)}(s) - H_{(M\ -out)}(s)$ and the sentences with the lowest scores are selected. According to literature results, this cross-entropy difference approach leads to good performance in the scope of data selection.

In the context of Machine Translation, Axelrod et al. [10] improved the cross-entropy difference based approach and proposed "bilingual cross-entropy difference" as a ranking function with the in-domain and the general- domain language models. The translation models obtained with 35k selected sentences outperformed the model trained with the all data of about 12 million sentences.

Duh et al. [11] employed the method of Axelrod et al. based on bilingual cross-entropy difference for data selection in the context of Machine Translation too. They further explored neural language model rather than the conventional n-gram language model.

Liu and al. [12] propose three data selection methods based on translation model and language model to rank the sentence pairs in the general-domain corpus: Data Selection with "Translation Model", Data Selection by "Combining Translation and Language model" and Data Selection by "Bidirectionally Combining Translation and Language Models". These methods are able to select high-quality domain-relevant sentence pairs.

For a task of building domain-adapted Statistical Machine Translation (SMT) systems, the authors [13] propose a data selection based on the "Edit distance". After investigating the individual model, a combination of three techniques "Edit distance, perplexity and cosine tf-idf" is proposed at both corpus level and model level. Comparative experiments are conducted on the Hong Kong law Chinese-English corpus.

However, in the context of radio broadcast shows transcription, Mezzoudj et al. [14] [15] adapted the method based on difference of cross-entropy using an in-domain and three out-domain corpora for multi-sources data selection. Despite the challenging situation, the use of many cross-entropy difference approaches lead to good performance in the scope of data selection for broadcast LVCSR. The final language model obtained after the best data selection approach has a smaller size (reduction by a factor 2/3) and leads in an improvement of 8.3 in terms of perplexity.

In [16], the authors used a semi-supervised recursive neural network to learn a vector space representation for huge bilingual data (Chinese-English) for a useful intrinsic data selection. A high-performance computing cluster with sixty 3.3-GHz Xeon E5-2670 cores (120 threads) is used in the considered experimentations which are not available in our case.

Recent works [17][18] have especially dealt with domain adaptation for Neural Machine Translation (NMT) by selecting and providing meta-information to the Neural Network at the sentence level. The considered technique allows to a model built from a diverse set of out-domain training data to produce in-

domain translations.

B. B. *Mean Square Difference Probability selection*

To contribute in solving these problems and improving the n-gram language modeling, we use and analysis two standard different sentence selection techniques based on random and difference cross-entropy criteria and we introduce a novel data selection criterion based on the Mean Square Difference Probabilities (MSDP).

The idea is to push the language model to choose the most appropriate vocabulary and sentence structure while using the information from all the domains to improve the modeling quality. So we try to find the minimum difference of the language models probabilities trained on the data which are close to the target task (in-domain data) and the general data (out-domain data).

This proposed novel data selection criterion is inspired from the mean square error (MSE) formula used in the Multi-layer perceptron (MLP) retro-propagation algorithm [19]. We recall that over the classification process using a neural network (or a simple MLP), the training set is considered as desired output (label). During supervised training, the mean square error function is used to minimize errors between the obtained output (y) and the desired output (l), using the equation 4:

$$MSE = \frac{1}{2}\sum_{i=1}^{N}(y_i - l_i)^2 \qquad (4)$$

The function is based on the principle of maximum likelihood on the output distribution.

Our goal is to explore and adapt this idea to the Mean Square Difference Probability (MSDP) criterion, by considering the in-domain data as the target output and the other sources as obtained (or general) outputs. So, for each sentence s from the used training corpus, we evaluated the MSDP between the log-probabilities of the two language models in-domain and out-domain, using the equation 5:

$$MSDP(s) = \frac{1}{2}(\log P_{(M\ -in\ -domain)}(s) - \log P_{(M\ -out-domain)}(s))^2 \qquad (5)$$

I. Experiments, Results and Discussion

A. *Used data and toolkit*

The data used for training the language models in this evaluation is a textual data corresponding to the manual transcriptions of the acoustic data from broadcast news and broadcast conversations (such as talk shows and interviews from both TV and radio emissions) noted "*Tr* corpus". This data are used during the evaluation campaign ETAPE [21] which is a French campaign of automatic transcription of radio broadcasting emissions. The manual transcription data *Tr,* which is expensive to produce, represents the in-domain data for our broadcast language model.

Two corpora are also used in the training step data, the "Web data" and "Gigaword" respectively which are considered as out-domain. The Webdata is a free corpus crawled from online Newspapers and TV site web, noted "Web corpus" . However, the French Gigaword [20] second edition is a huge and a standardized corpus for knowledge extraction and distributional semantics of the Linguistic Data Consortium, noted Gw corpus (see: https://catalog.ldc.upenn. edu/LDC2009T28).

Also, The textual data used for validation are taken from ETAPE [21], it is noted DevLM. The test data which are not used for building the language models id noted TestLM. The size of used training and test data in terms of sentences and words counts are shown in table 1 and table 2.

In these experimentations, the free SRILM toolkit [22][23] is used for training and evaluating all the language models. They are smoothed using the Kneser-Ney method [5]. The highest n-gram sequence length which was used for modeling was 3. As a standard practice, the individual language models of different data sources (in-domain and out-domain) are linearly interpolated, using EM algorithm, to obtain final language model.

TABLE I. Training Language Models data

Training Data	#Files	#Sentences	#Words
Transcription (Tr)	74	5 437 203	113 986 727
Webdata (Web)	293	16 590 162	334 057 000
Gigaword (Gw)	637	28 699 758	783 380 463

TABLE II. Validation and Test Language Models - ETAPE data

Validation and test Data	#Files	#Sentences	#Words
Validation (DevLM)	1	20 091	276 770
Test (TestLM)	1	7 551	85 191

B. *Data Selection on Web data*

First, we use the in-domain transcription data (*Tr*) and choose Web data as out-domain data (*Web*) for the language modeling. In this sub-section, different selections based on many criteria are conducted independently on the *Web* data: random selection (noted random) on the *Web* corpus and Difference cross-entropy (noted dXent) and Mean square difference probability (noted MSDP). Each time, language models trained on the *Tr* data and the selected sentences from *Web* data are interpolated and compared with the baseline language model.

The data selection based on the cross-entropy difference (dXent) estimated for all the Web sentences between the two balanced size language models M_{-Tr} (trained on Tr corpus with the size of about 114 M words) and $M_{-TinyWeb}$ (trained with about 114 M words extracted randomly from 334 M words of the Web data), using the equation 6:

$$dXent(s) = H_{(M\ -Tr)}(s) - H_{M\ -TinyWeb}(s) \qquad (6)$$

For each threshold applied on dXent, a language model is trained using the selected sentences. The perplexity obtained on the TestLM corpus is reported in the Figure 1 (curve red).

The data selection based on the mean square difference probabilities (MSDP) estimated for all the Web sentences between the two balanced size language models M_{-Tr} and $M_{-TinyWeb}$ using the equation 7:

$$MSDP(s) = \frac{1}{2}(\log P_{(M\ -Tr)}(s) - \log P_{M\ -TinyWeb}(s))^2 \qquad (7)$$

For each threshold applied on MSDP, a language model is trained using the selected sentences. The perplexity obtained on the TestLM corpus is reported in the Figure 1 (curve brown).

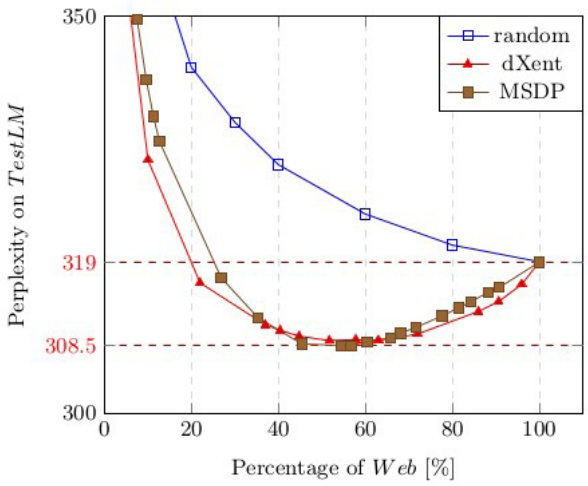

Fig.1. Perplexity on the TestLM corpus with respect to the percentage of data selected in Web data, using random selection and dXent selection and MSDP selection.

We notice that the interpolated baseline 3-gram language model trained on all the Gigaword has a perplexity of 319. Using the selected data by MSDP method and dXent methods (about 50% from *Gw*) lead us to train a better language model dedicated to broadcast task with a perplexity of 308.

C. Data Selection on Gigaword data

The same experiments was also done on the huge French Gigaword corpus (*Gw*), which is available via the LDC. So, we use the in-domain transcription data (*Tr*) and take the Gigaword data as out-domain data (*Gw*). The same selections are applied on the Gigaword data: random selection (noted random) on the Gw corpus and Difference cross-entropy selection dXent and Mean square difference probability selection MSDP.

The data selection based on the cross-entropy difference (dXent) estimated for all the Gw sentences between the two balanced size language models M_{-Tr} (trained on Tr corpus of about 114 M words) and $M_{-TinyGw}$ (trained with about 114 M words extracted randomly from 783 M words of the Gigaword data), using the equation 8:

$$dXent(s) = H_{(M_{-Tr})}(s) - H_{M_{-TinyGw}}(s) \quad (8)$$

For each threshold applied on dXent, a language model is trained using the set of the selected sentences in *Gw*. The perplexity obtained on the TestLM corpus is displayed in Figure 2 (curve red).

The data selection based on the mean square difference probabilities (MSDP) estimated for all the Web sentences between the two balanced size language models M_{-Tr} and $M_{-TinyGw}$ using the equation 9:

$$MSDP(s) = \frac{1}{2}(\log P_{(M_{-Tr})}(s) - \log P_{(M_{-TinyGw})}(s))^2 \quad (9)$$

For each threshold applied on MSDP, a language model is trained using the *Gw* selected sentences. The perplexity obtained on the TestLM corpus is reported in the Figure 2 (brown curve).

We notice that the second baseline 3-gram language model trained on all the Gigaword has a perplexity of 671. Using the selected data by MSDP method (about 22% from *Gw*) leads us to train a better language model dedicated to broadcast task with a perplexity of 532. However, the dXent method can select until only 2% of data from Gw, in order to obtain a language model with 455 of perplexity.

These results show that, the selection data based on Mean Square Difference Probability (MSDP) process leads to different results behavior relative to the different used data sources. In general, the MSDP selection method is successful and competitive with the state of the art represented by the dXent method and can, in some cases be the most successful, when the data are close to the specific task.

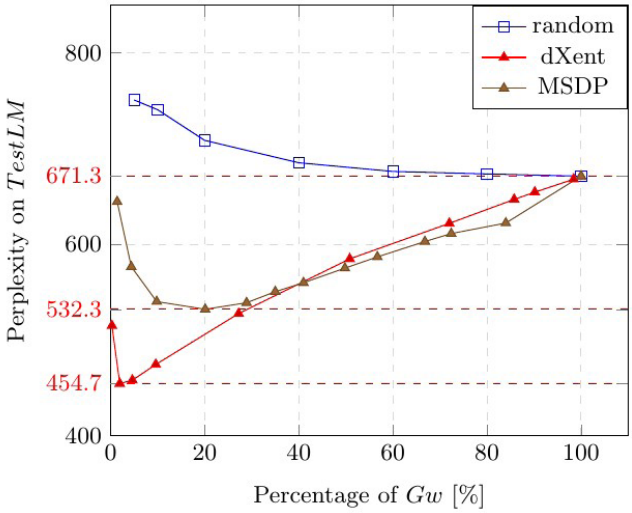

Fig.2. Perplexity on the TestLM corpus with respect to the percentage of data selected in *Gw* data, using random selection and dXent selection and (3) MSDP selection.

Also, if we compare the data selection results on the two corpora Web data (*Web*) and Gigaword (*Gw*) (see the figure 1 and the figure 2), we deduce that the Webdata are more close to the manual transcription data of broadcast and the data selection in Gw is more challenging for this specific task. This phenomenon can be partially due to the characteristics of Gw sentences which are very long vs. the Web sentences length.

II. Conclusion

This paper presents the results of a research on a suitable method for reducing the size of the language model and improves his quality. This experiment was a part of a research on the language modeling conducted in order to improve the language model quality and implementation for use within a Large Vocabulary Continuous Speech Recognition (LVCSR) system for French.

Three basic principles have been adopted and compared as data selection strategies. These methods are the random selection which is the weakest one, and the difference dXent which is considered as the best one in the state of the art, and our proposed MSDP method. The latter and the proposed method which is based on the principle of minimizing the language model probabilities difference trained on the general data and the desired target task data showed significant results.

The data selection and the training language models on a free Web data were interesting. However, the data selection on structured French Gigaword for improving language modeling is still a challenging task for broadcast language modeling.

It would be interesting to use large data from other general domain to improve the broadcast language models using other criterion. Also, we expect to extend the current study using some deep learning techniques for language modeling and feature selection methods, in future works..

References

[1] G. Saon, et J-T. Chien. Large-vocabulary continuous speech recognition systems: A look at some recent advances. IEEE Signal Processing Magazine, 2012, vol. 29, no 6, p. 18-33.

[2] J. T. Goodman, Joshua T. A bit of progress in language modeling. Computer Speech & Language, 2001, vol. 15, no 4, p. 403-434.

[3] 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.

[4] F. Mezzoudj, and A. Benyettou, "An empirical study of statistical language models: n-gram language models vs. neural network language models", Int. J. Innovative Computing and Applications, Vol. 9, No. 3, (in press).

[5] S. F. Chen, J. Goodman, "An Empirical Study of Smoothing Techniques for Language Modeling", Computer Speech and Language, 1999, pp.359-394.

[6] F. Mezzoudj, M. Loukam, and A. Benyettou, "On an empirical study of smoothing techniques for a tiny language model", in Proceedings of IPAC 15, ACM, 23–25 November, Batna, Algeria, 2015, pp.67–80.

[7] T. Brants, Popat, A.C., Xu, P., Och, F.J. and Dean, J. (2007) 'Large language models in machine translation', in Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Citeseer.

[8] A. Sethy, S. Narayanan and B. Ramabhadran. "Data driven approach for language model adaptation using stepwise relative entropy minimization,". In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '07), volume IV, 2007, pg. 177-180.

[9] 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.

[10] 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, pg. 355–362.

[11] K. Duh, G. Neubig, K. Sudoh, & H. Tsukada. Adaptation data selection using neural language models: Experiments in machine translation. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2013, Vol. 2, pp. 678-683.

[12] L. Liu, Y. Hong, H. Liu, X. Wang, & J. Yao, J. (2014). Effective selection of translation model training data. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) Vol. 2, pp. 569-573.

[13] L. Wang, D. F. Wong, L.S. Chao, Y. Lu, & J. Xing, (2014). "A systematic comparison of data selection criteria for smt domain adaptation". The Scientific World Journal, Hindawi Publishing Corporation Volume 2014, Article ID 745485, 10 pages.

[14] F. Mezzoudj. and D. Langlois, and D. Jouvet, and A. Benyettou: "Textual data selection for language modelling in the scope of automatic speech recognition". In International Conference on Natural Language and Speech Processing, Algeria, 2015. pp.28-33

[15] F. Mezzoudj, D. Langlois, D. Jouvet, and A. Benyettou. "Textual data selection for language modelling in the scope of automatic speech recognition". In Procedia Computer Science (2018). 10-APR-2018. pp. 55-64. DOI information : 10.1016/j.procs.2018.03.008

[16] D. Wong, Y. Lu, S. L. Chao. "Bilingual Recursive Neural Network Based Data Selection for Statistical Machine Translation". Knowledge-Based Systems. 108. 10.1016/j.knosys.2016.05.003. 2016.

[17] W. Chen, E. Matusov, S. Khadivi, and J.-T. Peter. "Guided alignment training for topic-aware neural machine translation". CoRR abs/1607.01628v1. 2016.

[18] C. Kobus, J. Crego, et J. Senellart. " Domain control for neural machine translation". arXiv preprint arXiv:1612.06140, 2016.

[19] H. Aly, K. Rady.: "Shannon Entropy and Mean Square Errors for speeding the convergence of Multilayer Neural Networks: A comparative approach". In Egyptian Informatics Journal 12, 197{209 (2011).

[20] A. Mendona, G. David, and D. Denise, "French gigaword second edition ," Web Download: https ://catalog.ldc.upenn.edu/LDC2011T10, 2009.

[21] 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.

[22] A. Stolcke, "SRILM - an extensible language modeling toolkit", Proceedings of ICSLP, vol. 2, Denver, USA, 2002, pp. 901-904.

[23] 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. 5. 2011.