Introduction of semantic model to help speech recognition

Stephane Level, Irina Illina, Dominique Fohr

Introduction of semantic model to help speech recognition

Stephane Level, Irina Illina, and Dominique Fohr

Université de Lorraine, CNRS, Inria, F-54000 Nancy, France
{irina.illina,dominique.fohr}@loria.fr

Abstract. Current Automatic Speech Recognition (ASR) systems mainly take into account acoustic, lexical and local syntactic information. Long term semantic relations are not used. ASR systems significantly decrease performance when the training conditions and the testing conditions differ due to the noise, etc. In this case the acoustic information can be less reliable. To help noisy ASR system, we propose to supplement ASR system with a semantic module. This module re-evaluates the N-best speech recognition hypothesis list and can be seen as a form of *adaptation in the context of noise*. For the words in the processed sentence that could have been poorly recognized, this module chooses words that correspond better to the semantic context of the sentence. To achieve this, we introduced the notions of a *context part* and *possibility zones* that measure the similarity between the semantic context of the document and the corresponding possible hypothesis. The proposed methodology uses two continuous representations of words: *word2vec* and *FastText*. We conduct experiments on the publicly available TED conferences dataset (TED-LIUM) mixed with real noise. The proposed method achieves a significant improvement of the word error rate (WER) over the ASR system without semantic information.

Keywords: Automatic speech recognition · Semantic context · Embeddings

1 Introduction

Despite constant efforts and some spectacular advances, the ability of a computer to recognize speech is still far from equaling that of humans. Current ASR systems significantly deteriorate performance when the conditions in which they are trained and those in which they are used differ. The causes of variability between these conditions can be the acoustic environment and / or the acquisition of the signal. Even if many approaches to compensate this variability have been proposed [18], the performance of an ASR system on a given word always depends on the distortion at the precise moment when this word was spoken.

Current ASR systems mainly take into account only acoustic (acoustic model), lexical and syntactic information (local n-gram language models). We suggest moving towards a *contextualization* of the ASR system. Indeed, lexical and semantic information is important for an ASR system to be efficient. Recently,

several researchers have proposed to use semantic information to improve the ASR performance. For example, exploring the topic and semantic context to enable the recovery of proper names [14], using a semantic language model based on the theory of frame semantics [2], assigning semantic category labels to entire utterances and re-ranking the N-best list of ASR [11]. [7] learns semantic grammar for the ASR system. In [5] authors combine information from the semantic parser and ASR's language model for re-ranking. In [4], a method for re-ranking black-box ASR hypotheses using an in-domain language model and semantic parser trained for a particular task is investigated.

In this article, we propose to complete the noisy ASR step by adding the semantic information in order to detect the words in the processed sentence that could have been poorly recognized and to investigate words of similar pronunciations that correspond better to the context. This semantic analysis re-evaluates (rescores) the N-best transcription hypotheses (N-best) and can be seen as a form of *dynamic adaptation in the specific context of noisy data*. Reevaluation is performed through a definition of context part and possibility zones. Semantic information is introduced using predictive continuous representations [3], [9]. These representations have proven to be effective for a series of natural language processing tasks [1]. The efficiency and the semantic properties of these representations motivate us to explore them for our task of ASR in mismatched conditions. We hope that in very noisy parts, the language model and the semantic model could remove the acoustic ambiguities in order to find the words spoken by the speaker. All our models are based on high-performance DNN technologies. Compared to the previous works using the rescoring of N-best list [12], [15], [16], we don't use several features, and we only rely on semantic information. Furthermore, the specificity of our approach is the use of the context part and the possibility zones of N-hypotheses list: semantic part represents the semantic information of the topic context of the document to recognize and possibility zone corresponds to the area where we want to find the words to be corrected. This allows us to give less importance to the words in the possibility zone which do not correspond to the context of the document, and to give low semantic score to the corresponding hypothesis.

2 Proposed methodology

2.1 Semantic model

An effective way to take into account semantic information is to re-evaluate (rescore) the best hypotheses of the ASR system. This system provides an acoustic score $P_{ac}(w)$ and a linguistic score $P_{lm}(w)$ for each word of the hypothesis sentence. The best sentence is the one that maximizes the probability of the word sequence:

$$\hat{W} = \arg\max_{h_i \in H} \prod_{w \in h_i} P_{ac}(w)^{\alpha} \cdot P_{lm}(w)^{\beta} \tag{1}$$

$\hat{W}$ is the recognized sentence (the end result); H is the set of N-best sentence hypotheses; h_i is the i-th sentence hypothesis; w is a hypothetical word. α and β

represent the weights of the acoustic and the language models. These weights are essential because acoustic scores and linguistic scores are not always normalized (they are often likelihoods and not probabilities).

We want to add semantic information to guide the recognition process. The most natural approach to integrating this information is to modify the calculation of the probability of the sequence of words in the following way:

$$\hat{W} = \arg\max_{h_i \in H} \prod_{w \in h_i} P_{ac}(w)^{\alpha} \cdot P_{lm}(w)^{\beta} \cdot P_{sem}(w)^{\gamma} \quad (2)$$

We added the semantic probability of each word: $P_{sem}(w)$. To have a good balance between the different models, we introduce a third weight γ to weigh the semantic information. It will be adjusted on a development corpus.

2.2 Definition of context part and possibility zones

To estimate the semantic probability, we propose to introduce the concepts of *context part* and *possibility zone*. A *context part* consists of words which are common to all the N-best hypotheses generated by the ASR. We assume that they are correct. This context part allows to extract semantic information of the topic context of the document or of the current part of the document to be recognized. The context part can contain several parts. A *zone of possibilities* is an area between the context parts. It is in this area that we want to find the words to be corrected. From the N-best hypotheses of a sentence, we extract *only one* context part and *one or more* possibility zones. Each zone can contain several words. Figure 1 illustrates these concepts on an example. Here, the 2-best hypotheses list is the following:

H1*: the cat eats the big fat mouse*
H2*: the cat bits the bigfoot mouse*

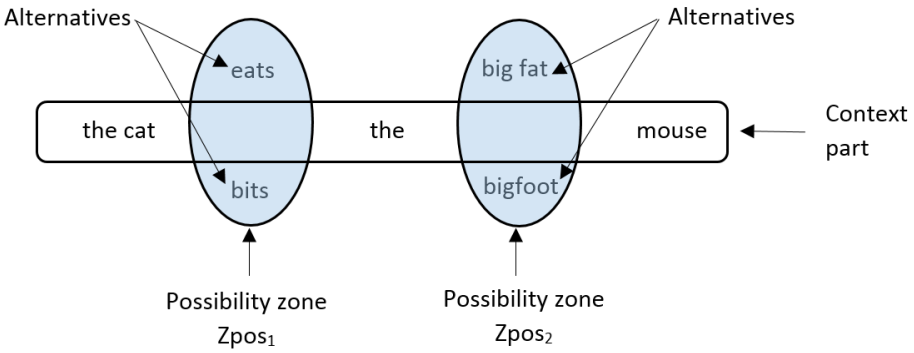

Fig. 1. Illustration of the context part and the possibility zones, as an example.

In this example, the context part Z_{cont} is composed of four words: $Z_{cont} = \{the, cat, the, mouse\}$. These are the words which are common to all the hypotheses and we assume that they are correct. Between these words, we define two possibility zones: the first is made up of two alternatives, eats and bits: $Z_{pos,1} = \{eats, bits\}$. The second is also made up of two alternatives: $Z_{pos,2} = \{bigfat, bigfoot\}$. One alternative corresponds to a choice in the possibility zone. We assume that the possibility zones correspond to the zones where the ASR hesitates between different solutions.

To obtain the context part, we use a dynamic programming algorithm which allows us to pair the hypotheses two by two in order to determine the words common to all the hypotheses. If the context part is empty, we don't study this sentence.

2.3 Semantic representation of the context part and the possibility zones

To take into account the semantics of the document, we propose to represent *each word of the N-best hypotheses by an embedding vector*. In our approach, we used *word2vec* [9] and *FastText* [3]. We compute an average embedding E_{cont} for the context part which is equal to the average of the embedding vectors of all the words in the context part. In the same way, we calculate an average embedding $E_{pos}(i, a_h)$ for i-th possibility zone of alternative a_h of hypothesis h as the average of the embedding vectors of all the words in this alternative of possibility zone. We use the angular similarity to estimate a semantic score between each possibility zone and the context part:

$$S_{sem}(E_{cont}, E_{pos}(i, a_h)) = 1 - \frac{\cos^{-1}\cos(E_{cont}, E_{pos}(i, a_h))}{\pi} \tag{3}$$

From the semantic representations of the context part and the possibility zones, we compute a semantic probability of a hypothesis h. **A semantic probability of a hypothesis** h $P_{sem}(h)$ is computed as follows:

$$P_{sem}(h) = \prod_{i=1}^{N_p} S_{sem}(E_{cont}, E_{pos}(i, a_h)) \tag{4}$$

where N_p is the number of possibility zones. We assume that the equation (2) can be approximated as follow:

$$\hat{H} = \arg\max_{h \in H} P_{ac}(h)^{\alpha} \cdot P_{lm}(h)^{\beta} \cdot P_{sem}(h)^{\gamma} \tag{5}$$

where $\hat{H}$ is the N-best list. The equation (5) is used to re-rank the N-best hypothesis list. For each hypothesis we compute the semantic score and associate it with acoustic and linguistic scores according to (5). The hypothesis obtaining the best score is considered as the recognized sentence.

3 Experiments

3.1 Corpus description

We used the publicly available TED-LIUM corpus [6], containing the recordings of the TED conferences. This corpus is well suited to our study because each conference is focused on a particular subject. We want to add the semantic module to improve the performance of our recognition system.

We used the partition of the TED corpus into a train, a development and a test corpus proposed in the TED-LIUM distribution: 452 hours for training, 8 conferences (496 sentences, 17926 words) for development and 11 conferences (1091 sentences, 27021 words) for testing.

3.2 Recognition system

Our recognition system is based on the *Kaldi* voice recognition toolbox [13]. We used TDNN triphone acoustic models, trained on the training part of TED-LIUM. The lexicon and language model was provided in the TED-LIUM distribution. The lexicon contains 150k words and the language model has 2 million 4-grams, learned from a textual corpus of 250 million words. We also performed the recognition using the RNNLM model [10]. We want to see if using more powerful language model (LM), the proposed semantic module can improve the ASR. As usual, we used the development set to choose the best parameter configuration and the test set to evaluate the proposed methods with this best configuration. We used the word error rate (WER) to measure the ASR performance.

The performance of our ASR system on TED-LIUM using n-gram LM is around 8 % of WER. We are not interested in noise-free conditions because in this case the acoustics allow to properly guide the recognition. This research work was carried out as part of an industrial project. This project concerns the recognition of speech in noisy condition, more precisely, in a fighter aircraft. To get closer to actual conditions, we added noise to the development and test sets: additive noise at 10 dB and 5dB SNR (noise of F16 from the NOISEX-92 corpus [17]).

3.3 Embeddings

We trained *word2vec* model on a text corpus of a billion words extracted from the *OpenWebText* corpus. The generated models have the size of 300 and model 700K words. As *FastText* model, we used the same embedding dimension. The advantage of *FastText* compared to *word2vec* is the taking into account of all possible words.

4 Experimental results

4.1 Overall results

Before performing the speech recognition evaluation, we wanted to investigate the impact of the semantic module alone on the search for the best sentence,

without using the acoustic and linguistic scores. For this, for a reference sentence text, we simulated the recognition errors by replacing a random word (or two successive words) of the reference sentence by one (or two) acoustically close word(s). This can be easily performed using a phonetic dictionary. In this way, we generated N-best hypotheses for the given sentence ($N = 10$). We performed this generation for every 496 sentences of the development set.

After N-best hypotheses generation, we used our semantic module to rank the 11 hypotheses (the 10 generated sentences plus the correct sentence) and we evaluated the number of errors corrected on the top hypothesis. Here, we did not use the acoustic and the language scores. For 496 sentences of the development set, the word2vec-based semantic module corrects about 67 % of simulated errors and the FastText semantic module corrects about 61 % of errors. We see that the long context embeddings alone succeed to correct the large number of errors. This shows that the proposed semantic module captures well the semantic information of a sentence.

Table 1 presents the WER for the development and the test sets for two noise condition (10dB and 5dB) and two language models (n-gram and RNNLM). The first line of results (method *Random*), corresponds to the random selection of the recognition result from the N-best hypotheses without using the semantic module. The second line, *Baseline*, corresponds to the speech recognition system without using the semantic module (standard ASR). The last line, *Oracle*, represents the maximum performance that can be obtained by searching in the N-best hypotheses: selection of the hypothesis which minimizes the WER for each sentence. The other lines of the table give the performance of the proposed approaches. At each case of the table, value between the parentheses corresponds to the recognition result using the RNNLM. From this table we can make the following observations.

Table 1. Recognition results in terms of WER (%). N-best hypotheses list of 50 hypotheses. TED-LIUM development and test sets, SNR of 10 dB and of 5 dB. n-gram LM and RNNLM (between the parentheses).

Method	SNR 10dB		SNR 5dB	
	Dev	Test	Dev	Test
Random	17.9 (14.8)	24.1 (21.5)	34.2 (29.8)	42.1 (39.3)
Baseline system	15.7 (12.3)	21.1 (17.7)	32.7 (28.2)	40.3 (37.1)
word2vec embedding	15.3 (12.0)	20.7 (17.6)	31.9 (**27.4**)	39.4 (36.4)
FastText	**15.2 (11.8)**	**20.5 (17.5)**	**31.8 (27.4)**	**39.2 (36.1)**
Oracle	9.6 (6.9)	12.8 (10.3)	25.4 (21.1)	30.5 (27.6)

The proposed semantic module outperforms the baseline system for all conditions and all evaluated embeddings. For example, on the test set, the semantic module with the *FastText* obtained an absolute improvement of 0.6 % for 10dB and n-gram LM (21.1 % WER versus 20.5 % WER) and 1.1 % for 5dB and n-gram LM (40.3 % versus 39.2 %) compared to the baseline system. This

represents 8 % of relative improvement for 10dB and about 11 % for 5dB in the reduction of the gap between the baseline and the oracle systems. For all datasets, noise levels and two language models the obtained improvements are significant (confidence interval is computed according to matched-pairs test [8]). This shows that the proposed semantic module is able to capture a significant proportion of the semantic information in the data.

The proposed embeddings give similar performances with a slight superiority of the *FastText* embedding. All these observation are valid for two experimented language models: n-gram and RNNLM.

4.2 Impact of hyperparameters

Figure 2 (left) shows the evolution of the WER according to the parameter γ (cf. equation (2)) for the development set, SNR of 5dB and n-gram LM. We observe that this parameter plays an important role. For too large values of γ (bigger than 300), the semantic information becomes dominant compared to the acoustic and linguistic information and the WER begins to increase. Therefore, the value of γ between 100 and 300 seems to be optimal. Figure 2 (right) reports the WER as a function of the N-best list size. We can see that 5 or 10 hypotheses are not enough. Using more than 25 hypotheses shows no further improvement.

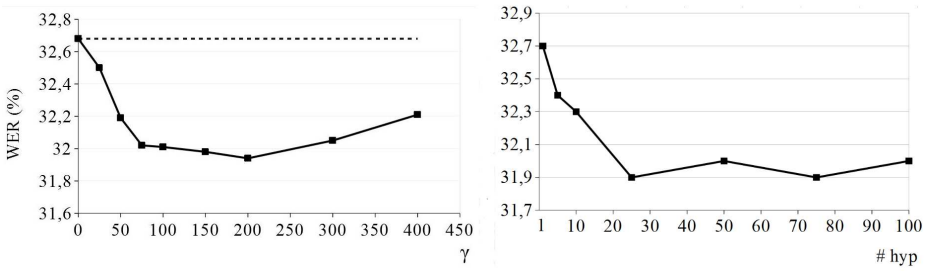

Fig. 2. Semantic module with word2vec embedding, TED-LIUM development set, SNR of 5dB. WER as a function of the semantic weight γ (left figure) and the N-best hypothesis number (right figure). The dotted line corresponds to the baseline result. n-gram LM.

5 Conclusion and discussion

In this article, we proposed a new approach of introducing semantic information for the performance improvement of a noisy ASR system. We investigated a new methodology for taking into account semantics through predictive representations that capture the semantic characteristics of words and their context.

The efficiency and the semantic properties of these representations motivate us to explore these representations for our task of speech recognition. We used *word2vec* and *FastText* embeddings. The semantic information is taken into account through the rescoring module of the N-best hypotheses of the recognition system. Semantic representations are applied to the context part and possibility zones. We evaluated our methodology on the corpus of TED-LIUM conferences with added real noise. The proposed methodology shows a better WER compared to the baseline system. This represents 8 % of relative im-provement for 10dB and about 10 % for 5dB in the reduction of the gap between the baseline and the oracle systems. These improvements are statistically significant. This observation is valid for the ASR with n-gram and with RNNLM.

It is important to note that in *word2vec* and *FastText* the word embedding is static and the words with multiple meanings are conflated into a single representation. In future work, we would like to investigate the dynamic BERT embedding. We will conduct a deep analysis of the performance of semantic module as a function of the noise characteristics (e.g., nonstationarity) and the uncertainty propagation in noisy environment to guide the rescoring.

6 Acknowledgments

The authors thank the DGA (Direction Générale de l'Armement, part of the French Ministry of Defence), Thales AVS and Dassault Aviation who are supporting the funding of this study and the "Man-Machine Teaming" scientific program in which this research project is taking place.

References

1. Baroni M., Dinu G., Kruszewski G.: Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 238–247 (2014)
2. Bayer A., Riccardi G.: Semantic Language Models for Automatic Speech Recognition. In: Proceedings of the IEEE Spoken Language Technology Workshop (SLT) (2014)
3. Bojanowski P., Grave E., Joulin A., Mikolov T., Enriching Word Vectors with Subword Information. In: Transactions of the Association for Computational Linguistics, pp.135–146 (2017)
4. Corona R., Thomason J., Mooney R.: Improving Black-box Speech Recognition using Semantic Parsing. In: Proceedings of the The 8th International Joint Conference on Natural Language Processing, pp.122–127 (2017)
5. Erdogan, H., Sarikaya, R., Chen, S., Gao, Y., Picheny, M.: Using semantic analysis to improve speech recognition performance. In: Computer Speech and Language, 19, pp. 321–343 (2005)
6. Fernandez H., Nguyen H., Ghannay S., Tomashenko N., Estève Y.: TED-LIUM 3: twice as much data and corpus repartition for experiments on speaker adaptation. In: Proceedings of SPECOM, pp. 18–22 (2018)

7. Gaspers J., Cimiano P., Wrede B.: Semantic parsing of speech using grammars learned with weak supervision. In: Proceedings of the HLT-NAACL, pp. 872-881 (2015)
8. Gillick L., Cox S.: Some Statistical Issues in the Comparison of Speech Recognition Algorithms, In Proceedings of ICASSP, v. 1, pp. 532-535 (1989)
9. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, 26, pp. 3111-3119 (2013)
10. Mikolov T., Kombrink S., Burget L., Cernocky J.-H., Khudanpur S.: Extensions of recurrent neural network language model. In: Proceedings of the ICASSP, pp. 5528–5531 (2011)
11. Morbini, F., Audhkhasi, K., Artstein, R., Van Segbroeck, M, Sagae, K., Georgiou P., Traum D., Narayanan S.: A reranking approach for recognition and classification of speech input in conversational dialogue systems. In: Proceedings of the Spoken Language Technology Workshop (SLT), IEEE, pp. 49-54 (2012)
12. Ogawa A., Delcroix M., Karita S., Nakatani T.: Rescoring N-best speech recognition list based on one-on-one hypothesis comparaison using encoder-classifier model. In: Proceedings of the ICASSP (2018)
13. Povey D., Ghoshal A., Boulianne G., Burget L., Glembek O., Goel N., Hannemann M., Motlicek P., Qian Y., Schwarz P., Silovsky J., Stemmer G., Vesely K.: The Kaldi Speech Recognition Toolkit. In: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (2011)
14. Sheikh I., Fohr D., Illina I., Linarès G.: Modelling Semantic Context of OOV Words in Large Vocabulary Continuous Speech Recognition. In: IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25 (3), pp. pp.598 - 610 (2017)
15. Shin J., Lee Y., Jung K.: Effective Sentence Scoring Method Using BERT for Speech Recognition. In: Proceedings of Machine Learning Research 101, pp 1081-1093 (2019)
16. Song Y., Jiangy D., Zhao X., Xuy Q., Wong R., Fany L., Yang Q.: L2RS: a learning-to-rescore mechanism for automatic speech recognition. arXiv:1910.11496v1 (2019)
17. Varga A., Steeneken H.: Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems, In: Speech Communication, Volume 12, Issue 3, pp. 247-251 (1993)
18. Zhang Z., Geiger J., Pohjalainen J., Mousa A., Jin W., Schuller B.: Deep Learning for Environmentally Robust Speech Recognition: An Overview of Recent Developments. In: ACM Transactions on Intelligent Systems and Technology, 9 (5) (2018)