Automatically Finding Semantically Consistent N-grams to Add New Words in LVCSR Systems

Gwénolé Lecorvé, Guillaume Gravier, Pascale Sébillot

AUTOMATICALLY FINDING SEMANTICALLY CONSISTENT *N*-GRAMS TO ADD NEW WORDS IN LVCSR SYSTEMS

Gwénolé Lecorvé[1,2,3], *Guillaume Gravier*[1,4], *Pascale Sébillot*[1,2,3]

[1]IRISA – [2]INSA de Rennes, [3]Université européenne de Bretagne, [4]CNRS
Campus universitaire de Beaulieu, 35042 Rennes Cedex, France

ABSTRACT

This paper presents a new method to automatically add *n*-grams containing out-of-vocabulary (OOV) words to a baseline language model (LM), where these *n*-grams are sought to be grammatically correct and to make sense according to the meaning of OOV words. First, this method consists in determining the word sequences, *i.e.*, *n*-grams, in which the usage of a given OOV word is the most semantically consistent. Then, conditional probabilities of these *n*-grams have to be computed. To do this, semantic relations between words are used to assimilate each OOV word to several equivalent in-vocabulary words. Based on these last words, *n*-grams from the baseline LM are re-used to find the word sequences to be added and to compute their probabilities. After augmenting the vocabulary and launching a recognition process, experiments show that our method results in WER improvements which are comparable to those obtained using a state-of-the-art open vocabulary LM.

***Index Terms*—** Automatic speech recognition, vocabulary adaptation, natural language processing, language modeling

1. INTRODUCTION

Large vocabulary continuous speech recognition (LVCSR) systems rely on a static finite vocabulary, which defines the set of all recognizable words, and on a statistical language model (LM), which gathers vocabulary word sequence (*n*-gram) probabilities. Hence, when an out-of-vocabulary (OOV) word is pronounced in a spoken document, the system cannot correctly transcribe this word and typically generates a sequence of acoustically close shorter words instead, thus decreasing recognition accuracy. A solution to this well-known problem is to adapt the vocabulary according to spoken documents to be transcribed. This vocabulary adaptation task can be seen as selecting new words to be added to the vocabulary before integrating these words into the system's components. This paper focuses on this last point while disregarding the new word selection problem.

Integrating an OOV word into a pre-existing system can be split into two tasks. On the one hand, it is necessary to augment the pronunciation dictionary by phonetizing the new word. This task can typically be achieved by using automatic grapheme-to-phoneme conversion tools. On the other hand, the new word has to be integrated into the system's language model in order to associate probabilities with transcript hypotheses containing the new word. To do this, a first approach consists in re-estimating the language model based on the augmented vocabulary, usually by interpolating counts or probabilities of *n*-grams containing the OOV word that have been directly observed from a general-purpose text corpus and from a small task-specific one [1, 2, 3]. However, OOV words are often rare words—even in a task-specific corpus—, and this approach is not optimal because only a few *n*-grams related to each OOV word to be added can be observed and their probabilities cannot be reliably estimated. Thus, after adaptation, a new word still frequently tends not to be recognized by the system when decoding a speech signal. A second, state-of-the-art, approach relies on hybrid LMs which mix probabilities over words and lexical class sequences. In these LMs, referred as *open vocabulary LMs*, each OOV word is replaced by its lexical class when computing the probability of a sequence containing such a word. The main interest of this approach is then to avoid any *n*-gram probability re-estimation step when adaptating the vocabulary. However, an important drawback is that lexical classes are chosen based only on one criterion, *e.g.*, the morphosyntactic information [4, 5] or some semantic aspects of OOV words [6]. Thus, the open vocabulary strategy disregards other linguistic features which could assist the system in discriminating OOV words of a same class while transcribing an utterance. Whereas named entities can typically be effectively accounted for by classes since they usually share common syntactic usages, other words are much more affected by this restriction of lexical classes since these words can be used in different specific ways depending on their various potential meanings. Hence, we will focus on OOV words aside named entities.

In this paper, we present a new method to automatically add *n*-grams containing OOV words to a baseline LM. This method is particularly original since it does not rely on any extra textual corpus where the OOV words could be found, and the *n*-grams added are sought to be semantically consistent, *i.e.*, they have to make sense according to the meaning of OOV words. The proposed strategy first consists in determining the word sequences, *i.e.*, *n*-grams, in which the usage of a given OOV word is the most semantically, as well as grammatically, consistent, and then in estimating conditional probabilities for these *n*-grams before integrating them into the baseline LM. To do this, semantic relations between words are used to assimilate each OOV word to several equivalent in-vocabulary (IV) words. Based on these IV words, *n*-grams from the baseline LM are re-used to find the word sequences to be added and to compute their probabilities.

This paper is organized as follows: our automatic scheme to find semantically consistent word sequences is exposed in Section 2 while Section 3 outlines how the probability of these sequences can be integrated into a pre-existing LM. Finally, Section 4 presents back-end experiments and compares the results of our method with those obtained using an open vocabulary LM.

2. FINDING CONSISTENT WORD SEQUENCES

The basic idea to find word sequences related to new words is that, within an utterance, some words can be replaced by some others without significantly changing the meaning of the utterance. For ex-

ample, the utterances "*the nightingales are singing*" and "*the owls are singing*" are semantically close since they both rely on the fact that birds can sing. Based on this principle, word sequences related to an OOV word to be added can be determined by relying on equivalent IV word sequences. This section first presents a formal equivalence relation between words and between word sequences. Then, a score is proposed to select equivalent word sequences which tend to be the most substantiated for each OOV word. Finally, an instantiation of these two points is exposed through the use of paradigmatic relations.

2.1. Word and *n*-gram equivalence relations

Within natural language, the usage of a word in an utterance depends on various information. On the one hand, it depends on its meaning with respect to the other words. For example, the semantic information of a given noun may favor some adjectives or verbs to come along with, while it may also prevent from using other words. Hence, some words can be replaced by others without changing the global meaning of an utterance, whereas some other replacements are clearly inconsistent. On the other hand, since natural language is constrained by grammatical rules, the correct usage of a word depends on its morphosyntactic information, *i.e.*, its grammatical category and its inflection.

Hence, each word w can be represented by its lemma[1] ℓ_w and its part-of-speech (POS) tag p_w, and two words w and v can be considered as equivalent if their lemmas are linked according to a semantic relation $\mathcal{R}$ and if their POS tags are equal, *i.e.*,:

$$w \equiv v \quad \Leftrightarrow \quad \ell_w \mathcal{R}\, \ell_v \text{ and } p_w = p_v \ , \tag{1}$$

where $\ell_w \mathcal{R}\, \ell_v$ means that ℓ_w is linked to ℓ_v according to $\mathcal{R}$.

By extension, two word sequences X and Y of length n are defined as equivalent if and only if:

$$X \equiv Y \quad \Leftrightarrow \quad \exists i \in [1..n], x_i \equiv y_i \text{ and } \forall j \neq i, x_j = y_j \ , \tag{2}$$

where x_i and y_i stand for the i-th word of X and Y respectively. Based on this equivalence principle, the goal is now to find out which sequences containing a new word have to be integrated into the baseline LM.

2.2. Selecting equivalent word sequences

Given an OOV word w, all the sequences AwB, where A and B are word sequences, are candidates to be integrated into the pre-existing LM. Since many of these possible sequences are irrelevant to the word w, a score $S(AwB)$ is defined to only select the few most linguistically substantiated. This score evaluates the frequency of each equivalent sequence AvB within a set Ω of pre-existing word sequences, weighted by the strength $r(\ell_w, \ell_v)$ of the link between the lemmas of w and v according to a given semantic relation $\mathcal{R}$. This can be formulated as follows:

$$S_n(AwB) = \sum_{v \text{ st. } w \equiv v} r(\ell_w, \ell_v) \times |AvB|_\Omega \ , \tag{3}$$

where n is the length of AwB and $|AvB|_\Omega$ stands for the number of times where AvB is observed in Ω. For each length n, the average value μ_n is computed based on all non-zero scores $S_n(.)$. Then, for each length n, only word sequences whose score is greater than μ_n are selected, whereas others are discarded. From a computational point of view, this whole scheme can efficiently be implemented using a sequencial scanning of Ω.

[1] A lemma is a canonical form of a word. For example, plural nouns are reduced to their singular form, conjugated verbs are reduced to their infinitive form, *etc.*

OOV word: JAMAICAN
French word: *jamaïcaines* (feminine, plural)
Related lemmas: afro-american, mixed, pop, techno, west indian
Found *n*-grams: jamaican musics, jamaican origins, jamaican and african, jamaican quarters, jamaican songs, his (her) jamaican roots, jamaican waters, jamaican and caribbean

OOV word: ONCOLOGISTS
French word: *cancérologues* (masculine or feminine, plural)
Related lemmas: biologist, cardiologist, doctor, lawyer, neurologist, paediatrician, psychiatrist, surgeon
Found *n*-grams: oncologists who specialize in, oncologists from hospitals, many oncologists, american oncologists, for the oncologists, most oncologists, oncologists and doctors, according to oncologists

OOV word: PARK (Verb)
French word: *garent* (3^{rd} person, plural)
Related lemmas: burn, crash, immobilize, station
Found *n*-grams: park on the, park in front of the, park along, permanently park, they park, double park, who park, park their vehicle

Table 1. Examples of automatically found word sequences for 3 OOV words.

2.3. Instantiation through paradigmatic relations

In our experiments, the set of word sequences Ω is made of all the n-grams of the pre-existing LM[2] and $\mathcal{R}$ has been instantiated by paradigmatic relations between lemmas. These relations show which lemmas appear in same lexical environments, while they do not necessarily appear together within these environments [7]. Typically, this results in automatically linking a word with its synonyms, hypernyms, antonyms, *etc.* In our work, these relations have been trained automatically from French newspaper archives. For each lemma ℓ_1, a context vector is computed from this large text corpus by sliding a fixed-size window (20 words) along the whole text. This context vector gathers the frequencies of words appearing with ℓ_1 within a same window. Then, only the 10 lemmas ℓ_2 whose context vector results in the highest cosine measure with the context vector of ℓ_1 are considered as linked with ℓ_1. At the same time, this cosine value is defined as the strength $r(\ell_1, \ell_2)$ of this link.

Following this scheme, Table 1 lists a few automatically found word sequences for 3 sample OOV words, presented along with their semantically related lemmas. For understanding purposes, these examples have been translated from French into English. However, since French contains word inflection forms which do not exist in English, genders, numbers and tenses of original French OOV words are precised. From these examples, it clearly appears that the word sequences found by our technique make sense and are grammatically consistent. Hence, their use while decoding a speech signal should be worthwhile to transcribe OOV words. Thus, these sequences are now considered as new n-grams for which one has to compute conditional probabilities before integrating them into a baseline LM.

[2] This represents an amount of about 5 millions 4-grams.

3. ASSIGNING PROBABILITIES

After finding semantically consistent *n*-grams for a new word, conditional probabilities must be computed in order to give a real existence to this word in a pre-existing LM. To do this, given a selected *n*-gram W, the joint probability $P(W)$ is first computed as:

$$P(W) = \frac{\sum_{V \text{ st. } W \equiv V} P(V)}{N_W} , \quad (4)$$

where $P(V)$ is the joint probability of an equivalent *n*-gram V in the LM, and N_W denotes the number of these equivalent *n*-grams. Unfortunately, sometimes, no *n*-gram can be found for a given OOV word because the latter cannot be associated with any other word according to the automatically trained paradigmatic relations. In this case, only a unigram can be added to the LM. However, the probability of this unigram cannot be estimated as in (4). To circumvent this problem, a default unigram probability is empirically set to 10^{-8}. Since this case is rather rare—this phenomenon only affects less than $20\,\%$ of our OOV words—, no specific attention has been paid to this probability. Hence, the default value used does not result from a tuning process. It only corresponds to a typical value for rare unigrams already listed in our back-end experiment LM. Then, using the Bayes law, conditional probabilities are computed, probability mass of modified LM histories are renormalized and backoff weights are rescored.

4. EXPERIMENTS AND RESULTS

Experiments are carried out on 3 hours of broadcast news shows coming from the French corpus ESTER [8]. These shows are split into a development set and a test set. As it can be seen from statistics of Table 2, the number of OOV words in these sets is rather limited. Hence, experiments presented in this section must be thought as still preliminary, and results should be carefully interpreted. Then, let us recall that only common words are considered while proper nouns have been discarded, hence the low OOV rates. For each dataset, remaining OOV words are added to the vocabulary and phonetized using the French grapheme-to-phoneme converter LIAPHON [9]. Assuming that the POS of the OOV words are known, new *n*-grams are found, their probabilities are computed and they are integrated into our 4-gram LM based on a 65K words vocabulary. A new transcript is finally generated.

To validate the performance of our vocabulary adaptation scheme, the WER of this new transcript is compared to results obtained using other strategies. First, oracle WERs have been computed. These rates correspond to the best theoretically obtainable WERs, *i.e.*, assuming all considered OOV words are well recognized. Another strategy consists in limiting inserted *n*-grams to the sole unigrams. This corresponds to the most basic way of integrating new words into the baseline LM. Finally, our technique is also compared to the state-of-the-art open vocabulary LM strategy in which the probability of a new word w_{oov} given an history h is computed as:

$$P(w_{oov}|h) = P(c_{oov}|h) \times P(w_{oov}|c_{oov}) , \quad (5)$$

where c_{oov} is the class of w_{oov}, and the term $P(w_{oov}|c_{oov})$ prevents from probability over-estimating for OOV words. In practice, a 4-gram LM based on 14 lexical classes corresponding to POS categories (verbs, nouns, adjectives, *etc.*) was built so that it results in the same baseline performances as those obtained using the closed vocabulary LM. Besides, the probability $P(w_{oov}|c_{oov})$ has been set to 10^{-5} after optimizing the WER on the development set.

	Development	Test
Number of OOV words	91	78
Number of OOV tokens	83	72
Number of words in the reference	20,080	20,190
OOV rate	0.45 %	0.39 %

Table 2. Statistics on OOV words in the development set and in the test set.

		Development	Test
Word only vocabulary	Initial vocabulary	23.32	22.36
	Oracle	22.69 (–0.63)	21.82 (–0.54)
	Unigrams	23.04 (–0.28)	22.21 (–0.15)
	N-grams	***22.83 (–0.49)***	***22.08 (–0.28)***
Word+class vocabulary	Initial vocabulary	23.37	22.48
	Oracle	22.75 (–0.62)	21.95 (–0.53)
	Augmented	**22.88 (–0.49)**	**22.19 (–0.29)**

Table 3. WERs measured on the development and on the test sets using a standard LM (Word only vocabulary) or using an open vocabulary LM (Word+class vocabulary). For both setups, baseline and oracle results are presented, as well as those obtained after augmenting the vocabulary. Considering the closed vocabulary LM, experiments have been carried out by considering only unigrams or by relying on automatically found *n*-grams. In parentheses, the absolute WER variation with respect to the baseline WER.

The WERs measured using each of these strategies for the development and the test sets are presented in Table 3, where our method is referred by the label "*N-grams*". While the absolute oracle WER decrease is of 0.6, it first appears that using only unigrams already leads to an absolute gain of 0.3. This is not surprising since OOV words are usually long words, which makes them acoustically unambiguous and easily recognizable by the sole acoustic models. This case is illustrated by example #1 of Table 4. Second, it appears that our technique leads to an absolute WER decrease of 0.5 on the development set. This improvement is similar to the one returned by the open vocabulary LM. The same trends can be observed on the test set, though global improvements are a bit lower. Especially, our method still leads to the same improvement as when using an open vocabulary LM. This is all the more interesting since these absolute WER decreases of 0.3 are both statistically significant according to a paired t-test and to a Wilcoxon test[3].

To evaluate the impact of the different strategies on added OOV words, WERs on these sole words have been measured. These rates, referred as WER_{oov}, are presented in Table 5. On the one hand, as previously highlighted, it appears that using unigrams already results in correctly transcribing up to $50\,\%$ of all the OOV words. On the other hand, results show that the open vocabulary LM leads to more frequently correctly transcribe OOV words. However, after precisely analyzing the *n*-grams and the transcripts generated using our technique, it appears that these slightly worse WER_{oov} results emanate from the absence of paradigmatic relations for some OOV words, thus leading to only consider unigrams. This problem could

[3]For our technique, p-values are of 2.7×10^{-5} and 2.2×10^{-5} for the paired t-test and for the Wilcoxon test, respectively, while they are of 1.7×10^{-6} when using the open vocabulary LM. For each test, the confidence level α was set to 0.05.

Example #1

Reference	these are **pan-islamist** movements
Baseline	these are LOINCLOTH ISLAMIST movements
Unigrams	these are **pan-islamist** movements
N-grams	*these are **pan-islamist** movements*
Open vocabulary	these are **pan-islamist** movements

Example #2

Reference	dozens of women dressed in blue **djellabas**
Baseline	dozens of women dressed in blue DJELLABA
Unigrams	dozens of women dressed in blue DJELLABA
N-grams	*dozens of women dressed in blue **djellabas***
Open vocabulary	dozens of women dressed in THE **djellabas**

Table 4. Comparison of 2 reference segments transcribed using different setups. These examples have been translated from French. Erroneous words are in uppercase, while OOV words are in bold.

		Development	Test
Word only vocabulary	Initial vocabulary	100	100
	Unigrams	50	66
	N-grams	*35*	*32*
Word+class vocabulary	Initial vocabulary	100	100
	Augmented	**21**	**21**

Table 5. WERs measured on the sole OOV words (WER_{oov}) for the same set-ups as in Table 3.

probably be avoided by relying on additional relations between lemmas. Nonetheless, when opposed to the similar WERs previously reported in Table 3, these WER_{oov} differences interestingly mean that, for each transcribed OOV word, our technique leads to more effectively transcribe its nearby words with respect to the open vocabulary LM. This point is illustrated by the example #2 of Table 4. By relying on paradigmatic relations, our technique associates the OOV word "djellabas" with other types of garment and, then, leads to add *n*-grams with adjectives refering to colors. This enables the system to produce the right transcription of the whole utterance, whereas the open vocabulary LM leads to correctly transcribe the word "djellabas" but it does not for the adjective "blue". Hence, our technique achieves a better integration of OOV words according to their meaning.

5. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a new method to automatically find semantically consistent *n*-grams containing OOV words in order to add these *n*-grams into a baseline LM within a vocabulary adaptation process. This strategy relies on the use of semantic relations between lemmas in order to assimilate OOV words to equivalent in-vocabulary words. By instantiating this general scheme through the use of automatically learned paradigmatic relations, we have shown that it results in the same WER improvements as when using the state-of-the-art open vocabulary strategy. Moreover, while open vocabulary LMs tend to consider all the words of a same class as having a same semantic role, the merit of our approach is to achieve a better semantic integration of OOV words. This conclusion is all the more interesting since our method remains based on a closed vocabulary LM, as those used in most current LVCSR systems.

Nonetheless, further improvement possibilities should be investigated. First, apart from paradigmatic relations, it could be interesting to rely on other relations between lemmas. Especially, it could be useful to consider relations between compounds and their constituent words. For example, an OOV word like "orthomyxovirus", for which no paradigmatic relation can easily be learned since this is a very rare word, could be automatically reduced to its simpler form "virus" by using word composition rules. Additionally, when only a few *n*-grams are automatically found, the relation $\mathcal{R}$ could be avoided so that the equivalence relation between words would rely on the sole morphosyntactic information. Hence, the integration of a minimum number of *n*-grams would be guaranteed for each OOV word. Then, further investigation should be done according to the integration of new *n*-gram probabilities into a pre-existing LM. Especially, future work should focus on the probability mass re-estimation problem of enriched LM histories in order to ensure the whole LM distribution not to degenerate when to many new *n*-grams are added. This is an arduous task since, when estimating the baseline LM, the probability masses are estimated through a complex process involving elaborate smoothing techniques. Finally, our strategy should be tested on a dataset with a larger number of OOV words. Similarly, it should be plugged with an automatic OOV word selection scheme in order to make the whole vocabulary adaptation process completely unsupervised. This could also be a good opportunity to assess the behavior of our technique when thousands of OOV words are added to the system.

6. REFERENCES

[1] Thomas Kemp and Alex Waibel, "Reducing the OOV rate in broadcast news speech recognition," in *Proc. of ICSLP*, 1998, pp. 1839–1842.

[2] Cira Martins, António Texeira, and João Neto, "Dynamic vocabulary adaptation for a daily and real-time broadcast news transcription system," in *Proc. of the Spoken Language Technology Workshop*, 2006, pp. 146–149.

[3] Hiroki Yamazaki, Koji Iwano, Koichi Shinoda, Sadaoki Furui, and Haruo Yokota, "Dynamic language model adaptation using presentation slides for lecture speech recognition," in *Proc. of Interspeech*, 2007, pp. 2349–2352.

[4] Alexandre Allauzen and Jean-Luc Gauvain, "Open vocabulary ASR for audiovisual document indexation," in *Proc. of ICASSP*, 2005, vol. 1, pp. 1013–1016.

[5] Cira Martins, António Texeira, and João Neto, "Automatic estimation of language model parameters for unseen words using morpho-syntactic contextual information," in *Proc. of Interspeech*, 2008, pp. 1602–1605.

[6] Grace Chung, Stephanie Seneff, Chao Wang, and Lee Hetherington, "A dynamic vocabulary spoken dialogue interface," in *Proc. of ICSLP*, 2004, pp. 1457–1460.

[7] Gregory Grefenstette, "Corpus-derived first, second and third-order word affinities," in *Proc. of EURALEX*, 1994, pp. 279–290.

[8] Sylvain Galliano, Édouard Geoffrois, Djamel Mostefa, Khalid Choukri, Jean-François Bonastre, and Guillaume Gravier, "The ESTER phase II evaluation campaign for the rich transcription of French broadcast news," in *Proc. of Interspeech*, 2005, pp. 1149–1152.

[9] Frédéric Béchet, "LIA_PHON : un système complet de phonétisation de textes," *Traitement Automatique des Langues (TAL)*, vol. 42, no. 1, pp. 47–67, 2001.