USING NON-WORD LEXICAL UNITS IN AUTOMATIC SPEECH UNDERSTANDING

M. Peñagarikano, G. Bordel, A. Varona, K. López de Ipiña

Dpto. Electricidad y Electrónica
Universidad del País Vasco (UPV/EHU)
Lejona, Vizcaya, SPAIN

ABSTRACT

If the objective of a Continuous Automatic Speech Understanding system is not a speech-to-text translation, words are not strictly needed, and then the use of alternative lexical units (LUs) will bring us a new degree of freedom to improve the system performance. Consequently, we experimentally explore some methods to automatically extract a set of LUs from a Spanish training corpus and verify that the system can be improved in two ways: reducing the computational costs and increasing the recognition rates. Moreover, preliminary results point out that, even if the system target is a speech-to-text translation, using non-word units and post-processing the output to produce the corresponding word chain outperforms the word based system.*

1. INTRODUCTION

The use of words as lexical units (LUs) in Continuous Automatic Speech Understanding has been, basically, not questioned. Only very few recent papers deal in some way with alternative units [1][2][3][4]. The need for new units has been better seen from languages were the word concept is no clear (i.e. Chinese) or those were words are highly structured (i.e. German or, to a lesser extent, Spanish). We thought that the only reason to adopt the word is given by the fact that normally it is the element composing texts, and then word based Language Models can be learnt straightforwardly. Our interest for alternative LUs came from this field: Language Modelling. Using words as units, our models ignore important internal word structure and phrasal structures are modelised with a high cost. If the system objective is not a speech-to-text translation, words are not strictly needed, and then the use of alternative LUs will bring us a new degree of freedom to improve the system performance.

Once the word is questioned, many approaches can be adopted to investigate the alternatives. As the word is used as a connecting element between the phonetic knowledge and the syntactic-semantic-pragmatic knowledge (the Language Model), an adequate adaptation to both parts would lead to the best results. Nevertheless, the high computational costs suggest that we should solve the problem only partially, trying to obtain a performance improvement for the language model and observing if it results in an improvement of the whole system.

Next section describes different aspects we tested to automatically obtain LU sets from samples. The experiments carried out gave us some results that are briefly shown in section 3. The obtaining of the LUs was carried out on textual information attending to the perplexity given by the Language Model (3.1), whereas the whole system evaluation is made in terms of recognition rates (3.2).

2. OBTAINING THE ALTERNATIVE LEXICAL UNITS

To automatically obtain a set of LU from a database we propose a procedure based on the alteration of a predetermined set. Two algorithms are tried. The first one needs a criterion to generate the new units, and the second needs a criterion to evaluate the performance given by the altered sets. So, the following paragraphs are devoted to these four aspects:

- The Initial set of units.
- The algorithms.
- The generation of new units criterion.
- The evaluation criterion.

2.1 Initial set of units

The computational cost of the analysis for one utterance is linear to its length and at least quadratic to the number of LUs considered by the system (that is the case for a smoothed bigram model). So, if it were not that the recognition rates drop dramatically (for a fixed kind of LM), it would be worth using single phonemes as LUs because of the reduction of the quadratic dimension at the expense of the linear one. Obviously, the low recognition rate is due to the loss of the information about phoneme combinations contained in the word set.

So, we can start our search for a good LU set with the phoneme set. As the new units were generated, the recognition rate and the computational costs will grow. The hope is that at some point the performance will be more satisfactory that the word based system.

As we will see later, the criterion for new word generation is based on probabilistic considerations on the database. This fact made us realise that those words appearing only few times in the database had no chance to be formed from phonemes. So, three alternative initial sets were also tried: phonemes plus words appearing three or less times, twice, and once.

Two more initial sets have been tried: supposing that semantic constrains are good criteria to form LUs, we also tried a set of pseudo-morphemes (not exactly morphemes in the linguistic sense but a very close approximation), and attending to the idea that recognition rates can be improved at the expense of increasing the computational cost, we also used the word set as an initial set.

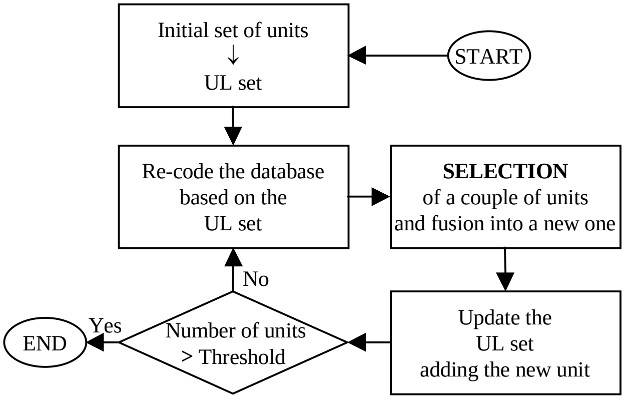
* This work has been partially supported by the Spanish CICYT under grant TIC-94-0210-E and by the UPV/EHU under grant UPV-224.310-EA036/97.

As a result, these are the initial sets tried:

- Phonemes
- Phonemes + words appearing once.
- Phonemes + words appearing once or twice.
- Phonemes + words appearing once, twice or three times
- Pseudo-Morphemes
- Words

2.2 Algorithms

Two algorithms had been applied. The first one implements a simple *Greedy* scheme consisting on the iterative generation of new LUs according to the selected criterion (see **Figure 1**). The success of this mechanism relies entirely on the appropriateness of the generation criterion.



Initial set of units
↓
UL set
START
Re-code the database based on the UL set
SELECTION of a couple of units and fusion into a new one
No
Update the UL set adding the new unit
END
Yes
Number of units > Threshold

Figure 1. The first algorithm to obtain the LU sets uses a *Greedy* approach.

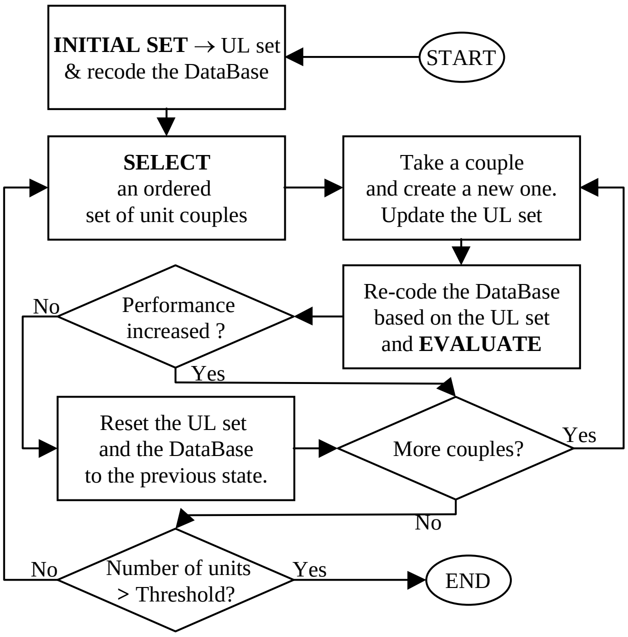
The second algorithm is taken from [6], where it is used to obtain phrasal structures. This algorithm tries a more intelligent evolution of the LU set assuring that the recognition rate monotonically decreases all through the execution. This is implemented as a *Local Search* scheme. The optimal implementation consists on trying, at each step, all the possible new LUs and selecting the one showing the highest reduction of the recognition rate. This strategy is computationally prohibitive, so a sub-optimal approach is implemented: a subset is determined, and iterativelly all the units improving the performance are accepted before a new subset is formed. The construction of these ordered subsets is based on the same criteria used in algorithm 1.

2.3 Generation of new units criteria

The new units are always generated by concatenation of two previously existing units. The selection of the two units to be joined is based on the maximisation of a predetermined function. Three functions were chosen to be tested. The simplest one is the frequency observed in the database. This approach was also adopted in [1]:

$$\mathbf{F}(u,v)=\frac{\mathbf{N}(u,v)}{N}$$

were u and v are the LUs, **F** the frequency function, **N** the count function and N the total number of LUs composing the sentences of the database.



INITIAL SET → UL set & recode the DataBase
START
SELECT an ordered set of unit couples
Take a couple and create a new one. Update the UL set
Re-code the DataBase based on the UL set and EVALUATE
No
Performance increased ?
Yes
Reset the UL set and the DataBase to the previous state.
More couples?
Yes
No
No
Number of units > Threshold?
Yes
END

Figure 2. The second algorithm to obtain the LU sets uses a “sub-optimal” *Local Search* approach.

A second function we tried was a correlation coefficient (**CC**) as it is proposed in [6] to select word phrases in a recognition task:

$$\mathbf{CC}(u,v)=\frac{\mathbf{N}(u,v)}{\mathbf{N}(u)+\mathbf{N}(v)}$$

Observing that low frequency LUs can present high **CC** values but they have low impact in the final performance of the system, we defined a modified CC function (**MCC**) diminishing these values:

$$\mathrm{M}\mathbf{CC}(u,v)=\mathbf{CC}(u,v)\ \mathbf{N}(u,v)=\frac{\mathbf{N}(u,v)^2}{\mathbf{N}(u)+\mathbf{N}(v)}$$

2.4 Evaluation criteria

In this work, we tried to better the recognition system performance by improving the Language Model for a fixed phonetic model. Hence, for Algorithm II we used a LM evaluation: the perplexity. A more realistic whole-system evaluation would be computationally unaffordable. Nevertheless, the UL sets obtained were evaluated via recognition rate of the whole system.

The Perplexity function as usually expressed to comparatively evaluate Language Models is not valid in this case. There is a dependency on the units used to compose the sentences, so it must be altered to be invariable to the units change. This can be straightforwardly accomplished by basing the evaluation on an invariable unit, in our case the phoneme:

$$\mathbf{PP}(M)\Big|_T=2^{\left[\frac{-1}{F}\sum_{i=1}^{N}\mathbf{log_2}\,P(W_i \mid M)\right]}$$

were **PP** is the perplexity function, M is the Language model, T is the test set, F and N are the number of phonemes and units composing the sentences respectively, and W_i is each unit.

3. EXPERIMENTAL RESULTS

The experiments were carried out over a task-oriented Spanish speech corpus consisting of 9309 sentences (93460 words, 531456 phonemes) and a vocabulary of 1284 words [8]. This corpus represents a set of queries to a Spanish geography database. This is a very specific task designed to test integrated systems (acoustic decoding + language modelling) in automatic speech understanding, which leads to a very low perplexity. To obtain the recognition rates, a test set of 600 utterances was used.

The acoustic models of the system were fixed, and the language modelling part was implemented by means of K-TLSS(S) (K-Testable Language in the Strict Sense, smoothed) which are a kind of Variable N-grams[7]. 8262 sentences were used for training and 1147 for test.

The whole combinatory of initial sets, algorithms and generation criteria was tested. All these experiments were carried out in textual form due to the nearly one to one mapping from letters to phonemes in Spanish. For the validation experiments with the whole recognition system, the previous best case was repeated based on phonemes.

3.1 Perplexity

The first conclusion of our experiments is that all the three generation criteria used leads to very similar performances and, even though the frequency criterion is the simplest one, it gives the best results. **Figure 3** is a sample of this behaviour, where the initial set is composed of pseudo-morphemes.

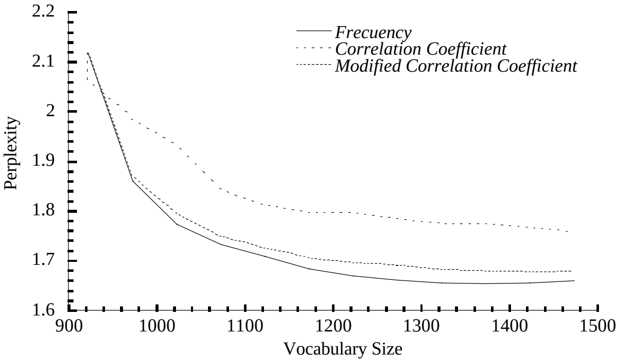


Figure 3. Perplexity reduction using the three criteria to grow the LU sets from an initial pseudo-morpheme set. Even though the frequency criterion is the simplest one, it gives the best results.

In relation with the algorithms, we also obtained very similar results. Obviously the local search strategy can not be worse than the greedy strategy, but the sub-optimal implementation leads to comparable performances. **Figure 4** shows this fact in the case of the phoneme initialisation.

From the study of the initial sets, first we see that the inclusion of low count words in the initial phoneme set has not positive effect (see **Figure 5**). We introduced these words because their low frequencies gave no chance to be formed from phonemes, but this same reason (their sparseness in the test set) makes their inclusion not justified.

So, we will focus on the three original initialisations: phonemes, pseudo-morphemes and words (**Figure 6**). Clearly some sets of units are better than the word set. For the phoneme and pseudo-morpheme initialisations there are a range of sets presenting better perplexities at lower vocabulary sizes. What is more, the method allows the selection of smaller vocabularies with some quality loss or better quality systems with some size increase. In this direction, the use of the word set as an initial set allows a rapid improvement in performance with a small increment in size.

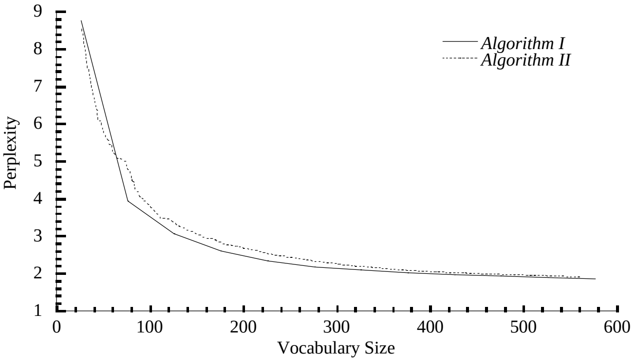


Figure 4. Both algorithms lead to similar results, so the much more economical *Greedy* strategy demonstrates to be good enough to our purposes.

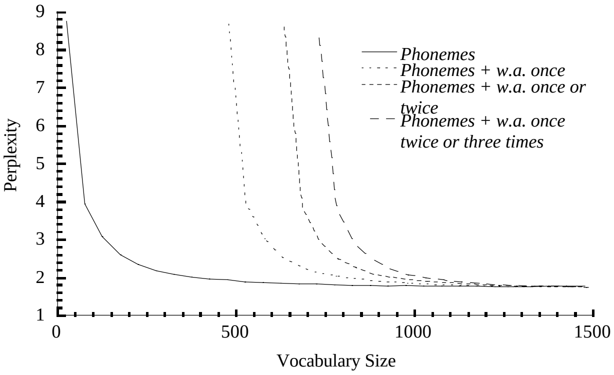


Figure 5. The explicit inclusion of low count words in the initial phoneme set has no positive effect. The inclusion is not justified because of the low impact of these words in the evaluation.

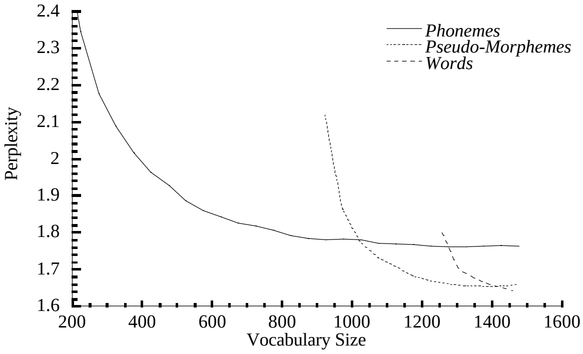


Figure 6. The sets formed from phonemes and pseudo-morphemes can improve the word-based model performance in terms of perplexity and size (related to the required computational effort). The application of the procedure to the word set can improve the model accuracy.

3.2 Recognition rate

As a first experiment to observe the quality of the whole recognition system using alternative LUs, we chosen two sets based on a pseudo-morpheme initialisation and the application of the algorithm I. As our objective was the use of these sets in the recognition system, the procedure was performed using the phonetic transcriptions of the database. The first set (UL1) is that which presents a similar perplexity to the word based model, and the second (UL2) is that with a similar vocabulary-phonetic-length (the length of the vocabulary in phonemes, which is directly related to the computational cost of the recognition process because it is the size of the quadratic axis).

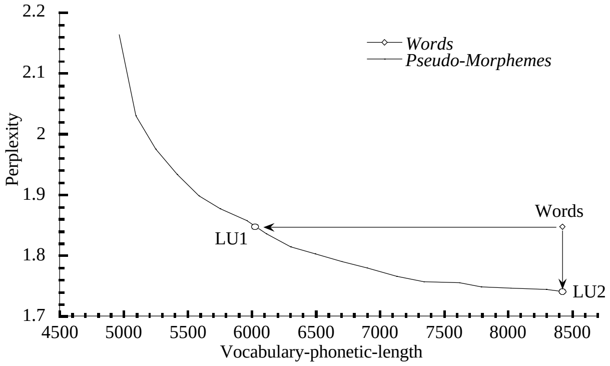


Figure 7. Two sets were chosen to build the whole recognition system. UL1 presents a similar perplexity for the Language Model to the word-based model. UL2 presents a similar vocabulary-phonetic-length to the word set.

The use of these two sets was tested through 600 uttered sentences, manually weighting the acoustic and language contributions for each case for the best value. The first observation was that the UL1 set did not fulfil our expectations. The reason is clear: the use of a real system, implementing a beam-search strategy can drastically reduce the quadratic aspect of the recognition procedure, and then, the improvement in computational cost is less than expected. **Table 1** shows that even widening the beam (increasing the average number of states, that is, increasing the computational cost), the word based performances are not achieved.

Table 1. Recognition rates (LURR for LUs and SRR for sentences) using words and the selected LU sets. The Beam-Search has been tuned to perform similar computational cost executions (average number of states-ANS). The word performances did not come to our expectations, but the sentence performances did.

Units	ANS	LURR%	SRR%
UL1	1274.89	80.20	40.00
WORDS	1014.26	85.59	44.17
UL2	941.15	76.84	50.00

The use of UL2 brings a hopeful result: the sentence recognition rate is clearly improved, so although the unit level results are worse, a post-processing for realignment from UL2 to words can improve the system.

In **Table 2** we show the result of a very simple automatic realignment where the only information used is the set of words. It seems not to be too difficult to obtain an alignment procedure preserving or even improving the sentence error rate (this is part of our actual work). For a set obtained from a word set as initial, the post-processing procedure would be trivial.

Table 2. For a transcriptor, the ULN2 units can be used applying a post-process. A simple alignment leads to recognition rates that are slightly higher than those obtained using words. So, it is reasonable that a more intelligent post-process will improve the performance significantly.

Units	ANS	LURR%	SRR%
WORDS	1014.26	85.59	44.17
WORDS from UL2	941.15	86.68	48.17

4. SUMMARY

Some alternatives had been tried in order to obtain sets of lexical units different from the word set to increase the performance of Automatic Speech Understanding systems. The experiments were carried out over a task oriented Spanish corpus. As a conclusion, we saw that a very simple strategy can be successful to this end. The use of a beam-search strategy reduces the expected benefits of the change of lexical units but the recognition rate is still increased.

5. REFERENCES

[1] Hwang K., “Vocabulary Optimization Based on Perplexity”. *International Conference on Acoustics, Speech and Signal Processing*, Munich, 1997, pp. 1419-1422 (vol 2).

[2] Yang K.-C., Ho T.-H., Chien L.-F., Lee L.-S., “Statistics-Based Segment Pattern Lexicon - A New Direction for Chinesse Language Modelling” *International Conference on Acoustics, Speech and Signal Processing*, Seattle, 1998, pp. 169-172 (vol 1).

[3] Geunter P., “Using Morphology towards better Large Vocabulary Speech Recognition Systems”, *International Conference on Acoustics, Speech and Signal Processing*, Detroit, 1995, pp. 445-448 (vol. 1).

[4] Mayfield L., Ries K., “An Automatic Method For Learning Japanese Lexicon for Recognition of Spontaneous Speech” *International Conference on Acoustics, Speech and Signal Processing*, Seattle, 1998, pp. 305-308 (vol 1).

[5] Klakow D., “Language-Model Optimization by Mapping of Corpora” *International Conference on Acoustics, Speech and Signal Processing*, Seattle, 1998, pp. 701-704 (vol 2).

[6] G. Riccardi, A.L. Gorin, A. Ljolje, M. Riley, “A spoken Language System for Automated Call Routing.” *International Conference on Acoustics, Speech and Signal Processing*, Munich, 1997.

[7] Bordel G., Varona A., Torres I. “K-TLSS(S) Language Models for Speech Recognition”. *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Munich, April 1997, pages 819-822.

[8] Diaz J. E., Rubio A. J., Peinado A. M., Segarra E., Prieto N., Casacuberta F., “Development of Task Oriented Spanish Speech Corpora”, *Proceedings of EUROSPEECH* 93.