SPEECH-TO-TEXT TRANSLATION BY A NON-WORD LEXICAL UNIT BASED SYSTEM.

Mikel Peñagarikano, Germán Bordel

Dpto. Electricidad y Electrónica
Universidad del País Vasco (UPV/EHU)
Lejona, Vizcaya, Spain
{mpenagar,german}@we.lc.ehu.es

ABSTRACT

Speech understanding applications where a word based output of the uttered sentence is not needed, can benefit from the use of alternative lexical units. Experimental results from these systems show that the use of non word lexical units bring us a new degree of freedom in order to improve the system performance (better recognition rate and lower size can be obtained in comparison to word based models). However, if the aim of the system is a speech-to-text translation, a post-processing stage must be included in order to convert the non-word sequences into word sentences. In this paper a technique to perform this conversion as well as an experimental test carried out over a task oriented Spanish corpus are reported. As a conclusion, we see that the whole speech-to-text system neatly outperforms the word-constrained baseline system.

1. INTRODUCTION

Only few recent papers deal in some way with alternative units to words in Language Modelling for Continuous Speech Understanding. The need for new units has been better seen from languages were the word concept is not clear (i.e. Chinese) [1], or those were words are highly structured (i.e. German or, to a lesser extent, Spanish) [2] [3] [4].

The Continuous Speech Recognition System, which was based on non-word lexical units (LU) that are automatically acquired from the same text samples used to learn the target language structure, was evaluated in terms of the learned units, and the experiments pointed out that the output had enough information to be post-processed and outperform the word based system (the same system constrained to use words) [5]. Therefore, a post-processing step for translating the recognized LU output into words is analyzed.

The starting point of the post-process is the output of the lexical unit based recognition system. A compromise between the translation post-process complexity and its efficiency is made in order to avoid the whole system overload (recognition + translation). Thus, complex translation algorithms are rejected while simpler ones are studied. Another constraint added to the post-process is the portability between different lexical unit sets. That is, the translation system must be able to automatically translate the lexical units into words, whatever the lexical units are (phoneme sequences, word sequences, or mixed sequences).

As a first approach to the problem, in the second section it is assumed that the system recognized sentences can be straightforwardly aligned from the LU phonetic transcription (phoneme string) into words, in such a way that there is at least one word sequence with the same phonetic transcription. In the third section, the problem of translating the recognized sentences that lead to phoneme strings which cannot be so aligned is studied. Results from experiments of speech-to-text translation are also given.

Those experiments have been carried out over a task-oriented Spanish speech corpus, where translation has been automatically learned from the recognition system output.

2. TRANSLATION OF ALIGNABLE STRINGS

The translation of those phoneme strings that can be aligned into words is divided into two steps. First, the combinatory set of all the possible word sequences is obtained and second, one out of them is selected.

2.1. Obtaining the word sequence combinatory set

In order to obtain all the possible word sequences that have the same phonetic transcription as the recognized LU sequence, a *backtracking* technique has been used. As shown in **Fig.1**, a structure has been created, which includes the phonetic transcription of the LU sequence as well as all the words starting with those phonemes and fitting the phoneme string. By using this structure, the whole combinatory word sequences that entirely match the phoneme string can be obtained.

This work has been partially supported by the UPV/EHU under grant UPV-224.310-EA036/97

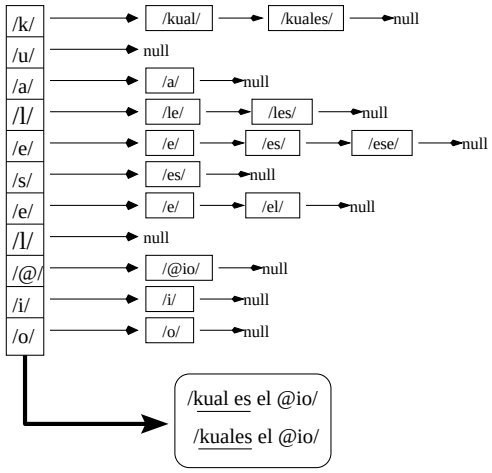

Figure 1: In order to obtain the word sequence combinatory set, a backtracking algorithm is applied to the structure that contains all the word matching any phonetic substring.

The combinatory set can contain more than one sequence since there may be different word sequences matching the same phoneme string. **Fig.1** shows the structure derived from a recognized sentence. The phoneme string obtained from the LU sentence is */kualesel@io/*, and there are two possible fitting word sentences (word are detached with spaces), */kual es el @io/* ("which(singular) is the river") and */kuales el @io/* ("which(plural) the river"). This is due to the fact that both word subsequences */kual es/* and */kuales/*, have the same transcription, so that any recognized phonetic string that matches */kuales/* will also match */kual es/*.

An algorithm has been tried in order to obtain this kind of subsecuences from the aligned word sequence. In **Fig.2** the output of an alignment is symbolically shown. Six word sequences are obtained. Some subsequences (those in white) are common to all the sentences, but some others (those in grey) are not.

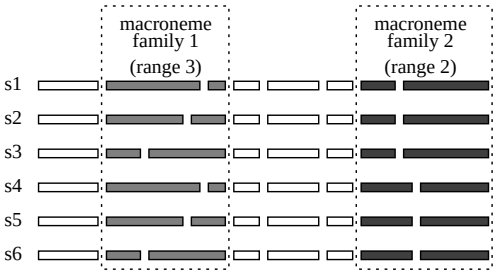

Figure 2: A family of macronemes consists of those word subsequences that are not common to all the sentences obtained from the LU output derived phoneme string.

We define *family* of *macronemes* as the set of shortest different word subsequences matching the same transcription in all the aligned sentences that have been derived from a phoneme string. According to this definition, the previously mentioned */kual es/* and */kuales/* are macronemes, and they make up a family.

We define *range* as the number of macronemes in a family. **Fig.3** shows three macroneme families. As seen previously, all the macronemes in a family share the same phonetic transcription. Note that the number of words making up the macronemes inside a family can differ.

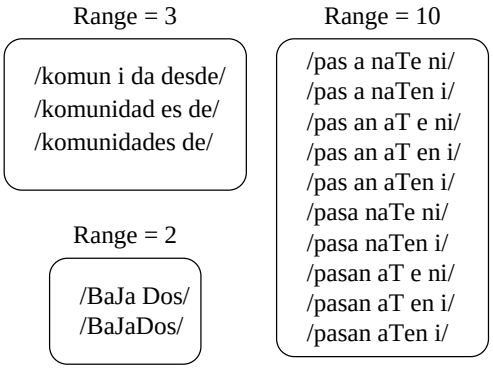

Figure 3: All the macronemes in a family share the same phonetic transcription. The number of words making up the macronemes inside a family can differ.

Given that the only differences in the aligned sentences are inside each appearing family, the total number of possible word sequences is equal to the product of the ranges of all the families:

$$N = \prod_{i=0}^{n} \rho(f_i) \quad (1)$$

with N being the total number of sentences, f_i a family of macronemes, n the number of families appearing in the phoneme string and $\rho(f_i)$ the range of f_i.

2.2. Word sequence selection

Once the word sequence combinatory set has been entirely obtained, the next stage consist of selecting one sequence out of them.

The procedure starts by finding out which macronemes appear in each aligned sentence. In order to do that, a search algorithm has been used, which relies on looking through the word sequence, from left to right, searching for the longest appearing macroneme (a list of all possible macronemes has been previously obtained). Once a word subsequence has been matched by a macroneme, the search begins again at the end of the matched subsequence.

When the macronemes have been extracted, the sentence can be described as a sequence of words and macronemes. Using the example in **Fig.2**, any aligned sentence can be expressed as:

$$s_i = (w_1^i, \ldots, w_6^i) = (w_{i_1}^i, m_1^i, w_{i_2}^i, w_{i_3}^i, w_{i_4}^i, m_2^i) \quad (2)$$

with w_j^i being the jth word and w_j^i the jth macroneme in the s_i sentence. Given that the word differences are only inside the macronemes:

$$w_{i_k}^i = w_{j_k}^j = w_k, \ \ \forall\, i, j \tag{3}$$

$$s_i = (w_1, m_1^i, w_2, w_3, w_4, m_2^i) \tag{4}$$

Thus, selecting one word sequence is equivalent to choosing one macroneme out of each family:

$$s_{sel} = (w_1, m_1^{sel}, w_2, w_3, w_4, m_2^{sel}) \tag{5}$$

In order to select the right sentence, stochastic macroneme models have been built, which predict the probability of each macroneme according to their left and right context on the sentence (the left/right context may be a word, the beginning of the sentence or the termination):

$$P_{m_i} = P(l_{m_i} m_i r_{m_i}) \tag{6}$$

$$s_i = \{m_1^i, m_2^i, \ldots, m_n^i\} \tag{7}$$

$$P(s_i) = \prod_{j=0}^{n} P_{m_j^i} \tag{8}$$

with P_{m_i} being the probability of the m_i macroneme constrained to its both left and right contexts, (l_{m_i}, r_{m_i}), and $P(s_i)$ the probability of the sentence s_i. The selected sentence will be the one that maximizes the refered probability:

$$P(s_{sel}) = \max_{i=1\ldots N} P(s_i) \tag{9}$$

Since only word-based text is needed for training the macroneme models, the same training set was used to train both language model and macroneme models. Even though macroneme models (3-gram models) might seem computationally expensives, they are not, since the number of macronemes and their contexts is quite acceptable (the number of obtained macronemes and trained contexts are given in the fifth section).

3. TRANSLATION OF NON-ALIGNABLE STRINGS

Since lexical units can be either bigger or smaller than words, there may be a recognized LU sequence that leads to a phoneme string which cannot be aligned. This problem has been approached by a recursive algorithm.

The recursive translation algorithm is shown in **Fig.4**. As a first step, the longest phoneme substring that can be aligned is searched. Once it has been found, the substring is translated as it has been described for the alignable strings in the former section. As a second step, the same translation algorithm is applied to the remaining substrings. The recursiveness ends up when the applied substring is entirely aligned (translated) or when there is no possible alignment, in which case this substring will be omited.

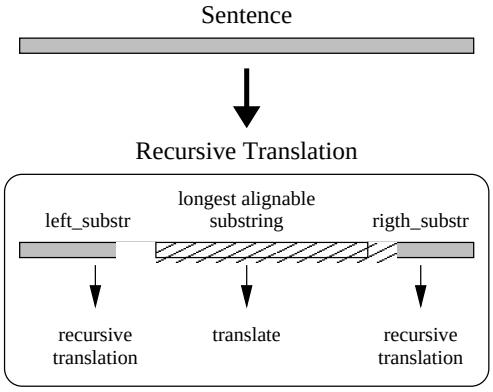

Figure 4: In the recursive translation algorithm, the longest alignable substring is searched and translated into words. the same algorithm is applied to the remaining substrings.

Even though this recursive algorithm obtains a word sequence from the recognized LU output, it leads to the loss of a portion of information, since the substrings that cannot be aligned are omited. Nevertheless, the simplicity of the algorithm makes it suitable for a pots-process translation. Note that the translation of alignable sentences presented in the former section was a particular case of the one described here. In fact, when the recursive translation is applied to a phoneme string that can be entirely aligned into words, the whole string will be translated and recursiveness will end up.

4. EXPERIMENTAL CONTEXT

4.1. Lexical Unit Based Recognition System

The experiments have been carried out over a task-oriented Spanish speech corpus consisting of 9.309 sentences (93.460 words, 531.456 phonemes) and a vocabulary of 1284 words [7]. This corpus represents a set of queries to a Spanish geographic 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.

The acoustic models were fixed and the language modelling part has been implemented by means of K-TLSS(S) (K-Testable Language in the Strict Sense, Smoothed) which are a kind of Variable N-grams [6]. 8.262 sentences have been used for training. The LUs have been automatically infered from the same text samples used to learn the target language structure, resulting in a vocabulary of 1210 units.

4.2. Translation System

600 utterances have been used to carry out the translation test. Since the macroneme vocabulary must be previously obtained in order to apply the translation algorithm, a *leaving K-out (K=50)* technique has been used so that an open test can be carried out. Therefore, 550 sentences have been used to obtain the macroneme vocabulary and the remaining 50 utterances have been translated into words. The obtained macroneme models have been trained from the same text samples used to learn the target language structure (8.262 sentences). The 550 and 50 sentence sets have been exchanged until the whole 600 sentences have been translated.

5. EXPERIMENTAL RESULTS

5.1. Macroneme extraction

Macroneme extraction average values are shown in **Table 1** (no absolute value can be given, due to the *leaving K-out* technique). In average, 117 macronemes were found, but only 49 of them appeared in the training corpus (many macronemes have no real sense at all). 1.411 different contexts were trained for those 49 macronemes from the training corpus.

Obtained macronemes	117
Trained macronemes	49
Trained contexts	1411

Table 1: Not all the obtained macronemes will appear in the training tes, since many of them have no real sense at all.

5.2. Recognition results

Word error rates (WER) and sentence error rates (SER) for the baseline system and our novel system (LU-recognition + translation) are shown in **Table 2**. Sentence error rate is also given for the LU based system.

System	WER%	SER%
WORD-based (baseline)	14.41	55.83
LU-based	—	50.00
LU-based + Translation	12.25	48.67

Table 2: The low sentence error rate of the LU based system has not only been maintained, but also reduced and, comparing the LU+translation system with the baseline system, both the word and sentence error rates have been significantly reduced (15% for the WER and 12.8% for the SER).

In spite of the translation post-process, the low sentence error rate of the LU based system has not only been maintained, but also reduced. This is because there are wrong recognized LU sentences that lead to a correct phonetic transcription which can be later translated into the right word sentence.

Comparing the LU+translation system with the baseline system, both the word and sentence error rates have been significantly reduced (15% for the WER and 12.8% for the SER).

6. CONCLUSIONS

The obtained results evidence that the whole speech-to-text system neatly outperforms the word-constrained recognition system. That is, there is no significant information loss at the translation stage, in such a way that an alternative lexical unit based recognition system can be used in order to build a speech-to-text translator.

These satisfactory results confirm, once again, the advantage of using alternative lexical units in automatic speech understanding.

7. REFERENCES

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

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

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

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

[5] Peñagarikano M., Bordel G., Varona A., López de Ipiña K., " Using Non-Word Lexical Units in Automatic Speech Understanding", *IEEE International Conference on Acoustics, Speech and Signal Processing*, Phoenix, 1999.

[6] 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.

[7] 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.