ISCA Archive
http://www.isca-speech.org/archive

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

AUTOMATICALLY DERIVING CATEGORIES FOR TRANSLATION*

Sergio Barrachina barrachi@inf.uji.es
Juan Miguel Vilar jvilar@inf.uji.es
Unidad Predepartamental de Informática
Universidad Jaume I
12071 Castellón (SPAIN)

10.21437/Eurospeech.1999-530

ABSTRACT

An adequate approach to speech translation for small to medium sized tasks is the use of subsequential transducers —a finite state model— as language model for a speech recognizer. These transducers can be automatically trained from sample corpora.

The use of manually defined categories improves the training of the subsequential transducers when the available data are scarce. These categories depend on the source and target languages we want to translate.

We introduce an automatic approach to derive categories that can be used in training subsequential transducers. This approach extends monolingual word clustering methods to the bilingual case using alignments obtained from statistical models. Experimental results indicate that the models trained with these categories have lower translation errors.

1 INTRODUCTION

Subsequential Transducers (SSTs) are adequate for medium sized tasks of speech input translation [1]. One of their advantages is the existence of efficient learning algorithms, like OSTIA [2]. Unfortunately, OSTIA and similar algorithms require large amounts of training data.

The use of manually generated bilingual categories can improve the translation results when the available data are scarce [3,4]. We present a method that automatically clusters words in classes from the data in the training corpus. These classes could be then used by the SST in the same way as the manual categories. We will show that the use of these classes improves the translation results of an SST.

2 THE CLASSES AND THEIR DERIVATION

There exists efficient monolingual clustering methods that group words in classes [5–8]. This clustering is generally achieved minimizing the perplexity of the resulting class n-gram model.

A first approximation to the problem could be to use one of these monolingual clustering methods in both source and target languages independently to obtain such classes. Unfortunately, it has been shown that:

- The mapping between source and target language tags might not been meaningful in a translation model: it is not evident that there should be a direct correspondence between parts of speech in two different languages [9].
- Word equivalence classes independently derived for two different languages are not always correlated: the class of a source language word will not always give much information about the class of the generated target language word [10]. Och and Weber propose an approach to compute bilingual correlated classes that consist of deriving word classes for the target language using a monolingual method and afterwards determining the word classes for the source language taking the other classes into account.

This means, that in order to use classes in a translation model it is desirable that there exists a strong correlation between those in the input and output languages.

Our objective will be to cluster pairs of target and source words, finding those sets of pairs in such a way that the elements in the same set would be interchangeable; i.e. supposing that $[y, x]$ and $[y', x']$ belong to the same set and the translation of a sentence containing x contains y, then if x is substituted for x' in the source sentence, the translation sentence should have y' in the place where y was.

To accomplish this, we follow these steps:

- The training corpus is aligned using a statistical model, like those in [11].
- The aligned corpus is transformed in a monolingual corpus by labelling the words of the input sentences with their translations.
- The clustering algorithm from [7] is applied to this new corpus.

*Work partially funded by the European Union (ESPRIT Project no. 30268) and by the Spanish C.I.C.Y.T. (project TIC-97-0745-CO2).

- An SST is obtained from the new monolingual corpus and the classes are expanded.

Now we can explain these steps in more detail using the following example that we have picked out from the training corpus:

por favor , tengo reservada una habitación .
I have booked a room .

2.1 Aligning the corpus

The corpus is aligned by means of a statistical method and the word alignment information is added to each sentence pair in the following manner:

por favor , tengo reservada una habitación .
I have (4) booked (5) a room (7) . (8)

The numbers in parentheses represent the alignment between target and source words (i.e. in this sentence pair, the word *have* is aligned with the fourth target word[1]: *tengo*). Note that the word *I* is not aligned with any input word —in IBM's terminology it is aligned with the empty word—.

To ensure the accuracy of the alignment information we perform two alignments: from target to source and viceversa. Only those words that are aligned in both directions are considered.

2.2 Generation of a new monolingual corpus

Every target sentence is then rewritten labelling each aligned target word with the corresponding source word:

I [have,tengo] [booked,reservada] a [room,habitación] [.,.]

We call *e-words* (for extended words) these pairs, and *e-corpus* to the resulting corpus. We consider that an *e-word* is a target word whose exact meaning is given by the source word aligned with it.

We have made experiments labelling the source sentences rather than the target ones, but the results were poorer than those obtained when labelling the target ones.

The next step is the use of a monolingual clustering algorithm on the *e-corpus*.

2.3 Clustering the *e-words*

To cluster the *e-words*, a preliminary mapping is made of the most frequent *e-words* to the first $N-1$ classes —N being the total number of classes—. The remaining *e-words* go to the last class.

Not all the words in the *e-corpus* are *e-words*. Some of the original target words weren't aligned with a source word and they remain unlabelled in the *e-corpus*. Each one of these words is assigned to a different class, in which it is the only member. These classes —that we call *non-movable* classes— are handled in a special way: they are used to compute the perplexity of the *e-corpus* but do not accept new members nor does the unique word they contain ever move to another class.

Once the initial mapping of *e-words* and words into *movable* and *non-movable* classes has been made, the initial training set perplexity is computed. Each *e-word* is then moved in turn to every *movable* class and the movement that most reduces the current perplexity is carried out. This process is repeated until a stop criterion is met.

In a more formal way, the clustering algorithm is:

```
set up initial mapping;
compute initial training set perplexity;
do
  for each e-word e in vocabulary
    remove e from its class;
    for all movable classes c
      compute the perplexity if e moves to c;
    assign e to the class with the best perplexity;
until a stopping criterion is met;
```

2.4 Training the SST

The original training corpus is rewritten with the classes obtained. Each word is substituted by the class of its correspondent *e-word*. This parallel bilingual corpus labelled with classes, have sentence pairs of the form:

por favor , C36 C44 una C1 C0
I C36 (4) C44 (5) a C1 (7) C0 (8)

An SST is trained using this last corpus. This SST is finally expanded to the word level substituting each class by the words in it as in [1].

3 EXPERIMENTAL RESULTS

The experiments on this section have been carried out on the Spanish-English *Traveler Task Corpus* [3]. This corpus aims at covering usual sentences that are frequently needed in typical scenarios by a traveler visiting a foreign country whose language he or she does not speak. Clearly, for these situations a word to word translation is not feasible even for the simplest sentences.

The approach proposed has been tested using:

- The IBM model 2 [11] with smoothing techniques for computing the alignments [12].
- The OMEGA [13] algorithm for training the SSTs.
- Error Correcting Parsing [14] for translating the sentences.

To evaluate the *e-cluster* algorithm, we have trained SSTs with 1,000 to 10,000 sentences, using 25 to 200 classes. The test was done over 3,000 sentences unseen in training.

[1] The punctuation marks are considered words.

Table 1: Some of the classes obtained with 5,000 training sentences and 125 classes.

[double,doble] [single,individual] [single,sencilla]
[double,dobles] [single,individuales] [quiet,tranquilas]
[pardon,cómo] [when,cuándo] [where,dónde] [please,por] [who,quién]
[bath,aseo] [minibar,bar] [safe,caja] [shower,ducha] [telephone,teléfono] [tv,tele] [tv,televisión]
[okay,acuerdo] [okay,conforme] [okay,correcto] [how,cuánta] [how,cuánto] [sorry,disculpe] [hello,hola] [no,no] [yes,sí] [sorry,perdone] [sorry,perdón]

Some of the classes obtained are shown in Table 1.

As it can be seen from the first two classes, the way in which the bilingual classes are built allows the same target word (i.e. *double* and *single*) to belong to different classes. The differentiation arises from which source word generated them (i.e. one class seems to group some characteristics for a room and the other one the characteristics for more than one room).

The third class shows that the target word *tv* appears twice in the same class: one as a translation of the Spanish word *televisión* and another as a translation of another Spanish word with the same meaning: *tele*.

Obviously, not all the obtained classes are meaningful. On the other hand, some words such as names, family names, months, days of the week, and so on are typically clustered in different classes.

In Table 2 the sentence error rate (SER) and the word error rate (WER) of the translation of the test sentences is presented. The use of the automatically derived classes reduces both the SER and the WER.

With 5,000 training sentences and considering the optimum number of classes for this experiment (125 classes), the WER drops from 11.03% using OMEGA to 7.69% when the class information is added to the OMEGA algorithm. In the same way, SER is reduced from 64.4% to 41.8%.

With 10,000 training sentences and considering the optimum number of classes for this experiment (150 classes), the WER drops from 8.32% using OMEGA to 6.27% when the class information is added to the OMEGA algorithm. In the same way, SER is reduced from 54.47% to 34.03%.

The Figure 1 represents the evolution of WER with the number of training pair samples for different number of classes. The SSTs produced are poor if the number of classes is small (in comparison with the vocabulary size). But, when the number of classes is sufficiently large (more than a lower limit —i.e. 75 classes—), the WER using classes is lower than without using them.

Table 2: Sentence error rate (SER) and word error rate (WER) for different number of classes. The first line corresponds to OMEGA without classes.

	SER (%)		WER (%)	
n. classes	5,000	10,000	5,000	10,000
-	64.40	54.47	11.03	8.32
25	70.17	70.33	16.04	17.65
50	47.80	52.27	9.02	8.82
75	46.23	37.67	8.71	6.63
100	42.60	40.53	7.69	7.00
125	41.80	34.80	7.69	6.34
150	43.20	34.03	7.89	6.27
175	46.50	35.40	8.53	6.57
200	48.77	36.20	9.05	6.42

Although the number of classes has to be manually fixed for the algorithm, the WER remains more or less at the same level when the number of classes varies over a wide range, so this is not critical.

4 CONCLUSION

Automatic methods for deriving classes can be employed in bilingual clustering when the training material is scarce in order to improve the performance of subsequential transducers learning algorithms.

We plan to extend this approach in different ways: improving the alignment information; applying other monolingual clustering methods; and grouping consecutive target words when they are aligned with the same source word.

REFERENCES

[1] J. C. Amengual, J. M. Benedí, K. Beulen, F. Casacuberta, A. Castaño, A. Castellanos, V. M. Jiménez, D. Llorens, A. Marzal, H. Ney, F. Prat, E. Vidal, and J. M. Vilar. Speech translation based on automatically trainable finite-state models. In *Eurospeech 97*, volume 3, pages 1439–1442, Rhodes (Greece).

[2] José Oncina, Pedro García, and Enrique Vidal. Learning subsequential transducers for pattern recognition interpretation tasks. In *Transactions on Pattern Analysis and Machine Intelligence*, volume 15. IEEE, 1993.

[3] J. C. Amengual, J. M. Benedí, F. Casacuberta, A. Castaño, A. Castellanos, D. Llorens, A. Marzal, F. Prat, E. Vidal, and J. M. Vilar. Using categories in the EuTrans system. In *Proceedings of the Spoken Language Translation Workshop*, pages 44–53, Madrid (Spain), 1997. Association of Computational

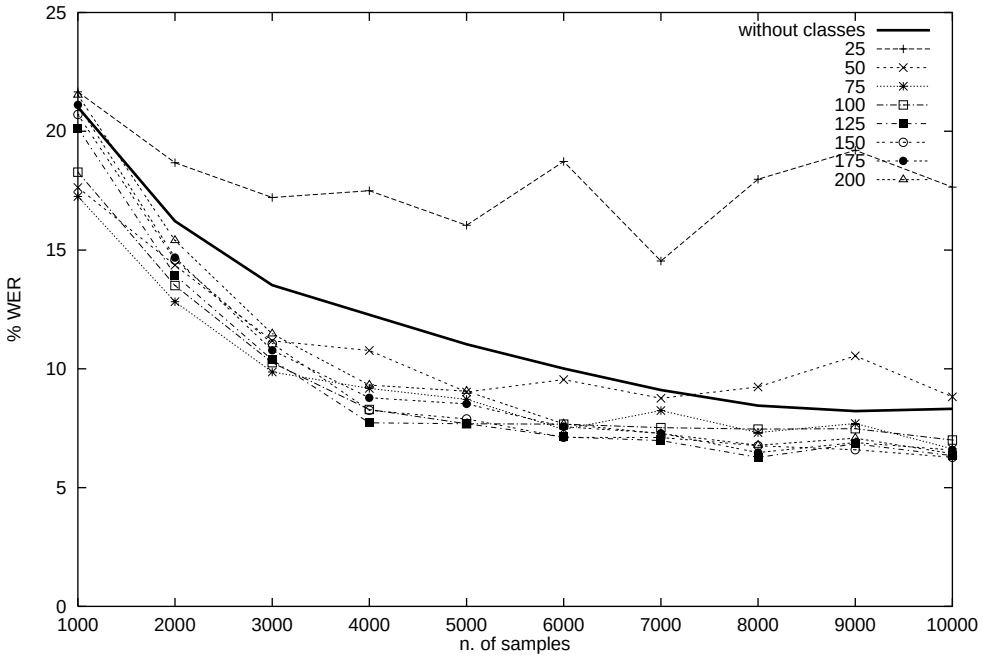

Figure 1: Translation WER for different number of classes and different training sizes.

Linguistics and European Network in Language and Speech.

[4] J. M. Vilar, A. Marzal, and E. Vidal. Learning language translation in limited domains using finite-state models: Some extensions and improvements. In *EuroSpeech 95*, pages 1231–1234, Madrid (Spain).

[5] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.

[6] Frederick Jelinek, Robert Mercer, and Salim Roukos. Classifying words for improved statistical language models. In *Proceedings of the ICASSP'90*, pages 621–624, Albuquerque, NM (USA).

[7] Reinhard Kneser and Hermann Ney. Improved clustering techniques for class-based statistical language modelling. In *Proceedings of the Eurospeech'93*, pages 973–976, Berlin (Germany).

[8] Sven Martin, Jörg Liermann, and Hermann Ney. Algorithms for bigram and trigram word clustering. In *Proceedings of the Eurospeech'95*, Berlin (Germany).

[9] Pascale Fung and Dekai Wu. Coerced markov models for cross-lingual lexical-tag relations. In *Proceedings of the TMI'95*, pages 240–255, Leuven (Belgium).

[10] Franz Josef Och and Hans Weber. Improving statistical natural language translations with categories and rules. In *Proceedings of COLING'98*, Montreal (Canada).

[11] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[12] Ismael García-Varea, Francisco Casacuberta, and Hermann Ney. An interative, dp-based search algorithm for statistical machine translation. In *Proceedings of the ICSLP'98*, volume 4, pages 1135–1138, Sydney (Australia).

[13] Juan Miguel Vilar. *Aprendizaje de Traductores Subsecuenciales para su Empleo en Tareas de Dominio Restringido*. PhD thesis, Dpto. de Sistemas Informáticos y Computación. Univ. Politécnica de Valencia, Valencia (Spain), 1998.

[14] Juan C. Amengual, José M. Benedí, Francisco Casacuberta, Asunción Castaño, Antonio Castellanos, David Llorens, Andrés Marzal, Federico Prat, Enrique Vidal, and Juan M. Vilar. Error correcting parsing for text-to-text machine translation using finite state models. In *Proceedings of the TMI'97*, pages 135–142, Santa Fe, NM (USA).