An Efficient Method for Determining Bilingual Word Classes

Franz Josef Och
Lehrstuhl für Informatik VI
RWTH Aachen - University of Technology
Ahornstraße 55
52056 Aachen
GERMANY
och@informatik.rwth-aachen.de

Abstract

In statistical natural language processing we always face the problem of sparse data. One way to reduce this problem is to group words into equivalence classes which is a standard method in statistical language modeling. In this paper we describe a method to determine bilingual word classes suitable for statistical machine translation. We develop an optimization criterion based on a maximum-likelihood approach and describe a clustering algorithm. We will show that the usage of the bilingual word classes we get can improve statistical machine translation.

1 Introduction

Word classes are often used in language modelling to solve the problem of sparse data. Various clustering techniques have been proposed (Brown et al., 1992; Jardino and Adda, 1993; Martin et al., 1998) which perform automatic word clustering optimizing a maximum-likelihood criterion with iterative clustering algorithms.

In the field of statistical machine translation we also face the problem of sparse data. Our aim is to use word classes in statistical machine translation to allow for more robust statistical translation models. A naive approach for doing this would be the use of mono-lingually optimized word classes in source and target language. Unfortunately we can not expect these independently optimized classes to be correspondent. Therefore mono-lingually optimized word classes do not seem to be useful for machine translation (see also (Fung and Wu, 1995)). We define *bilingual word clustering* as the process of forming corresponding word classes suitable for machine translation purposes for a pair of languages using a parallel training corpus.

The described method to determine bilingual word classes is an extension and improvement of the method mentioned in (Och and Weber, 1998). Our approach is simpler and computationally more efficient than (Wang et al., 1996).

2 Monolingual Word Clustering

The task of a statistical language model is to estimate the probability $Pr(w_1^N)$ of a sequence of words $w_1^N = w_1 \ldots w_N$. A simple approximation of $Pr(w_1^N)$ is to model it as a product of bigram probabilities: $Pr(w_1^N) = \prod_{i=1}^{N} p(w_i|w_{i-1})$. If we want to estimate the bigram probabilities $p(w|w')$ using a realistic natural language corpus we are faced with the problem that most of the bigrams are rarely seen. One possibility to solve this problem is to partition the set of all words into equivalence classes. The function C maps words w to their classes $C(w)$. Rewriting the corpus probability using classes we arrive at the following probability model $p(w_1^N|C)$:

$$p(w_1^N|C) := \prod_{i=1}^{N} p(C(w_i)|C(w_{i-1})) \cdot p(w_i|C(w_i)) \quad (1)$$

In this model we have two types of probabilities: the transition probability $p(C|C')$ for class C given its predecessor class C' and the membership probability $p(w|C)$ for word w given class C.

To determine the optimal classes $\hat{C}$ for a given number of classes M we perform a maximum-likelihood approach:

$$\hat{C} = \arg\max_{C} p(w_1^N|C) \quad (2)$$

We estimate the probabilities of Eq. (1) by relative frequencies: $p(C|C') := n(C|C')/n(C')$, $p(w|C) = n(w)/n(C)$. The function $n(\cdot)$ provides the frequency of a uni- or bigram in the training corpus. If we insert this into Eq. (2) and apply the negative logarithm and change the summation order we arrive at the following optimization

criterion LP_1 (Kneser and Ney, 1991):

$$\begin{aligned} LP_1(C,n) &= -\sum_{C,C'} h(n(C|C')) \\ &\quad +2\sum_{C} h(n(C)) \qquad (3) \\ \hat{C} &= \arg\min_{C} LP_1(C,n). \qquad (4) \end{aligned}$$

The function $h(n)$ is a shortcut for $n \cdot \log(n)$.

It is necessary to fix the number of classes in C in advance as the optimum is reached if every word is a class of its own. Because of this it is necessary to perform an additional optimization process which determines the number of classes. The use of leaving-one-out in a modified optimization criterion as in (Kneser and Ney, 1993) could in principle solve this problem.

An efficient optimization algorithm for LP_1 is described in section 4.

3 Bilingual Word Clustering

In bilingual word clustering we are interested in classes $\mathcal{F}$ and $\mathcal{E}$ which form partitions of the vocabulary of two languages. To perform bilingual word clustering we use a maximum-likelihood approach as in the monolingual case. We maximize the joint probability of a bilingual training corpus (e_1^I, f_1^J):

$$\begin{aligned} (\hat{\mathcal{E}},\hat{\mathcal{F}}) &= \arg\max_{\mathcal{E},\mathcal{F}} p(e_1^I, f_1^J|\mathcal{E},\mathcal{F}) \qquad (5) \\ &= \arg\max_{\mathcal{E},\mathcal{F}} p(e_1^I|\mathcal{E}) \cdot p(f_1^J|e_1^I;\mathcal{E},\mathcal{F}) \qquad (6) \end{aligned}$$

To perform the maximization of Eq. (6) we have to model the monolingual a priori probability $p(e_1^I|\mathcal{E})$ and the translation probability $p(f_1^J|e_1^I;\mathcal{E},\mathcal{F})$. For the first we use the class-based bigram probability from Eq. (1).

To model $p(f_1^J|e_1^I;\mathcal{E},\mathcal{F})$ we assume the existence of an alignment a_1^J. We assume that every word f_j is produced by the word e_{a_j} at position a_j in the training corpus with the probability $p(f_j|e_{a_j})$:

$$p(f_1^J|e_1^I) = \prod_{j=1}^{J} p(f_j|e_{a_j}) \qquad (7)$$

The word alignment a_1^J is trained automatically using statistical translation models as described in (Brown et al., 1993; Vogel et al., 1996). The idea is to introduce the unknown alignment a_1^J as hidden variable into a statistical model of the translation probability $p(f_1^J|e_1^I)$. By applying the EM-algorithm we obtain the model parameters. The alignment a_1^J that we use is the Viterbi-Alignment of an HMM alignment model similar to (Vogel et al., 1996).

By rewriting the translation probability using word classes, we obtain (corresponding to Eq. (1)):

$$p(f_1^J|e_1^I;\mathcal{E},\mathcal{F}) = \prod_{j=1}^{J} p(\mathcal{F}(f_j)|\mathcal{E}(e_{a_j})) \cdot p(f_j|\mathcal{F}(f_j)) \qquad (8)$$

The variables F and E denote special classes in $\mathcal{F}$ and $\mathcal{E}$. We use relative frequencies to estimate $p(F|E)$ and $p(f|F)$:

$$\begin{aligned} p(F|E) &= n_t(F|E) / \left(\sum_F n_t(F|E)\right) \\ p(f|F) &= n_t(f) / \left(\sum_f n_t(f|F)\right) \\ &= n_t(f) / \left(\sum_E n_t(F|E)\right) \end{aligned}$$

The function $n_t(F|E)$ counts how often the words in class F are aligned to words in class E. If we insert these relative frequencies into Eq. (8) and apply the same transformations as in the monolingual case we obtain a similar optimization criterion for the translation probability part of Eq. (6). Thus the full optimization criterion for bilingual word classes is:

$$\begin{aligned} &-\sum_{E,E'} h(n(E|E')) - \sum_{E,F} h(n_t(F|E)) \\ &\qquad +2\sum_{E} h(n(E)) \\ &+\sum_{F} h(\sum_{E} n_t(F|E)) + \sum_{E} h(\sum_{F} n_t(F|E)) \end{aligned}$$

The two count functions $n(E|E')$ and $n_t(F|E)$ can be combined into one count function $n_g(X|Y) := n(X|Y)+n_t(X|Y)$ as for all words f and all words e and e' holds $n(f|e) = 0$ and $n_t(e|e') = 0$. Using the function n_g we arrive at the following optimization criterion:

$$\begin{aligned} LP_2((\mathcal{E},\mathcal{F}),n_g) &= -\sum_{X,X'} h(n_g(X|X')) + \\ &\quad \sum_{X} h(n_{g,1}(X)) + \sum_{X} h(n_{g,2}(X)) \qquad (9) \end{aligned}$$

$$(\hat{\mathcal{E}},\hat{\mathcal{F}}) = \arg\min_{\mathcal{E},\mathcal{F}} LP_2((\mathcal{E},\mathcal{F}),n_g) \qquad (10)$$

Here we defined $n_{g,1}(X) = \sum_{X'} n_g(X|X')$ and $n_{g,2}(X) = \sum_{X'} n_g(X'|X)$. The variable X runs over the classes in $\mathcal{E}$ and $\mathcal{F}$. In the optimization process it cannot be allowed that words of

INPUT: Parallel corpus (e_1^I, f_1^J) and number of classes in $\mathcal{E}$ and $\mathcal{F}$.
Determine the word alignment a_1^J.
Get some initial classes $\mathcal{E}$ and $\mathcal{F}$.
UNTIL convergence criterion is met:
  FOR EACH word e:
    FOR EACH class E:
      Determine the change of $LP((\mathcal{E},\mathcal{F}),n_g)$ if e is moved to E.
    Move e to the class with the largest improvement.
  FOR EACH word f:
    FOR EACH class F:
      Determine the change of $LP((\mathcal{E},\mathcal{F}),n_g)$ if f is moved to F.
    Move f to the class with the largest improvement.
OUTPUT: Classes $\mathcal{E}$ and $\mathcal{F}$.

Figure 1: Word Clustering Algorithm.

different languages occur in one class. It can be seen that Eq. (3) is a special case of Eq. (9) with $n_{g,1} = n_{g,2}$.

Another possibility to perform bilingual word clustering is to apply a two-step approach. In a first step we determine classes $\hat{\mathcal{E}}$ optimizing only the monolingual part of Eq. (6) and secondly we determine classes $\hat{\mathcal{F}}$ optimizing the bilingual part (without changing $\hat{\mathcal{E}}$):

$$\hat{\mathcal{E}} = \arg\min_{\mathcal{E}} \mathrm{LP}_2(\mathcal{E}, n) \quad (11)$$

$$\hat{\mathcal{F}} = \arg\min_{\mathcal{F}} \mathrm{LP}_2((\hat{\mathcal{E}}, \mathcal{F}), n_t). \quad (12)$$

By using these two optimization processes we enforce that the classes $\hat{\mathcal{E}}$ are mono-lingually 'good' classes and that the classes $\hat{\mathcal{F}}$ correspond to $\hat{\mathcal{E}}$. Interestingly enough this results in a higher translation quality (see section 5).

4 Implementation

An efficient optimization algorithm for LP_1 is the exchange algorithm (Martin et al., 1998). For the optimization of LP_2 we can use the same algorithm with small modifications. Our starting point is a random partition of the training corpus vocabulary. This initial partition is improved iteratively by moving a single word from one class to another. The algorithm to determine bilingual classes is depicted in Figure 1.

If only one word w is moved between the partitions C and C' the change $LP(C, n_g) - LP(C', n_g)$ can be computed efficiently looking only at classes C for which $n_g(w, C) > 0$ or $n_g(C, w) > 0$. We define M_0 to be the average number of seen predecessor and successor word classes. With the notation I for the number of iterations needed for convergence, B for the number of word bigrams, M for the number of classes and V for the vocabulary

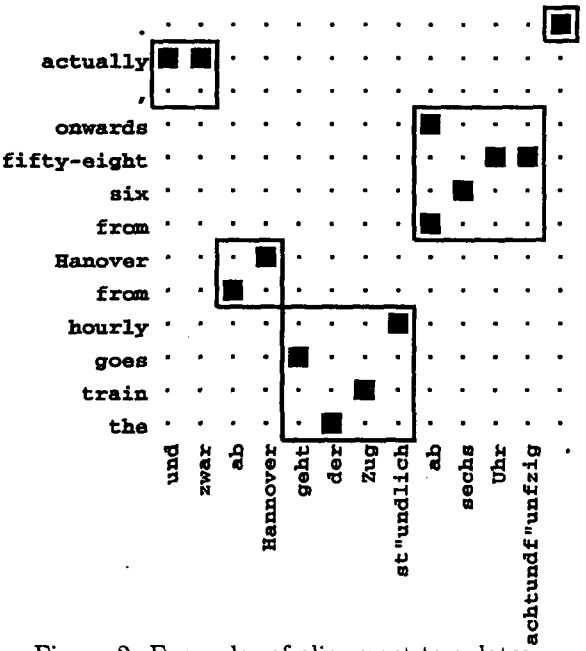

Figure 2: Examples of alignment templates.

size the computational complexity of this algorithm is roughly $I \cdot (B \cdot \log_2 (B/V) + V \cdot M \cdot M_0)$. A detailed analysis of the complexity can be found in (Martin et al., 1998).

The algorithm described above provides only a local optimum. The quality of the resulting local optima can be improved if we accept a short-term degradation of the optimization criterion during the optimization process. We do this in our implementation by applying the optimization method *threshold accepting* (Dueck and Scheuer, 1990) which is an efficient simplification of *simulated annealing*.

Table 1: The EuTrans-I corpus.

		Spanish	English
Train:	Sentences	10 000	
	Words	97 131	99 292
	Vocabulary Size	686	513
Test:	Sentences	2 996	
	Words	35 023	35 590
	Bigr. Perplexity	-	5.2

Table 2: The EuTrans-II corpus.

		German	English
Train:	Sentences	16 226	
	Words	266 080	299 945
	Vocabulary Size	39 511	25 751
Test:	Sentences	187	
	Words	2 556	2 853
	Bigr. Perplexity	-	157

5 Results

The statistical machine-translation method described in (Och and Weber, 1998) makes use of bilingual word classes. The key element of this approach are the *alignment templates* (originally referred to as translation rules) which are pairs of phrases together with an alignment between the words of the phrases. Examples of alignment templates are shown in Figure 2. The advantage of the alignment template approach against word-based statistical translation models is that word context and local re-orderings are explicitly taken into account.

The alignment templates are automatically trained using a parallel training corpus. The translation of a sentence is done by a search process which determines the set of alignment templates which optimally cover the source sentence. The bilingual word classes are used to generalize the applicability of the alignment templates in search. If there exists a class which contains all cities in source and target language it is possible that an alignment template containing a special city can be generalized to all cities. More details are given in (Och and Weber, 1998; Och and Ney, 1999).

We demonstrate results of our bilingual clustering method for two different bilingual corpora (see Tables 1 and 2). The EuTrans-I corpus is a subtask of the "Traveller Task" (Vidal, 1997) which is an artificially generated Spanish-English corpus. The domain of the corpus is a human-to-human communication situation at a reception desk of a hotel. The EuTrans-II corpus is a natural German-English corpus consisting of different text types belonging to the domain of tourism: bilingual Web pages of hotels, bilingual touristic brochures and business correspondence. The target language of our experiments is English.

Table 3: Example of bilingual word classes (corpus EuTrans-I, method BIL-2).

```
E1: how it pardon what when where which
who why
E2: my our
E3: today tomorrow
E4: ask call make
E5: carrying changing giving looking
moving putting sending showing waking
E6: full half quarter
S1: c'omo cu'al cu'ando cu'anta d'onde
dice dicho hace qu'e qui'en tiene
S2: ll'eveme mi mis nuestra nuestras
nuestro nuestros s'ubanme
S3: hoy ma nana mismo
S4: hacerme ll'ameme ll'amenos llama
llamar llamarme llamarnos llame p'idame
p'idanos pedir pedirme pedirnos
pida pide
S5: cambiarme cambiarnos despertarme
despertarnos llevar llevarme llevarnos
subirme subirnos usted ustedes
S6: completa cuarto media menos
```

We compare the three described methods to generate bilingual word classes. The classes MONO are determined by monolingually optimizing source and target language classes with Eq. (4). The classes BIL are determined by bilingually optimizing classes with Eq. (10). The classes BIL-2 are determined by first optimizing mono-lingually classes for the target language (English) and afterwards optimizing classes for the source language (Eq. (11) and Eq. (12)).

For EuTrans-I we used 60 classes and for EuTrans-II we used 500 classes. We chose the number of classes in such a way that the final performance of the translation system was optimal. The CPU time for optimization of bilingual word classes on an Alpha workstation was under 20 seconds for EuTrans-I and less than two hours for EuTrans-II.

Table 3 provides examples of bilingual word classes for the EuTrans-I corpus. It can be seen that the resulting classes often contain words that are similar in their syntactic and semantic functions. The grouping of words with a different

Table 4: Perplexity (PP) of different classes.

Corpus	MONO	BIL	BIL-2
EuTrans-I	2.13	**1.78**	1.80
EuTrans-II	13.2	**9.3**	9.8

Table 5: Average ϵ-mirror of different classes.

Corpus	MONO	BIL	BIL-2
EuTrans-I	3.5	**2.6**	**2.6**
EuTrans-II	2.2	**1.8**	2.0

meaning like `today` and `tomorrow` does not imply that these words should be translated by the same Spanish word, but it does imply that the translations of these words are likely to be in the same Spanish word class.

To measure the quality of our bilingual word classes we applied two different evaluation measures:

1. Average ϵ-mirror size (Wang et al., 1996): The ϵ-mirror of a class E is the set of classes which have a translation probability greater than ϵ. We use $\epsilon = 0.05$.

2. The perplexity of the class transition probability on a bilingual test corpus:

$$\exp\left(J^{-1} \cdot \sum_{j=1}^{J} \max_i \log\left(p\left(C\left(f_j\right)|C\left(e_i\right)\right)\right)\right)$$

Both measures determine the extent to which the translation probability is spread out. A small value means that the translation probability is very focused and that the knowledge of the source language class provides much information about the target language class.

Table 4 shows the perplexity of the obtained translation lexicon without word classes, with monolingual and with bilingual word classes. As expected the bilingually optimized classes (BIL, BIL-2) achieve a significantly lower perplexity and a lower average ϵ-mirror than the mono-lingually optimized classes (MONO).

The tables 6 and 7 show the translation quality of the statistical machine translation system described in (Och and Weber, 1998) using no classes (WORD) at all, mono-lingually, and bilingually optimized word classes. The translation system was trained using the bilingual training corpus without any further knowledge sources. Our evaluation criterion is the word error rate (WER) — the minimum number of insertions/deletions/substitutions relative to a reference translation.

Table 6: Word error rate (WER) and average alignment template length (AATL) on EuTrans-I.

Method	WER [%]	AATL
WORD	6.31	2.85
MONO	5.64	5.03
BIL	5.38	4.40
BIL-2	4.76	5.19

Table 7: Word error rate (WER) and average alignment template length (AATL) on EuTrans-II.

Method	WER [%]	AATL
WORD	64.3	1.36
MONO	63.5	1.74
BIL	63.2	1.53
BIL-2	62.5	1.54

As expected the translation quality improves using classes. For the small EuTrans-I task the word error rates reduce significantly. The word error rates for the EuTrans-II task are much larger because the task has a very large vocabulary and is more complex. The bilingual classes show better results than the monolingual classes MONO. One explanation for the improvement in translation quality is that the bilingually optimized classes result in an increased average size of used alignment templates. For example the average length of alignment templates with the EuTrans-I corpus using WORD is 2.85 and using BIL-2 it is 5.19. The longer the average alignment template length, the more context is used in the translation and therefore the translation quality is higher.

An explanation for the superiority of BIL-2 over BIL is that by first optimizing the English classes mono-lingually, it is much more probable that longer sequences of classes occur more often thereby increasing the average alignment template size.

6 Summary and future works

By applying a maximum-likelihood approach to the joint probability of a parallel corpus we obtained an optimization criterion for bilingual word classes which is very similar to the one used in monolingual maximum-likelihood word clustering. For optimization we used the exchange algorithm. The obtained word classes give a low translation lexicon perplexity and improve the quality of sta-

tistical machine translation.

We expect improvements in translation quality by allowing that words occur in more than one class and by performing a hierarchical clustering.

Acknowledgements This work has been partially supported by the European Community under the ESPRIT project number 30268 (EuTrans).

References

P. F. Brown, V. J. Della Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

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

G. Dueck and T. Scheuer. 1990. Threshold accepting: A general purpose optimization algorithm appearing superior to simulated annealing. *Journal of Computational Physics*, 90(1):161–175.

Pascale Fung and Dekai Wu. 1995. Coerced markov models for cross-lingual lexical-tag relations. In *The Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 240–255, Leuven, Belgium, July.

M. Jardino and G. Adda. 1993. Automatic Word Classification Using Simulated Annealing. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, volume 2, pages 41–44, Minneapolis.

R. Kneser and H. Ney. 1991. Forming Word Classes by Statistical Clustering for Statistical Language Modelling. In *1. Quantitative Linguistics Conference.*

R. Kneser and H. Ney. 1993. Improved Clustering Techniques for Class-Based Statistical Language Modelling. In *European Conference on Speech Communication and Technology*, pages 973–976.

Sven Martin, Jörg Liermann, and Hermann Ney. 1998. Algorithms for bigram and trigram word clustering. *Speech Communication*, 24(1):19–37.

Franz Josef Och and Hermann Ney. 1999. The alignment template approach to statistical machine translation. To appear.

Franz Josef Och and Hans Weber. 1998. Improving statistical natural language translation with categories and rules. In *Proceedings of the 35th Annual Conference of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pages 985–989, Montreal, Canada, August.

Enrique Vidal. 1997. Finite-state speech-to-speech translation. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, volume 1, pages 111–114.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *COLING '96: The 16th Int. Conf. on Computational Linguistics*, pages 836–841, Copenhagen, August.

Ye-Yi Wang, John Lafferty, and Alex Waibel. 1996. Word clustering with parallel spoken language corpora. In *Proceedings of the 4th International Conference on Spoken Language Procesing (ICSLP'96)*, pages 2364–2367.