To appear in *TMI-95, Sixth International Conf. on Theoretical and Methodological Issues in Machine Translation*, Leuven, Belgium: Jul. 95.

Coerced Markov Models for Cross-Lingual Lexical-Tag Relations

Pascale Fung[*] and *Dekai Wu*[†]

Technical Report HKUST-CS95-24
May 1995

[*]*Columbia University*
Computer Science Department
New York, NY 10027
USA
pascale@cs.columbia.edu

[†]*HKUST*
Department of Computer Science
University of Science & Technology
Clear Water Bay, Hong Kong
dekai@cs.ust.hk

Abstract

We introduce the *Coerced Markov Model* (CMM) to model the relationship between the lexical sequence of a source language and the tag sequence of a target language, with the objective of constraining search in statistical transfer-based machine translation systems. CMMs differ from standard hidden Markov models in that state sequence assignments can take on values coerced from external sources. Given a Chinese sentence, a CMM can be used to predict the corresponding English tag sequence, thus constraining the English lexical sequence produced by a translation model. The CMM can also be used to score competing translation hypotheses in N-best models. Three fundamental problems for CMM designed are discussed. Their solutions lead to the training and testing stages of CMM.

The Hong Kong University of Science & Technology
Technical Report Series
Department of Computer Science

1 Introduction

The analysis, transfer, and synthesis paradigm for machine translation is readily amenable to statistical methods (Brown *et al.* 1993). The transfer stage exploits mapping knowledge about various linguistic relationships between the source and target languages; statistical information is readily incorporated at this stage. Typical kinds of mapping relations include sentence-to-sentence, word-to-word, or part-of-speech (POS) tags to tags. Statistical algorithms use probabilities to model the word-to-word lexical relations between a pair of sentences in the source and target languages (Brown *et al.* 1993; Dagan *et al.* 1993; Dagan & Church 1994; Fung & McKeown 1994; Wu & Xia 1994; Fung 1995). These probabilities help in the transfer stage to constrain or prune the search for an optimal sequence of translated words. Linguistic information such as part-of-speech has also been found to be useful for constraining this search. (Chang & Chen 1994; Papageorgiou *et al.* 1994).

In this paper we investigate an underutilized source of constraints, namely, the mapping between words in the source language and parts-of-speech in the target language. Such information would also constrain search in the translation model. We believe this mapping relation can be automatically learned from bilingual corpora. However, to our knowledge no such attempt has been made, perhaps due to the modeling difficulties in the problem. We introduce a *Coerced Markov Model* (CMM) representation that accommodates mapping relations between source-words and target-tags in a statistical framework.

Although there has been work on mapping between source language tags and target language tags (Chang & Chen 1994; Papageorgiou *et al.* 1994), this mapping might not be meaningful or sufficiently helpful for translation. In the most common scenario, texts of both languages are tagged by their respective POS taggers. A tag-to-tag mapping between the two languages is obtained from the tagged text. However, most part-of-speech classes are determined by humans according to the linguistic knowledge in that particular language. It is not evident that there should be a direct correspondence between POS classes in two different languages, especially in language pairs which do not share any common root such as English and Chinese. The relationship we derive from English and Chinese part-of-speech mapping is therefore not necessarily a good constraint for translation search.

On the other hand, source language *words* are capable of giving much more discriminative information about target tags than source tags are. Moreover, a reliable tagger for source languages such as Chinese may not be available in the first place. We propose to capture the correlation between source words and target tags with the Coerced Markov Model. As we discuss below, CMMs are a modified variant of discrete, first-order, hidden Markov models such that the state sequence is determined by coercion from some second state sequence from outside the model.

One application of the CMM is that it can predict the English tag sequence corresponding to a given Chinese sentence. This tag sequence can be used as a pruning constraint on the search of the transfer model for the production of an English lexical sequence.

Since a transfer model produces an English translation sentence by choosing the individual English words corresponding to the individual words in the Chinese sentence, it can produce a number of translation hypotheses. An alternative application of CMM is to provide a measure of the goodness of the hypotheses.

In the following sections, we first define the CMM formalism, and then describe its training and testing stages.

2 Coerced Markov Models

Markov chains are widely used for characterizing parametric random processes. The basic theory of hidden Markov models (HMMs) was proposed by Baum & Petrie (1966) and Baum & Egon (1967). It was later adapted by Baker (1975) and Jelinek *et al.* (1975) for processing speech signals. The fundamental assumption of using a Markov model for a linguistic mapping (in our case between words in one language and tags in the other language) is that the mapping is a stochastic process and its parameters are estimable.

A Markov chain describes the changes of states of a system. For example, at time t, the system is in state b, it changes to state a at time $t+1$, then there is a state transition from b to a with certain probability. First-order Markov chains assume the probability of a state depends only on its preceding state, i.e.,

$$P[q_{t+1} = a | q_t = b, q_{t-1} = c, q_{t-2} = d, \ldots] = P[q_{t+1} = a | q_t = b]$$

At each given state there is an associated output. This output can be continuous, such as the spectral signal of speech in a speech recognition system, or discrete, such as the identity of an individual word within a sentence. If we regard the mapping between Chinese word sequences

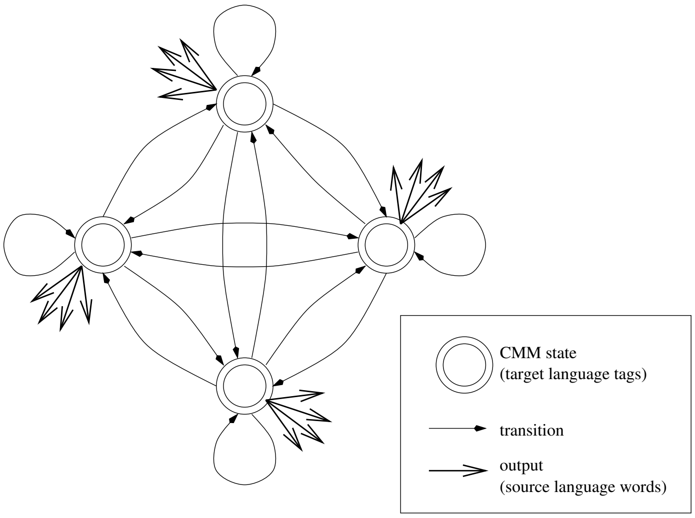

Figure 1: Example of a four-state CMM

and the tag sequence of its corresponding English translation as a stochastic process, the Coerced Markov Model for the process is discrete.

A Markov model is *hidden* if its states are not deterministically observable. Given an observation sequence, the underlying states are non-deterministic. Hidden Markov models are typically used in speech processing, where the underlying states do not actually correspond to explicit entities (such as phonemes or words). CMM states are also non-deterministic and therefore hidden because the same output sequence can be generated from different state sequences given a particular model.

For our application in Chinese-English translation, the CMM is *coercing* English tags into Chinese language modeling. In a CMM, English tags cannot just follow the rules in English language models; they must also consider the fact that they are now "partners" of Chinese words that also follow their own rules. The CMM is modeling the "adaptation" of English tags to Chinese word order. This is a step beyond monolingual language modeling such as word N-gram or class N-gram computation. The CMM's purpose, and its strength, is to model *cross-lingual* class N-grams. An example of a four-state CMM is illustrated in Figure 1.

Formally, we define the following elements:

1. the set of states $\mathcal{N}$: Hidden Chinese states, with coerced English tag class values

2. the set of observable symbols $\mathcal{M}$: Chinese words

We choose to associate states with POS classes, and outputs with lexical items, in order to obtain the most modeling and discriminative power with the CMM. If we had chosen individual Chinese lexical items to be the states, there would be many cases where word v never follows word w, where both v and w are entries in the lexicon. There would be many restrictions on state transitions, making the model neither flexible nor powerful. Instead, it is more reasonable to use POS classes as the states of a CMM because these classes have some linguistic significance. In addition, since most POS classes can follow any other POS class given a large enough corpus, there can be a transition between any two given states. This means that the CMM is an ergodic model with null transitions between only very few states (e.g., from DT to VB), which makes the CMM potentially more flexible. Having fixed the states N to be POS classes in one language, it follows that the observable symbols M should be the lexical items in the other language.

Once we have defined the nature of $\mathcal{N}$ and $\mathcal{M}$, we have to choose which language $\mathcal{N}$ and $\mathcal{M}$ should come from. We choose $\mathcal{N}$ to be the English POS classes, because English taggers are readily available and there has been some consensus as to the basic POS classes. On the other hand, due to the short history of Chinese NLP, Chinese tagging is still under research and there is still a lack of a general paradigm for Chinese POS classes. Having fixed the POS classes $\mathcal{N}$ to be English, it follows that $\mathcal{M}$ is the set of Chinese lexical items.

Referring again to Figure 1, each state in the CMM corresponds to an English POS class. For our experiments, we use Brill's (1993) tagger which uses $N = 106$ English tag classes. Given any two states, there is a weighted transition going in either direction from one to the other. Each state can also transit into itself. An array of possible Chinese output words with different weights is associated with each state.

The next three sections of this paper discuss methods and experiments for three fundamental problems of CMMs:

1. **Estimation:** Given a CMM (i.e., its topology), estimate its parameters so as to best describe an observed training sequence.

2. **Path recovery:** Given a CMM, its parameters, and a test observation sequence, determine the optimal hidden state sequence. Can be used to *suggest* constraints on translation hypotheses.

3. **Scoring:** Given a CMM and its parameters, determine the probabilistic score of a sequence of states. Can be used to *score* translation hypotheses.

It may be helpful, in order to understand these three problems, to note a certain parallel between them and the three fundamental problems of HMM (Rabiner & Juang 1993), although the cross-lingual coercion leads to substantial differences. We will see that problem (1) is the parameter estimation process for a CMM, and that problems (2) and (3) can be used for two different translation applications that each yield an experimental evaluation.

3 Estimation

In this section we describe how we estimate

1. the transition probabilities $A = a_{ij}$
2. the output probabilities $B = b_j(k)$

given a word-aligned parallel corpus. Remember that the objectives of training the CMM are, first, to best model the stochastic process of Chinese word sequences co-occuring with their English tag counterparts, and second, to supply the most useful constraints possible to help prune the search process in a statistical transfer model.

Transition probabilities We first describe how to compute the transition probabilities a_{ij} where i and j are any two states in CMM.

To use an example, the Chinese sentence

這些安排可加強我們日後維持金融穩定的能力。

has the English alignment

These arrangements enhance our ability to maintain monetary stability.

with their POS tags as shown in Table 1. The tag sequence $(<>, DT, NNS, <>, VB, PRP\$, <>, <>, VB, JJ, NN, <>, NN, .)$ contains $<>$ as null tags since there is no English word alignment to the Chinese word at that position. According to this Chinese sentence and its aligned English words, there is a transition from the initial state to DT, DT to NNS, NNS to $<>$, and so forth. Here, the English tag sequence is *coerced* into modeling the Chinese word sequence. If our training data had this single sentence only, then we would get a total of 13 transitions and each transition probability would be a_{ij} = 1/13.

Table 1: Training data format

Chinese word	Alignment position	English word	English POS	Transitions
</s>				
這些	1	These	DT	</s>,DT
安排	2	arrangements	NNS	DT, NNS
可				NNS, <>
加強	4	enhance	VB	<>, VB
我們	5	our	PRP$	VB, PRP$
日				PRP$, <>
後				<>, <>
維持	8	maintain	VB	<>, VB
金融	9	monetary	JJ	VB, JJ
穩定	10	stability	NN	JJ, NN
的				NN, <>
能力	6	ability	NN	<>, NN
。	16	.	.	NN, .

The null tag state comes from the particular phenomenon in Chinese/English translations where many Chinese words are not aligned to any English words due to a relatively large linguistic difference between the two languages. We believe these null alignments give highly unreliable information. In our experiments we penalize the transitions into and out of the null state by assigning a very low probability to them. The final transition probabilities are converted into the logarithmic form for computational purpose.

In general, since the probabilities are less than one, their logarithms are negative numbers; therefore we take the negative log probabilities for computation.

Thus the formula for transition probabilities is:

$$a_{ij} = -\ln\left(\frac{\sum \text{number of transitions from i to j}}{\text{total number of transitions}} + \xi\right)$$

where ξ is a small *flooring* parameter to avoid undefined log probabilities.

Output probabilities Next, we have to compute the output probabilities $b_j(k)$ of the CMM.

The CMM is a discrete Markov model in which the observable output is in the set $\mathcal{M}$ of Chinese words. For example, in the sentence in Table 1, the Chinese word 穩定 is aligned to *stability* which is tagged as NN; this means in the state NN, the output k=穩定 occurs once here. In other context, the same Chinese word could be aligned to *stable* which would be tagged as an adjective JJ. The probability of $b_{NN}(k)$ depends on how often 穩定 is observed when its corresponding English word is tagged as NN.

Using a similar form as for transition probabilities, the output probabilities are estimated as:

$$b_j(k) = -\ln\left(\frac{\text{number of Chinese word k observed when in state j}}{\text{total number of Chinese word observed in state j}} + \xi\right)$$

An Experimental Setup We used the HKUST Chinese-English Parallel corpus (Wu 1994) to train our CMM. To prepare a training corpus in the required format, we carried out the following steps:

1. **Sentence align the corpus** into Chinese-English sentence pairs by a length-based method (Wu 1994).

2. **Tokenize the Chinese sentences** by segmenting substrings of Chinese characters into individual words (this was necessary since Chinese text does not have word delimiters). We used a Viterbi tagger with a statistically augmented dictionary (Fung & Wu 1994; Wu & Fung 1994).

3. **Tag the English sentences** by using a corpus-based POS tagger (Brill 1993),

4. **Compute the English word alignment** to the individual Chinese words using an estimation-maximization model (Wu & Xia 1994),

5. **Filter the training corpus** by applying criteria described in (Wu 1995).

We obtained a total of 1885 Chinese sentences with aligned English words and English POS tags as our training corpus. An example of the training corpus format is shown in the first four columns of Table 1.

Using this training data, we estimated the CMM parameters as follows:

1. **Compute initial probabilities** π_i: $1 \leq i \leq N$
2. **Compute transition probabilities** a_{ij}: there were 1,969 null transitions probabilities out of a total of 11,236 transitions.
3. **Compute output probabilities** $b_j(k)$: $1 \leq j \leq N, 1 \leq k \leq M$

Table 2: Test sentence

Chinese word	Alignment position	English word	English POS	Transitions
</s>				
我們	1	We	PRP	</s>,PRP
將	2	will	MD	PRP,MD
為	3	provide	VB	MD,VB
老人	22	aged	JJ	VB,JJ
增	7	additional	JJ	JJ,JJ
設			<>	JJ,<>
5	8	5	CD	<>,CD
0	9	0	CD	CD,CD
0	10	0	CD	CD,CD
0	11	0	CD	CD,CD
個			<>	CD,<>
護理	14	care	NN	<>,NN
安老院	19	homes	NNS	NN,NNS
和	18	and	CC	NNS,CC
安	16	attention	NN	CC,NN
老院	17	homes	NNS	NN,NNS
名額	12	places	NNS	NNS,NNS
。	23	.	.	NNS,.

4 Optimal path recovery

We used two different evaluation methods corresponding to the solutions of problem (2) and problem (3) in CMM design. The first evaluation was to produce a English tag sequence given a Chinese sentence.

We used a Chinese sentence from the corpus which was not included in the training set as the test sample. The Chinese and its corresponding English aligned words and their tags are shown in Table 2. Against this correct tag sequence, we compared the output computed as follows.

To find the solution for predicting the best state sequence (i.e., English tag sequence $\mathbf{q} = (q_1, q_2, \ldots, q_C)$) from the observation sequence (i.e., the Chinese sentence $\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_C)$) of length C, we use a Viterbi algorithm (Viterbi 1967; Forney 1973) where the transition probabilities a_{ij} and output probabilities $b_j(\mathbf{o}_c)$ are in the negative logarithmic form:

- **Initialization**

$$\begin{aligned} \delta_1(i) &= \pi_i + b_i(\mathbf{o}_1) \quad \text{where } 1 \leq i \leq N \text{ and } \pi_i = \Pr(\text{initial state} = i) \\ \psi_1(i) &= 0 \qquad\qquad\quad \text{where } 1 \leq i \leq N \end{aligned}$$

- **Recursion**

$$\begin{aligned} \delta_c(j) &= \min_{1\leq i\leq N}[\delta_{c-1}(i) + a_{ij}] + b_j(\mathbf{o}_c) \\ \psi_c(j) &= \operatorname*{argmin}_{1\leq i\leq N}[\delta_{c-1}(i) + a_{ij}] \qquad \text{where } 2 \leq c \leq C, 1 \leq j \leq N \end{aligned}$$

- **Termination**

$$\begin{aligned} \text{Viterbi score } P* &= \min_{1\leq i\leq N}[\delta_C(i)] \\ \text{state sequence } q_C* &= \operatorname*{argmin}_{1\leq i\leq N}[\delta_C(i)] \end{aligned}$$

- **Path reconstruction**

$$q_c* = \psi c + 1(q_{c+1}*)$$

The state sequence obtained is compared to the tag sequence in the corpus as follows:

Viterbi tag sequence	PRP	MD	IN	NN	JJ	<>	CD	CD	CD	CD	NNS	NN	:	CC	<>	:	<>	.
Corpus tag sequence	PRP	MD	VB	JJ	JJ	<>	CD	CD	CD	CD	<>	NN	NNS	CC	NN	NNS	NNS	.
Mismatchings			*	*							*		*		*	*	*	

We can see that our tag sequence output corresponds mostly to the original one. All the mismatchings are due to either the Chinese word not being found in the lexicon or there being no English word alignment for a Chinese word. This illustrates the fact that the CMM can generate English tags from Chinese words when the Chinese word was correctly segmented and found in the lexicon. However, when we actually apply the CMM to constrain a translation model, we can easily deal with these two cases by applying a null CMM constraint default, i.e.:

if *$Word_c$ not found in lexicon* or *no English word alignment*
  $P[Word_e|Word_c]$ = *translation model probability*
else
  $P[Word_e|Word_c] = P[\text{CMM}(Tag_e|Word_e)]$ + *translation model probability*

5 Scoring Translation Hypotheses

Another way to use the CMM for translation lies in the solution to problem (3): given an English state sequence, we use the CMM to assign it a score. The score is useful because statistical machine translation models can generate a number of translation hypothesis sentences, which can then be compared using the CMM scores. This is analogous to the N-best method employed for speech recognition, which has been found to be more optimal than taking only the 1-best hypothesis sentence (Schwartz & Chow 1990); except in this case the hypothesis sentences are translations.

Given a hypothesized English sentence $\mathbf{E} = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_E)$ with length E, we obtain a tag sequence $q = (q_1, q_2, \ldots, q_E)$ in the following way:

$$
\begin{aligned}
\delta(1) &= a_{1,q_1} \\
\delta(i) &= \delta(i-1) + a_{q_{i-1}q_i} + b_{q_i}(\mathbf{o}_i) \\
\text{Score } P* &= \delta(E)/E \\
\text{where } \mathbf{o}_i &= \text{the Chinese word aligned to the English word } \mathbf{e}_i \\
\text{and } E &= \text{length of the English sentence}
\end{aligned}
$$

For our test, we manually generated a list of 13-best translation hypotheses according to the Chinese words in the following sentence:

我們將為老人增設5000個老人院和安老院名額。

Since the Chinese character sequence can be segmented in different ways into word sequences, the total number of Chinese words in a sentence can be different. For each Chinese sentence with a particular length, we manually generate an alignment English word to the individual Chinese words. Some Chinese words can be aligned to multiple English words leading to multiple hypotheses. Each of these hypothetical sentences is tagged by Brill's tagger. We score the tag sequence of each hypothesis by summing the logarithmic transition probabilities from one tag to the following one, normalized by the length of the sentence. The English hypotheses with their tag sequences, sorted by CMM scores, are shown in Table 3. The lowest score indicates the best translation. The best candidate was chosen to be *We will provide the aged an additional 5000 home and attention home places.* which is indeed the reference translation for the sentence in the original corpus.

Note that CMM scoring cannot choose between two sequences which differ only in their lexical items but not tag sequences. For example, sequence (9) and (10) differ only by their final word—*places* versus *seats*. These two words are both tagged as NNS, therefore the scores for (9) and (10) are the same. However, this lexical choice is obviously a problem of English language modeling, and we can hope that the synthesis part of the statistical translation model will make an intelligent decision between the two.

Table 3: 13-best translation hypotheses and their CMM scores

1:	*11.36*	*We will provide the aged an additional 5000 home and attention home places .* *PRP MD VB DT JJ DT JJ CD NN CC NN NN NNS .*
2:	*11.93*	*We will provide old people in addition 5000 old people home and attention home places .* *PRP MD VB JJ NNS IN NN CD JJ NNS NN CC NN NN NNS .*
3:	*11.98*	*We will for the old people increase 5000 old people homes and attention attention homes places .* *PRP MD IN DT JJ NNS NN CD JJ NNS NNS CC NN NN NNS NNS .*
4:	*12.15*	*We will for the aged an additional 5000 home and attention attending home places .* *PRP MD IN DT JJ DT JJ CD NN CC NN VBG NN NNS .*
5:	*12.22*	*We will for the aged add 5000 home and attention home places .* *PRP MD IN DT JJ VB CD NN CC NN NN NNS .*
6:	*12.34*	*We will provide the aged additional 5000 home and attention home places .* *PRP MD VB DT JJ JJ CD NN CC NN NN NNS .*
7:	*12.77*	*We will for the aged increase 5000 aged people home and caring and attention home places .* *PRP MD IN DT JJ NN CD VBN NNS NN CC NN CC NN NN NNS .*
8:	*12.83*	*We will provide the aged an additional 5000 aged home and attention home places .* *PRP MD VB DT JJ DT JJ CD VBN NN CC NN NN NNS .*
9:	*12.93*	*We will provide the aged increasing 5000 old people home and attention attention home places .* *PRP MD VB DT JJ NN CD JJ NNS NN CC NN NN NN NNS .*
10:	*12.93*	*We will provide the aged increasing 5000 old people home and attention attention home seats .* *PRP MD VB DT JJ NN CD JJ NNS NN CC NN NN NN NNS .*
11:	*13.12*	*We will provide the aged an additional 5000 the aged home and attention home places .* *PRP MD VB DT JJ DT JJ CD DT JJ NN CC NN NN NNS .*
12:	*13.37*	*We will provide the aged adding 5000 aged home and attention home place .* *PRP MD VB DT JJ NN CD VBN NN CC NN NN NN .*
13:	*13.40*	*We will for the aged addition 5000 home and caring attending old people home places .* *PRP MD IN DT JJ NN CD NN CC VBG VBG JJ NNS NN NNS .*

6 Directions

We now plan to investigate how to better model the null states. Since there are many null alignments of Chinese words to English, we would like to develop a more powerful model by looking at the classes of Chinese words that typically have null alignments or other patterns for these alignments.

A single English POS class was used to represent a state in the CMM in our experiments. In the future, we hope to experiment with more complex Markov assumptions. Tri-POS models, for example, are widely used for monolingual language modeling, and we believe that their inclusion can render CMMs more powerful as well.

Finally, we have employed predefined English POS classes for training our CMM. It would be interesting to investigate how different POS class definitions can affect the CMM's performance.

7 Conclusion

We have seen that the Coerced Markov Model is effective in modeling the relationship between the lexical sequence of a sentence in one language and part-of-speech sequence in its translated version. The model coerces the English tag sequence into modeling Chinese word sequence structure, and can be seen as a form of cross-lingual language modeling.

We have formally specified the CMM states, transitions, and output symbols. A method was given for estimating its parameters from a word-aligned training corpus, corresponding to the solution to the first fundamental problem of CMMs. We have shown two applications to improving the statistical transfer model, corresponding to the solutions of fundamental problems (2) and (3) of CMMs: first, we showed that CMM can predict an English tag sequence given a Chinese sentence, providing tag constraints to the search of best English lexical sequence as translation; second, we showed that CMM scoring of a N-best list of translation hypotheses can help to select the best one.

References

BAKER, J.K. 1975. The Dragon system–An overview. *IEEE Trans. Acoustics, Speech, Signal Proc.*, 23(1):24–29.

BAUM, L.E. & J.A. EGON. 1967. An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology. *Bull. Amer. Meteorol. Soc.*, 73:360–363.

BAUM, L.E. & T. PETRIE. 1966. Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat.*, 37:1554–1563.

BRILL, ERIC, 1993. *A corpus-based approach to language learning*. University of Pennsylvania dissertation.

BROWN, P.F., S.A. DELLA PIETRA, V.J. DELLA PIETRA, & R.L. MERCER. 1993. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

CHANG, JYUN-SHENG & HUEY-CHYUN CHEN. 1994. Using partically aligned parallel text and part-of-speech information in word alignment. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, 16–23, Columbia, Maryland.

DAGAN, IDO & KENNETH W. CHURCH. 1994. Termight: Identifying and translating technical terminology. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, 34–40, Stuttgart, Germany.

DAGAN, IDO, KENNETH W. CHURCH, & WILLIAM A. GALE. 1993. Robust bilingual word alignment for machine aided translation. In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, 1–8, Columbus, Ohio.

FORNEY, G.D. 1973. The Viterbi algorithm. In *Proceedings of the IEEE*, volume 61, 268–278.

FUNG, PASCALE. 1995. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proceedings of the 33rd Annual Conference of the Association for Computational Linguistics*, Boston, Massachusettes. To appear.

FUNG, PASCALE & KATHLEEN MCKEOWN. 1994. Aligning noisy parallel corpora across language groups: Word pair feature matching by dynamic time warping. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, 81–88, Columbia, Maryland.

FUNG, PASCALE & DEKAI WU. 1994. Statistical augmentation of a Chinese machine-readable dictionary. In *Proceedings of the 2nd Annual Workshop on Very Large Corpora*, Kyoto, Japan.

JELINEK, F., L.R. BAHL, & R.L. MERCER. 1975. Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Trans. Information Theory*, 21:250–256.

PAPAGEORGIOU, H., L. CRANIAS, & S. PIPERIDIS. 1994. Automatic alignmnet in parallel corpora. In *Proceedings of the 32nd Annual Conference of the Association for Computational Linguistics Student Session*, 331–333, Las Cruces, New Mexico.

RABINER, LAWRENCE & BING-HWANG JUANG. 1993. *Fundamentals of speech recognition*. Signal Processing Series. Englewood Cliffs, NJ: Prentice-Hall.

SCHWARTZ, RICHARD & YEN-LU CHOW. 1990. The N-best algorithm: An efficient and exact procedure for finding the N most likely sentence hypotheses. In *Proceedings of ICASSP 90*, volume S2.12, 81–84, Albuquerque, New Mexico.

VITERBI, A.J. 1967. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transaction on Information Theory*, 13:260–269.

WU, DEKAI. 1994. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *Proceedings of the 32nd Annual Conference of the Association for Computational Linguistics*, 80–87, Las Cruces, New Mexico.

WU, DEKAI. 1995. Grammarless extraction of phrasal translation examples from parallel texts. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, Leuven, Belgium. To appear.

WU, DEKAI & PASCALE FUNG. 1994. Improving Chinese tokenization with linguistic filters on statistical lexical acquisition. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, 180–181, Stuttgart, Germany.

WU, DEKAI & XUANYIN XIA. 1994. Learning an English-Chinese lexicon from a parallel corpus. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, 206–213, Columbia, Maryland.