COERCED MARKOV MODELS FOR CROSS-LINGUAL LEXICAL-TAG RELATIONS

Pascale Fung

Computer Science Department

Columbia University

pascale@cs.columbia.edu

CURRENT MAILING ADDRESS:

Dept. of Electrical and Electronic Engineering

University of Science & Technology

Clear Water Bay, Hong Kong

Dekai Wu

Department of Computer Science

University of Science & Technology

Clear Water Bay, Hong Kong

dekai@cs.ust.hk

Summary

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 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.

1. INTRODUCTION

The analysis, transfer, and synthesis paradigm for machine translation is readily amenable to statistical methods (Brown *et al.* 1993). Since the transfer stage is designed to exploit mapping knowledge about different linguistic relationships between the source and target languages, statistical information can be incorporated into this stage. Typical types of mapping relations include sentence to sentence, word to word, or part-of-speech (POS) tags to tags. Statistical algorithms generally model the word to word lexical relations between a pair of sentences in the target and source languages with probabilities (Brown *et al.* 1993; Dagan *et al.* 1993; Dagan & Church 1994; Fung & McKeown 1994; Wu & Xia 1994; Fung 1994). 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 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 the mapping relations can be automatically learned from bilingual corpora. However, to our knowledge no such attempt has been made, perhaps 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 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 human 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 particular case 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 constraint to the pruned 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(HMM) was proposed by Baum & Petrie (1966); Baum & Egon (1967) as early as the 1960s. It was later adapted by Baker (1975); 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 a, it changes to state b at time $t+1$, then there is a state transition from a to b with certain probability. First order Markov chain assumes the probabilistic dependency of a state is 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 a given state, there is an output from the state. This output can be continuous, such as the spectral signal of speech in a speech recognition system, or discrete, such as the weather condition of a meteorology system. If we regard the mapping between Chinese word sequences 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 Model(HMM) are typically used in speech processing where the underlying states of a model do not correspond to something explicit such as a phoneme or a word. 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 other words, the CMM says that English tags cannot just follow the rules in English language models, they also have to consider the fact that they are now "partners" of Chinese words which also have their own rules. CMM is modeling the "adaptation" of English tags to Chinese word orders. This is a step beyond monolingual language modeling such as word

N-gram or class N-gram computation. CMM's purpose and strength is to model *cross-lingual* class N-grams. An example of a four-state CMM is illustrated in Figure 1.

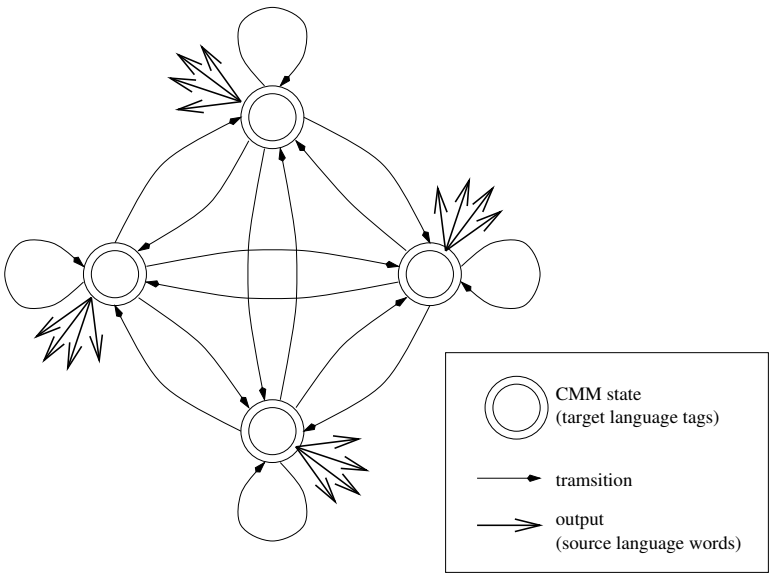

Figure 1: Example of a four-state CMM

Formally, we define the following elements:

1. the state variable N: Hidden Chinese states, with coerced English tag class values

2. the observable symbol M: Chinese words

Our definition of N and M is to optimize the both the modeling and the discriminative power of the CMM. If we had chosen individual lexical items to be the states, there would be a lot of cases where word v never follows word w given any v, w in the dictionary. There would be many restrictions on which state interconnects with which and makes our model neither flexible nor powerful. Instead, it is logical to use POS classes as the states of a CMM because these classes have some physical (or rather, linguistic) significance. In addition, since presumably any POS class can follow any other POS class given a large enough corpus, there can be interconnection between any two given states. This means that the CMM is an ergodic model. This makes CMM potentially

more flexible and powerful. It follows that M, the observable symbols should be the lexical items in the other language.

Once we have defined the nature of N and M, we have to choose which language N and M should come from. We choose N to be the English POS classes, because English POS taggers are readily available and there has been enough agreement in the field as to which once the basic "good" English POS classes are. On the other hand, due to the short history of Chinese natural language processing, most Chinese taggers are still under research, and there is still a lack of a general paradigm for POS classes determination in Chinese. It follows that M is the set of Chinese words in the dictionary.

Referring again to Figure 1, each state in the CMM corresponds to an English POS class. For our experiments, we use Brill (1993)'s tagger of $N = 106$ 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. The output from a state is an array of Chinese words with different weights.

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 have defined the state and output symbols of CMM, now we need to train the parameters of CMM. In this section, we 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 initial state to DT, DT to NNS, NNS to $<>$,.... 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.

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, the logs become negative numbers, therefore we take the negation of the logarithm probabilities for computation.

So 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 number used for *flooring*, i.e. all zero transition probabilities are padded with this small number so we do not get undefined log probabilities.

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

CMM is a discrete Markov Model in which the observable output is in the set M of Chinese words in the dictionary. For example, in the sentence above in Table 1, the Chinese word 穩定 is aligned to *stability* which is tagged as NN, that 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.

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, .

Using a similar negative logarithmic form with flooring as transition probabilities, the output probabilities we get are:

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

An Experimental Setup We use the HKUST Chinese-English Parallel corpus to train our CMM. To prepare a training corpus into 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. The Chinese text did not have word delimiters and it was necessary to **tokenize** strings of Chinese characters into individual words. We used a Viterbi tagger with statistically augmented dictionary (Fung & Wu 1994; Wu & Fung 1994).

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

4. **English word alignment** to the individual Chinese words were found by using a Estimation-Maximization model(Wu & Xia 1994)

5. **Filtering of the training corpus** was done 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 trained CMM as follows:

1. **Compute initial probabilities** π_i: $1 \leq i \leq N$

2. **Compute transition probabilities** a_{ij}:there are 1969 null transitions probabilities out of a total of 11 236 transitions.

3. **Compute output probabilities** $b_j(k)$: $1 \leq j \leq N, 1 \leq k \leq M$

4. OPTIMAL PATH RECOVERY

We use two different evaluation methods corresponding to the solutions of problem(2) and problem(3) in CMM design. Evaluation one was to produce a English tag sequence from a Chinese sentence.

We use 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.

We use the Chinese part of sentence in Table 2 as input to this test, and compare the output to the English tag sequence. To find the solution for predicting the best state sequence, i.e. English tag sequence $q = (q_1, q_2, \ldots, q_C)$ from the observation sequence, i.e. the Chinese sentence $O = (o_1, o_2, \ldots, o_C)$ of length C, we use a Viterbi algorithm(Viterbi 1967; Forney 1973) as the

following, considering that transition probabilities a_{ij} and output probabilities $b_j(o_c)$ are in the negative logarithmic form:

- **Initialization**

$$\begin{aligned} \delta_1(i) &= \pi_i + b_i(O_1) \\ \text{where} \quad 1 \leq i \leq N & \\ \text{where} \quad \pi_i &= \text{probability of } i \text{ being the initial state} \\ \psi_1(i) &= 0 \\ \text{where} \quad 1 \leq i \leq N & \end{aligned}$$

- **Recursion**

$$\begin{aligned} \delta_c(j) &= \min_{1 \leq i \leq N} [\delta_{c-1}(i) + a_{ij}] + b_j(bfO_c) \\ \psi_c(j) &= \arg \min_{1 \leq i \leq N} [\delta_{c-1}(i) + a_{ij}] \\ \text{where} & \quad 2 \leq c \leq C, 1 \leq j \leq N \end{aligned}$$

- **Termination**

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

- **Path reconstruction**

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

Table 2: Test sentence

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

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 dictionary or there being no English word alignment for a Chinese word. This illustrates the fact that CMM can generate English tags from Chinese words when Chinese word was correctly segmented and found in the dictionary. However, when we actually apply 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 dictionary || no English word alignment**
  $P[Word_e|Word_c]$ = *translation model probability*;
else
  $P[Word_e|Word_c]$ = $P[CMM(Tag_e|Word_e)]$ + *translation model probability* ;

5. SCORING TRANSLATION HYPOTHESES

Another way of using CMM for translation is in the solution to problem(3): given an English state sequence, we try to score it by CMM. Statistical machine translation models can generate a number of translation hypotheses sentences according to the EM-based transfer model.This is analogous to the N-best algorithm used for speech recognition and is found to be more optimal in choosing the best candidate sentence(). This sentence would be the best translation in our case.

Given a hypothesis English sentence $g = (g_1, g_2, \ldots, g_E)$ with length E, we obtain a tag sequence $q = (q_1, q_2, \ldots, q_E)$ by the following way:

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

For our test, we manually generated a list of 13-best translation hypotheses according to the Chinese words in the 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 sentence 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, 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 age and additional 5000 home and care 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.

6. DISCUSSION

A problem of CMM which might deserve more research is how to better model the null states. Since there are many null alignments of Chinese words to English, one can try to come up with a more powerful model by looking at the classes of Chinese words which typically have null alignments or other patterns for these alignments.

We used a single English POS class to represent a state in the CMM, it would be worth experimenting with a more complex state such as POS bigrams. POS bigrams are a feature of monolingual language modeling and their inclusion can possibly render CMM more powerful.

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

11.364408	*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 .
11.932087	*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 .
11.982643	*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 .
12.153794	*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 .
12.219560	*We will for the aged add 5000 home and attention home places .*
	PRP MD IN DT JJ VB CD NN CC NN NN NNS .
12.342510	*We will provide the aged additional 5000 home and attention home places .*
	PRP MD VB DT JJ JJ CD NN CC NN NN NNS .
12.766230	*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 .
12.827581	*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 .
12.928034	*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 .
12.928034	*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 .
13.120893	*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 .
13.371675	*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 .
13.402670	*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 .

We have used a predefined English POS classes for training our CMM, it might be worthwhile to investigate how different POS class definitions can affect CMM.

Finally, we would like to point out that there is another similarity between CMM and HMM which is that CMM can also be regarded as a generator of observations: given N,M, A,B, and a sequence of English tags as input, CMM can generate a observation sequence of Chinese words. This would seem to be an application in English to Chinese translation.

7. CONCLUSION

We have seen that the Coerced Markov Model is effective in modeling the relationship between 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 a 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, 1994. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In submission.

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.

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 submission.

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.