Trainable Coarse Bilingual Grammars for Parallel Text Bracketing

Dekai Wu
HKUST
Department of Computer Science
University of Science & Technology
Clear Water Bay, Hong Kong
dekai@cs.ust.hk

Abstract

We describe two new strategies to automatic bracketing of parallel corpora, with particular application to languages where prior grammar resources are scarce: (1) coarse bilingual grammars, and (2) unsupervised training of such grammars via EM (expectation-maximization). Both methods build upon a formalism we recently introduced called *stochastic inversion transduction grammars*. The first approach borrows a coarse monolingual grammar into our bilingual formalism, in order to transfer knowledge of one language's constraints to the task of bracketing the texts in both languages. The second approach generalizes the inside-outside algorithm to adjust the grammar parameters so as to improve the likelihood of a training corpus. Preliminary experiments on parallel English-Chinese text are supportive of these strategies.

1 Introduction

A number of empirical studies have found bracketing to be a useful type of corpus annotation (e.g., Pereira & Schabes 1992; Black *et al.* 1993). Bracketed corpora have been available for some time in English, and to some extent other European languages, the best-known example being perhaps the Penn Treebank (Marcus 1991). However, at present bracketed corpora for Chinese are unknown, as is the case for many other languages. Moreover, even for better-studied languages, *parallel* bracketed texts are scarce.

The problem of bracketing such corpora is the focus of two new strategies described in this paper. The strategies build upon *stochastic inversion transduction grammars* (SITGs), a formalism that we have been developing for bilingual language modeling. Numerous experiments have shown parallel bilingual corpora to provide a rich source of constraints for statistical analysis (e.g., Brown *et al.* 1990; Gale & Church 1991; Gale *et al.* 1992; Church 1993; Brown *et al.* 1993; Dagan *et al.* 1993; Fung & Church 1994; Wu & Xia 1994; Fung & McKeown 1994). SITGs are a generalization of context-free grammars that have several desirable properties for parallel corpus analysis; a brief summary of these properties is given in Section 2.

Our first strategy is to expropriate a very simple, coarse monolingual grammar of English as the backbone for a bilingual English-Chinese SITG, which is then used for bracketing parallel text. The effect of this is to transfer knowledge of English syntactic constraints (or more precisely, probabilistic preferences) to the bilingual task. This is discussed in Section 3.

Our second strategy is to apply an unsupervised training algorithm to tune the probabilistic parameters of the SITG. For this purpose we have devised an EM-based algorithm, a bilingual generalization of the inside-outside method, that iteratively improves the likelihood of the training corpus. This is discussed in Section 4.

It is important to stress at the outset that a *parallel bracketed* corpus is different from a bracketed parallel corpus. The latter is simply a parallel corpus in which both halves have been independently bracketed. In contrast, in a parallel bracketed corpus, the bracketed sub-constituents are themselves parallel in the sense that explicit matching relationships are designated between sub-constituents of each half. This is a much more interesting kind of annotation if it can be accomplished, especially for machine translation applications.

2 Stochastic Inversion Transduction Grammars

In Wu (1995b) we define an *inversion tranduction grammar* (ITG) formalism for bilingual language modeling, i.e., modeling of two languages (referred to as L_1 and L_2) simultaneously. The description here is necessarily brief; for further details the reader is referred to Wu (1995a, 1995b).

An ITG is a context-free grammar that generates output on two separate streams, together with a matching that associates the corresponding tokens and constituents of each stream. The formalism also differs from standard context-free grammars in that the concatenation operation, which is implicit in any production rule's right-hand side, is replaced with two kinds of concatenation with either *straight* or *inverted* orientation. Thus, the following are two distinct productions in an ITG:

$$\begin{array}{lcl} \mathrm{C} & \rightarrow & [\mathrm{A\ B}] \\ \mathrm{C} & \rightarrow & \langle \mathrm{A\ B} \rangle \end{array}$$

Consider each nonterminal symbol to stand for a pair of matched strings, so that for example (A_1, A_2) denotes the string-pair generated by A. The operator $[\,]$ performs the "usual" pairwise concatenation so that $[AB]$ yields the string-pair (C_1, C_2) where $C_1 = A_1B_1$ and $C_2 = A_2B_2$. But the operator $\langle\rangle$ concatenates constituents on output stream 1 while reversing them on stream 2, so that $C_1 = A_1B_1$ but $C_2 = B_2A_2$. The inverted concatenation operator permits the extra flexibility needed to accommodate many kinds of word-order variation between source and target languages. Since inversion is permitted at any level of rule expansion, a derivation may intermix productions of either orientation within the parse tree. More on the ordering flexibility will be said later.

There are also lexical productions of the form:

$$\mathrm{A} \rightarrow x/y$$

where x and y are symbols of languages L_1 and L_2, respectively. Either or both x and y may take the special value ϵ denoting an empty string, allowing a symbol of either language to have no counterpart in the other language by being matched to an empty string. We call x/ϵ an L_1-singleton and ϵ/y an L_2-singleton.

Parsing, in the context of ITGs, means to take as input a sentence-*pair* rather than a sentence, and to output a parse tree that imposes a shared hierarchical structuring on both sentences. For example, Figure 1 shows a parse tree for an English-Chinese sentence translation. The English is read in the usual depth-first left-to-right order, but for the Chinese, a horizontal line means the right subtree is traversed before the left, so that the following sentence pair is generated:

(1) a. [[[The Authority]$_{\text{NP}}$ [will [[be accountable]$_{\text{VV}}$ [to [the [[Financial Secretary]$_{\text{NN}}$]$_{\text{NNN}}$]$_{\text{NP}}$]$_{\text{PP}}$]$_{\text{VP}}$]$_{\text{VP}}$]$_{\text{SP}}$ ./。]$_{\text{S}}$

b. [[[管理局]$_{\text{NP}}$ [將會 [[向 [[[財政 司]$_{\text{NN}}$]$_{\text{NNN}}$]$_{\text{NP}}$]$_{\text{PP}}$ [負責]$_{\text{VV}}$]$_{\text{VP}}$]$_{\text{VP}}$]$_{\text{SP}}$ ./。]$_{\text{S}}$

Alternatively, we can show the common structure of the two sentences more compactly using bracket notation with the aid of the $\langle\rangle$ operator:

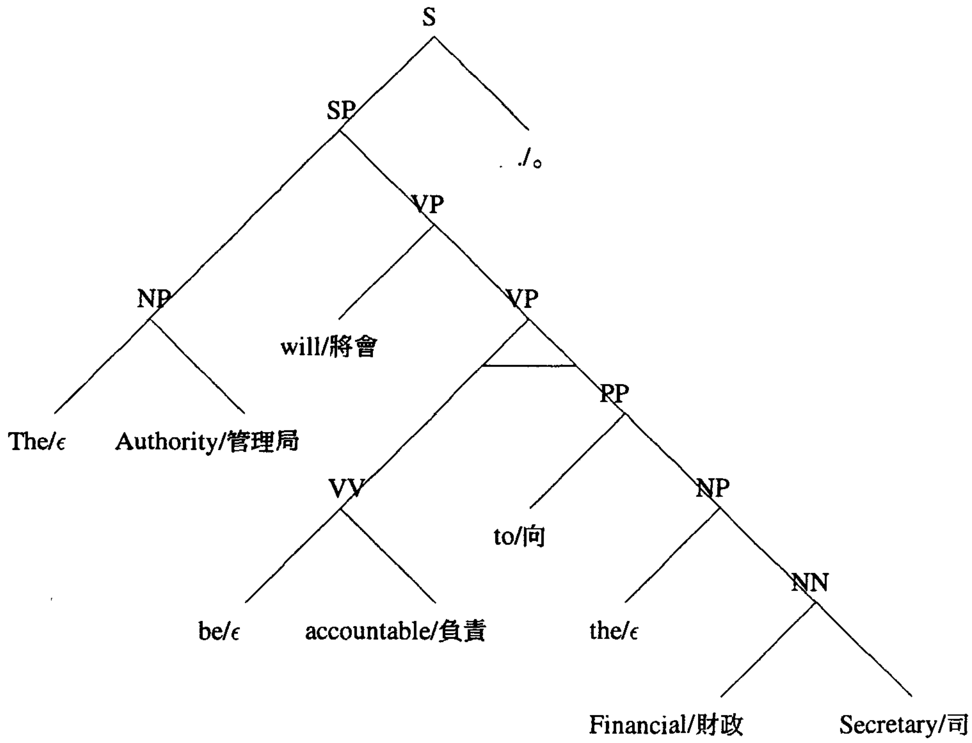

Figure 1: Inversion transducer parse tree.

(2) [[[The/ε Authority/管理局]$_{\mathrm{NP}}$ [will/將會 $\langle$[be/ε accountable/負責]$_{\mathrm{VV}}$ [to/向 [the/ε [[Financial/財政 Secretary/司]$_{\mathrm{NN}}$]$_{\mathrm{NNN}}$]$_{\mathrm{NP}}$]$_{\mathrm{PP}}$ $\rangle_{\mathrm{VP}}$]$_{\mathrm{VP}}$]$_{\mathrm{SP}}$ ./。]$_{\mathrm{S}}$

where the horizontal line from Figure 1 corresponds to the $\langle\rangle$ level of bracketing.

A *stochastic inversion transduction grammar* is an ITG where a probability is associated with each production, subject to the constraint that

$$\sum_{1\le j,k\le N} (a_{i\to[j\,k]} + a_{i\to\langle j\,k\rangle}) + \sum_{\substack{1\le x\le W_1\\ 1\le y\le W_2}} b_i(x,y) = 1$$

where $a_{i\to[j\,k]} = P(i \to [jk]|i)$, $b_i(x,y) = P(i \to x/y|i)$, W_1 and W_2 are the vocabulary sizes of the two languages, and N is the number of nonterminal categories.

Under the stochastic formulation, the objective of parsing is to find the maximum-likelihood parse for a given sentence pair. A general algorithm for this is given in Wu (1995b).

The following convenient theorem is proved in Wu (1995b), which indicates that any ITG can be converted to a normal form, where all productions are either lexical productions or binary-fanout productions:

Theorem 1 *For any inversion transduction grammar G, there exists an equivalent inversion transduction grammar G′ in which every production takes one of the following forms:*

$$\begin{array}{lllllllll} S & \to & \epsilon/\epsilon & A & \to & x/\epsilon & A & \to & [B\,C] \\ A & \to & x/y & A & \to & \epsilon/y & A & \to & \langle B\,C\rangle \end{array}$$

$$\begin{array}{llll} \mathrm{A} & \xrightarrow{a} & [\mathrm{A}\ \mathrm{A}] & \\ \mathrm{A} & \xrightarrow{a} & \langle \mathrm{A}\ \mathrm{A} \rangle & \\ \mathrm{A} & \xrightarrow{b_{ij}} & u_i/v_j & \text{for all } i,j \text{ English-Chinese lexical translations} \\ \mathrm{A} & \xrightarrow{b_{i\epsilon}} & u_i/\epsilon & \text{for all } i \text{ English vocabulary} \\ \mathrm{A} & \xrightarrow{b_{\epsilon j}} & \epsilon/v_j & \text{for all } j \text{ Chinese vocabulary} \end{array}$$

Figure 2: A simple constituent-matching ITG.

The algorithms in this paper assume that ITGs are in this normal form, with one slight relaxation. Lexical productions of the form A $\rightarrow x/y$ may generate multiple-word sequences, i.e., x and y may each be more than one word. This does not affect the generative power, but allows probabilities to be placed on collocation translations. The form is called *lexical normal form.*

ITGs impose two desirable classes of constraints on the space of possible matchings between sentences. *Crossing constraints* prohibit arrangements where the matchings between subtrees cross each another, unless the subtrees' immediate parent constituents are also matched to each other. Aside from linguistic motivations stemming from the compositionality principle, this constraint is important for computational reasons, to avoid exponential bilingual matching times. *Fanout constraints* limit the number of direct sub-constituents of any single constituent, i.e., the number of subtrees whose matchings may cross at any level. We have shown that ITGs inherently permit nearly free matchings for fanouts up to four, with strong constraints thereafter creating a rapid falloff in the proportion of matchings permitted (Wu 1995a). This characteristic gives ITGs just the right degree of flexibility needed to map syntactic structures interlingually.

3 Coarse Bilingual Grammars

Because the expressiveness of ITGs naturally constrains the space of possible matchings in a highly appropriate fashion, the possibility arises that the information supplied by a word-translation lexicon alone may be adequately discriminating to match constituents, without language-specific monolingual grammars for the source and target languages, simply by bringing the ITG constraints to bear in tandem with lexical matching. That is, the bilingual SITG parsing algorithm can perform constituent identification and matching using only a generic, language-independent bracketing grammar.

Several earlier experiments (Wu 1995a) tested out variants of this hypothesis, using generic SITGs similar to the one shown in Figure 2, which employs only one nonterminal category. The first two productions are sufficient to generate all possible matchings of ITG expressiveness (this follows from the normal form theorem). The remaining productions are all lexical. Productions of the A $\rightarrow u_i/v_j$ form list all word translations found in the translation lexicon, and the others list all potential singletons without corresponding translations. Thus, a parser with this grammar can build a bilingual parse tree for any possible ITG matching on a pair of input sentences.

Probabilities on the grammar are placed as follows. The b_{ij} distribution encodes the English-Chinese translation lexicon with degrees of probability on each potential word translation. A small ϵ-constant can be chosen for the probabilities $b_{i\epsilon}$ and $b_{\epsilon j}$, so that the optimal matching resorts to these productions only when it is otherwise impossible to match the singletons. The result is that the maximum-likelihood parser selects the parse tree that best meets the combined lexical translation preferences, as expressed by the b_{ij} probabilities.

Performance, as reported in Wu (1995a), was encouraging, with precision on automatically-filtered

```
Chinese:  他 們 這 樣 做 十 分 正 確 。
English:  They are right to do so .

A[
  They/他們
  A<
    are/*
    right/正確
    to/*
    */這樣
    do/做
    */十分
    so/*
  >A
  ./。
]A
```

Figure 3: A problematic sentence pair with a generic bracketing grammar.

sentence pairs in the 80% range with the aid of supporting heuristics. However, there are of course inherent limitations of any approach that relies entirely on crossing- and fanout-constrained lexical matching. In particular, if the sub-constituents of any constituent appear in the same order in both languages, lexical matchings do not provide the discriminative leverage to identify the sub-constituent boundaries. This applies to both straight and inverted orientations; an example with inverted orientation is shown in Figure 3. In such cases, specific grammatical information about one or both of the languages is needed.

Grammatical information is far less easily available for Chinese than for English, however, with respect to part-of-speech lexicons as well as grammars. The SITG formalism offers another possibility: the generic bracketing grammar can be replaced with a context-free backbone designed for English.

It is critical under this approach that the English grammar be reasonably robust. It should also avoid being too specific, since to be effective at bracketing, its structure must accomodate Chinese to a reasonably broad extent. For these reasons it is best to employ a simple, coarse grammar, with fallback productions that simulate the generic bracketing grammar when the English productions are too inflexible.

As before, the lexical productions will constitute the bulk of the rules set. However, we can now distinguish between different part-of-speech nonterminals. Different part-of-speech nonterminals may generate the same words. We can accomodate the fact that no Chinese part-of-speech lexicon is available with noninformative distributions as follows:

1. The conditional distribution over $L \xrightarrow{b_{i\epsilon}} u_i/\epsilon$ productions is estimated from the frequencies for each English part-of-speech L.

2. The conditional distribution over $L \xrightarrow{b_{ij}} u_i/v_j$ productions is estimated from the frequencies for the English part-of-speech L uniformly distributed over the set of matching Chinese words.

3. The conditional distribution over $L \xrightarrow{b_{\epsilon j}} \epsilon/v_j$ productions is uniformly distributed over the Chinese vocabulary.

$$\begin{array}{llll}
\mathrm{S0} & \rightarrow & [\mathrm{S0\ S0}] \mid \langle \mathrm{S0\ S0} \rangle & \text{start symbol} \\
\mathrm{S0} & \rightarrow & [\mathrm{S\ S0}] \mid \langle \mathrm{S\ S0} \rangle & \\
\mathrm{S0} & \rightarrow & [\mathrm{N1\ V2}] \mid \langle \mathrm{N1\ V2} \rangle & \\
\mathrm{S0} & \rightarrow & [\mathrm{N1\ V1}] \mid \langle \mathrm{N1\ V1} \rangle & \\
\mathrm{S0} & \rightarrow & [\mathrm{N1\ VB}] \mid \langle \mathrm{N1\ VB} \rangle & \\
\mathrm{S0} & \rightarrow & [\mathrm{NP\ V2}] \mid \langle \mathrm{NP\ V2} \rangle & \\
\mathrm{S0} & \rightarrow & [\mathrm{NP\ V1}] \mid \langle \mathrm{NP\ V1} \rangle & \\
\mathrm{S0} & \rightarrow & [\mathrm{NP\ VB}] \mid \langle \mathrm{NP\ VB} \rangle & \\
\mathrm{V2} & \rightarrow & [\mathrm{V2\ PP}] \mid \langle \mathrm{V2\ PP} \rangle & \text{ditransitive verb phrases} \\
\mathrm{V2} & \rightarrow & [\mathrm{V1\ PP}] \mid \langle \mathrm{V1\ PP} \rangle & \\
\mathrm{V2} & \rightarrow & [\mathrm{VB\ PP}] \mid \langle \mathrm{VB\ PP} \rangle & \\
\mathrm{V2} & \rightarrow & [\mathrm{V1\ N1}] \mid \langle \mathrm{V1\ N1} \rangle & \\
\mathrm{V2} & \rightarrow & [\mathrm{V1\ NP}] \mid \langle \mathrm{V1\ NP} \rangle & \\
\mathrm{V1} & \rightarrow & [\mathrm{VB\ N1}] \mid \langle \mathrm{VB\ N1} \rangle & \text{transitive verb phrases} \\
\mathrm{V1} & \rightarrow & [\mathrm{V0\ N1}] \mid \langle \mathrm{V0\ N1} \rangle & \\
\mathrm{V1} & \rightarrow & [\mathrm{VB\ NP}] \mid \langle \mathrm{VB\ NP} \rangle & \\
\mathrm{V1} & \rightarrow & [\mathrm{V0\ NP}] \mid \langle \mathrm{V0\ NP} \rangle & \\
\mathrm{V0} & \rightarrow & [\mathrm{VB\ V0}] \mid \langle \mathrm{VB\ V0} \rangle & \text{verb sequences} \\
\mathrm{PP} & \rightarrow & [\mathrm{IN\ N1}] \mid \langle \mathrm{IN\ N1} \rangle & \text{prepositional phrases} \\
\mathrm{PP} & \rightarrow & [\mathrm{IN\ NP}] \mid \langle \mathrm{IN\ NP} \rangle & \\
\mathrm{N1} & \rightarrow & [\mathrm{N1\ PP}] \mid \langle \mathrm{N1\ PP} \rangle & \text{noun phrases} \\
\mathrm{N1} & \rightarrow & [\mathrm{NP\ PP}] \mid \langle \mathrm{NP\ PP} \rangle & \\
\mathrm{N1} & \rightarrow & [\mathrm{DT\ N0}] \mid \langle \mathrm{DT\ N0} \rangle & \\
\mathrm{N1} & \rightarrow & [\mathrm{DT\ NN}] \mid \langle \mathrm{DT\ NN} \rangle & \\
\mathrm{N0} & \rightarrow & [\mathrm{NN\ N0}] \mid \langle \mathrm{NN\ N0} \rangle & \text{complex nominals}
\end{array}$$

Figure 4: Syntactic productions of a stochastic constituent-matching ITG.

Because the grammar is coarse while the lexicon is fine, the approach retains the previous approach's high sensitivity to lexical matching constraints.

It is interesting to constrast this method with the "parse-parse-match" approaches that have been reported recently for producing parallel bracketed corpora (Sadler & Vendelmans 1990; Kaji *et al.* 1992; Matsumoto *et al.* 1993; Cranias *et al.* 1994; Grishman 1994). "Parse-parse-match" methods first bracket a parallel corpus by parsing each half individually using a monolingual grammar.[1] Heuristic procedures are subsequently used to select a matching between the bracketed constituents across sentence-pairs. These approaches can encounter difficulties with incompatibilities between the monolingual grammars used to parse the texts. The grammars will usually be of unrelated origins, not designed to make interlingual matching easy. Furthermore, how to deal with ambiguities presents another serious problem. Most sentences in the corpus will have multiple possible parses. In a pure "parse-parse-match" approach, however, the monolingual parsers must arbitrarily select one bracketing with which to annotate the corpus. The resulting parse may be incompatible with the parse chosen for the other half of the sentence-pair, causing a matching error even though some alternative parse might in fact been compatible.

The coarse bilingual grammar approach proposed here solves these problems by choosing the parse

[1] Of course, this assumes that adequate grammars are available for both languages, contrary to our present assumptions.

S	→	u_i/v_j	miscellaneous
S	→	u_i/ϵ	
S	→	ϵ/v_j	
VB	→	u_i/v_j	verbs, auxiliary verbs
VB	→	u_i/ϵ	
VB	→	ϵ/v_j	
NN	→	u_i/v_j	nouns, adjectives
NN	→	u_i/ϵ	
NN	→	ϵ/v_j	
NP	→	u_i/v_j	pronouns
NP	→	u_i/ϵ	
NP	→	ϵ/v_j	
IN	→	u_i/v_j	prepositions
IN	→	u_i/ϵ	
IN	→	ϵ/v_j	
DT	→	u_i/v_j	determiners
DT	→	u_i/ϵ	
DT	→	ϵ/v_j	

Figure 5: Lexical productions of a stochastic constituent-matching ITG.

structure for both sentences simultaneously with the interlingual constituent matching criteria. The weighting of the bracketing constraints and matching constraints is probabilistic. Even if a sentence pair's translations truly contain structural mismatches that are beyond syntactic accounts, the soft constraint optimization permits graceful degradation in the bilingual parse. The parser will attempt to match those constituents for which a partial decomposition and matching can be found, parsing the rest largely according to the English grammar backbone.

More sophisticated "parse-parse-match" procedures postpone ambiguity resolution until the matching stage (Kaji *et al.* 1992; Matsumoto *et al.* 1993; Grishman 1994). This tactic bears closer resemblance to our approach, but still requires *ad hoc* heuristics to determine exactly how the matching task influences the monolingual parsès that are chosen. On the other hand, the present framework incorporates all these aspects within a single probabilistic optimization.

Another alternative approach discussed in Wu (1995b) is to first use a monolingual grammar to bracket only the English half of the text, followed by a SITG parallel bracketing procedure constrained by the English brackets. However, this hybrid approach is subject to the same incompatibility and ambiguity problems that arise for pure "parse-parse-match" procedures; thus the proposed coarse bilingual grammar approach is superior for the same reasons given above.

For our experiments, we employed the grammar shown in Figures 4 and 5, with only 50 syntactic productions and 13 nonterminal categories, including 6 part-of-speech categories. Each syntactic production occurs in both straight and inverted orientations, to model ignorance of the ordering tendencies of the corresponding Chinese constituents. The part-of-speech categories were designed by conflating categories in the Brown corpus tagset, under the following general principle: categories should be as broad as possible, while still maintaining reasonable discriminativeness for bracketing structure. Thus, notice that adjectives and nouns are conflated, since complex nominal phrases have largely similar parse structures regardless of the difference between adjective and noun labels. Similarly, all verbs including auxiliaries are grouped to allow simple tail-recursive compounding. The S category (not to be confused with the start symbol S0) is a placeholder for miscellaneous items including punctuation and adverbs, and functions as a fallback category similar to the A nonterminal in the generic bracketing grammars.

Probabilities were placed on the syntactic productions uniformly, but all inverted productions were

```
他 們 這 樣 做 十 分 正 確 。
They are right to do so .


S0[
  S0[
    They/他們 NP
    V1[
      V0[
        are/* VB
        V0<
          right/正確 VB
          V0[
            to/* VB
            V0[
              */這樣 VB
              V0[
                do/做 VB
                */十分 V0
              ]V0
            ]V0
          ]V0
        >V0
      ]V0
      so/* NP
    ]V1
  ]S0
  ./。 S0
]S0
```

```
管 理 局 將 會 向 財 政 司 負 責 。
The Authority will be accountable
  to the Financial Secretary .

S0[
  S0[
    N1<
      N1[
        The/* DT
        Authority/局 NN
      ]N1
      */管理 PP
    >N1
    V1[
      will/將會 VB
      N1<
        N1[
          be/* DT
          accountable/負責 NN
        ]N1
        PP[
          to/* IN
          N1[
            the/* DT
            N0[
              */向 NN
              N0[
                Financial/財政司 NN
                Secretary/* N0
              ]N0
            ]N0
          ]N1
        ]PP
      >N1
    ]V1
  ]S0
  ./。 S0
]S0
```

Figure 6: Sample outputs with a coarse bilingual grammar.

assigned a slightly smaller probability in order to break ties in favor of straight matchings. Probabilities were placed on the lexical productions as discussed above, with the following additional provisions. The translation lexicon was automatically learned from the HKUST English-Chinese Parallel Bilingual Corpus via statistical sentence alignment (Wu 1994) and statistical Chinese word and collocation extraction (Fung & Wu 1994; Wu & Fung 1994), followed by an EM word-translation learning procedure (Wu & Xia 1994). The latter stage gives us the lexical translation probabilities. The translation lexicon contained approximately 6,500 English words and 5,500 Chinese words, and was *not* manually corrected for this experiment, having about 86% translation accuracy. The English part-of-speech lexicon with relative frequencies was derived from the English portion of our corpus as tagged by Brill's (1993) tagger.

Our preliminary experiments show improved parsing behavior in general, compared to generic bracketing grammars. Examples of the output are shown in Figure 6. The latter example shows problematic behavior on the example given earlier in Figure 3 of sentence pairs without sufficient ordering discrimination. Although an attempt is made in this case to fit the English constraints, the main difficulty is that the translation "so/這樣" was missing from the automatically-learned lexicon; also, the simple grammar lacks infinitival clauses.

4 An EM Algorithm for Training SITGs

An unavoidable consequence of using more structured, complex grammars—coarse though they may be—is that the bilingual matching process becomes more sensitive to the syntactic production probabilities than under the earlier generic bracketing grammar approaches. Performance therefore suffers if the probabilities are not appropriate, a serious problem given that the syntactic production probabilities above are manually, and arbitrarily, set to be uniform.

It therefore becomes desirable to find means to tune the syntactic production probabilities automatically, so as to be optimal with respect to some training data set. Note that we do not expect the parallel training corpus to be parsed or otherwise syntactically annotated beforehand. To this end we present an EM (expectation-maximization) algorithm for iteratively improving the syntactic production parameters of a SITG, according to a likelihood criterion. The method is a generalization of the inside-outside algorithm for SCFG estimation (Baker 1979; Lari & Young 1990).

A few notational preliminaries: we will denote the sentence pairs by $(\mathbf{E}, \mathbf{C})$ where the English sentence $\mathbf{E} = \mathbf{e}_1, \ldots, \mathbf{e}_T$ and the corresponding Chinese sentence $\mathbf{C} = \mathbf{c}_1, \ldots, \mathbf{c}_V$ are vectors of observed symbols (that is, lexemes or words). As an abbreviation we write $\mathbf{e}_{s..t}$ for the sequence of words $\mathbf{e}_{s+1}, \mathbf{e}_{s+2}, \ldots, \mathbf{e}_t$, and similarly for $\mathbf{c}_{u..v}$. It will be convenient to use a 4-tuple of the form $q = (s, t, u, v)$ to identify each node of the parse tree, where the substrings $\mathbf{e}_{s..t}$ and $\mathbf{c}_{u..v}$ both derive from the node q. Denote the nonterminal label on $q = (s, t, u, v)$ by ℓ_q or ℓ_{stuv}, with the convention that $\ell_{stuv} = 0$ means that $\mathbf{e}_{s..t}$ and $\mathbf{c}_{u..v}$ are not derived from a single common nonterminal.

The *inside probabilities*, defined as:

$$(1) \qquad \beta_{stuv}(i) \;=\; P[i \overset{*}{\Rightarrow} \mathbf{e}_{s..t}/\mathbf{c}_{u..v} | \ell_{stuv} = i, \Phi]$$

are computed recursively as follows.

1. **Basis**

$$(2) \qquad \beta_{ttvv}(i) \;=\; 0 \qquad 0 \leq t \leq T, 0 \leq v \leq V$$

$$(3) \qquad \beta^0_{stuv}(i) \;=\; b_i(\mathbf{e}_{s..t}/\mathbf{c}_{u..v}) \qquad \begin{cases} 0 \leq s \leq t \leq T, 0 \leq u \leq v \leq V, \\ \qquad (t-s)(v-u) \neq 0 \end{cases}$$

2. **Recursion**

$$\begin{array}{lrcl}
(4) & \beta_{stuv}(i) & = & \beta^{[]}_{stuv}(i) + \beta^{\langle\rangle}_{stuv}(i) + \beta^{0}_{stuv}(i) \\
(5) & \beta^{[]}_{stuv}(i) & = & \sum\limits_{\substack{1\le j\le N\\ 1\le k\le N\\ s\le S\le t\\ u\le U\le v\\ (S-s)(t-S)+(U-u)(v-U)\neq 0}} a_{i\to[jk]}\, \beta_{sSuU}(j)\, \beta_{StUv}(k) \\
(6) & \beta^{\langle\rangle}_{stuv}(i) & = & \sum\limits_{\substack{1\le j\le N\\ 1\le k\le N\\ s\le S\le t\\ u\le U\le v\\ (S-s)(t-S)+(U-u)(v-U)\neq 0}} a_{i\to\langle jk\rangle}\, \beta_{sSUv}(j)\, \beta_{StuU}(k)
\end{array}$$

Subsequent to the inside computation, the *outside probabilities*, defined as:

$$(7) \qquad \alpha_{stuv}(i) = P[\mathrm{S} \overset{*}{\Rightarrow} \mathbf{e}_{0..s} i \mathbf{e}_{t..T} / \mathbf{c}_{0..u} i \mathbf{c}_{v..V}, \ell_{stuv} = i | \Phi]$$

are also computed recursively:

1. **Basis**

$$\begin{array}{lrcll}
(8) & \alpha_{0,T,0,V}(i) & = & \begin{cases} 1 & \text{if } i = \mathrm{S} \\ 0 & \text{otherwise} \end{cases} & \\
(9) & \alpha_{ttvv}(i) & = & 0 & 0 \le t \le T, 0 \le v \le V
\end{array}$$

2. **Recursion**

$$\begin{array}{lrcl}
(10) & \alpha_{stuv}(i) & = & \alpha^{[]}_{stuv}(i) + \alpha^{\langle\rangle}_{stuv}(i) \\
(11) & \alpha^{[]}_{stuv}(i) & = & \sum\limits_{\substack{1\le j\le N\\ 1\le k\le N\\ 0\le S\le s\\ 0\le U\le u\\ (s-S)(u-U)\neq 0}} \alpha_{StUv}(j)\, a_{j\to[ki]}\, \beta_{SsUu}(k) + \sum\limits_{\substack{1\le j\le N\\ 1\le k\le N\\ t\le S\le T\\ v\le U\le V\\ (S-t)(U-v)\neq 0}} \alpha_{sSuU}(j)\, a_{j\to[ik]}\, \beta_{tSvU}(k) \\
(12) & \alpha^{\langle\rangle}_{stuv}(i) & = & \sum\limits_{\substack{1\le j\le N\\ 1\le k\le N\\ 0\le S\le s\\ v\le U\le V\\ (s-S)(U-v)\neq 0}} \alpha_{StuU}(j)\, a_{j\to\langle ki\rangle}\, \beta_{SsvU}(k) + \sum\limits_{\substack{1\le j\le N\\ 1\le k\le N\\ t\le S\le T\\ 0\le U\le u\\ (S-t)(u-U)\neq 0}} \alpha_{sSUv}(j)\, a_{j\to\langle ik\rangle}\, \beta_{tSUu}(k)
\end{array}$$

The estimation procedure for adjusting the model parameter set Φ is defined in terms of the inside and outside probabilities. We begin by considering for each nonterminal the probability of its use in a derivation of the observed sentence-pair:

$$P[i \text{ used} \mid \mathrm{S} \overset{*}{\Rightarrow} \mathbf{E}/\mathbf{C}, \Phi] = \frac{\sum_{s=0}^{T}\sum_{t=s}^{T}\sum_{u=0}^{V}\sum_{v=u}^{V} P[\mathrm{S} \overset{*}{\Rightarrow} \mathbf{E}/\mathbf{C} | \ell_{stuv} = i, \Phi]}{P[\mathrm{S} \overset{*}{\Rightarrow} \mathbf{E}/\mathbf{C} | \Phi]} \tag{13}$$

$$= \frac{\sum_{s=0}^{T}\sum_{t=s}^{T}\sum_{u=0}^{V}\sum_{v=u}^{V} \alpha_{stuv}(i)\beta_{stuv}(i)}{P[\mathrm{S} \overset{*}{\Rightarrow} \mathbf{E}/\mathbf{C} | \Phi]} \tag{14}$$

The probability of using each straight production rule in a derivation of the observed sentence-pair is:

$$P[i \rightarrow [jk] \text{ used} \mid \mathrm{S} \overset{*}{\Rightarrow} \mathbf{E}/\mathbf{C}, \Phi] = \sum_{s=0}^{T}\sum_{t=s}^{T}\sum_{u=0}^{V}\sum_{v=u}^{V} P[i \Rightarrow [jk] \overset{*}{\Rightarrow} \mathbf{e}_{s..t}/\mathbf{c}_{u..v} \mid \mathrm{S} \overset{*}{\Rightarrow} \mathbf{E}/\mathbf{C}, \Phi] \tag{15}$$

$$= \frac{\sum_{s=0}^{T}\sum_{t=s}^{T}\sum_{u=0}^{V}\sum_{v=u}^{V}\sum_{S=s}^{t}\sum_{U=u}^{v} a_{i\rightarrow[jk]}\alpha_{stuv}(i)\beta_{sSuU}(j)\beta_{StUv}(k)}{P[\mathrm{S} \overset{*}{\Rightarrow} \mathbf{E}/\mathbf{C} | \Phi]} \tag{16}$$

Similarly for each inverted production rule:

$$P[i \rightarrow \langle jk \rangle \text{ used} \mid \mathrm{S} \overset{*}{\Rightarrow} \mathbf{E}/\mathbf{C}, \Phi] = \sum_{s=0}^{T}\sum_{t=s}^{T}\sum_{u=0}^{V}\sum_{v=u}^{V} P[i \Rightarrow \langle jk \rangle \overset{*}{\Rightarrow} \mathbf{e}_{s..t}/\mathbf{c}_{u..v} \mid \mathrm{S} \overset{*}{\Rightarrow} \mathbf{E}/\mathbf{C}, \Phi] \tag{17}$$

$$= \frac{\sum_{s=0}^{T}\sum_{t=s}^{T}\sum_{u=0}^{V}\sum_{v=u}^{V}\sum_{S=s}^{t}\sum_{U=u}^{v} a_{i\rightarrow\langle jk \rangle}\alpha_{stuv}(i)\beta_{sSUv}(j)\beta_{StuU}(k)}{P[\mathrm{S} \overset{*}{\Rightarrow} \mathbf{E}/\mathbf{C} | \Phi]} \tag{18}$$

By definition, the syntactic production probabilities are:

$$a_{i\rightarrow[jk]} = P[i \rightarrow [jk] \text{ used} \mid i \text{ used}, \mathrm{S} \overset{*}{\Rightarrow} \mathbf{E}/\mathbf{C}, \Phi] \tag{19}$$

$$a_{i\rightarrow\langle jk \rangle} = P[i \rightarrow \langle jk \rangle \text{ used} \mid i \text{ used}, \mathrm{S} \overset{*}{\Rightarrow} \mathbf{E}/\mathbf{C}, \Phi] \tag{20}$$

Substitution yields a re-estimation procedure for $\hat{A}$:

$$\hat{a}_{i\rightarrow[jk]} = \frac{\sum_{s=0}^{T}\sum_{t=s}^{T}\sum_{u=0}^{V}\sum_{v=u}^{V}\sum_{S=s}^{t}\sum_{U=u}^{v} a_{i\rightarrow[jk]}\alpha_{stuv}(i)\beta_{sSuU}(j)\beta_{StUv}(k)}{\sum_{s=0}^{T}\sum_{t=s}^{T}\sum_{u=0}^{V}\sum_{v=u}^{V} \alpha_{stuv}(i)\beta_{stuv}(i)} \tag{21}$$

$$\hat{a}_{i\rightarrow\langle jk \rangle} = \frac{\sum_{s=0}^{T}\sum_{t=s}^{T}\sum_{u=0}^{V}\sum_{v=u}^{V}\sum_{S=s}^{t}\sum_{U=u}^{v} a_{i\rightarrow\langle jk \rangle}\alpha_{stuv}(i)\beta_{sSUv}(j)\beta_{StuU}(k)}{\sum_{s=0}^{T}\sum_{t=s}^{T}\sum_{u=0}^{V}\sum_{v=u}^{V} \alpha_{stuv}(i)\beta_{stuv}(i)} \tag{22}$$

The behavior of a typical training run is shown in Figure 7. The relative movement of the log likelihood is what is important here. The absolute magnitudes are not meaningful since they are largely determined by the fixed lexical translation probabilities. What is significant is that due to the relatively small number of parameters being trained, convergence is achieved within two or three iterations. (The rise in perplexity afterwards is caused by numerical error on overtrained parameters; we terminate training as soon as this occurs.)

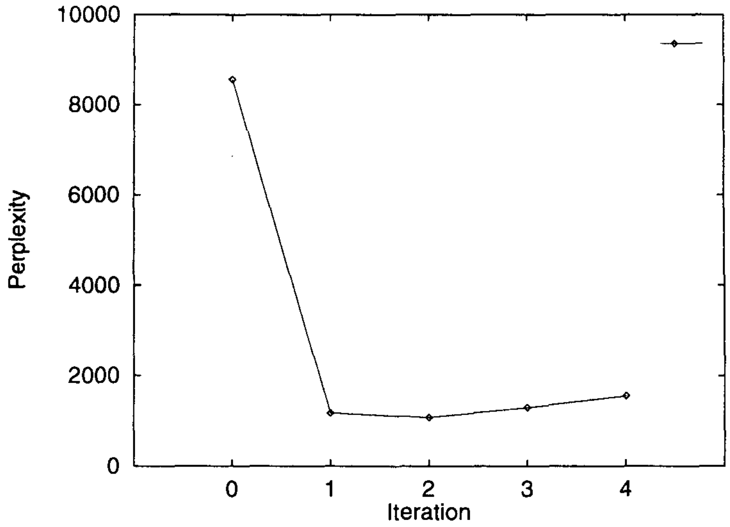

Figure 7: Perplexity on successive training iterations.

5 Conclusion

We have described two new approaches to automatic bracketing of parallel corpora, which are particularly applicable to languages where grammar resources are scarce. The methods—coarse bilingual grammars expropriated from monolingual grammars, with EM parameter estimation—are grounded upon a firm theoretical model, and preliminary experiments show promising behavior. The training method does not require syntactically annotated parallel corpora, which are difficult to obtain. We are presently conducting more quantitative evaluations of the bracketing performance improvement.

6 Acknowledgements

I would like to thank Xuanyin Xia and Eva Wai-man Fong for data conversion assistance.

References

BAKER, JAMES K. 1979. Trainable grammars for speech recognition. In *Speech Communication Papers for the 97th Meeting of the Acoustic Society of America*, ed. by D. H. Klatt & J. J. Wolf, 547–550.

BLACK, EZRA, ROGER GARSIDE, & GEOFFREY LEECH (eds.). 1993. *Statistically-driven computer grammars of English: The IBM/Lancaster approach*. Amsterdam: Editions Rodopi.

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

BROWN, PETER F., JOHN COCKE, STEPHEN A. DELLAPIETRA, VINCENT J. DELLAPIETRA, FREDERICK JELINEK, JOHN D. LAFFERTY, ROBERT L. MERCER, & PAUL S. ROOSSIN. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):29–85.

BROWN, PETER F., STEPHEN A. DELLAPIETRA, VINCENT J. DELLAPIETRA, & ROBERT L. MERCER. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

CHURCH, KENNETH W. 1993. Char-align: A program for aligning parallel texts at the character level. In *Proceedings of the 31st Annual Conference of the Association for Computational Linguistics*, 1–8, Columbus, OH.

CRANIAS, LAMBROS, HARRIS PAPAGEORGIOU, & STELIOS PEPERIDIS. 1994. A matching technique in example-based machine translation. In *Proceedings of the Fifteenth International Conference on Computational Linguistics*, 100–104, Kyoto.

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*, 1–8, Columbus, OH.

FUNG, PASCALE & KENNETH W. CHURCH. 1994. K-vec: A new approach for aligning parallel texts. In *Proceedings of the Fifteenth International Conference on Computational Linguistics*, 1096–1102, Kyoto.

FUNG, PASCALE & KATHLEEN MCKEOWN. 1994. Aligning noisy parallel corpora across language groups: Word pair feature matching by dynamic time warping. In *AMTA-94, 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 Second Annual Workshop on Very Large Corpora*, 69–85, Kyoto.

GALE, WILLIAM A. & KENNETH W. CHURCH. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Conference of the Association for Computational Linguistics*, 177–184, Berkeley.

GALE, WILLIAM A., KENNETH W. CHURCH, & DAVID YAROWSKY. 1992. Using bilingual materials to develop word sense disambiguation methods. In *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, 101–112, Montreal.

GRISHMAN, RALPH. 1994. Iterative alignment of syntactic structures for a bilingual corpus. In *Proceedings of the Second Annual Workshop on Very Large Corpora*, 57–68, Kyoto.

KAJI, HIROYUKI, YUUKO KIDA, & YASUTSUGU MORIMOTO. 1992. Learning translation templates from bilingual text. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, 672–678, Nantes.

LARI, K. & S. J. YOUNG. 1990. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language*, 4:35–56.

MARCUS, MITCHELL. 1991. The automatic acquisition of linguistic structure from large corpora: An overview of work at the university of pennsylvania. In *Working Notes from the Spring Symposium on Machine Learning of Natural Language and Ontology*, 123–125, Stanford University, Stanford, CA. AAAI.

MATSUMOTO, YUJI, HIROYUKI ISHIMOTO, & TAKEHITO UTSURO. 1993. Structural matching of parallel texts. In *Proceedings of the 31st Annual Conference of the Association for Computational Linguistics*, 23–30, Columbus, OH.

PEREIRA, FERNANDO & YVES SCHABES. 1992. Inside-outside reestimation from partially bracketed corpora. In *Proceedings of the 30th Annual Conference of the Association for Computational Linguistics*, 128–135, Newark, DE.

SADLER, VICTOR & RONALD VENDELMANS. 1990. Pilot implementation of a bilingual knowledge bank. In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, 449–451, Helsinki.

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. 1995a. 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. 1995b. Stochastic inversion transduction grammars, with application to segmentation, bracketing, and alignment of parallel corpora. In *Proceedings of IJCAI-95, Fourteenth International Joint Conference on Artificial Intelligence*, Montreal. To appear.

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

WU, DEKAI & XUANYIN XIA. 1994. Learning an English-Chinese lexicon from a parallel corpus. In *AMTA-94, Association for Machine Translation in the Americas*, 206–213, Columbia, Maryland.