Using Brackets to Improve Search for Statistical Machine Translation

Dekai WU **Cindy NG**

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

Abstract

We propose a method to improve search time and space complexity in statistical machine translation architectures, by employing linguistic bracketing information on the source language sentence. It is one of the advantages of the probabilistic formulation that competing translations may be compared and ranked by a principled measure, but at the same time, optimizing likelihoods over the translation space dictates heavy search costs. To make statistical architectures practical, heuristics to reduce search computation must be incorporated. An experiment applying our method to a prototype Chinese-English translation system demonstrates substantial improvement.

1 Introduction

The work we discuss here is embedded within the SILC project at HKUST (Wu 1994; Fung & Wu 1994; Wu & Fung 1994; Wu & Xia 1995; Wu 1995a; Wu 1995b; Wu 1995c) which focuses on problems of *machine translation learning*. We are developing machine learning techniques to bear upon the shortage of adequate knowledge resources for natural language analysis, particularly for Chinese where there is relatively little previous computational linguistics research from which to draw. It is one of our objectives to investigate the suitability for Chinese of the statistical translation model originally proposed by IBM (Brown *et al.* 1990; Brown *et al.* 1993) for Indo-European languages. Henceforth we will therefore use "Chinese" to refer to the source language and "English" to refer to the target language, reflecting the prototype SILC system.

An inherent characteristic of the basic IBM stochastic channel model is the large search space, due to the wide range of distortions that must be allowed in order to successfully transfer sentences of one language to the other. The underlying generative model maps target-language strings into source-language strings (i.e., in the reverse direction from translation). During translation, a maximum likelihood target-language string is sought for the input source-language string, according to Bayes' formula:

$$\operatorname*{argmax}_{\mathbf{e}} \Pr(\mathbf{e}|\mathbf{c}) = \operatorname*{argmax}_{\mathbf{e}} \alpha \Pr(\mathbf{c}|\mathbf{e}) \Pr(\mathbf{e}) \tag{1}$$

The distortion operations in the channel model are chosen to permit sufficient flexibility to map English strings into Chinese translations that have greatly different word order. (It is a simplifying assumption of the model that the only sentence translations considered are those where the majority of words can be translated by lexical substitution.) The scheme admits many implausible mappings along with the legitimate translations, but thereby gains robustness. During the recognition process, legitimate translations will be selected so long as the implausible mappings have lower likelihoods.

The IBM model employs an $A*$ search strategy on the space of translation hypotheses using incremental hypothesis expansion. The distance-to-goal heuristic is not admissible but reasonable estimates can be made yielding good performance. This approach arguably provides the highest possible accuracy assuming that no additional information is available.

In reality, however, additional information *can* usually be made available. The method we propose here exploits one such type of information, namely, that a preprocessing stage can be used to annotate the input source-language sentence with a syntactic bracketing. We will not dwell on the bracketing method here; numerous approaches for automatic bracketing have been developed, including strategies employing full grammars, local patterns, and information-theoretic metrics. Work on Chinese parsing (Jiang 1985; Zhou & Chang 1986; Lum & Pun 1988; Lee & Hsu 1991; Lee *et al.* 1992) would be particularly applicable here.

2 Baseline Translation Model

The translation system employs two main sets of learned parameters corresponding to the two factors on the right side of Equation 1: the *language model* and the *translation model*.

Parameters for the translation model consist of (1) translation probabilities $\Pr(c|e)$ which describe bilingual lexical correspondences in terms of the probability that a given English word e translates into a Chinese word c, and (2) alignment probabilities $\Pr(a_j|j, l, m)$ which crudely describe word order variation in terms of the probability that a word in position j of a length-m Chinese sentence corresponds to a word in position a_j of a corresponding length-l English translation. The translation and alignment probabilities are automatically estimated by an iterative expectation-maximization algorithm (Wu & Xia 1995), using as training data a parallel bilingual corpus containing parliamentary transcripts from the Hong Kong Legislative Council which are available in both English and Chinese versions. The size of the training corpus was approximately 17.9Mb of raw English text and 9.6Mb of corresponding raw Chinese translation, or about 3 million English words, and approximately 3.2 million Chinese words (under certain Chinese segmentation assumptions). Since these proceedings were not originally available in machine-analyzable form, it was necessary to carry out data conversion and reformatting using manual and automatic processing, and then to perform automatic sentence alignment (Wu 1994).

Parameters for the English language model, on the other hand, were estimated from a much larger monolingual corpus to reduce sparse data problems. About 280Mb of text from the Wall Street Journal were used to to obtain a bigram model with the parameters are $\Pr(e_i|e_{i-1})$, under a vocabulary restriction to match the translation lexicon.

Given the parameters, translation of a test sentence in Chinese is performed by a search to solve Equation 1. In our baseline system, we employ a beam search algorithm, a variation of A* with a thresholded agenda width.

3 Incorporating Bracketing Constraints

In the baseline model, the coupling between words of the test sentence is ignored. The search process considers each of the input tokens as an individual word. In reality, however, often there exist known relations between individual words, as for example in (一些 病人 要求 安排 宿舍 。), where 一些 病人 is a noun phrase in which 一些 is a measuring element to describe 病人. Thus we would not expect the translations of these two tokens to be separated far apart in the target output. Again, in (我們 必須 照顧 自己 的 利益 。), we consider (自己 的 利益) a phrase to be translated as a unit.

The search strategy we propose accepts any available bracketing information, full or partial. The bracketing information is used to partition the search in divide-and-conquer fashion. Innermost constituents are translated first, then assembled compositionally into larger constituents. Within any level of bracketing, an $A*$ search is performed. The merits of the bracket-guided search strategy can be summarized as follows:

1. *Use of divide-and-conquer.* The problem of finding a complete English translation is recursively decomposed into sub-problems of finding translations of substrings.

2. *Independence of syntactic knowledge.* While it is true that the bracketing preprocessor may utilize syntactic knowledge, such knowledge is not used by the search algorithm itself. Moreover, the brackets do not carry syntactic category labels. Thus if alternative non-syntactic (e.g., statistical) bracketing strategies are available, the proposed algorithm can be deployed *without* any grammar.

3. *Preservation of robustness.* The spirit of the statistical approach with respect to robustness is preserved. At one extreme, given a complete bracketing of the input sentence, the solution of the sub-problems immediately yields the solution to the original problem. At the other extreme, if no brackets are given (or equivalently, each individual input token is bracketed by itself), the algorithm simply degenerates into the baseline model. In between the extremes, the search is guided heuristically as in the baseline model.

Our search algorithm dictates that nodes in the lower levels (those with higher level numbers) of the tree of $\mathbf{c}$ must be processed before nodes in the higher levels. In Figure 1, we have five subtrees labeled S_1, S_2, S_3, S_4, and S (which is the whole sentence). subtree S_4 is processed first, followed arbitrarily by S_1, S_2, or S_3. If we assume the subtrees S_1 and then

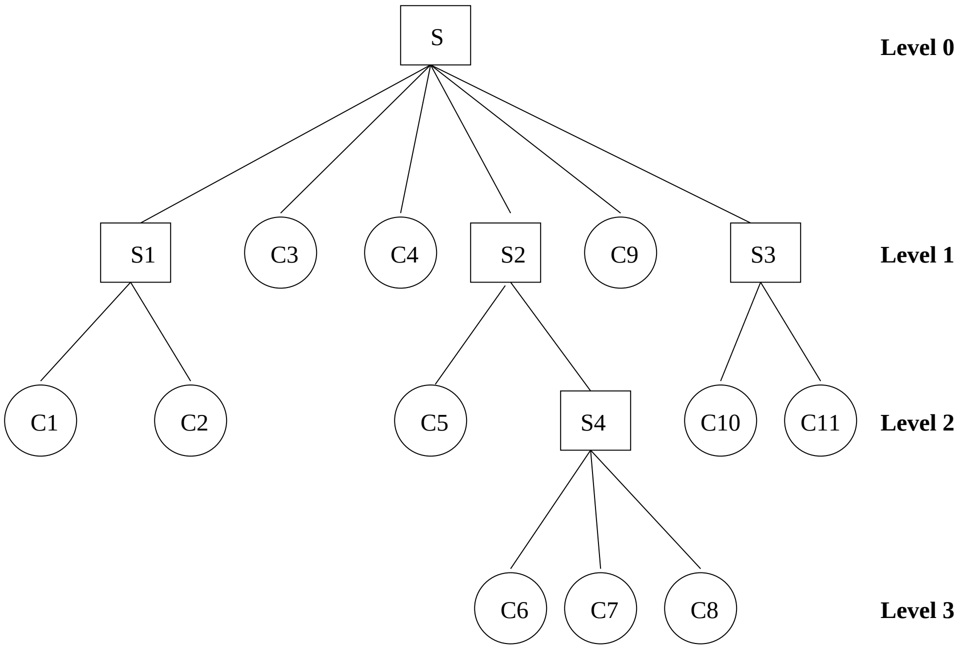

Figure 1: Example bracket structure of a test sentence **c**.

S_3 are processed next, the intermediate result will be as shown in Figure 2, where P_1, P_2, and P_3 hold English substrings. Thus at any point during the search, a subtree may consist of:

1. *Chinese tokens only.* In this case, the sub-search is identical to that in the baseline system.

2. *English substrings only.* All lexical translations have been made; it may still remain to align the English substrings.

3. *A mixture of Chinese tokens and English substrings.* This is analogous to a partial hypothesis in the baseline model where some of the English words have been translated. As above, the English substrings may still need to be aligned. In addition the Chinese tokens must still be translated and aligned. We impose an additional assumption: the available English substrings are aligned prior to continuing the search on Chinese token translations.

The search algorithm follows the general schema below:

- While unprocessed nodes in the Chinese tree remain, choose an unprocessed subtree S_i at the deepest remaining level, and replace S_i with its translation computed as follows:

    1. Create hypothesis nodes in the search tree representing alternative target lengths l for the output English phrases $\mathbf{P}$ that might be translations of S_i.

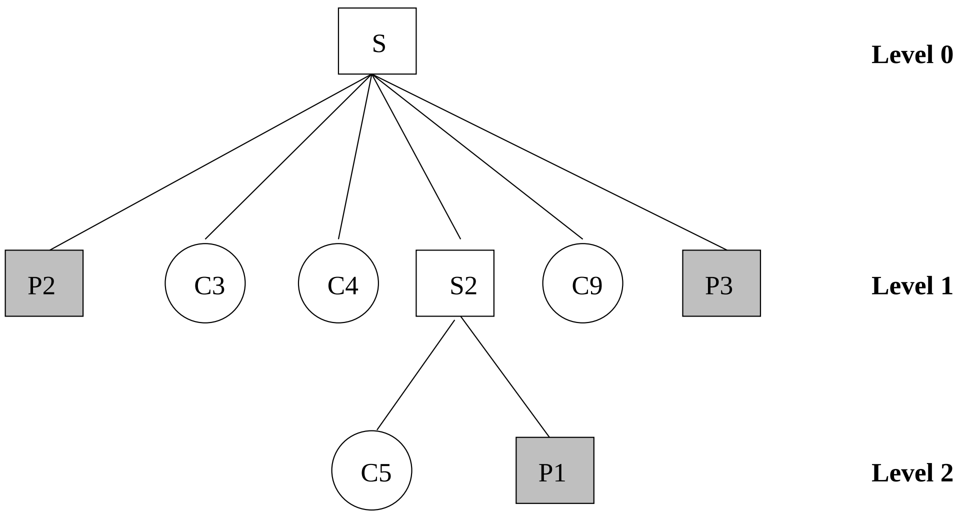

Figure 2: Bracket structure of an intermediate sentence translation hypothesis, where subtrees S1, S3, and S4 of Figure 1 have been processed.

2. Arrange the search order of any previously computed English substrings under S_i according to their length-normalized joint probability $g = \Pr(\mathbf{e}) \Pr(\mathbf{c}, \mathbf{a}|\mathbf{e})$.
3. While any previously computed English substrings of the subtree remain to be processed:
   (a) Let $\mathbf{p}^*$ be the remaining English substring with largest value of g. Expand the hypothesis space to include the set of hypotheses that include $\mathbf{p}^*$ (each hypothesis corresponds to mapping $\mathbf{p}^*$ to a different location in $\mathbf{P}$). Calculate $\tilde{w}$ for each hypothesis.
4. (At this point the subtree consists of Chinese tokens only.) Initialize a set of hypotheses using the translation probabilities: for each Chinese word c_j in S_i, find all English words e such that $\Pr(c_j|e)$ is non-zero. Arrange their search order according to their $\Pr(c_j|e)$ value.
5. While any Chinese tokens remain to be processed:
   (a) Expand the hypothesis with the maximum remaining $\Pr(c_j|e)$ value. Generate subhypotheses that associate alternative positions a_j for the English word e. Calculate $\tilde{w}$ for each hypothesis.
6. (At this point all Chinese in the subtree has been eliminated.) For each hypothesis:
   (a) While empty positions in the output string remain:
      i. Fill in the empty positions using the bigram probabilities $\Pr(e_i|e_{i-1})$ from the language model, and calculate $\tilde{w}$.

4 Experiments

We have tested our model with both natural test cases (from the Hong Kong Hansard) as well as synthetic ones. The synthetic cases are artifically constructed using the natural corpus vocabulary. Only noun phrases and verb phrases were bracketed, using the following simple pattern templates:

- *NP*.

  1. two consecutive nouns, e.g. 中國 人民; or
  2. an adjective + a noun, e.g. 實際 撥款; or
  3. two nouns with the word 的 in between, e.g. 中國 的 利益; or
  4. an adjective + a NP, e.g. 實際 撥款 日期; or
  5. two NPs with the word 的 in between, e.g. 一些 問題 的 主要 原因

  In addition, each of the above NP forms allows insertion of a *measuring phrase* of the form "(specifier) + (number) + (unit)" where the parentheses denote optionality.

- *VP*.

  1. a verb + a noun, e.g. 安排 宿舍; or
  2. a verb + a NP, e.g. 解決 青年 的 問題

As a measure of efficiency, the average number of nodes in the search tree for each strategy was recorded. Table 1 shows the average number of nodes in the search tree expanded per test case for both the baseline and bracketing strategies, with a significant reduction in the search cost. Two example test sentences are shown in the Appendix. For both the cases with and without bracketing on each test sentence input, the top five output candidate translations are shown, along with their log probabilities.

Corpus Test Cases			**Synthetic Test Cases**		
Baseline	Bracketing	% reduction	Baseline	Bracketing	% reduction
443819	309860	30.2	434351	346702	20.2

Table 1: Average number of nodes in the search tree per bracketed test case

In addition to improving efficiency, the bracketing strategy simultaneously achieves higher accuracy as summarized in the tables below. The correctness criteria for the two sets of test cases are a bit different, as the outputs from the synthetic set do not have any reference translations to serve as an evaluation standard.

For the natural test cases from the corpus, a translation is considered:

1. *Correct* if it is exactly the same as the translation made in the bilingual corpus, or conveys the same meaning as that in the bilingual corpus;
2. *Partially Correct* if it conveys more or less the same meaning as that in the bilingual corpus and is grammatically incorrect;
3. *Not Correct* otherwise.

Category	**Baseline**	**Percent**	**Bracketing**	**Percent**
Correct	11	25.6	16	37.2
Partial Correct	18	41.9	14	32.6
Not Correct	14	32.5	13	30.2
Total	43	100.0	43	100.0

Table 2: Results with test cases from corpus

For the synthetic test cases, a translation is considered:

1. *Correct* if it is an acceptable translation judged by a human evaluator;
2. *Partially Correct* if it conveys part of the meaning of the original sentence;
3. *Not Correct* otherwise.

Category	**Baseline**	**Percent**	**Bracketing**	**Percent**
Correct	10	25.7	13	33.3
Partial Correct	21	53.8	20	51.3
Not Correct	8	20.5	6	15.4
Total	39	100.0	39	100.0

Table 3: Results with synthetic test cases

5 Conclusion

In most systems only partial bracketing information will be available since full-coverage grammars are not robust. The degree of bracketing affects performance as follows. A minimally-bracketed sentence, where there is only one pair of brackets enclosing the entire sentence, reduces to the original $A*$ search. On the other hand, a fully-bracketed sentence offers the least room for variation in the translation hypotheses, and dictates clausal translation at every level of the phrase structure. Thus speed will be maximally enhanced, but robustness will be minimized. Because of these properties, it is best to bias the bracketer conservatively, i.e., to commit to a pair of brackets only when certain.

This study underlines the effectiveness of combining linguistic analysis with statistical corpus-based techniques for practical applications such as machine translation. A conservative use of linguistic analysis improves both speed and accuracy, while maintaining the robustness and broad coverage of statistical methods.

References

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.

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.

Jiang, Y.P. 1985. Chinese Parsing: An Initial Exploration at LRC. *Computer Processing of Chinese and Oriental Languages*, 2(2):127–138.

Lee, H.J., J.C. Dai, & Y.S. Chang. 1992. Parsing Chinese Nominalizations based on HPSG. *Computer Processing of Chinese and Oriental Languages*, 6(2):143–158.

Lee, H.J. & P.R. Hsu. 1991. Parsing Chinese Sentences in a Unification-based Grammar. *Computer Processing of Chinese and Oriental Languages*, 5(3-4):271–284.

Lum, B. & K.H. Pun. 1988. On Parsing Complex Noun Phrases in a Chinese Sentence. In *1988 International Conference on Computer Processing of Chinese and Oriental Languages. Proceedings*, 470–474.

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. An algorithm for simultaneously bracketing parallel texts by aligning words. In *Proceedings of the 33rd Annual Conference of the Association for Computational Linguistics*, 244–251, Cambridge, Massachusetts.

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. 1995c. Trainable coarse bilingual grammars for parallel text bracketing. In *Proceedings of the Third Annual Workshop on Very Large Corpora*, 69–81, Cambridge, Massachusetts.

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. 1995. Large-scale automatic extraction of an English-Chinese lexicon. *Machine Translation*. To appear.

ZHOU, J.Y. & S.K. CHANG. 1986. A Methodology for Deterministic Chinese Parsing. *Computer Processing of Chinese and Oriental Languages*, 2(3):139–161.

Appendix

Sentence 1, unbracketed: (主席) (先生) (,) (我) (有) (這些) (數字) (。)

1. -1.95992 Sir(先生) ,(,) I(我))(NULL) These(這些) figures(數字) exist(有) .(。)
2. -2.03140 Sir(先生) ,(NULL) my(我) remarks(NULL) ,(,) were(有) these(這些) figures(數字) .(。)
3. -2.05858 Sir(先生) ,(NULL) my(我) remarks(NULL) ,(,) there(有) such(這些) figures(數字) .(。)
4. -2.06033 Sir(先生) ,(,) my(我) main(NULL) duty(有) </s>(NULL) These(這些) figures(數字) .(。)
5. -2.06463 Sir(先生) ,(,) my(我) colleagues(NULL) have(有) .(NULL) These(這些) figures(數字) .(。)

Sentence 1, bracketed: ((主席) (先生)) (,) (我) ((有) ((這些) (數字))) (。)

1. -1.40294 Mr(先生) Deputy(NULL) President(主席) ,(,) I(我) have(有) these(這些) figures(數字) .(。)
2. -1.53207 Sir(先生) ,(,) these(這些) figures(數字) I(我) have(有) .(。)
3. -1.63287 Mr(先生) Deputy(NULL) President(主席) I(我) have(有) these(這些) figures(數字) .(。)
4. -1.63503 Sir(主席) I(我) have(有) these(這些) figures(數字) .(。)
5. -1.68973 Mr(先生) Deputy(NULL) President(主席) I(我) have(,) these(這些) figures(數字) .(。)

Sentence 2, unbracketed: (環境保護) (並) (不) (等於) (清潔) (運動) (。)

1. -6.61285 tantamount(等於) no(並) need(NULL) clean(清潔) and(NULL) environmental(環境保護) campaign(運動) .(。)
2. -6.61400 tantamount(等於) clean(清潔) and(NULL) not(並) contravene(NULL) environmental(環境保護) campaign(運動) .(。)
3. -6.97496 does(並) not(不) to(NULL) clean(清潔) tantamount(等於) environmental(環境保護) campaign(運動) .(。)
4. -7.18585 does(並) no(不) need(NULL) clean(清潔) tantamount(等於) environmental(環境保護) campaign(運動) .(。)
5. -7.20299 no(並) clean(清潔) and(NULL) not(不) tantamount(等於) environmental(環境保護) campaign(運動) .(。)

Sentence 2, bracketed: (環境保護) (並) (不) ((等於) ((清潔) (運動))) (。)

1. -4.38424 EPD(環境保護) should(NULL) not(不) tantamount(等於) to(NULL) clean(清潔) campaigns(運動) .(。)
2. -4.40295 EPD(環境保護) should(NULL) not(不) tantamount(等於) to(NULL) clean(清潔) campaign(運動) .(。)
3. -4.41738 Protection(環境保護) is(NULL) not(不) tantamount(等於) to(NULL) clean(清潔) campaigns(運動) .(。)
4. -4.42583 EPD(環境保護) should(NULL) not(不) tantamount(等於) clean(清潔) up(NULL) campaigns(運動) .(。)
5. -4.43609 Protection(環境保護) is(NULL) not(不) tantamount(等於) to(NULL) clean(清潔) campaign(運動) .(。)