Inversion Transduction Grammar Constraints for Mining Parallel Sentences from Quasi-Comparable Corpora*

Dekai Wu[1] and Pascale Fung[2]

[1] Human Language Technology Center, HKUST,
Department of Computer Science
[2] Department of Electrical and Electronic Engineering,
University of Science and Technology, Clear Water Bay, Hong Kong
dekai@cs.ust.hk, pascale@ee.ust.hk

Abstract. We present a new implication of Wu's (1997) Inversion Transduction Grammar (ITG) Hypothesis, on the problem of retrieving truly parallel sentence translations from large collections of highly *non*-parallel documents. Our approach leverages a strong language universal constraint posited by the ITG Hypothesis, that can serve as a strong inductive bias for various language learning problems, resulting in both efficiency and accuracy gains. The task we attack is highly practical since non-parallel multilingual data exists in far greater quantities than parallel corpora, but parallel sentences are a much more useful resource. Our aim here is to mine truly parallel sentences, as opposed to comparable sentence pairs or loose translations as in most previous work. The method we introduce exploits Bracketing ITGs to produce the first known results for this problem. Experiments show that it obtains large accuracy gains on this task compared to the expected performance of state-of-the-art models that were developed for the less stringent task of mining comparable sentence pairs.

1 Introduction

Parallel sentences are a relatively scarce but extremely useful resource for many applications including cross-lingual retrieval and statistical machine translation. Parallel sentences, or *bi-sentences* for short, can be exploited for a wealth of applications ranging from mining term translations for cross-lingual applications, to training paraphrase models and inducing structured terms for indexing, query processing, and retrieval.

Unfortunately, far more is available in the way of monolingual data. High-quality parallel corpora are currently largely limited to specialized collections of government (especially UN) and certain newswire collections, and even then relatively few bi-sentences are available in tight sentence-by-sentence translation.

* This work was supported in part by the Hong Kong Research Grants Council through grants RGC6083/99E, RGC6256/00E, DAG03/04.EG09, and RGC6206/03E.

R. Dale et al. (Eds.): IJCNLP 2005, LNAI 3651, pp. 257–268, 2005.

Increasingly sophisticated methods for extracting loose translations from non-parallel monolingual corpora—and in particular, what have been called *comparable sentence pairs*—have also recently become available. But while loose translations by themselves already have numerous applications, truly parallel sentence translations provide invaluable types of information for the aforementioned types of mining and induction, which cannot easily be obtained from merely loose translations or comparable sentence pairs. In particular, truly parallel bi-sentences are especially useful for extracting more precise syntactic and semantic relations within word sequences.

We present a new method that exploits a novel application of *Inversion Transduction Grammar* or *ITG* expressiveness constraints (Wu 1995 [1], Wu 1997 [2]) for mining monolingual data to obtain tight sentence translation pairs, yielding accuracy significantly higher than previous known methods. We focus here on very non-parallel *quasi-comparable* monolingual corpora, which are available in far larger quantities but are significantly more difficult to mine than either noisy parallel corpora or comparable corpora. The majority of previous work has concerned *noisy parallel corpora* (sometimes imprecisely also called "comparable corpora"), which contain non-aligned sentences that are nevertheless mostly bilingual translations of the same document. More recent work has examined *comparable corpora*, which contain non-sentence-aligned, non-translated bilingual documents that are topic-aligned. Still relatively few methods attempt to mine quasi-comparable corpora, which contain far more heterogeneous, very non-parallel bilingual documents that could be either on the same topic (in-topic) or not (off-topic).

Our approach is motivated by a number of desirable characteristics of ITGs, which historically were developed for translation and alignment purposes, rather than mining applications of the kind discussed in this paper. The ITG Hypothesis posits a strong language universal constraint that can act as a strong inductive bias for various language learning problems, resulting in both efficiency and accuracy gains. Specifically, the hypothesis asserts that sentence translation between any two natural languages can be accomplished within ITG expressiveness (subject to certain conditions). So-called *Bracketing ITGs* (BITG) are particularly interesting in certain applications such as the problem we consider here, because they impose ITG constraints in language-independent fashion, and do not require any language-specific linguistic grammar. (As discussed below, Bracketing ITGs are the simplest form of ITGs, where the grammar uses only a single, undifferentiated non-terminal.)

The key modeling property of bracketing ITGs that is most relevant to the task of identifying parallel bi-sentences is that they assign strong preference to candidate sentence pairs in which nested constituent subtrees can be recursively aligned with a minimum of constituent boundary violations. Unlike language-specific linguistic approaches, however, the shape of the trees are driven in unsupervised fashion by the data. One way to view this is that the trees are hidden explanatory variables. This not only provides significantly higher robustness than more highly constrained manually constructed grammars, but also makes

the model widely applicable across languages in economical fashion without a large investment in manually constructed resources.

Moreover, for reasons discussed by Wu [2], ITGs possess an interesting intrinsic combinatorial property of permitting roughly up to four arguments of any frame to be transposed freely, but not more. This matches suprisingly closely the preponderance of linguistic verb frame theories from diverse linguistic traditions that all allow up to four arguments per frame. Again, this property falls naturally out of ITGs in language-independent fashion, without any hardcoded language-specific knowledge. This further suggests that ITGs should do well at picking out translation pairs where the order of up to four arguments per frame may vary freely between the two languages. Conversely, ITGs should do well at rejecting candidates where (1) too many words in one sentence find no correspondence in the other, (2) frames do not nest in similar ways in the candidate sentence pair, or (3) too many arguments must be transposed to achieve an alignment—all of which would suggest that the sentences probably express different ideas.

Various forms of empirical confirmation for the ITG Hypothesis have emerged recently, which quantitatively support the qualitative cross-linguistic characteristics just described across a variety of language pairs and tasks. Zens and Ney (2003) [3] show that ITG constraints yield significantly better alignment coverage than the constraints used in IBM statistical machine translation models on both German-English (Verbmobil corpus) and French-English (Canadian Hansards corpus). Zhang and Gildea (2004) [4] found that unsupervised alignment using Bracketing ITGs produces significantly lower Chinese-English alignment error rates than a syntactically supervised tree-to-string model [5]. Zhang and Gildea (2005) [6] show that lexicalized ITGs can further improve alignment accuracy. With regard to translation rather than alignment accuracy, Zens *et al.* (2004) [7] show that decoding under ITG constraints yields significantly lower word error rates and BLEU scores than the IBM constraints. Chiang (2005) [8] obtains significant BLEU score improvements via unsupervised induction of hierarchical phrasal bracketing ITGs. Such results partly motivate the work we discuss here.

We will begin by surveying recent related work and reviewing the formal properties of ITGs. Subsequently we describe the architecture of our new method, which relies on multiple stages so as to balance efficiency and accuracy considerations. Finally we discuss experimental results on a quasi-comparable corpus of Chinese and English from the topic detection task.

2 Recent Approaches to Mining Non-parallel Corpora

Recent work (Fung and Cheung 2004 [9]; Munteanu *et al.* 2004 [10]; Zhao and Vogel 2002 [11]) on extracting bi-sentences from comparable corpora is largely based on finding on-topic documents first through similarity matching and time alignment.

However, Zhao and Vogel used a corpus of Chinese and English versions of news stories from the Xinhua News agency, with "roughly similar sentence order

of content". This corpus can be more accurately described as a noisy parallel corpus. Munteanu *et al.* used comparable corpora of news articles published within the same 5-day window. In both cases, the corpora contain documents on the same matching topics; unlike our present objective of mining quasi-comparable corpora, these other methods assume corpora of on-topic documents.

Munteanu *et al.* first identify on-topic document pairs by looking at publication date and word overlap, then classify all sentence pairs as being parallel or not parallel, using a maximum entropy classifier trained on parallel corpora. In contrast, the method we will propose identifies candidate sentence pairs without assuming that publication date information is available, and then uses the ITG constraints to automatically find parallel sentence pairs without requiring any training.

It is also difficult to relate Munteanu *et al.*'s work to our present objective because they do not directly evaluate the quality of the extracted bi-sentences (they instead look at performance of their machine translation application); however, as with Fung and Cheung, they noted that the sentences extracted were not truly parallel on the whole.

In this work, we aim to find parallel sentences from much more heterogenous, very non-parallel quasi-comparable corpora. Since many more multilingual text collections available today contain documents that do not match documents in the other language, we propose finding more parallel sentences from off-topic documents, as well as on-topic documents. An example is the TDT corpus, which is an aggregation of multiple news sources from different time periods.

3 Inversion Transduction Grammars

Formally, within the expressiveness hierarchy of transduction grammars, the ITG level of expressiveness has highly unusual intrinsic properties as seen in Figure 1. Wu [2] showed that the ITG class is an equivalence class of subsets of syntax-directed transduction grammars or SDTGs (Lewis and Stearns 1968 [12]), equivalently defined by meeting any of the following three conditions: (1) all rules are of rank 2, (2) all rules are of rank 3, or (3) all rules are either of *straight* or *inverted* orientation (and may have *any* rank). Ordinary unrestricted SDTGs allow any permutation of the symbols on the right-hand side to be specified when translating from the input language to the output language. In contrast, ITGs only allow two out of the possible permutations. If a rule is straight, the order of its right-hand symbols must be the same for both languages (just as in a *simple SDTG* or *SSDTG*). On the other hand, if a rule is inverted, then the order is left-to-right for the input language and right-to-left for the output language. Since inversion is permitted at any level of rule expansion, a derivation may intermix productions of either orientation within the parse tree. The ability to compose multiple levels of straight and inverted constituents gives ITGs much greater expressiveness than might seem at first blush, as indicated by the growing body of empirical results mentioned earlier.

A simple example may be useful to fix ideas. Consider the following pair of parse trees for sentence translations:

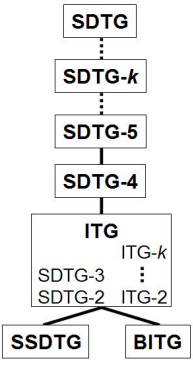

Fig. 1. The ITG level of expressiveness constitutes a surprisingly broad equivalence class within the expressiveness hierarchy of transduction grammars. The simple monolingual notion of "context-free" is too coarse to adequately categorize the bilingual case of transduction grammars. The expressiveness of a transduction grammar depends on the maximum rank k of rules, i.e., the maximum number of nonterminals on the right-hand-side. SDTG-k is always more expressive than SDTG-(k-1), except for the special case of the ITG class which includes both SDTG-2 and SDTG-3. In contrast, for monolingual CFGs, expressiveness is not affected by rank, as shown by the existence of a binary Chomsky normal form for any CFG. A binary normal form exists for ITGs but not SDTGs.

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

[[[管理局]$_{\text{NP}}$ [将会 [[向 [[[财政 司]$_{\text{NN}}$]$_{\text{NNN}}$]$_{\text{NP}}$]$_{\text{PP}}$ [负责]$_{\text{VV}}$]$_{\text{VP}}$]$_{\text{VP}}$]$_{\text{SP}}$ 。]$_{\text{S}}$

Even though the order of constituents under the inner VP is inverted between the languages, an ITG can capture the common structure of the two sentences. This is compactly shown by writing the parse tree together for both sentences with the aid of an $\langle\rangle$ angle bracket notation marking parse tree nodes that instantiate rules of inverted orientation:

[[[The/εAuthority/管理局]$_{\text{NP}}$ [will/将会 ⟨[be/εaccountable/负责]$_{\text{VV}}$ [to/向 [the/ε [[Financial/财政Secretary/司]$_{\text{NN}}$]$_{\text{NNN}}$]$_{\text{NP}}$]$_{\text{PP}}$ ⟩$_{\text{VP}}$]$_{\text{VP}}$]$_{\text{SP}}$./。]$_{\text{S}}$

In a weighted or stochastic ITG (SITG), a weight or a probability is associated with each rewrite rule. Following the standard convention, we use a and b to denote probabilities for syntactic and lexical rules, respectively. For example, the probability of the rule NN $\stackrel{0.4}{\rightarrow}$ [A N] is $a_{\text{NN}\rightarrow[\text{A N}]} = 0.4$. The probability of a lexical rule A $\stackrel{0.001}{\rightarrow}$ x/y is $b_{\text{A}}(x, y) = 0.001$. Let W_1, W_2 be the vocabulary sizes of the two languages, and $\mathcal{N} = \{A_1, \ldots, A_N\}$ be the set of nonterminals with indices $1, \ldots, N$.

Polynomial-time algorithms are possible for various tasks including translation using ITGs, as well as bilingual parsing or *biparsing*, where the task is to build the highest-scored parse tree given an input bi-sentence.

For present purposes we can employ the special case of Bracketing ITGs, where the grammar employs only one single, undistinguished "dummy" nonterminal category for any non-lexical rule. Designating this category A, a Bracketing ITG has the following form (where, as usual, lexical transductions of the form $A \rightarrow e/f$ may possibly be singletons of the form $A \rightarrow e/\epsilon$ or $A \rightarrow \epsilon/f$).

$$\begin{aligned} A &\rightarrow [AA] \\ A &\rightarrow \langle AA \rangle \\ A &\rightarrow \epsilon, \epsilon \\ A &\rightarrow e_1/f_1 \\ &\cdots \\ A &\rightarrow e_{\mathrm{i}}/f_{\mathrm{j}} \end{aligned}$$

Broadly speaking, Bracketing ITGs are useful when we wish to make use of the structural properties of ITGs discussed above, without requiring any additional linguistic information as constraints. Since they lack differentiated syntactic categories, Bracketing ITGs merely constrain the *shape* of the trees that align various nested portions of a sentence pair. The only linguistic knowledge used in Bracketing ITGs is the purely lexical set of collocation translations. Nevertheless, the ITG Hypothesis implies that biparsing truly parallel sentence pairs with a Bracketing ITG should typically yield high scores. Conversely, some non-parallel sentence pairs could be ITG-alignable, but any significant departure violating constituent boundaries will be downgraded.

As an illustrative example, in the models employed by most previous work on mining bi-sentences from non-parallel corpora, the following pair of sentences (found in actual data arising in our experiments below) would receive an inappropriately high score, because of the high lexical similarity between the two sentences:

Chinese president Jiang Zemin arrived in Japan today for a landmark state visit .
江泽民 将 是 到 日本 做 国事访问 的 首 位 中国 国家 主席 。
(*Jiang Zemin will be the first Chinese national president to pay a state vist to Japan.*)

However, the ITG based model is sensitive enough to the differences in the constituent structure (reflecting underlying differences in the predicate argument structure) so that our experiments show that it assigns a low score. On the other hand, the experiments also show that it successfully assigns a high score to other candidate bi-sentences representing a true Chinese translation of the same English sentence, as well as a true English translation of the same Chinese sentence.

4 Candidate Generation

An extremely large set of pairs of monolingual sentences from the quasi-comparable monolingual corpora will need to be scanned to obtain a useful

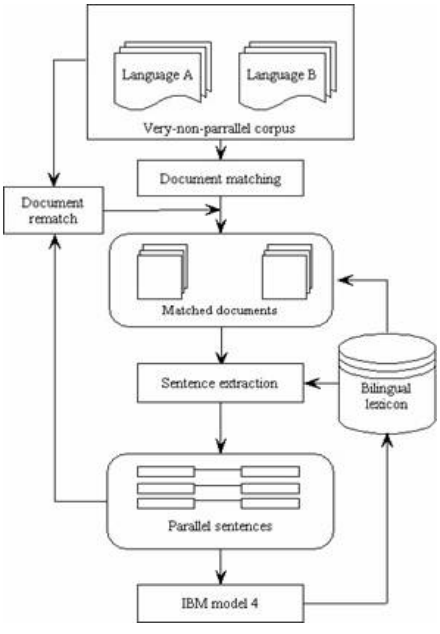

Fig. 2. Candidate generation overview. The iterative bootstrapping algorithm first mines loosely parallel sentence pairs from quasi-comparable corpora that contain both on-topic and off-topic documents. In a preprocessing step, documents that are believed to be on the same topic according to their similarity score are extracted, then "parallel" pairs are mined from these matched documents. The extracted sentences are used to bootstrap the entire process iteratively in two ways: (1) they are used to update a bilingual lexicon, which is then used again to reprocess the documents to be matched again; (2) any document pairs that are found to contain at least one "parallel" sentence pairs are considered to be on-topic, and added to the matched document set. Note that step (2) adds to the on-topic document set certain document pairs that are not considered to be on-topic by document matching scores.

number of parallel sentences, since obviously, the overwhelming majority of the n^2 possible sentence pairs will *not* be parallel. It is infeasible to run the ITG biparsing algorithm on n^2 candidate sentence pairs. Therefore a multi-stage algorithm is needed that first generates likely candidates using faster heuristics, and then biparses the candidates to obtain the final high-precision results.

We base our candidate generation on a method that Fung and Cheung (2004) developed for extracting loose translations (comparable sentence pairs) from quasi-comparable corpora [9], as shown in Figure 2. We selected this model because it produces the highest known accuracy on that task.

Figure 3 outlines the algorithm in greater detail. In the following sections, we describe the document pre-processing step followed by each of the subsequent iterative steps of the algorithm.

1. **Initial document matching**
   For all documents in the comparable corpus D:
   - Gloss Chinese documents using the bilingual lexicon (Bilex)
   - For every pair of glossed Chinese document and English documents:
     - compute *document similarity* => S(i,j)
   - Obtain all matched bilingual document pairs whose S(i,j) > threshold1 => D2
2. **Sentence matching**
   For each document pair in D2:
   - For every pair of glossed Chinese sentence and English sentence:
     - compute *sentence similarity* => S2(i,j)
   - Obtain all matched bilingual sentence pairs whose S2(i,j) > threshold2 => C1
3. **EM learning of new word translations**
   For all bilingual sentences pairs in C1, do:
   - Compute *translation lexicon probabilities* of all bilingual word pairs =>S3(i,j)
   - Obtain all bilingual word pairs previously unseen in Bilex and whose S3(i,j) > threshold3 => L1, and update Bilex
   - Compute *sentence alignment scores* => S4; if S4 does not change then return C1 and L1, otherwise continue
4. **Document re-matching**
   - Find all pairs of glossed Chinese and English documents which contain parallel sentences (anchor sentences) from C1 => D3
   - Expand D2 by finding documents similar to each of the document in D2
   - D2 := D3
5. **Goto 2 if termination criterion not met**

Fig. 3. Candidate generation algorithm

Document preprocessing. The documents are word segmented with the Linguistic Data Consortium (LDC) Chinese-English dictionary 2.0. The Chinese document is then glossed using all the dictionary entries. When a Chinese word has multiple possible translations in English, it is disambiguated using an extension of Fung *et al.*'s (1999) method [13].

Initial document matching. The aim of this step is to roughly match the Chinese-English documents pairs that are on-topic, in order to extract parallel sentences from them. Following previous work, cosine similarity between document vectors is used to judge whether a bilingual document pair is on-topic or off-topic.

Both the glossed Chinese document and English are represented in word vectors, with term weights. Pair-wise similarities are calculated for all possible Chinese-English document pairs, and bilingual documents with similarities above a certain threshold are considered to be comparable. Comparable documents are often on-topic.

Sentence matching. All sentence pair combinations within the on-topic documents are considered next in the selection process. Each sentence is again represented as word vectors. For each extracted document pair, pair-wise cosine similarities are calculated for all possible Chinese-English sentence pairs. Sentence pairs above a set threshold are considered parallel and extracted from the documents. Since cosine similarity is computed on translated word pairs within the sentence pairs, the better our bilingual lexicon is, the more accurate the sentence similarity will be. In the following section, we discuss how to find new word translations.

EM lexical learning from matched sentence pairs. This step updates the bilingual lexicon according to the intermediate results of parallel sentence extraction. New bilingual word pairs are learned from the extracted sentence pairs based on an EM learning method. In our experience any common method can be used for this purpose; for the experiments below we used the GIZA++ [14] implementation of the IBM statistical translation lexicon Model 4 of Brown *et al.* (1993) [15].

This model is based on the conditional probability of a source word being generated by the target word in the other language, based on EM estimation from aligned sentences. Zhao and Vogel (2002) showed that this model lends itself to adaptation and can provide better vocabulary coverage and better sentence alignment probability estimation [11]. In our work, we use this model on the intermediate results of parallel sentence extraction, i.e., on a set of aligned sentence pairs that may or may not truly correspond to each other.

We found that sentence pairs with high alignment scores are not necessarily more similar than others. This might be due to the fact that EM estimation at each intermediate step is not reliable, since we only have a small amount of aligned sentences that are truly parallel. The EM learner is therefore weak when applied to bilingual sentences from very non-parallel quasi-comparable corpora.

Document re-matching. This step implements a "find-one-get-more" principle, by augmenting the earlier matched documents with document pairs that are found to contain at least one parallel sentence pair. We further find other documents that are similar to each of the monolingual documents found. The algorithm then iterates to refine document matching and parallel sentence extraction.

Convergence. The IBM model parameters, including sentence alignment score and word alignment scores, are computed in each iteration. The parameter values eventually stay unchanged and the set of extracted bi-sentence candidates also converges to a fixed size. The iteration then terminates and returns the last set of bilingual sentence pairs as the generated candidate sentences.

5 ITG Scoring

The ITG model computes scores upon the set of candidates generated in the preceding stage. A variant of the approach used by Leusch *et al.* (2003) [16] allows us to forego training to estimate true probabilities; instead, rules are simply given unit weights. This allows the scores computed by ITG biparsing to be interpreted as a generalization of classical Levenshtein string edit distance, where inverted block transpositions are also allowed. Even without probability estimation, Leusch *et al.* found excellent correlation with human judgment of similarity between translated paraphrases.

As mentioned earlier, biparsing for ITGs can be accomplished efficiently in polynomial time, rather than the exponential time required for classical SDTGs. The biparsing algorithm employs a dynamic programming approach described by Wu [2]. The time complexity of the algorithm in the general case is $\Theta\left(T^3V^3\right)$ where T and V are the lengths of the two sentences. This is a factor of V^3 more

than monolingual chart parsing, but has turned out to remain quite practical for corpus analysis, where parsing need not be real-time.

6 Experiments

Method. For our experiments we extracted the bi-sentences from a very non-parallel, quasi-comparable corpus of TDT3 data which consists of transcriptions of news stories from radio and TV broadcasts in both English and Chinese channels during the period 1998-2000. This corpus contained approximately 290,000 English sentences and 110,000 Chinese sentences. This yields over 30 billion possible sentence pairs, so a multi-stage approach is clearly necessary.

Experience showed that the lexicon learned in the candidate generation stage, while adequate for candidate generation, is not of sufficient quality for biparsing due to the non-parallel nature of the training data. However, any translation lexicon of reasonable accuracy can be used. For these experiments we employed the LDC Chinese-English dictionary 2.0.

To conduct as blind an evaluation as possible, an independent annotator separately produced gold standard labels for a random sample of approximately 300 of the top 2,500 candidate sentence pairs proposed by the generation stage. The annotator was instructed to accept any semantically equivalent translations, including non-literal ones. Inspection had shown that sentence pair candidates longer than about 15 words were practically never truly parallel translations, so these were a priori excluded by the sampling in order to ensure that precision/recall scores would be more meaningful.

Results. Under our method any desired tradeoff between precision and recall can be obtained. Therefore, rather than arbitrarily setting a threshold, we are interested in evaluation metrics that can show whether the ITG model is highly effective at any desired tradeoff points. Thus, we assess the contribution of ITG ranking by computing standard uninterpolated average precision scores used to evaluate the effectiveness of ranking methods. Specifically, in this case, this is the expected value of precision over the rank positions of the correctly identified truly parallel bi-sentences:

$$\text{uninterpolated average precision} = \frac{1}{\mid T \mid} \sum_{i \in T} \text{precision at rank}\,(i) \qquad (1)$$

where T is the set of correctly identified bi-sentences.

Our method yielded an uninterpolated average precision of 64.7%. No direct comparison of this figure is possible since previous work has focused on the rather different objectives of mining noisy parallel or comparable corpora to extract comparable sentence pairs and loose translations. However, we can understand the improvement by comparing against scores obtained using the cosine-based lexical similarity metric which is typical of the majority of previous methods for mining non-parallel corpora, including that of Fung and Cheung (2004)[9]. Evaluating the ranking produced under this more typical score yielded

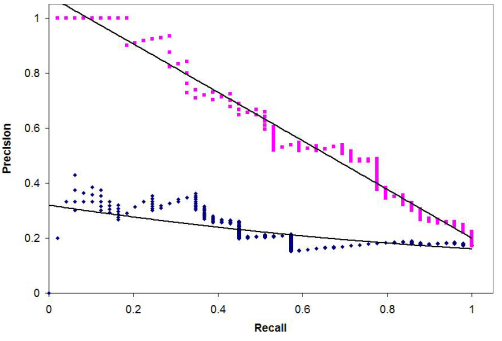

Fig. 4. Precision-recall curves for the ITG model (upper curve) versus traditional cosine model (lower curve); see text

an uninterpolated average precision of 24.6%. This suggests that the ITG based method could produce significant accuracy gains if applied to many of the existing non-parallel corpus mining methods.

Figure 4 compares precision versus recall curves obtained with rankings from the ITG model compared with the more traditional cosine lexical similarity model. The graph reveals that at all levels, much higher precision can be obtained using the ITG model. Up to 20% recall, the ITG ranking produces bi-sentences with perfect precision; in contrast, the cosine model produces 30% precision. Even at 50% recall, the ITG ranked bi-sentences have above 65% precision, as compared with 21% for the cosine model.

As can be seen from the following examples of extracted bi-sentences (shown with rough word glosses), the ITG constraints are able to accommodate nested inversions accounting for the cross-linguistic differences in constituent order:

It is time to break the silence.
现在 呢 , 是 打破 沉默 的 时候 了 。
(*Now* topical , *is break silence* genitive *time* aspectual .)

I think that's what people were saying tonight.
我 认为 这 是 人们 今晚 所 说 的话 。
(*I think this is people today by say* genitive *words* .)

If the suspects are convicted, they will serve their time in Scotland.
如果 两 名 嫌疑 人 被 判 有罪 , 就 得 在 苏格兰 服刑 。
(*If two* classifier *suspected person* bei-particle *sentence guilty, then must in Scotland serve time* .)

7 Conclusion

We have introduced a new method that exploits generic bracketing Inversion Transduction Grammars giving the first known results for the new task of mining truly parallel sentences from very non-parallel quasi-comparable corpora.

The method takes the strong language universal constraint posited by the ITG Hypothesis as an inductive bias on the bi-sentence extraction task which we anticipate will become a key stage in unsupervised learning for numerous more specific models. Experiments show that the method obtains large accuracy gains on this task compared to the performance that could be expected if state-of-the-art models for the less stringent task of mining comparable sentence pairs were applied to this task instead. From a practical standpoint, the method has the dual advantages of neither requiring expensive training nor requiring language-specific grammatical resources, while producing high accuracy results.

References

1. Wu, D.: An algorithm for simultaneously bracketing parallel texts by aligning words. In: ACL-95, Cambridge, MA (1995)
2. Wu, D.: Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. Computational Linguistics **23** (1997)
3. Zens, R., Ney, H.: A comparative study on reordering constraints in statistical machine translation. In: ACL-03, Sapporo (2003) 192–202
4. Zhang, H., Gildea, D.: Syntax-based alignment: Supervised or unsupervised? In: COLING-04, Geneva (2004)
5. Yamada, K., Knight, K.: A syntax-based statistical translation model. In: ACL-01, Toulouse, France (2001)
6. Zhang, H., Gildea, D.: Stochastic lexicalized inversion transduction grammar for alignment. In: ACL-05, Ann Arbor (2005) 475–482
7. Zens, R., Ney, H., Watanabe, T., Sumita, E.: Reordering constraints for phrase-based statistical machine translation. In: COLING-04, Geneva (2004)
8. Chiang, D.: A hierarchical phrase-based model for statistical machine translation. In: ACL-05, Ann Arbor (2005) 263–270
9. Fung, P., Cheung, P.: Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and em. In: EMNLP-2004, Barcelona (2004)
10. Munteanu, D.S., Fraser, A., Marcu, D.: Improved machine translation performance via parallel sentence extraction from comparable corpora. In: NAACL-04. (2004)
11. Zhao, B., Vogel, S.: Adaptive parallel sentences mining from web bilingual news collections. In: IEEE Workshop on Data Mining. (2002)
12. Lewis, P.M., Stearns, R.E.: Syntax-directed transduction. Journal of the Association for Computing Machinery **15** (1968) 465–488
13. Fung, P., Liu, X., Cheung, C.S.: Mixed-language query disambiguation. In: ACL-99, Maryland (1999)
14. Och, F.J., Ney, H.: Improved statistical alignment models. In: ACL-2000, Hong Kong (2000)
15. Brown, P.F., DellaPietra, S.A., DellaPietra, V.J., Mercer, R.L.: The mathematics of statistical machine translation. Computational Linguistics **19** (1993) 263–311
16. Leusch, G., Ueffing, N., Ney, H.: A novel string-to-string distance measure with applications to machine translation evaluation. In: MT Summit IX. (2003)