INVERSION TRANSDUCTION GRAMMAR COVERAGE OF ARABIC-ENGLISH WORD ALIGNMENT FOR TREE-STRUCTURED STATISTICAL MACHINE TRANSLATION

Dekai Wu *Marine Carpuat* *Yihai Shen*

Human Language Technology Center
HKUST, Department of Computer Science and Engineering, Hong Kong
{dekai,marine,shenyh}@cs.ust.hk

ABSTRACT

We present the first known direct measurement of word alignment coverage on an Arabic-English parallel corpus using inversion transduction grammar constraints. While direct measurements have been reported for several European and Asian languages, to date no results have been available for Arabic or any Semitic language despite much recent activity on Arabic-English spoken language and text translation. Many recent syntax based statistical MT models operate within the domain of ITG expressiveness, often for efficiency reasons, so it has become important to determine the extent to which the ITG constraint assumption holds. Our results on Arabic provide further evidence that ITG expressiveness appears largely sufficient for core MT models.

Index Terms— Speech communication, Natural language interfaces, Natural languages, Formal languages, Languages, Transducers, Hidden Markov Models

1. INTRODUCTION

Several recent empirical studies have been directed at measuring the extent to which word alignments between parallel training corpora fall within the constrained expressiveness of inversion transduction grammars or ITGs [2, 3, 4, 5, 6]. Measurements on French-English and German-English were reported by Zens and Ney [7]. Additional measurements on a smaller scale using manual alignments for Chinese-English, Romanian-English, Hindi-English, Spanish-English, as well as French-English were reported by Wellington *et al.* [8].

It is important to know whether adequate coverage of word alignment is preserved under ITG constraints, because an increasing number of tree-structured statistical machine translation models for both spoken language and text translation rely on assumptions that formally reduce to ITG expressiveness [9, 10, 11, 12, 13, 14, 15, 16, 17, 18]. Often the appeal of assuming ITG constraints is for the sake of the computational feasibility benefits, since ITG constraints significantly reduce combinatorial complexity and facilitate various polynomial time alignment and decoding algorithms with lower order than other statistical MT approaches. At the same time, ITG constraints tend to improve translation accuracy.

ITG constrained alignments can be utilized either during training for hierarchical phrasal aligned example extraction [19], and/or during testing for translation decoding [4, 20]. Many syntax based statistical MT models employ manually constructed or automatically learned rules/patterns/templates that are similar to those in traditional transfer-based MT models as well as modern example-based MT models—these representations generally being formally equivalent to syntax-directed transduction grammars or SDTGs [21, 22], which have recently also been referred to as synchronous context-free grammars. Unconstrained SDTGs exceed the expressiveness of ITGs, and necessitate much more expensive alignment and/or decoding algorithms. Yet the binarization algorithm of Zhang *et al.* [23] has been applied to several such SDTG-level syntax based statistical MT systems at ISI [24, 25] so as to simplify them from SDTG down to ITG expressiveness (by eliminating any rules that cannot be converted down to a binary rank SDTG, which yields an ITG as discussed below) thus allowing significantly more efficient and thorough search leading to higher BLEU scores. The results of Zens and Ney [7] show that ITG constraints yield significantly better alignment coverage than the constraints used in IBM statistical machine translation models both on German-English (Verbmobil corpus) and on French-English (Canadian Hansards corpus). Zhang and Gildea [26] find that unsupervised alignment using Bracketing ITGs produces significantly lower Chinese-English alignment error rates than a syntactically supervised tree-to-string model (Yamada and Knight [27]). With regard to translation rather than alignment accuracy, Zens *et al.* [28] show that decoding under ITG constraints yields significantly lower word error rates and BLEU scores than the IBM constraints. Chiang [29] and Vilar and Vidal [30] also obtain improved alignment and translation performance by imposing ITG constraints on their models.

This work was supported in part by DARPA GALE contract HR0011-06-C-0023, and by the Hong Kong Research Grants Council (RGC) research grants RGC6083/99E, RGC6256/00E, and DAG03/04.EG09. Thanks to the 2005 Language Engineering Workshop at Johns Hopkins University [1] and Mona Diab, Nizar Habash, Yaser Al-Onaizan, Fatiha Sadat, and Rayan Naal.

Such empirical results, over a range of different languages and task metrics, can be seen as indirect evidence suggesting that the strong restrictive inductive bias imposed by ITG constraints appears to yield both efficiency and accuracy gains.

However, to date no direct measurements of the extent to which Arabic (or any Semitic language) alignments can be covered have been reported, in spite of the large amount of recent effort on Arabic-English machine translation.

In this paper, therefore, we report the first known direct measurements of word alignment coverage under ITG constraints over a large Arabic-English parallel corpus.

2. ITG CONSTRAINTS

For reasons discussed by Wu [5], 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. This property emerges naturally from ITGs in language-independent fashion, without any hardcoded language-specific knowledge.

Formally, within the expressiveness hierarchy of transduction grammars, the ITG level of expressiveness has highly unusual intrinsic properties as seen in Fig. 1. Wu [5] showed that the ITG class is an equivalence class of subsets of syntax-directed transduction grammars or SDTGs (Lewis and Stearns 1968 [21]), 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.

Some examples may be useful to fix ideas. Following Wu [5], in general the leaves of an ITG parse tree are phrase-to-phrase alignments rather than simple "word" alignments. We will omit any constituent categories, since nonterminals are irrelevant to the question of whether ITG parse trees can cover the phrase alignments. Consider the word-aligned English-Arabic sentence translation pair shown in Fig. 2 (using Buckwalter romanization of Arabic). Even though the subject and verb constituents are inverted between the languages—with the *would be set* in English's SVO order versus the *ttHdd* in Arabic's VSO order—an ITG can nevertheless capture the common structure of the two sentences. This is shown by the parse tree in Fig. 2 which generates both the English and

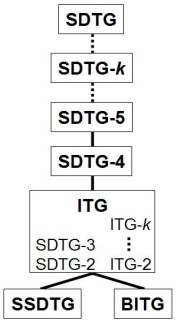

Fig. 1. The ITG level of expressiveness constitutes a surprisingly broad equivalence class within the expressiveness hierarchy of transduction grammars. The simple monolingual notion of "context-free" is too coarse to adequately categorize the bilingual case of transduction grammars. The expressiveness 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.

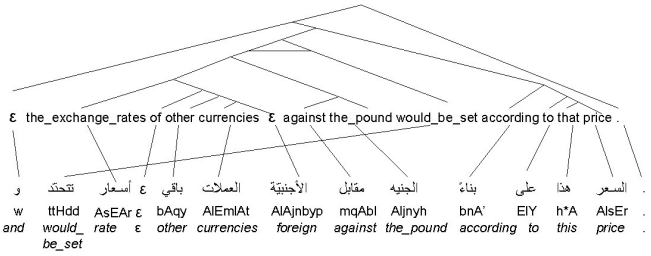

Fig. 2. Example ITG parse tree generating a pair of English and Arabic sentences with the word order distortion (permutation) shown by the alignments. Instances of inverted rules are marked with a horizontal bar.

Arabic sentences. The same parse tree can also be written more compactly with the aid of an $\langle\rangle$ angle bracket notation marking parse tree nodes that instantiate rules of inverted orientation:

> [[ϵ/w [$\langle$[[The_exchange_rates/AsEAr [of/ϵ [other/bAqy [currencies/AlEmlAt ϵ/AlAjnbyp]]]] [against/mqAbl the_pound/Aljnyh]] would_be_set/ttHdd$\rangle$ [according/bnA' [to/ElY [that/h*A price/AlsEr]]]]] .]

As shown in Fig. 3, the ability to compose multiple levels of straight and inverted constituents gives ITGs much greater expressiveness to accommodate messy distortions than might seem at first blush.

On the other hand, (2, 4, 1, 3) permutations of the type in Fig. 4 lie outside the expressiveness of ITG constraints. It is the proportion of such occurrences, which have been shown to be relatively infrequent in other language pairs, that we aim to measure for Arabic-English.

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 take word-aligned Arabic-English parallel sentence pairs and use the algorithm of Wu [5] to attempt to biparse each sentence while obeying the given word alignments, using the special case of Bracketing ITGs [2] 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 lexical transductions of the form $A \rightarrow e/f$ may possibly be singletons of the form $A \rightarrow e/\epsilon$ or $A \rightarrow \epsilon/f$).

$$\begin{array}{lcllcl} A & \rightarrow & [AA] & A & \rightarrow & e_1/f_1 \\ A & \rightarrow & \langle AA \rangle & & \ldots & \\ & & & A & \rightarrow & e_{\mathrm{i}}/f_{\mathrm{j}} \end{array}$$

Since they lack differentiated syntactic categories, Bracketing ITGs merely constrain the *shape* of the trees that align various nested portions of a sentence pair, thereby determining what word order distortions (permutations) can be accommodated. No linguistic knowledge is contained in a Bracketing ITGs other than the purely lexical set of phrase translations.

3. EXPERIMENTAL SETUP

Following the experimental setup of Zens and Ney [7], we use automatically generated IBM Model 4 word alignments of a parallel corpus, and investigate whether these alignments are consistent with the Bracketing ITG constraints.

We test the ITG hypothesis on parallel corpora typically used to train statistical Arabic-English MT systems. We combine news data from the Arabic-English Parallel News and the News Translation corpora, with parallel United Nations data. Our test corpus contains 180,413 sentence pairs.

Before performing word alignment, both sides of the parallel corpus are tokenized. English tokenization is performed using the standard script distributed with the Penn Treebank, which essentially separates words and punctuation. In contrast, Arabic is a morphologically rich language and requires more sophisticated processing.

Tokenization and lemmatization are performed using the ASVMT Arabic morphological analysis toolkit [31]. An Arabic word is typically formed of a stem, and possibly affixes and clitics. Affixes are inflectional markers for tense, gender and/or number, while the clitics include some prepositions, conjunctions, determiners, etc. Tokenization, which consists of separating those syntactic units, is the first step of processing in ASVMT. This is followed by lemmatization which, in ASVMT, refers to a normalization step where the tokens coming from stems that were modified when agglutinated are converted back to their original form.

After preprocessing, all sentence pairs are word aligned. Following Zens and Ney [7], we use GIZA++ [32] to align

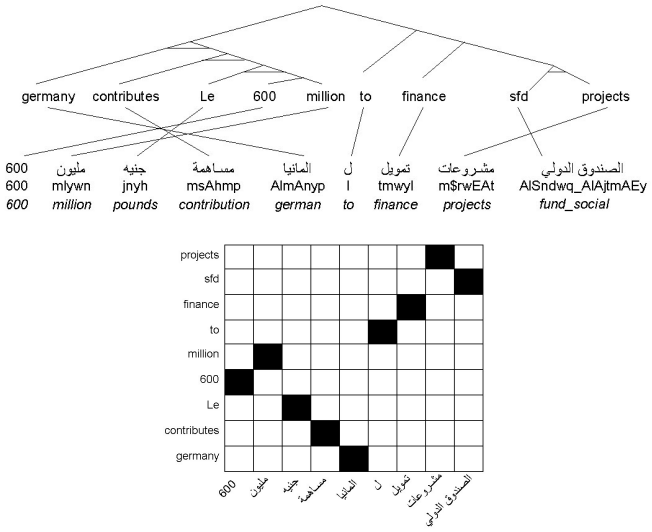

Fig. 3. ITG parse tree generating a more complex permutation via composition of straight and inverted constituents, along with an alternative matrix view of the alignment permutation.

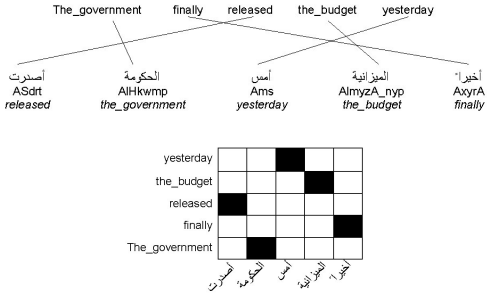

Fig. 4. Example permutation outside ITG expressiveness.

the corpus in both translation directions using IBM model 4.

4. RESULTS

Our results on Arabic-English indicate a reasonable level of coverage within ITG expressiveness and therefore support the hypothesis that Arabic may benefit from the efficiency of models within the ITG expressiveness class.

As shown in Table 1, the coverage of Arabic-English alignments by ITG expresiveness compares consistently with previous results for German and especially French [7]. As expected for non-European languages, alignment with Arabic is more difficult and the sentence coverage is slightly lower than that obtained for French-English. A key factor in this difference appears to be Arabic morphological complexity.

Zens and Ney [7] found that modifying Bracketing ITGs to handle split lexical transductions significantly helped coverage for both language pairs, especially for French-English where such transductions are very frequent: for instance, the very frequent English word *not* translates as *ne ... pas*. Making use of split lexical transductions improved coverage from 81.3% to 96.1% for the French-English alignments. Simi-

Table 1. Alignment coverage within Bracketing ITG expressiveness. Zens and Ney's [7] results on German and French are provided for reference.

Source-target language pair	*Sentence-level coverage*	*Word-level coverage*
English-German	91.6%	-
German-English	87.0%	-
English-French	81.3%	-
French-English	73.6%	-
English-Arabic	76.6%	97.0%
Arabic-English	70.9%	96.2%

larly, we anticipate that well-adapted split lexical transductions could improve coverage for Arabic to a comparable degree.

5. CONCLUSION

We have presented the first direct measurement of word alignment coverage for Arabic-English translation under ITG constraints. The results prove similar to those for previous European languages, with a small drop consistent with the slightly greater linguistic divergence of Arabic from English. Given the relatively small difference from previous French-English measurements, and in light of the success of the ITG constraint for improving both translation speed and accuracy on European languages, this first result on Arabic supports the hypothesis that Arabic-English translation may indeed benefit from the efficiency of models within the ITG expressiveness class.

6. REFERENCES

[1] Andrea Burbank, Marine Carpuat, Stephen Clark, Markus Dreyer, Pamela Fox, Declan Groves, Keith Hall, Mary Hearne, I. Dan Melamed, Yihai Shen, Andy Way, Ben Wellington, and Dekai Wu, "Final report of the 2005 Language Engineering Workshop on Statistical Machine Translation by Parsing," http://www.clsp.jhu.edu/ws2005/groups/statistical/report.html, 2005.

[2] Dekai Wu, "An algorithm for simultaneously bracketing parallel texts by aligning words," in *ACL-95*, Cambridge, MA, Jun 1995, pp. 244–251.

[3] Dekai Wu, "Trainable coarse bilingual grammars for parallel text bracketing," in *3rd Annual Workshop on Very Large Corpora (WVLC-3)*, Cambridge, MA, Jun 1995, pp. 69–81.

[4] Dekai Wu, "A polynomial-time algorithm for statistical machine translation," in *ACL-96*, Santa Cruz, CA, Jun 1996.

[5] Dekai Wu, "Stochastic inversion transduction grammars and bilingual parsing of parallel corpora," *Computational Linguistics*, vol. 23, no. 3, pp. 377–404, Sep 1997.

[6] Dekai Wu, "Alignment," in *Handbook of Natural Language Processing*, Robert Dale, Herman Moisl, and Harold Somers, Eds., pp. 415–458. Marcel Dekker, New York, 2000.

[7] Richard Zens and Hermann Ney, "A comparative study on reordering constraints in statistical machine translation," in *ACL-2003*, Sapporo, Aug 2003, pp. 192–202.

[8] Benjamin Wellington, Sonjia Waxmonsky, and I. Dan Melamed, "Empirical lower bounds on the complexity of translational equivalence," in *COLING/ACL-2006*, Sydney, July 2006, pp. 977–984.

[9] Liang Huang, Hao Zhang, and Daniel Gildea, "Machine translation as lexicalized parsing with hooks," in *International Workshop on Parsing Technologies (IWPT'05)*, Vancouver, 2005.

[10] Hao Zhang and Daniel Gildea, "Stochastic lexicalized inversion transduction grammar for alignment," in *ACL-2005*, Ann Arbor, Jun 2005, pp. 475–482.

[11] Colin Cherry and Dekang Lin, "Soft syntactic constraints for word alignment through discriminative training," in *COLING/ACL-2006*, Sydney, July 2006, pp. 105–112.

[12] Masaaki Nagata, Kuniko Saito, Kazuhide Yamamoto, and Kazuteru Ohashi, "A clustered global phrase reordering model for statistical machine translation," in *COLING/ACL-2006*, Sydney, July 2006, pp. 713–720.

[13] Joan Andreu Sánchez and José Miguel Benedí, "Stochastic inversion transduction grammars for obtaining word phrases for phrase-based statistical machine translation," in *Workshop on Statistical Machine Translation (WMT-2006)*, New York, June 2006, pp. 130–133.

[14] Christoph Tillmann and Tong Zhang, "A discriminative global training algorithm for statistical mt," in *COLING/ACL-2006*, Sydney, July 2006, pp. 721–728.

[15] Taro Watanabe, Hajime Tsukada, and Hideki Isozaki, "Left-to-right target generation for hierarchical phrase-based translation," in *COLING/ACL-2006*, Sydney, July 2006, pp. 777–784.

[16] Deyi Xiong, Qun Liu, and Shouxun Lin, "Maximum entropy based phrase reordering model for statistical machine translation," in *COLING/ACL-2006*, Sydney, July 2006, pp. 521–528.

[17] Hao Zhang and Daniel Gildea, "Inducing word alignments with bilexical synchronous trees," in *COLING/ACL-2006*, Sydney, July 2006, pp. 953–960.

[18] Hao Zhang and Daniel Gildea, "Efficient search for inversion transduction grammar," in *EMNLP-2006*, Sydney, July 2006, pp. 224–231.

[19] Dekai Wu, "Grammarless extraction of phrasal translation examples from parallel texts," in *Sixth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-95)*, Leuven, Belgium, Jul 1995, pp. 354–372.

[20] Dekai Wu and Hongsing Wong, "Machine translation with a stochastic grammatical channel," in *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*, Montreal, Aug 1998.

[21] P. M. Lewis and R. E. Stearns, "Syntax-directed transduction," *Journal of the Association for Computing Machinery*, vol. 15, pp. 465–488, 1968.

[22] Alfred V. Aho and Jeffrey D. Ullman, "Syntax-directed translations and the pushdown assembler," *Journal of Computer and System Sciences*, vol. 3, no. 1, pp. 37–56, 1969.

[23] Hao Zhang, Liang Huang, Daniel Gildea, and Kevin Knight, "Synchronous binarization for machine translation," in *HLT/NAACL-2006*, New York, June 2006, pp. 256–263.

[24] Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer, "Scalable inference and training of context-rich syntactic translation models," in *COLING/ACL-2006*, Sydney, July 2006, pp. 961–968.

[25] Daniel Marcu, Wei Wang, Abdessamad Echihabi, and Kevin Knight, "Spmt: Statistical machine translation with syntactified target language phrases," in *EMNLP-2006*, Sydney, July 2006, pp. 44–52.

[26] Hao Zhang and Daniel Gildea, "Syntax-based alignment: Supervised or unsupervised?," in *COLING-04*, Geneva, Aug 2004.

[27] Kenji Yamada and Kevin Knight, "A syntax-based statistical translation model," in *ACL-2001*, Toulouse, France, Jul 2001.

[28] Richard Zens, Hermann Ney, Taro Watanabe, and Eiichiro Sumita, "Reordering constraints for phrase-based statistical machine translation," in *COLING-04*, Geneva, Aug 2004.

[29] David Chiang, "A hierarchical phrase-based model for statistical machine translation," in *ACL-2005*, Ann Arbor, Jun 2005, pp. 263–270.

[30] Juan Miguel Vilar and Enrique Vidal, "A recursive statistical translation model," in *ACL-2005 Workshop on Building and Using Parallel Texts*, Ann Arbor, Jun 2005, pp. 199–207.

[31] Mona Diab, "Documentation for the Arabic SVM Toolkit," http://www.cs.columbia.edu/mdiab/, 2005.

[32] Franz Josef Och and Hermann Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–52, 2003.