Empirical lower bounds on translation unit error rate for the full class of inversion transduction grammars

Anders Søgaard
Center for Language Technology
University of Copenhagen
soegaard@hum.ku.dk

Dekai Wu
Human Language Technology Center
Hong Kong Univ. of Science and Technology
dekai@cs.ust.hk

Abstract

Empirical lower bounds studies in which the frequency of alignment configurations that cannot be induced by a particular formalism is estimated, have been important for the development of syntax-based machine translation formalisms. The formalism that has received most attention has been inversion transduction grammars (ITGs) (Wu, 1997). All previous work on the coverage of ITGs, however, concerns parse failure rates (PFRs) or sentence level coverage, which is not directly related to any of the evaluation measures used in machine translation. Søgaard and Kuhn (2009) induce lower bounds on translation unit error rates (TUERs) for a number of formalisms, incl. normal form ITGs, but not for the full class of ITGs. Many of the alignment configurations that cannot be induced by normal form ITGs can be induced by unrestricted ITGs, however. This paper estimates the difference and shows that the average reduction in lower bounds on TUER is 2.48 in absolute difference (16.01 in average parse failure rate).

1 Introduction

The first stage in training a machine translation system is typically that of aligning bilingual text. The quality of alignments is in that case of vital importance to the quality of the induced translation rules used by the system in subsequent stages. In string-based statistical machine translation, the alignment space is typically restricted by the n-grams considered in the underlying language model, but in syntax-based machine translation the alignment space is restricted by very different and less transparent structural contraints. While it is easy to estimate the consequences of restrictions to n-grams of limited size, it is less trivial to estimate the consequences of the structural constraints imposed by syntax-based machine translation formalisms. Consequently, much work has been devoted to this task (Wu, 1997; Zens and Ney, 2003; Wellington et al., 2006; Macken, 2007; Søgaard and Kuhn, 2009).

The task of estimating the consequences of the structural constraints imposed by a particular syntax-based formalism consists in finding what is often called "empirical lower bounds" on the coverage of the formalism (Wellington et al., 2006; Søgaard and Kuhn, 2009). Gold standard alignments are constructed and queried in some way as to identify complex alignment configurations, or they are parsed by an all-accepting grammar such that a parse failure indicates that no alignment could be induced by the formalism.

The assumption in this and related work that enables us to introduce a meaningful notion of alignment capacity is that simultaneously recognized words are aligned (Wu, 1997; Zhang and Gildea, 2004; Wellington et al., 2006; Søgaard and Kuhn, 2009). As noted by Søgaard (2009), this definition of alignment has the advantageous consequence that candidate alignments can be singled out by mere inspection of the grammar rules. It also has the consequence that alignments are transitive (Goutte et al., 2004), since simultaneity is transitive.

While previous work (Søgaard and Kuhn, 2009) has estimated empirical lower bounds for normal form ITGs at the level of translation units (TUER), or cepts (Goutte et al., 2004), defined as maximally connected subgraphs in alignments, nobody has done this for the full class of ITGs. What is important to understand is that while normal form ITGs can induce the same class of translations as the full class of ITGs, they do *not* induce the same class of alignments. They do not, for ex-

Proceedings of the 11th International Conference on Parsing Technologies (IWPT), pages 33–36,
Paris, October 2009. ©2009 Association for Computational Linguistics

ample, induce discontinuous translation units (see Sect. 3). Sect. 2 briefly presents some related results in the literature. Some knowledge about formalisms used in machine translation is assumed.

2 Related work

Aho and Ullman (1972) showed that 4-ary synchronous context-free grammars (SCFGs) could not be binarized, and Satta and Peserico (2005) showed that the hiearchy of SCFGs beyond ternary ones does not collapse; they also showed that the complexity of the universal recognition problem for SCFGs is NP-complete. ITGs on the other hand has a $\mathcal{O}(|G|n^6)$ solvable universal recognition problem, which coincides with the unrestricted alignment problem (Søgaard, 2009). The result extends to decoding in conjunction with a bigram language model (Huang et al., 2005).

Wu (1997) introduced ITGs and normal form ITGs. ITGs are a notational variant of the subclass of SCFGs such that all indexed nonterminals in the source side of the RHS occur in the same order or exactly in the inverse order in the target side of the RHS. It turns out that this subclass of SCFGs defines the same set of translations that can be defined by binary SCFGs. The different forms of production rules are listed below with the more restricted normal form production rules in the right column, with $\phi \in (N \cup \{e/f \mid e \in T^*, f \in T^*\})^*$ (N nonterminals and T terminals, as usual). The RHS operator [ ] preserves source language constituent order in the target language, while $\langle\ \rangle$ reverses it.[1]

$$\begin{array}{lll|lll} A & \rightarrow & [\phi] & A & \rightarrow & [BC] \\ A & \rightarrow & \langle\phi\rangle & A & \rightarrow & \langle BC\rangle \\ & & & A & \rightarrow & e/f \end{array}$$

Several studies have adressed the alignment capacity of ITGs and normal form ITGs. Zens and Ney (2003) induce lower bounds on PRFs for normal form ITGs. Wellington et al. (2006) induce lower bounds on PRFs for ITGs. Søgaard and Kuhn (2009) induce lower bounds on TUER for normal form ITGs and more expressive formalisms for syntax-based machine translation. No one has, however, to the best our knowledge induced lower bounds on TUER for ITGs.

3 Experiments

As already mentioned empirical lower bounds studies differ in four important respects, namely wrt.: (i) whether they use hand-aligned or automatically aligned gold standards, (ii) the level at which they count failures, e.g. sentence, alignment or translation unit level, (iii) whether they interpret translation units disjunctively or conjunctively, and (iv) whether they induce the lower bounds (a) by running an all-accepting grammar on the gold standard data, (b) by logical characterization of the structures that can be induced by a formalism, or (c) by counting the frequency of complex alignment configurations. The advantage of (a) and (b) is that they are guaranteed to find the highest possible lower bound on the gold standard data, whereas (c) is more modular (formalism-independent) and actually tells us what configurations cause trouble.

(i) In this study we use hand-aligned gold standard data. It should be obvious why this is preferable to automatically aligned data. The only reason that some previous studies used automatically aligned data is that hand-aligned data are hard to come by. This study uses the data also used by Søgaard and Kuhn (2009), which to the best of our knowledge uses the largest collection of hand-aligned parallel corpora used in any of these studies. (ii) Failures are counted at the level of translation units as argued for in the above, but supplemented by parse failure rates for completeness. (iii) Since we count failures at the level of translation units, it is natural to interpret them conjunctively. Otherwise we would in reality count failures at the level of alignments. (iv) We use (c).

The conjunctive interpretation of translation units was also adopted by Fox (2002) and is motivated by the importance of translation units and discontinuous ones in particular to machine translation in general (Simard and colleagues, 2005; Ayan and Dorr, 2006; Macken, 2007; Shieber, 2007). In brief,

$$\text{TUER} = 1 - \frac{2|S_U \cap G_U|}{|S_U| + |G_U|}$$

where G_U are the translation units in the gold standard, and S_U the translation units produced by the system. This evaluation measure is related to consistent phrase error rate (CPER) introduced in Ayan and Dorr (2006), except that it does not only consider contiguous phrases.

[1] One reviewer argues that our definition of full ITGs is *not* equivalent to the definition in Wu (1997), which, in the reviewer's words, allows "at most one lexical item from each language". Sect. 6 of Wu (1997), however, explicitly encourages lexical elements in rules to have more than one lexical item in many cases.

3.1 Data

The characteristics of the hand-aligned gold standard parallel corpora used are presented in Figure 1. The Danish-Spanish text is part of the Copenhagen Dependency Treebank (Parole), English-German is from Pado and Lapata (2006) (Europarl), and the six combinations of English, French, Portuguese and Spanish are documented in Graca et al. (2008) (Europarl).

3.2 Alignment configurations

The full class of ITGs induces many alignment configurations that normal form ITGs do not induce, incl. discontinuous translation units (DTUs), i.e. translation units with at least one gap, double-sided DTUs, i.e. DTUs with both a gap in the source side and a gap in the target side, and multigap DTUs with arbitrarily many gaps (as long as the contents in the gap are either respect the linear order of the source side or the inverted order).

ITGs do *not* induce (i) inside-out alignments, (ii) cross-serial DTUs, (iii) what is called the “bonbon” configuration below, and (iv) multigap DTUs with mixed order in the target side. The reader is referred to Wu (1997) for discussion of inside-out alignments. (ii) and (iii) are explained below.

3.2.1 Induced configurations

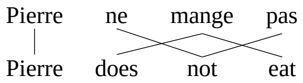
DTUs are easily induced by unrestricted ITG productions, while they cannot be induced by productions in normal form. The combination of the production rules $A \rightarrow [\epsilon/\textit{ne } B \textit{ nothing/pas}]$ and $B \rightarrow [\textit{change/modifie}]$, for example, induces a DTU with a gap in the French side for the pair of substrings $\langle$*change nothing, ne modifie pas*$\rangle$.

Multigap DTUs with up to three gaps are frequent (Søgaard and Kuhn, 2009) and have shown to be important for translation quality (Simard and colleagues, 2005). While normal form ITGs do not induce multigap DTUs, ITGs induce a particular subclass of multigap DTUs, namely those that are constructed by linear or inverse interpolation.

3.2.2 Non-induced configurations

Inside-out alignments were first described by Wu (1997), and their frequency has been a matter of some debate (Lepage and Denoual, 2005; Wellington et al., 2006; Søgaard and Kuhn, 2009).

Cross-serial DTUs are made of two DTUs non-contiguous to the same side such that both have material in the gap of each other. **Bonbons** are similar, except the DTUs are non-contiguous to different sides, i.e. D has a gap in the source side that contains at least one token in E, and E has a gap in the target side that contains at least one token in D. Here’s an example of a bonbon configuration from Simard et al. (2005):



Multigap DTUs with mixed transfer are, as already mentioned multigap DTUs with crossing alignments from material in two distinct gaps.

3.3 Results

The lower bounds on TUER for the full class of ITGs are obtained by summing the ratios of inside-out alignments, cross-serial DTUs, bonbons and mixed order multigap DTUs, subtracting any overlap between these classes of configurations. The lower bounds on TUER for normal form ITGs sum ratios of inside-out aligments and DTUs subtracting any overlap. Figure 1 presents the ratio ($\times 100$), and Figure 2 presents the induced lower bounds on the full class of ITGs and normal form ITGs. Any two configurations differ *on all translation units* in order to count as two distinct configurations in these statistics. Otherwise a single translation unit could be removed to simplify two or more configurations.

4 Discussion

The usefulness of alignment error rate (AER) (Och and Ney, 2000) has been questioned lately (Fraser and Marcu, 2007); most importantly, AER does not always seem to correlate with translation quality. TUER is likely to correlate better with translation quality, since it by definition correlates with CPER (Ayan and Dorr, 2006). No large-scale experiment has been done yet to estimate the strength of this correlation.

Our study also relies on the assumption that simulatenously recognized words are aligned in bilingual parsing. The relationship between parsing and alignment can of course be complicated in ways that will alter the alignment capacity of ITG and its normal form; on some definitions the two formalisms may even become equally expressive.

5 Conclusion

It was shown that the absolute reduction in average lower bound on TUER is 2.48 for the full class of ITGs over its canonical normal form. For PRF, it is 16.01.

	Snts	TUs	IOAs	DTUs	CDTUs	Bonbons	MIX-DTUs
Da-Sp	926	6441	0.56	9.16	0.81	0.16	0.23
En-Fr	100	869	0.23	2.99	0.12	0.23	0.23
En-Ge	987	17354	1.75	5.55	0.45	0.05	0.79
En-Po	100	783	0.26	2.17	0.00	0.00	0.38
En-Sp	100	831	0.48	1.32	0.00	0.00	0.36
Po-Fr	100	862	0.23	3.13	0.58	0.00	0.46
Po-Sp	100	882	0.11	0.90	0.00	0.00	0.00
Sp-Fr	100	914	0.11	2.95	0.55	0.00	0.22

Figure 1: Characteristics of the parallel corpora and frequency of configurations ($\frac{n}{\text{TUs}} \times 100$).

	ITGs				**NF-ITGs**			
	LB-TUER	LB-PFR	Ovlp(TUs)	Ovlp(Snts)	LB-TUER	PFR	Ovlp(TUs)	Ovlp(Snts)
Da-Sp	1.58	10.37	11	10	8.54	40.50	76	32
En-Fr	0.69	6.00	1	1	2.88	22.00	3	2
En-Ge	2.75	47.32	49	42	5.24	69.30	357	236
En-Po	0.64	5.00	0	0	2.43	19.00	0	0
En-Sp	0.84	7.00	0	0	1.80	15.00	0	0
Po-Fr	1.04	9.00	2	2	3.36	24.00	0	0
Po-Sp	0.11	1.00	1	1	0.90	8.00	1	1
Sp-Fr	0.77	7.00	1	1	3.06	23.00	0	0
AV	1.05	11.59			3.53	27.60		

Figure 2: Induced lower bounds for ITGs and normal form ITGs (NF-ITGs). LB-TUER lists the lower bounds on TUER. LB-PFR lists the lower bounds on parse failure rates. Finally, the third and fourth columns list configuration overlaps at the level of translation units, resp. sentences.

References

Alfred Aho and Jeffrey Ullman. 1972. *The theory of parsing, translation and compiling*. Prentice-Hall.

Necip Ayan and Bonnie Dorr. 2006. Going beyond AER. In *COLING-ACL'06*, Sydney, Australia.

Heidi Fox. 2002. Phrasal cohesion and statistical machine translation. In *EMNLP'02*, Philadelphia, PA.

Alexander Fraser and Daniel Marcu. 2007. Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3):293–303.

Cyril Goutte, Kenji Yamada, and Eric Gaussier. 2004. Aligning words using matrix factorisation. In *ACL'04*, Barcelona, Spain.

Joao Graca, Joana Pardal, Luísa Coheur, and Diamantino Caseiro. 2008. Building a golden collection of parallel multi-language word alignments. In *LREC'08*, Marrakech, Morocco.

Liang Huang, Hao Zhang, and Daniel Gildea. 2005. Machine translation as lexicalized parsing with hooks. In *IWPT'05*, pages 65–73, Vancouver, BC.

Yves Lepage and Etienne Denoual. 2005. Purest ever example-based machine translation. *Machine Translation*, 19(3–4):251–282.

Lieve Macken. 2007. Analysis of translational correspondence in view of sub-sentential alignment. In *METIS-II*, pages 9–18, Leuven, Belgium.

Franz Och and Hermann Ney. 2000. A comparison of alignment models for statistical machine translation. In *COLING'00*, Saarbrücken, Germany.

Sebastian Padó and Mirella Lapata. 2006. Optimal constituent alignment with edge covers for semantic projection. In *ACL-COLING'06*, Sydney, Australia.

Giorgio Satta and Enoch Peserico. 2005. Some computational complexity results for synchronous context-free grammars. In *HLT-EMNLP'05*, Vancouver, BC.

Stuart Shieber. 2007. Probabilistic synchronous tree-adjoining grammars for machine translation. In *SSST'07*, pages 88–95, Rochester, NY.

Michel Simard and colleagues. 2005. Translating with non-contiguous phrases. In *HLT-EMNLP'05*, Vancouver, BC.

Anders Søgaard and Jonas Kuhn. 2009. Empirical lower bounds on alignment error rates in syntax-based machine translation. In *NAACL-HLT'09, SSST-3*, Boulder, CO.

Anders Søgaard. 2009. On the complexity of alignment problems in two synchronous grammar formalisms. In *NAACL-HLT'09, SSST-3*, Boulder, CO.

Benjamin Wellington, Sonjia Waxmonsky, and Dan Melamed. 2006. Empirical lower bounds on the complexity of translational equivalence. In *ACL'06*, pages 977–984, Sydney, Australia.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.

Richard Zens and Hermann Ney. 2003. A comparative study on reordering constraints in statistical machine translation. In *ACL'03*, Sapporo, Japan.

Hao Zhang and Daniel Gildea. 2004. Syntax-based alignment: supervised or unsupervised? In *COLING'04*, pages 418–424, Geneva, Switzerland.