Can inversion transduction grammars generate hand alignments?

Anders Søgaard
University of Copenhagen
Njalsgade 140–2
DK-2300 Copenhagen
soegaard@hum.ku.dk

Abstract

The adequacy of inversion transduction grammars (ITGs) has been widely debated, and the discussion's crux seems to be whether the search space is inclusive enough (Zens and Ney, 2003; Wellington et al., 2006; Søgaard and Wu, 2009). Parse failure rate when parses are constrained by word alignments is one metric that has been used, but no one has studied parse failure rates of the full class of ITGs on representative hand aligned corpora. It has also been noted that ITGs in Chomsky normal form induce strictly less alignments than ITGs (Søgaard and Wu, 2009). This study is the first study that directly compares parse failure rates for this subclass and the full class of ITGs.

1 Introduction

The adequacy of grammar-based machine translation formalisms is sometimes empirically evaluated by running all-accepting grammars on large amounts of automatically aligned text (Zens and Ney, 2003). What is studied is called alignment capacity (Søgaard and Wu, 2009) or translation equivalence modeling (Zhang et al., 2008), i.e. a formalism's ability to generate observed alignments or translation equivalences; and the study is closely related to the study of translation model search spaces (Zens and Ney, 2003; Dreyer et al., 2007). All-acccepting grammars are simply grammars that contain all possible rules that can be expressed in a formalism. A grammar generates an aligned sentence pair if it can generate the two sentences in such a way that all aligned words are generated simultaneously (Wu, 1997). What is studied is thus: Can an all-accepting grammar generate the aligned sentence pairs observed in a text? The metric in these studies is parse failure rate (PFR) or its inverse, i.e. the number of sentence pairs that can be generated over the total number of sentence pairs.

Most alignment capacity studies use automatically aligned text, since hand-aligned text is hard to come by. Recently three important data sets have been released (Padó and Lapata, 2006; Graca et al., 2008; Buch-Kromann et al., 2010). Our experiments include 12 hand-aligned parallel texts of varying size.

Our main contribution is evaluating the empirical adequacy of inversion transduction grammars (ITGs) (Wu, 1997), a popular grammar-based machine translation formalism, on these data sets. It has been noted that the alignment capacity of the full class of ITGs extends that of the class of ITGs that are in Chomsky normal form (NF-ITGs) (Søgaard and Wu, 2009), i.e. while the normal form *is* a normal form in the sense that it does not alter the generative capacity in terms of sentence pairs, the normal form restrictions *do* exclude certain alignment configurations. Consequently, we compare the adequacy of *both* ITGs and NF-ITGs.

It is shown that while ITGs are more adequate than local reordering models (and in many cases also to IBM models; cf. Zens and Ney (2003) and Dreyer et al. (2007)), hand alignments are very hard to generate. While 1-PFR is >60% for most data sets, ITGs and NF-ITGs fit four of our data sets rather poorly: 1-PFR is less than 50% for three of the data sets involving Danish, and for English-German.

Sect. 2 briefly summarizes related work. Sect. 3 introduces a novel algorithm for simulating an all-

accepting grammar. Finally, Sect. 4 presents our experiments.

2 Related work

Unlike other studies that have studied the adequacy of ITG's alignment capacity, Wellington et al. (2006) used hand-aligned data in their studies. Hand alignments contain fewer errors than automatic alignments, are supposed to reflect translational equivalence more closely and will remain relevant regardless of improvements in technology for automatic word alignments. All the parallel data used in the experiments of Wellington et al. (2006) have English as one of the two languages. This of course biases their study a bit. On the other hand, translations to or from English are easier to come by, and more hand-aligned texts are available.

Our experiments include a total of 12 data sets, out of which five are translations to or from English. Wellington et al. (2006) use a total of five data sets, one of which is also used in our experiments below (Canadian Hansard). The data sets in their study are of about the same size as those used in ours. They consider a total of 1427 sentence pairs, whereas we consider a total of 2852 sentence pairs.

The methodology of Wellington et al. (2006) also differs from ours in one important respect, namely how incomplete coverage of multiword translation units is counted. The authors count multiword translation units in what they refer to as a *disjunctive* manner, i.e. if at least one link in every unit is generated, the alignment configuration is counted in as having been generated. So for instance if all words in a source sentence are aligned to all words in the target sentence, it only takes producing a single link to "generate" the alignment configuration.

In our experiments, we count coverage of translation units in a "conjunctive" manner, i.e. all links in every translation unit must be generated before the overall alignment configuration can be said to have been generated. See Søgaard and Kuhn (2009) for a number of arguments for measuring alignment capacity in terms of exact matches of translation units.

Søgaard and Wu (2009) show that ITG and NF-ITG generate different classes of alignment configurations. This is in a way surprising, since the two formalisms are equivalent in terms of generative capacity. The reason for the apparent paradox is of course that ITGs only align words that are simultaneously generated (Wu, 1997). Consequently, two ITG derivations of the same sentence pair may induce different alignment configurations. Zens and Ney (2003) and Wellington et al. (2006) introduce weaker normal form conditions in their studies.

Søgaard and Wu (2009) consider some of the same data sets used in our experiments, but their approach is very different. They identify alignment configurations that cannot be generated by ITGs or NF-ITGs, e.g. inside-out alignments or discontinuous translation units, and simply count their occurrences in the parallel corpora. This has the advantage that they provide some error analysis on the fly, e.g. they can immediately see the specific impact of inside-out alignments on error rates. On the other hand, the lower bounds that they provide on PFRs, are very conservative lower bounds. Our more agressive search show that the lower bounds on PFRs that they induce can be increased by 15-25%.

3 Alignment validation

Algorithms that compose constituents out of word-to-word links and try to find constituents that cover the entire sentence pairs have been used in similar studies (Wu, 1997; Zens and Ney, 2003; Wellington et al., 2006). This process seems to have no established name in the literature, but we refer to it as *alignment validation*, i.e. checking if an alignment is valid wrt. a formalism in the sense that it can be generated by an all-accepting grammar. Since we measure coverage in a "conjunctive" manner, alignment validation is a bit more complicated than in related work. Our input constituents are possibly discontinuous translation units.

The following alignment validation algorithm (which can no doubt be optimized) was used in our experiments. The subprocedure in Figure 1, which is called by the overall procedure described below, takes two parse charts, i.e. two matrices m, m' with $i < j$ for all $i \in m$ and $m[i] = j$ (resp., m'), a derivation step counter c, a string position p and a variable x with values $\{0, 1\}$ as input and controls a chart-based parsing algorithm. The subprocedure *complete* simply checks if there is a constituent that covers the entire span on both sides. Note that since there is no normal form assumption our parsing algorithm has to scan the chart twice before it can return a failure (lines 21–

24). The subprocedure *check_rule* is left out for brevity. It checks that the application of the rule adding two new constituents $\langle i, j\rangle$ and $\langle i', j'\rangle$ to the charts is possible and that it does not violate the alignment configuration, i.e. that the charts do not contain unvalidated links in these spans. Our naïve implementation of this procedure has asymptotic complexity $\mathcal{O}(n^8)$, since it needs to search for the maximal covered spans on both sides.

The subprocedure is embedded in the overall algorithm in Figure 2 which outputs the number of parsed sentence pairs, i.e. the number of sentence pairs where the alignment configuration can be generated.

The Boolean variable *nf* is set to be true if normal form conditions are imposed, i.e. this means that translations units must be continuous. The call in line 5 adds all continuous translation units.

4 Experiments

This section describes our experiments, incl. the data sets used, the metric used in evaluation and the results obtained on the data sets.

4.1 Data Sets

The characteristics of the hand-aligned parallel texts used are presented in Figure 3.

The parallel texts that involve Danish are part of the Copenhagen Dependency Treebank (Buch-Kromann et al., 2010), based on translations of the balanced Parole corpus, 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). We use the 200 sentences standard training section of the Canadian Hansard data set for supervised word alignment.

4.2 Alignment reachability

Similarly to Zens and Ney (2003), we use the inverse of PFR as metric. We refer to this below as *alignment reachability*, in analogy to translation reachability. Each experiment thus applies the above algorithm to a set of hand-aligned sentence pairs. The algorithm either reaches an alignment, which means that the alignment *can* be generated, or it does not, which mean that the alignment is beyond the expressive power of the formalism in question. Parse failure rate is the number of failures over the total number of hand-aligned sentence pairs, whereas alignment reachability is the number of reached alignments over the total number of hand-aligned sentence pairs.

4.3 Results

We compare our upper bounds on alignment reachability for ITG and NF-ITG to the configuration-based upper bounds obtained in Søgaard and Wu (2009). We also introduce a simpler baseline system, namely ITG without inverse production rules. Such a system is a generalization of local reordering models (LR) such as MJ-1 and MJ-2 (Kumar and Byrne, 2005) whose expressivity is studied by Dreyer et al. (2007).

Our results are presented in Figure 4. Our baseline generates a superset of the alignments that can be generated by MJ-1 and MJ-2. For the full class of ITGs the error rate is on average increased by more than 25% compared to the bounds presented in Søgaard and Wu (2009). For normal form grammars, the increase is about 15%.

A very interesting observation is that the difference in coverage between ITG and NF-ITG measured in terms of true PFRs is much smaller than when estimated in a configuration-based manner. While the results in Søgaard and Wu (2009) indicate that the normal form proposed in Wu (1997) drastically reduces the empirical adequacy of ITGs when evaluated on hand alignments, i.e. by more than 10% on average over the data sets in Graca et al. (2008), our results show that decrease in coverage is moderate ($<$1.5%).

It is also clear from our results that only using the Canadian Hansard in this type of studies leads to a significant bias. This data set does contain complex alignment configurations for which rules with up to five nonterminals and 18 terminals in the right-hand side are required (Zhang et al., 2008), but they are relatively infrequent.

Finally our results seem to suggest that hand-alignments are not much more complex than automatically generated alignments. Zens and Ney (2003) estimate alignment reachability for GIZA-aligned Canadian Hansard data and report comparable coverage. In one direction, NF-ITGs cover 81.3% of the alignments; in the other direction, they cover 73.6. The average is close to our 76.98% alignment reachability on the manually constructed alignments. On the other hand, they report much higher scores than we do for something that remains a subset of the full class of ITGs. Interestingly, they show that coverage is

```
1: begin chart_parse(m, m', c, p, x) :
2: c++
3: if complete(m, m') then
4:    return true
5: else
6:    rule_applied ← false
7:    for i = p to length(m) + 1 do
8:       for j ∈ m[i] do
9:          for i' = 1 to length(m') + 1 do
10:            for j' ∈ m[i'] do
11:               if check_rule(m, m', i, j, i', j') then
12:                  m[i][j]+ = [c]
13:                  m[i'][j']+ = [c]
14:                  rule_applied ← true
15:                  return chart_parse(m, m', c, p, 1)
16:               end if
17:            end for
18:         end for
19:      end for
20:   end for
21:   if (not rule_applied) and x = 1 then
22:      return chart_parse(m, m', c, 1, 0)
23:   else if (not rule_applied) and x = 0 then
24:      return false
25:   end if
26: end if
```

Figure 1: Subprocedure in our alignment validation algorithm.

```
1: for ⟨s, s'⟩ ∈ T do
2:    m ← matrix(s)
3:    m' ← matrix(s')
4:    if (not nf) or continuous(⟨s, s'⟩) then
5:       ⟨m, m', c⟩ ← add_continuous(m, m', s, s', 0)
6:       if chart_parse(m, m', c, 1, 1) then
7:          parsed++
8:       end if
9:    end if
10: end for
11:
12: print parsed
```

Figure 2: Our alignment validation algorithm.

	Sentences	Links
Da-De	266	1314
Da-It	26	1386
Da-Ru	33	833
Da-Sp	966	8944
En-Fr	100	1279
En-Ge	987	23243
En-Po	100	1198
En-Sp	100	1198
Po-Fr	100	1290
Po-Sp	100	1189
Sp-Fr	100	1303
Total	2852	43086

Figure 3: Characteristics of the data sets used in our experiments.

	NF-ITG	SW09(NF)	ITG	SW09	LR
En-Fr	65.00	78.00	68.00	94.00	32.00
En-Po	65.00	81.00	67.00	95.00	25.00
En-Sp	73.00	85.00	74.00	93.00	30.00
Po-Fr	63.00	76.00	63.00	91.00	44.00
Po-Sp	80.00	92.00	81.00	99.00	53.00
Sp-Fr	68.00	77.00	68.00	93.00	51.00
AV	69.00	81.50	70.17	94.17	-
Da-De(25)	47.62	-	49.35	-	-
Da-It(25)	60.00	-	60.00	-	-
Da-Ru(25)	47.05	-	47.05	-	-
Da-Sp(25)	30.68	*59.50	35.54	*89.63	-
En-Ge(15)	38.97	*30.70	45.13	*52.68	-
Hansard(15)	76.98	-	81.75	-	-

Figure 4: Alignment reachability scores for NF-ITG, ITG and (an upper bound on) local reordering models, compared to results in Søgaard and Wu (2009). Sentence length cut-off given in parentheses. * means that results are incomparable to those in Søgaard and Wu (2009), because different cut-offs were used.

also considerably better than that of the IBM models.

5 Conclusion

The status of alignments in machine translation is widely debated (Fraser and Marcu, 2007), but so are other metrics such as BLEU. Alignment reachability is related to BLEU oracle computation (Dreyer et al., 2007), but in a very indirect way. Our studies nevertheless show that there are translational equivalences that are very hard to capture in computational search spaces. While the ITG translation search space seems better fit than many other models, as indicated by our experiment as well as experiments cited above, there are still many alignment configurations that are beyond reach. Capturing those configurations may not be necessary from a practical point of view, but it may nevertheless be worth considering other ways of balancing expressivity and efficiency.

References

Matthias Buch-Kromann, Jürgen Wedekind, and Jakob Elming. 2010. The Copenhagen Danish–English Dependency Treebank. To appear.

Markus Dreyer, Keith Hall, and Sanjeev Khudanpur. 2007. Comparing reordering constraints for SMT using efficient BLEU oracle computation. In *The 1st Workshop on Syntax and Structure in Statistical Translation, North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT) 2007*, New York, NY.

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

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.

Shankar Kumar and William Byrne. 2005. Local phrase reordering models for statistical machine translation. In *HLT-EMNLP*, Vancouver, Canada.

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

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 and Dekai Wu. 2009. Empirical lower bounds on alignment error rates for the full class of inversion transduction grammars. In *Proceedings of the 11th International Conference on Parsing Technologies*, Paris, France.

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, Daniel Gildea, and David Chiang. 2008. Extracting synchronous grammar rules from word-level alignments in linear time. In *Coling*, Manchester, England.