Textual Entailment Recognition Based on Inversion Transduction Grammars

Dekai Wu[1]

Human Language Technology Center
HKUST
Department of Computer Science
University of Science and Technology, Clear Water Bay, Hong Kong
dekai@cs.ust.hk

Abstract

The PASCAL Challenge's textual entailment recognition (RTE) task presents intriguing opportunities to test various implications of the strong language universal constraint posited by Wu's (1995, 1997) Inversion Transduction Grammar (ITG) hypothesis. The ITG Hypothesis provides a strong inductive bias, and has been repeatedly shown empirically to yield both efficiency and accuracy gains for numerous language acquisition tasks. Since the RTE challenge abstracts over many tasks, it invites meaningful analysis of the ITG Hypothesis across tasks including information retrieval, comparable documents, reading comprehension, question answering, information extraction, machine translation, and paraphrase acquisition. We investigate two new models for the RTE problem that employ simple generic Bracketing ITGs. Experimental results show that, even in the absence of any thesaurus to accommodate lexical variation between the Text and the Hypothesis strings, surprisingly strong results for a number of the task subsets are obtainable from the Bracketing ITG's structure matching bias alone.

1 Introduction

The *Inversion Transduction Grammar* or *ITG* formalism, which historically was developed in the context of translation and alignment, hypothesizes strong expressiveness restrictions that constrain paraphrases to vary word order only in certain allowable nested permutations of arguments—even across different languages (Wu, 1997). The textual entailment recognition (RTE) challenge provides opportunities for meaningful analysis of the ITG Hypothesis across a broad range of application domains.

The strong inductive bias imposed by the ITG Hypothesis has been repeatedly shown empirically to yield both efficiency and accuracy gains for numerous language acquisition tasks, across a variety of language pairs and tasks. Zens and Ney (2003) 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) 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, 2001). With regard to translation rather than alignment accuracy, Zens *et al.* (2004) show that decoding under ITG constraints yields significantly lower word error rates and BLEU scores than the IBM constraints.

The present studies on the RTE challenge are motivated by the following observation: the empirically demonstrated suitability of ITG paraphrasing constraints across languages should hold, if anything, even more strongly in the monolingual case.

The simplest class of ITGs, *Bracketing ITGs*, are particularly interesting in applications like the RTE challenge, because they impose ITG constraints in language-independent fashion, and in the simplest case do not require any language-specific linguistic grammar or training. In Bracketing ITGs, the grammar uses only a single, undifferentiated non-terminal (Wu, 1995). The key modeling property of Bracketing ITGs that is most relevant to the RTE challenge is that they assign strong preference to candidate Text-Hypothesis 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 con-

[1]The author would like to thank the Hong Kong Research Grants Council (RGC) for supporting this research in part through grants RGC6083/99E, RGC6256/00E, and DAG03/04.EG09, and Marine Carpuat for invaluable assistance in preparing the datasets and stoplist.

structed resources.

Formally, ITGs can be defined as the restricted subset of syntax-directed transduction grammars or SDTGs Lewis and Stearns (1968) where all of the rules are either of *straight* or *inverted* orientation. Ordinary 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 language. 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.

Moreover, for reasons discussed by Wu (1997), 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 emerges naturally from ITGs in language-independent fashion, without any hardcoded language-specific knowledge. This further suggests that ITGs should do well at picking out Text-Hypothesis pairs where the order of up to four arguments per frame may vary freely between the two strings. Conversely, ITGs should do well at rejecting pairs 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.

As an illustrative example, in common similarity models, 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 bisentences representing a true Chinese translation of the same English sentence, as well as a true English translation of the same Chinese sentence.

We investigate two new models for the RTE problem that employ simple generic Bracketing ITGs, both with and without a stoplist. The experimental results show that, even in the absence of any thesaurus to accommodate lexical variation between the Text and the Hypothesis strings, surprisingly strong results for a number of the task subsets are obtainable from the Bracketing ITG's structure matching bias alone.

2 Experimental Method

Each Text-Hypothesis pair of the test set was scored via the biparsing algorithm described in Wu and Fung (2005) which is essentially similar to the dynamic programming approach of Wu (1997). As mentioned earlier, biparsing for ITGs can be accomplished efficiently in polynomial time, rather than the exponential time required for classical SDTGs.

The ITG scoring model can also be seen as a variant of the approach described by Leusch *et al.* (2003), which allows us to forego training to estimate true probabilities; instead, rules are simply given unit weights (with caveats discussed in the Results section). The ITG scores can 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.

We evaluated two different versions of the Bracketing ITG based RTE models.

In the basic version, all words of the vocabulary are included among the lexical transductions, allowing exact word matches between the Text and the Hypothesis.

The second version excludes a list of 172 words from a stoplist from the lexical transductions. The motivation for this model was to discount the effect of words such as "the" or "of" since, more often than not, they could be irrelevant to the RTE task.

No significant training was performed with the available development sets. Rather, the aim was to establish foundational baseline results, to see in this first round of RTE experiments what results could be obtained with the simplest versions of the ITG models.

The RTE test set consists of 300 Text-Hypothesis string pairs, selected from various sources by human collectors. Each string pair is labeled according to the task category that the data was drawn from. These labels divide the data into seven task subsets, which we analyze individually below. While the collectors were attempting to build a representative dataset, it is difficult to make claims about

distributional neutrality, due to the arbitrary nature of the example selection process.

3 Results

Across all subsets overall, the basic model produced a confidence-weighted score of 54.97% (better than chance at the 0.05 level). All examples were labeled, so precision, recall, and f-score are equivalent; the accuracy was 51.25%.

Surprisingly, the stoplisted model produced worse results. The overall confidence-weighted score was 53.61%, and the accuracy was 50.50%. We discuss the reasons below in the context of specific subsets.

As one might expect, the Bracketing ITG models performed better on the subsets more closely approximating the tasks for which Bracketing ITGs were designed: comparable documents (CD), paraphrasing (PP), and information extraction (IE). We will discuss some important caveats on the machine translation (MT) and reading comprehension (RC) subsets. The subsets least close to the Bracketing ITG models are information retrieval (IR) and question answering (QA).

3.1 Comparable Documents (CD)

The CD task definition can essentially be characterized as recognition of noisy word-aligned sentence pairs. Among all subsets, CD is perhaps closest to the noisy word alignment task for which Bracketing ITGs were originally developed, and indeed produced the best results for both of the Bracketing ITG models. The basic model produced a confidence-weighted score of 79.88% (accuracy 71.33%), while the stoplisted model produced an essentially unchanged confidence-weighted score of 79.83% (accuracy 70.00%).

The results on the RTE Challenge datasets closely reflect the larger-scale findings of Wu and Fung (2005), who demonstrate that an ITG based model yields far more accurate extraction of parallel sentences from quasi-comparable non-parallel corpora than previous state-of-the-art methods. Wu and Fung's results also use the evaluation metric of uninterpolated average precision (i.e., confidence-weighted score).

Note also that we believe the results here are artificially lowered by the absence of any thesaurus, and that significantly further improvements would be seen with the addition of a suitable thesaurus, for reasons discussed below under the MT subsection.

3.2 Paraphrase Acquisition (PP)

The PP task is also close to the task for which Bracketing ITGs were originally developed. For the PP task, the basic model produced a confidence-weighted score of 57.26% (accuracy 56.00%), while the stoplisted model produced a lower confidence-weighted score of 51.65% (accuracy 52.00%). Unlike the CD task, the greater importance of function words in determining equivalent meaning between paraphrases appears to cause the degradation in the stoplisted model.

The effect of the absence of a thesaurus is much stronger for the PP task as opposed to the CD task. Inspection of the datasets reveals much more lexical variation between paraphrases, and shows that cases where lexis does not vary are generally handled accurately by the Bracketing ITG models. The MT subsection below discusses why a thesaurus should produce significant improvement.

3.3 Information Extraction (IE)

The IE task presents a slight issue of misfit for the Bracketing ITG models, but yielded good results anyhow. The basic Bracketing ITG model attempts to align all words/collocations between the two strings. However, for the IE task in general, only a substring of the Text should be aligned to the Hypothesis, and the rest should be disregarded as "noise". We approximated this by allowing words to be discarded from the Text at little cost, by using parameters that impose only a small penalty on null-aligned words from the Text. (As a reasonable first approximation, this characterization of the IE task ignores the possibility of modals, negation, quotation, and the like in the Text.)

Despite the slight modeling misfit, the Bracketing ITG models produced good results for the IE subset. The basic model produced a confidence-weighted score of 59.92% (accuracy 55.00%), while the stoplisted model produced a lower confidence-weighted score of 53.63% (accuracy 51.67%). Again, the lower score of the stoplisted model appears to arise from the greater importance of function words in ensuring correct information extraction, as compared with the CD task.

3.4 Machine Translation (MT)

One exception to expectations is the machine translation subset, a task for which Bracketing ITGs were developed. The basic model produced a confidence-weighted score of 34.30% (accuracy 40.00%), while the stoplisted model produced a comparable confidence-weighted score of 35.96% (accuracy 39.17%).

However, the performance here on the machine translation subset cannot be directly interpreted, for two reasons.

First, the task as defined in the RTE Challenge datasets is not actually crosslingual machine translation, but rather evaluation of monolingual comparability between an automatic translation and a gold standard human translation. This is in fact closer to the problem of defining a good MT evaluation metric, rather than MT itself. Leusch *et al.* (2003 and personal communication) found that

Bracketing ITGs as an MT evaluation metric show excellent correlation with human judgments.

Second, no translation lexicon or equivalent was used in our model. Normally in translation models, including ITG models, the translation lexicon accommodates lexical ambiguity, by providing multiple possible lexical choices for each word or collocation being translated. Here, there is no second language, so some substitute mechanism to accommodate lexical ambiguity would be needed.

The most obvious substitute for a translation lexicon would be a monolingual thesaurus. This would allow matching synonomous words or collocations between the Text and the Hypothesis. Our original thought was to incorporate such a thesaurus in collaboration with teams focusing on creating suitable thesauri, but time limitations prevented completion of these experiments. Based on our own prior experiments and also on Leusch *et al.*'s experiences, we believe this would bring performance on the MT subset to excellent levels as well.

3.5 Reading Comprehension (RC)

The reading comprehension task is similar to the information extraction task. As such, the Bracketing ITG model could be expected to perform well for the RC subset. However, the basic model produced a confidence-weighted score of just 49.37% (accuracy 47.14%), and the stoplisted model produced a comparable confidence-weighted score of 47.11% (accuracy 45.00%).

The primary reason for the performance gap between the RC and IE domains appears to be that RC is less news-oriented, so there is less emphasis on exact lexical choices such as named entities. This puts more weight on the importance of a good thesaurus to recognize lexical variation. For this reason, we believe the addition of a thesaurus would bring performance improvements similar to the case of MT.

3.6 Information Retrieval (IR)

The IR task diverges significantly from the tasks for which Bracketing ITGs were developed. The basic model produced a confidence-weighted score of 43.14% (accuracy 46.67%), while the stoplisted model produced a comparable confidence-weighted score of 44.81% (accuracy 47.78%).

Bracketing ITGs seek structurally parallelizable substrings, where there is reason to expect some degree of generalization between the frames (heads and arguments) of the two substrings from a lexical semantics standpoint. In contrast, the IR task relies on unordered keywords, so the effect of argument-head binding cannot be expected to be strong.

3.7 Question Answering (QA)

The QA task is extremely free in the sense that questions can differ significantly from the answers in both syntactic structure and lexis, and can also require a significant degree of indirect complex inference using real-world knowledge. The basic model produced a confidence-weighted score of 33.20% (accuracy 40.77%), while the stoplisted model produced a significantly better confidence-weighted score of 38.26% (accuracy 44.62%).

Aside from adding a thesaurus, to properly model the QA task, at the very least the Bracketing ITG models would need to be augmented with somewhat more linguistic rules that include a proper model for *wh-* words in the Hypothesis, which otherwise cannot be aligned to the Text. In the Bracketing ITG models, the stoplist appears to help by normalizing out the effect of the *wh-* words.

4 Conclusion

The most serious omission in our experiments with Bracketing ITG models was the absence of any thesaurus model, allowing zero lexical variation between the Text and Hypothesis. This forced the models to rely entirely on the Bracketing ITG's inherent tendency to optimize structural match between hypothesized nested argument-head substructures. What we find highly interesting is the perhaps surprisingly large effect obtainable from this structure matching bias alone, which already produces good results on a number of the subsets.

We plan to remedy the absence of a thesaurus as the obvious next step. This can be expected to raise performance significantly on all subsets.

References

Gregor Leusch, Nicola Ueffing, and Hermann Ney. A novel string-to-string distance measure with applications to machine translation evaluation. In *Machine Translation Summit*, New Orleans, 2003.

P. M. Lewis and R. E. Stearns. Syntax-directed transduction. *Journal of the Association for Computing Machinery*, 15:465–488, 1968.

Dekai Wu and Pascale Fung. Inversion Transduction Grammar constraints for mining parallel sentences from quasi-comparable corpora. In *Forthcoming*, 2005.

Dekai Wu. An algorithm for simultaneously bracketing parallel texts by aligning words. In *33rd Annual Meeting of the Association for Computational Linguistics Conference (ACL-95)*, Cambridge, MA, Jun 1995. Association for Computational Linguistics.

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

Kenji Yamada and Kevin Knight. A syntax-based statistical translation model. In *39th Annual Meeting of the Association for Computational Linguistics Conference (ACL-01)*, Toulouse, France, 2001. Association for Computational Linguistics.

Richard Zens and Hermann Ney. A comparative study on reordering constraints in statistical machine translation. pages 192–202, Hong Kong, August 2003.

Richard Zens, Hermann Ney, Taro Watanabe, and Eiichiro Sumita. Reordering constraints for phrase-based statistical machine translation. In *Proceedings of COLING*, Geneva, August 2004.

Hao Zhang and Daniel Gildea. Syntax-based alignment: Supervised or unsupervised? In *Proceedings of COLING*, Geneva, August 2004.