Recognizing Paraphrases and Textual Entailment using 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

We present first results using paraphrase as well as textual entailment data to test the language universal constraint posited by Wu's (1995, 1997) Inversion Transduction Grammar (ITG) hypothesis. In machine translation and alignment, the ITG Hypothesis provides a strong inductive bias, and has been shown empirically across numerous language pairs and corpora to yield both efficiency and accuracy gains for various language acquisition tasks. Monolingual paraphrase and textual entailment recognition datasets, however, potentially facilitate closer tests of certain aspects of the hypothesis than bilingual parallel corpora, which simultaneously exhibit many irrelevant dimensions of cross-lingual variation. We investigate this using simple generic Bracketing ITGs containing no language-specific linguistic knowledge. Experimental results on the MSR Paraphrase Corpus show that, even in the absence of any thesaurus to accommodate lexical variation between the paraphrases, an uninterpolated average precision of at least 76% is obtainable from the Bracketing ITG's structure matching bias alone. This is consistent with experimental results on the Pascal Recognising Textual Entailment Challenge Corpus, which show surpisingly strong results for a number of the task subsets.

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 (Wu, 1997). The ITG Hypothesis has been more extensively studied across different languages, but newly available paraphrase datasets provide intriguing opportunities for meaningful analysis of the ITG Hypothesis in a monolingual setting.

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. For example, 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.

We are conducting a series of investigations 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 monolingual case allows in some sense closer testing of various implications of the ITG hypothesis, without irrelevant dimensions of variation arising from other cross-lingual phenomena.

Asymmetric textual entailment recognition (RTE) datasets, in particular the Pascal Recognising Textual Entailment Challenge Corpus (Dagan *et al.*, 2005), provide testbeds that abstract over many tasks, including information retrieval, comparable documents, reading comprehension, question answering, information extraction, machine translation, and paraphrase acquisition.

At the same time, the emergence of paraphrasing datasets presents an opportunity for complementary experiments on the task of recognizing symmetric bidirectional entailment rather than asymmetric directional entailment. In particular, for this study we employ the MSR Paraphrase Corpus (Quirk *et al.*, 2004).

[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 and Yihai Shen for invaluable assistance in preparing the datasets and stoplist.

Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, pages 25–30,
Ann Arbor, June 2005.

2 Inversion Transduction Grammars

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.

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

> [[[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/ε Authority/管 理 局]$_{\text{NP}}$ [will/将 会 ⟨[be/ε accountable/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 $1, \ldots, N$.

Wu (1997) also showed that ITGs can be equivalently be defined in two other ways. First, ITGs can be defined as the restricted subset of SDTGs where all rules are of rank 2. Second, ITGs can also be defined as the restricted subset of SDTGs where all rules are of rank 3.

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 $A \rightarrow \epsilon/f$).

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

The simplest class of ITGs, *Bracketing ITGs*, are particularly interesting in applications like paraphrasing, 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 paraphrase recognition is that they assign strong preference to candidate paraphrase 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 (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 paraphrase 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 a model for the paraphrase recognition problem that employ simple generic Bracketing ITGs. The experimental results show that, even in the absence of any thesaurus to accommodate lexical variation between the two strings, the Bracketing ITG's structure matching bias alone produces a significant improvement in average precision.

3 Scoring Method

All words of the vocabulary are included among the lexical transductions, allowing exact word matches between the two strings of any candidate paraphrase pair.

Each candidate pair of the test set was scored via the ITG biparsing algorithm, which employs a dynamic programming approach as follows.Let the input English sentence be $\mathbf{e}_1, \ldots, \mathbf{e}_T$ and the corresponding input Chinese sentence be $\mathbf{c}_1, \ldots, \mathbf{c}_V$. As an abbreviation we write $\mathbf{e}_{s..t}$ for the sequence of words $\mathbf{e}_{s+1}, \mathbf{e}_{s+2}, \ldots, \mathbf{e}_t$, and similarly for $\mathbf{c}_{u..v}$; also, $\mathbf{e}_{s..s} = \epsilon$ is the empty string. It is convenient to use a 4-tuple of the form $q = (s, t, u, v)$ to identify each node of the parse tree, where the substrings $\mathbf{e}_{s..t}$ and $\mathbf{c}_{u..v}$ both derive from the node q. Denote the nonterminal label on q by $\ell(q)$. Then for any node $q = (s, t, u, v)$, define

$$\delta_q(i) = \delta_{stuv}(i) = \max_{\text{subtrees of} q} P[\text{subtree of} q, \ell(q) = i, i \stackrel{*}{\Rightarrow} \mathbf{e}_{s..t}/\mathbf{c}_{u..v}]$$

as the maximum probability of any derivation from i that successfully parses both $\mathbf{e}_{s..t}$ and $\mathbf{c}_{u..v}$. Then the best parse of the sentence pair has probability $\delta_{0,T,0,V}(\mathrm{S})$.

The algorithm computes $\delta_{0,T,0,V}(\mathrm{S})$ using the following recurrences. Note that we generalize argmax to the case where maximization ranges over multiple indices, by making it vector-valued. Also note that $[\,]$ and $\langle\rangle$ are simply constants, written mnemonically. The condition $(S-s)(t-S)+(U-u)(v-U) \neq 0$ is a way to specify that the substring in one but not both languages may be split into an empty string ϵ and the substring itself; this ensures that the recursion terminates, but permits words that have no match in the other language to map to an ϵ instead.

1. **Initialization**

$$\begin{aligned}
\delta_{t-1,t,v-1,v}(i) &= b_i(\mathbf{e}_t/\mathbf{c}_v), & 1 \le t \le T \\
& & 1 \le v \le V \\
\delta_{t-1,t,v,v}(i) &= b_i(\mathbf{e}_t/\epsilon), & 1 \le t \le T \\
& & 0 \le v \le V \\
\delta_{t,t,v-1,v}(i) &= b_i(\epsilon/\mathbf{c}_v), & 0 \le t \le T \\
& & 1 \le v \le V
\end{aligned}$$

2. **Recursion** For all i, s, t, u, v such that $\left\{ \begin{array}{c} 1 \le i \le N \\ 0 \le s < t \le T \\ 0 \le u < v \le V \\ t-s+v-u > 2 \end{array} \right.$

$$\begin{aligned}
\delta_{stuv}(i) &= \max[\delta^{[]}_{stuv}(i), \delta^{\langle\rangle}_{stuv}(i)] \\
\theta_{stuv}(i) &= \left\{ \begin{array}{ll} [\,] & \text{if } \delta^{[]}_{stuv}(i) \ge \delta^{\langle\rangle}_{stuv}(i) \\ \langle\rangle & \text{otherwise} \end{array} \right.
\end{aligned}$$

where

$$\delta^{[]}_{stuv}(i) = \max_{\substack{1 \le j \le N \\ 1 \le k \le N \\ s \le S \le t \\ u \le U \le v \\ (S-s)(t-S)+(U-u)(v-U) \neq 0}} a_{i \to [jk]}\, \delta_{sSuU}(j)\, \delta_{StUv}(k)$$

$$\begin{bmatrix} \iota^{[]}_{stuv}(i) \\ \kappa^{[]}_{stuv}(i) \\ \sigma^{[]}_{stuv}(i) \\ \upsilon^{[]}_{stuv}(i) \end{bmatrix} = \operatorname*{argmax}_{\substack{1 \le j \le N \\ 1 \le k \le N \\ s \le S \le t \\ u \le U \le v \\ (S-s)(t-S)+(U-u)(v-U) \neq 0}} a_{i \to [jk]}\, \delta_{sSuU}(j)\, \delta_{StUv}(k)$$

$$\delta^{\langle\rangle}_{stuv}(i) = \max_{\substack{1 \le j \le N \\ 1 \le k \le N \\ s \le S \le t \\ u \le U \le v \\ (S-s)(t-S)+(U-u)(v-U) \neq 0}} a_{i \to \langle jk \rangle}\, \delta_{sSUv}(j)\, \delta_{StuU}(k)$$

$$\begin{bmatrix} \iota^{\langle\rangle}_{stuv}(i) \\ \kappa^{\langle\rangle}_{stuv}(i) \\ \sigma^{\langle\rangle}_{stuv}(i) \\ \upsilon^{\langle\rangle}_{stuv}(i) \end{bmatrix} = \operatorname*{argmax}_{\substack{1 \le j \le N \\ 1 \le k \le N \\ s \le S \le t \\ u \le U \le v \\ (S-s)(t-S)+(U-u)(v-U) \neq 0}} a_{i \to \langle jk \rangle}\, \delta_{sSUv}(j)\, \delta_{StuU}(k)$$

3. **Reconstruction** Initialize by setting the root of the parse tree to $q_1 = (0, T, 0, V)$ and its nonterminal label to $\ell(q_1) = \text{S}$. The remaining descendants in the optimal parse tree are then given recursively for any $q = (s, t, u, v)$ by:

$$\textsc{Left}(q) = \begin{cases} \text{NIL} & \text{if } t-s+v-u \leq 2 \\ (s, \sigma_q^{[]}(\ell(q)), u, v_q^{[]}(\ell(q))) & \text{if } \theta_q(\ell(q)) = [\,] \\ (s, \sigma_q^{\langle\rangle}(\ell(q)), v_q^{\langle\rangle}(\ell(q))) & \text{if } \theta_q(\ell(q)) = \langle\rangle \end{cases}$$

$$\textsc{Right}(q) = \begin{cases} \text{NIL} & \text{if } t-s+v-u \leq 2 \\ (\sigma_q^{[]}(\ell(q)), t, v_q^{[]}(\ell(q)), v) & \text{if } \theta_q(\ell(q)) = [\,] \\ (\sigma_q^{\langle\rangle}(\ell(q)), t, u, v_q^{\langle\rangle}(\ell(q))) & \text{if } \theta_q(\ell(q)) = \langle\rangle \end{cases}$$

$$\ell(\textsc{Left}(q)) = \iota_q^{\theta_q(\ell(q))}(\ell(q))$$

$$\ell(\textsc{Right}(q)) = \kappa_q^{\theta_q(\ell(q))}(\ell(q))$$

As mentioned earlier, biparsing for ITGs can be accomplished efficiently in polynomial time, rather than the exponential time required for classical SDTGs. The result in Wu (1997) implies that for the special case of Bracketing ITGs, the time complexity of the algorithm 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.

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. 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.

4 Experimental Results—Paraphrase Recognition

Our objective here was to isolate the effect of the ITG constraint bias. No training was performed with the available development sets. Rather, the aim was to establish foundational baseline results, to see in this first round of paraphrase recognition experiments what results could be obtained with the simplest versions of the ITG models.

The MSR Paraphrase Corpus test set consists of 1725 candidate paraphrase string pairs, each annotated for semantic equivalence by two or three human collectors. Within the test set, 66.5% of the examples were annotated as being semantically equivalent. The corpus was originally generated via a combination of automatic filtering methods, making it difficult to make specific claims about distributional neutrality, due to the arbitrary nature of the example selection process.

The ITG scoring model produced an uninterpolated average precision (also known as confidence weighted score) of 76.1%. This represents an improvement of roughly 10% over the random baseline. Note that this improvement can be achieved with no thesaurus or lexical similarity model, and no parameter training.

5 Experimental Results—Textual Entailment Recognition

The experimental procedure for the monolingual textual entailment recognition task is the same as for paraphrase recognition, except that one string serves as the Text and the other serves as the Hypothesis.

Results on the textual entailment recognition task are consistent with the above paraphrase recognition results. For the PASCAL RTE challenge datasets, across all subsets overall, the 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%.

For the RTE task we also investigated a second variant of the model, in which a list of 172 words from a stoplist was excluded 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.

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).

5.1 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.

5.2 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.

5.3 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.

5.4 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.

5.5 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.

5.6 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.

5.7 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.

6 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 two strings of a candidate paraphrase pair (or Text and Hypothesis, in the case of textual entailment recognition). 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 paraphrasing as well as a number of the RTE 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.

Wu and Fung (2005) also discuss how to obtain any desired tradeoff between precision and recall. This would be another interesting direction to pursue in the context of recognizing paraphrases or textual entailment.

Finally, using the development sets to train the parameters of the Bracketing ITG model would improve performance. It would only be feasible to tune a few basic parameters, however, given the small size of the development sets.

References

Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *PASCAL Proceedings of the First Challenge Workshop—Recognizing Textual Entailment*, pages 1–8, Southampton, UK, April 2005.

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.

C. Quirk, C. Brockett, and W. B. Dolan. Monolingual machine translation for paraphrase generation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, Barcelona, June 2004. SIGDAT, Association for Computational Linguistics.

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.