The Hiero Machine Translation System: Extensions, Evaluation, and Analysis

David Chiang, Adam Lopez, Nitin Madnani, Christof Monz, Philip Resnik, Michael Subotin
Institute for Advanced Computer Studies (UMIACS)
University of Maryland, College Park, MD 20742, USA
{dchiang,alopez,nmadnani,christof,resnik,msubotin}@umiacs.umd.edu

Abstract

Hierarchical organization is a well known property of language, and yet the notion of hierarchical structure has been largely absent from the best performing machine translation systems in recent community-wide evaluations. In this paper, we discuss a new hierarchical phrase-based statistical machine translation system (Chiang, 2005), presenting recent extensions to the original proposal, new evaluation results in a community-wide evaluation, and a novel technique for fine-grained comparative analysis of MT systems.

1 Introduction

Hierarchical organization is a well known property of language, and yet the notion of hierarchical structure has, for the last several years, been absent from the best performing machine translation systems in community-wide evaluations. Statistical phrase-based models (e.g. (Och and Ney, 2004; Koehn et al., 2003; Marcu and Wong, 2002)) characterize a source sentence f as a flat partition of non-overlapping subsequences, or "phrases", $\bar{f}_1 \cdots \bar{f}_J$, and the process of translation involves selecting target phrases $\bar{e}_i$ corresponding to the $\bar{f}_j$ and modifying their sequential order. The need for some way to model aspects of syntactic behavior, such as the tendency of constituents to move together as a unit, is widely recognized—the role of syntactic units is well attested in recent systematic studies of translation (Fox, 2002; Hwa et al., 2002; Koehn and Knight, 2003), and their absence in phrase-based models is quite evident when looking at MT system output. Nonetheless, attempts to incorporate richer linguistic features have generally met with little success (Och et al., 2004a).

Chiang (2005) introduces Hiero, a hierarchical phrase-based model for statistical machine translation. Hiero extends the standard, non-hierarchical notion of "phrases" to include nonterminal symbols, which permits it to capture both word-level and phrase-level reorderings within the same framework. The model has the formal structure of a synchronous CFG, but it does not make any commitment to a linguistically relevant analysis, and it does not require syntactically annotated training data. Chiang (2005) reported significant performance improvements in Chinese-English translation as compared with Pharaoh, a state-of-the-art phrase-based system (Koehn, 2004).

In Section 2, we review the essential elements of Hiero. In Section 3 we describe extensions to this system, including new features involving named entities and numbers and support for a fourfold scale-up in training set size. Section 4 presents new evaluation results for Chinese-English as well as Arabic-English translation, obtained in the context of the 2005 NIST MT Eval exercise. In Section 5, we introduce a novel technique for fine-grained comparative analysis of MT systems, which we employ in analyzing differences between Hiero's and Pharaoh's translations.

2 Hiero

Hiero is a stochastic synchronous CFG, whose productions are extracted automatically from unannotated parallel texts, and whose rule probabilities form a log-linear model learned by minimum-error-rate training; together with a modified CKY beam-search decoder (similar to that of Wu (1996)). We describe these components in brief below.

Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 779–786, Vancouver, October 2005. ©2005 Association for Computational Linguistics

$$\begin{aligned}
&S \to \langle S_{\boxed{1}} X_{\boxed{2}}, S_{\boxed{1}} X_{\boxed{2}} \rangle \\
&S \to \langle X_{\boxed{1}}, X_{\boxed{1}} \rangle \\
&X \to \langle \text{yu } X_{\boxed{1}} \text{ you } X_{\boxed{2}}, \text{have } X_{\boxed{2}} \text{ with } X_{\boxed{1}} \rangle \\
&X \to \langle X_{\boxed{1}} \text{ de } X_{\boxed{2}}, \text{the } X_{\boxed{2}} \text{ that } X_{\boxed{1}} \rangle \\
&X \to \langle X_{\boxed{1}} \text{ zhiyi}, \text{one of } X_{\boxed{1}} \rangle \\
&X \to \langle \text{Aozhou}, \text{Australia} \rangle \\
&X \to \langle \text{shi}, \text{is} \rangle \\
&X \to \langle \text{shaoshu guojia}, \text{few countries} \rangle \\
&X \to \langle \text{bangjiao}, \text{diplomatic relations} \rangle \\
&X \to \langle \text{Bei Han}, \text{North Korea} \rangle
\end{aligned}$$

Figure 1: Example synchronous CFG

2.1 Grammar

A *synchronous CFG* or *syntax-directed transduction grammar* (Lewis and Stearns, 1968) consists of pairs of CFG rules with aligned nonterminal symbols. We denote this alignment by coindexation with boxed numbers (Figure 1). A derivation starts with a pair of aligned start symbols, and proceeds by rewriting pairs of aligned nonterminal symbols using the paired rules (Figure 2).

Training begins with phrase pairs, obtained as by Och, Koehn, and others: GIZA++ (Och and Ney, 2000) is used to obtain one-to-many word alignments in both directions, which are combined into a single set of refined alignments using the "final-and" method of Koehn et al. (2003); then those pairs of substrings that are exclusively aligned to each other are extracted as phrase pairs.

Then, synchronous CFG rules are constructed out of the initial phrase pairs by subtraction: every phrase pair $\langle \bar{f}, \bar{e} \rangle$ becomes a rule X → $\langle \bar{f}, \bar{e} \rangle$, and a phrase pair $\langle \bar{f}, \bar{e} \rangle$ can be subtracted from a rule X → $\langle \gamma_1 \bar{f} \gamma_2, \alpha_1 \bar{e} \alpha_2 \rangle$ to form a new rule X → $\langle \gamma_1 X_{\boxed{i}} \gamma_2, \alpha_1 X_{\boxed{i}} \alpha_2 \rangle$, where i is an index not already used. Various filters are also applied to reduce the number of extracted rules. Since one of these filters restricts the number of nonterminal symbols to two, our extracted grammar is equivalent to an inversion transduction grammar (Wu, 1997).

2.2 Model

The model is a log-linear model (Och and Ney, 2002) over synchronous CFG derivations. The weight of a derivation is $P_{LM}(e)^{\lambda_{LM}}$, the weighted language model probability, multiplied by the product of the weights of the rules used in the derivation. The weight of each rule is, in turn:

$$(1) \qquad w(X \to \langle \gamma, \alpha \rangle) = \prod_i \phi_i(X \to \langle \gamma, \alpha \rangle)^{\lambda_i}$$

where the ϕ_i are features defined on rules. The basic model uses the following features, analogous to Pharaoh's default feature set:

- $P(\gamma \mid \alpha)$ and $P(\alpha \mid \gamma)$
- the lexical weights $P_w(\gamma \mid \alpha)$ and $P_w(\alpha \mid \gamma)$ (Koehn et al., 2003);[1]
- a phrase penalty exp(1);
- a word penalty $\exp(l)$, where l is the number of terminals in α.

The exceptions to the above are the two "glue" rules, which are the rules with left-hand side S in Figure 1. The second has weight one, and the first has weight $w(S \to \langle S_{\boxed{1}} X_{\boxed{2}}, S_{\boxed{1}} X_{\boxed{2}} \rangle) = \exp(-\lambda_g)$, the idea being that parameter λ_g controls the model's preference for hierarchical phrases over serial combination of phrases.

Phrase translation probabilities are estimated by relative-frequency estimation. Since the extraction process does not generate a unique derivation for each training sentence pair, a distribution over possible derivations is hypothesized, which gives uniform weight to all initial phrases extracted from a sentence pair and uniform weight to all rules formed out of an initial phrase. This distribution is then used to estimate the phrase translation probabilities.

The lexical-weighting features are estimated using a method similar to that of Koehn et al. (2003). The language model is a trigram model with modified Kneser-Ney smoothing (Chen and Goodman, 1998), trained using the SRI-LM toolkit (Stolcke, 2002).

[1] This feature uses word alignment information, which is discarded in the final grammar. If a rule occurs in training with more than one possible word alignment, Koehn et al. take the maximum lexical weight; Hiero uses a weighted average.

$$\begin{aligned}
\langle \mathrm{S}_{\boxed{1}}, \mathrm{S}_{\boxed{1}}\rangle &\Rightarrow \langle \mathrm{S}_{\boxed{2}}\mathrm{X}_{\boxed{3}}, \mathrm{S}_{\boxed{2}}\mathrm{X}_{\boxed{3}}\rangle \\
&\Rightarrow \langle \mathrm{S}_{\boxed{4}}\mathrm{X}_{\boxed{5}}\mathrm{X}_{\boxed{3}}, \mathrm{S}_{\boxed{4}}\mathrm{X}_{\boxed{5}}\mathrm{X}_{\boxed{3}}\rangle \\
&\Rightarrow \langle \mathrm{X}_{\boxed{6}}\mathrm{X}_{\boxed{5}}\mathrm{X}_{\boxed{3}}, \mathrm{X}_{\boxed{6}}\mathrm{X}_{\boxed{5}}\mathrm{X}_{\boxed{3}}\rangle \\
&\Rightarrow \langle \text{Aozhou } \mathrm{X}_{\boxed{5}}\mathrm{X}_{\boxed{3}}, \text{Australia } \mathrm{X}_{\boxed{5}}\mathrm{X}_{\boxed{3}}\rangle \\
&\Rightarrow \langle \text{Aozhou shi } \mathrm{X}_{\boxed{3}}, \text{Australia is } \mathrm{X}_{\boxed{3}}\rangle \\
&\Rightarrow \langle \text{Aozhou shi } \mathrm{X}_{\boxed{7}} \text{ zhiyi, Australia is one of } \mathrm{X}_{\boxed{7}}\rangle \\
&\Rightarrow \langle \text{Aozhou shi } \mathrm{X}_{\boxed{8}} \text{ de } \mathrm{X}_{\boxed{9}} \text{ zhiyi, Australia is one of the } \mathrm{X}_{\boxed{9}} \text{ that } \mathrm{X}_{\boxed{8}}\rangle \\
&\Rightarrow \langle \text{Aozhou shi yu } \mathrm{X}_{\boxed{1}} \text{ you } \mathrm{X}_{\boxed{2}} \text{ de } \mathrm{X}_{\boxed{9}} \text{ zhiyi, Australia is one of the } \mathrm{X}_{\boxed{9}} \text{ that have } \mathrm{X}_{\boxed{2}} \text{ with } \mathrm{X}_{\boxed{1}}\rangle
\end{aligned}$$

Figure 2: Example partial derivation of a synchronous CFG.

The feature weights are learned by maximizing the BLEU score (Papineni et al., 2002) on held-out data, using minimum-error-rate training (Och, 2003) as implemented by Koehn. The implementation was slightly modified to ensure that the BLEU scoring matches NIST's definition and that hypotheses in the *n*-best lists are merged when they have the same translation and the same feature vector.

3 Extensions

In this section we describe our extensions to the base Hiero system that improve its performance significantly. First, we describe the addition of two new features to the Chinese model, in a manner similar to that of Och et al. (2004b); then we describe how we scaled the system up to a much larger training set.

3.1 New features

The LDC Chinese-English named entity lists (900k entries) are a potentially valuable resource, but previous experiments have suggested that simply adding them to the training data does not help (Vogel et al., 2003). Instead, we placed them in a supplementary phrase-translation table, giving greater weight to phrases that occurred less frequently in the primary training data. For each entry $\langle f, \{e_1, \ldots, e_n\}\rangle$, we counted the number of times $c(f)$ that f appeared in the primary training data, and assigned the entry the weight $\frac{1}{c(f)+1}$, which was then distributed evenly among the supplementary phrase pairs $\{\langle f, e_i\rangle\}$. We then created a new model feature for named entities. When one of these supplementary phrase pairs was used in translation, its feature value for the named-entity feature was the weight defined above, and its value in the other phrase-translation and lexical-weighting features was zero. Since these scores belonged to a separate feature from the primary translation probabilities, they could be reweighted independently during minimum-error-rate training.

Similarly, to process Chinese numbers and dates, we wrote a rule-based Chinese number/date translator, and created a new model feature for it. Again, the weight given to this module was optimized during minimum-error-rate training. In some cases we wrote the rules to provide multiple uniformly-weighted English translations for a Chinese phrase (for example, 八日 (*bari*) could become "the 8th" or "on the 8th"), allowing the language model to decide between the options.

3.2 Scaling up training

Chiang (2005) reports on experiments in Chinese-English translation using a model trained on 7.2M+9.2M words of parallel data.[2] For the NIST MT Eval 2005 large training condition, considerably more data than this is allowable. We chose to use only newswire data, plus data from Sinorama, a Taiwanese news magazine.[3] This amounts to almost 30M+30M words. Scaling to this set required reducing the initial limit on phrase lengths, previously fixed at 10, to avoid explosive growth of

[2] Here and below, the notation "$X + Y$ words" denotes X words of foreign text and Y words of English text.

[3] From Sinorama, only data from 1991 and later were used, as articles prior to that were translated quite loosely.

the extracted grammar. However, since longer initial phrases can be beneficial for translation accuracy, we adopted a variable length limit: 10 for the FBIS corpus and other mainland newswire sources, and 7 for the HK News corpus and Sinorama. (During decoding, limits of up to 15 were sometimes used; in principle these limits should all be the same, but in practice it is preferable to tune them separately.)

For Arabic-English translation, we used the basic Hiero model, without special features for named entities or numbers/dates. We again used only the newswire portions of the allowable training data; we also excluded the Ummah data, as the translations were found to be quite loose. Since this amounted to only about 1.5M+1.5M words, we used a higher initial phrase limit of 15 during both training and decoding.

4 Evaluation

Figure 1 shows the performance of several systems on NIST MT Eval 2003 Chinese test data: Pharaoh (2004 version), trained only on the FBIS data; Hiero, with various combinations of the new features and the larger training data.[4] This table also shows Hiero's performance on the NIST 2005 MT evaluation task.[5] The metric here is case-sensitive BLEU.[6]

Figure 2 shows the performance of two systems on Arabic in the NIST 2005 MT Evaluation task: DC, a phrase-based decoder for a model trained by Pharaoh, and Hiero.

System	Features	Train	MT03	MT05
Pharaoh	standard	FBIS	0.268	
Hiero	standard	FBIS	0.288	
Hiero	standard	full	0.329	
Hiero	+nums, names	full	0.339	0.300

Table 1: Chinese results. (BLEU-4; MT03 case-insensitive, MT05 case-sensitive)

System	Train	MT05
DC	full	0.399
Hiero	full	0.450

Table 2: Arabic results. (BLEU-4; MT03 case-insensitive, MT05 scores case-sensitive.

5 Analysis

Over the last few years, several automatic metrics for machine translation evaluation have been introduced, largely to reduce the human cost of iterative system evaluation during the development cycle (Lin and Och, 2004; Melamed et al., 2003; Papineni et al., 2002). All are predicated on the concept of *n*-gram matching between the sentence hypothesized by the translation system and one or more *reference translations*—that is, human translations for the test sentence. Although the motivations and formulae underlying these metrics are all different, ultimately they all produce a single number representing the "goodness" of the MT system output over a set of reference documents. This facility is valuable in determining whether a given system modification has a positive impact on overall translation performance. However, the metrics are all holistic. They provide no insight into the specific competencies or weaknesses of one system relative to another.

Ideally, we would like to use automatic methods to provide immediate diagnostic information about the translation output—*what* the system does well, and what it does poorly. At the most general level, we want to know how our system performs on the two most basic problems in translation—word translation and reordering. Unigram precision and recall statistics tell us something about the performance of an MT system's internal translation dictionaries, but nothing about reordering. It is thought that higher order *n*-grams correlate with the reordering accuracy of MT systems, but this is again a holistic metric.

What we would really like to know is how well the system is able to capture systematic reordering patterns in the input, which ones it is successful with, and which ones it has difficulty with. Word *n*-grams are little help here: they are too many, too sparse, and it is difficult to discern general patterns from them.

[4] The third line, corresponding to the model without new features trained on the larger data, may be slightly depressed because the feature weights from the fourth line were used instead of doing minimum-error-rate training specially for this model.

[5] Full results are available at `http://www.nist.gov/speech/tests/summaries/2005/mt05.htm`. For this test, a phrase length limit of 15 was used during decoding.

[6] For this task, the translation output was uppercased using the SRI-LM toolkit: essentially, it was decoded again using an HMM whose states and transitions are a trigram language model of cased English, and whose emission probabilities are reversed, i.e., probability of cased word given lowercased word.

5.1 A New Analysis Method

In developing a new analysis method, we are motivated in part by recent studies suggesting that word reorderings follow general patterns with respect to syntax, although there remains a high degree of flexibility (Fox, 2002; Hwa et al., 2002). This suggests that in a comparative analysis of two MT systems, it may be useful to look for syntactic patterns that one system captures well in the target language and the other does not, using a syntax based metric.

We propose to summarize reordering patterns using part-of-speech sequences. Unfortunately, recent work has shown that applying statistical parsers to ungrammatical MT output is unreliable at best, with the parser often assigning unreasonable probabilities and incongruent structure (Yamada and Knight, 2002; Och et al., 2004a). Anticipating that this would be equally problematic for part-of-speech tagging, we make the conservative choice to apply annotation only to the reference corpus. Word *n*-gram correspondences with a reference translation are used to infer the part-of-speech tags for words in the system output.

First, we tagged the reference corpus with parts of speech. We used MXPOST (Ratnaparkhi, 1996), and in order to discover more general patterns, we map the tag set down after tagging, e.g. NN, NNP, NNPS and NNS all map to NN. Second, we computed the frequency $\mathit{freq}(t_i \ldots t_j)$ of every possible tag sequence $t_i \ldots t_j$ in the reference corpus. Third, we computed the correspondence between each hypothesis sentence and *each* of its corresponding reference sentences using an approximation to maximum matching (Melamed et al., 2003). This algorithm provides a list of *runs* or contiguous sequences of words $e_i \ldots e_j$ in the reference that are also present in the hypothesis. (Note that runs are order-sensitive.) Fourth, for each recalled *n*-gram $e_i \ldots e_j$, we looked up the associated tag sequence $t_i \ldots t_j$ and incremented a counter $\mathit{recalled}(t_i \ldots t_j)$. Finally, we computed the recall of tag patterns, $R(t_i \ldots t_j) = \mathit{recalled}(t_i \ldots t_j)/\mathit{freq}(t_i \ldots t_j)$, for all patterns in the corpus.

By examining examples of these tag sequences in the reference corpus and their hypothesized translations, we expect to gain some insight into the comparative strengths and weaknesses of the MT systems' reordering models. (An interactive platform for this analysis is demonstrated by Lopez and Resnik (2005).)

5.2 Chinese

We performed tag sequence analysis on the Hiero and Pharaoh systems trained on the FBIS data only. Table 3 shows those *n*-grams for which Hiero and Pharaoh's recall differed significantly ($p < 0.01$). The numbers shown are the ratio of Hiero's recall to Pharaoh's. Note that the *n*-grams on which Hiero had better recall are dominated by fragments of prepositional phrases (in the Penn Treebank tagset, prepositions are tagged IN or TO).

Our hypothesis is that Hiero produces English PPs better because many of them are translated from Chinese phrases which have an NP modifying an NP to its right, often connected with the particle 的 (*de*). These are often translated into English as PPs, which modify the NP to the left. A correct translation, then, would have to reorder the two NPs. Notice in the table that Hiero recalls proportionally more *n*-grams as *n* increases, corroborating the intuition that Hiero should be better at longer-distance reorderings.

Investigating this hypothesis qualitatively, we inspected the first five occurrences of the *n*-grams of the first type on the list (JJ NN IN DT NN). Of these, we omit one example because both systems recalled the *n*-gram correctly, and one because they differed only in lexical choice (Hiero matched the 5-gram with one reference sentence, Pharaoh with zero). The other three examples are shown below (H = Hiero, P = Pharaoh):

(2)

联合国	安全	理事会	的	五个	常任
UN	security	council	of	five	permanent
理事	国都				
member	countries-all				

R1. five permanent members of the UN Security Council

H. the five permanent members of the un security council

P. the united nations security council permanent members of the five countries

10.00	JJ NN IN DT NN	3.14	DT NN IN DT NN	1.46	DT NN PU	1.09	CD
7.00	IN NN TO	3.00	IN DT NN PU	1.44	IN DT JJ	1.07	VB
5.50	IN DT NN NN PU NN	2.50	NN TO NN	1.42	NN IN DT	1.06	NN NN
5.50	IN DT NN NN PU NN NN	2.03	DT JJ NN IN	1.41	IN DT NN	1.06	IN
4.50	NN JJ NN PU	1.95	IN NN PU	1.37	PU CC	1.05	NN
4.50	NN IN DT JJ	1.77	IN NN CD	1.34	IN CD	0.61	RB CD
4.00	VB CD IN DT	1.74	DT NN IN NN	1.32	JJ NN PU	0.21	TO VB PR
3.67	IN DT NN NN PU	1.70	JJ NN IN	1.30	IN NN	0.18	PU RB CD
3.50	NN IN DT NN NN	1.55	VB IN DT	1.29	NN IN	0.12	NN CD TO NN
3.30	NN IN DT NN	1.46	NN IN NN	1.18	NN PU	0.12	CD TO NN

Table 3: Chinese-English POS *n*-grams on which Hiero and Pharaoh had significantly different recall, arranged by recall ratio. Ratio > 1 indicates tag sequences that Hiero matched more frequently.

(3) 伊拉克 危机 的 最 新 发展
Iraq crisis of most new development

R1. the latest development on the Iraqi crisis

H. the latest development on the Iraqi crisis

P. on the iraqi crisis, the latest development

(4) 今年 上 半年
this-year upper half-year

R1. the first half of this year

H. the first half of this year

P. the first half of

All three of these examples involve an NP modifying an NP to its right; two with the particle 的 (*de*) and one without. In all three cases Hiero reorders the NPs correctly; Pharaoh preserves the Chinese word order in two cases, but in the third, for reasons not understood, drops the modifying NP.

The *n*-grams on which Hiero did worse than Pharaoh mostly involved numbers; here a pattern is not as easily discernible, but there are several cases where Hiero makes errors in translating numbers (neither system in this comparison used the dedicated number translator). For the *n*-gram TO VB PR, it seems Hiero has a tendency to delete possessive pronouns (PR, abbreviated from PRP$).

5.3 Arabic

Initial inspection of the *n*-grams on which Hiero showed significantly higher recall in the Arabic-English task suggested that here, too, better translation of nominal phrases may be at play. We investigated this conjecture further by examining several *n*-gram sets with the highest recall ratios. Some of them on closer inspection turned out to conflate different structural patterns, and provided little interpretable information. However, the 8 sentences in the *n*-gram list IN DT JJ JJ showed a degree of structural consistency. The list contained 6 instances where Hiero performed better in translating a complex NP or PP, one instance in which DC performed better in translating a complex PP, and one case in which they both performed equally poorly. Below we show two examples of phrases on which Hiero performed better, and the one example on which its hierarchical approach produced undesirable results (H = Hiero, D = DC).

(5) Al wjwd Al EskrY Al AmYrkY fY Al
the presence the military the American in the
mnTqp
region

R1. the American military presence in the region

H. the american military presence in the region

D. the military presence in the region

(6) AltY tSnEhA Al $rkp Al
which manufactures-them the company the
kwrYp Al jnwbYp
Korean the Southern

R1. which are manufactured by the South Korean company

H. which are manufactured by the south korean company

D. which are manufactured by the company , the south korean

8.00	WR DT NN	2.38	DT JJ JJ NN	1.46	JJ NN NN		
8.00	PR NN IN DT	2.08	CC JJ NN	1.43	JJ JJ	1.02	NN
7.00	DT PU	2.01	PR VB	1.35	IN DT JJ	0.47	NN CD PU CD NN NN
6.00	DT NN NN PO	2.00	TO DT NN NN	1.24	VB IN	0.47	NN CD PU CD NN NN NN
5.00	IN DT JJ JJ	1.80	NN PU WD	1.21	NN VB	0.47	NN CD PU CD NN NN NN PU
4.67	DT NN IN VB	1.80	NN IN DT JJ NN	1.20	NN IN DT	0.45	NN CD PU CD NN
2.89	NN NN NN VB	1.77	NN IN DT JJ	1.17	PR	0.29	NN CD NN
2.73	PR VB IN	1.76	JJ JJ NN	1.10	JJ NN	0.27	NN CD NN CD
2.56	NN PU WD VB	1.74	VB CD	1.08	NN NN	0.09	NN CD NN PU
2.45	JJ CC JJ NN	1.68	NN NN VB	1.07	IN DT		

Table 4: Arabic-English POS *n*-grams on which Hiero and DC had significantly different recall, arranged by recall ratio. Ratio > 1 indicates tag sequences that Hiero matched more frequently.

(7) swq Al EqArAt fY Akbr mdYnp
market the real-estate in largest city
SnAEYp SYnYp $AnghAY
industrial Chinese Shanghai

R2. The real estate market in the largest Chinese industrial city , Shanghai

H. chinese real estate market in the largest industrial city shanghai

D. real estate market in the largest chinese industrial city shanghai

In the last example we see that Hiero mistakenly identified the adjective “Chinese” as modifying the highest head of the first NP in the apposition.

The style of Arabic newswire tends strongly towards the verb-initial word order in the main clause. Based on our inspection of the *n*-gram collection NN VB, we were also able to note that Hiero performed noticeably better in reordering the subject and main verb to produce idiomatic English translations. Although in this set the differences in the recall for the NN VB bigram were influenced by many different translation issues, reordering the subject and main verbs was the only structural pattern that recurred consistently throughout the set, appearing in 8 of the 29 relevant sentences.

(8) wqAl Al bYAn An
and-said the statement that

R1. The statement said

H. the statement said that

D. said a statement that

(9) AEln ms&wl fY Al Amm Al mtHdp
announced official in the nations the united
An
that

R1. A United Nations official announced that

H. the united nations official announced that

D. an official in the united nations that

Looking at the bottom of the list, we find more examples of how Hiero’s reordering behavior sometimes backfires. These *n*-grams seem primarily to be parts of bylines, where Hiero has a tendency to reformat the date, whereas DC keeps the original format, matching more often.

(10) mAnYlA 26 YnAYr
Manila 26 January

R3. Manila 26 January

H. manila , january 26

P. manila 26 january

6 Conclusions

The work reported in this paper extends the original treatment of Hiero (Chiang, 2005) by evaluating an improved version in a community-wide exercise for Chinese-English and Arabic-English translation, and by introducing a novel analysis technique for comparing MT systems’ output. The evaluation results provide strong evidence that the approach gains performance from its hierarchical extensions to phrase-based translation. The analysis of part-of-speech tag sequences provides a way to perform finer-grained comparison of system output, pinpointing phenomena for which the systems differ significantly.

Acknowledgements

We would like to thank Philipp Koehn for the use of the Pharaoh system. This research was supported in part by ONR MURI Contract FCPO.810548265 and Department of Defense contract RD-02-5700.

References

Stanley F. Chen and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Harvard University Center for Research in Computing Technology.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 263–270.

Heidi J. Fox. 2002. Phrasal cohesion and statistical machine translation. In *Proceedings of EMNLP 2002*, pages 304–311.

Rebecca Hwa, Philip Resnik, Amy Weinberg, and Okan Kolak. 2002. Evaluating translational correspondence using annotation projection. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 392–399.

Philipp Koehn and Kevin Knight. 2003. Feature-rich statistical translation of noun phrases. In *Proceedings of the 41st Annual Meeting of the ACL*, pages 311–318.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL 2003*, pages 127–133.

Philipp Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Proceedings of AMTA 2004*, pages 115–124.

P. M. Lewis II and R. E. Stearns. 1968. Syntax-directed transduction. *Journal of the ACM*, 15:465–488.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the ACL*, pages 606–613.

Adam Lopez and Philip Resnik. 2005. Pattern visualization for machine translation output. In *Proceedings of HLT/EMNLP 2005*. Demonstration session.

Daniel Marcu and William Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of EMNLP 2002*, pages 133–139.

I. Dan Melamed, Ryan Green, and Joseph P. Turian. 2003. Precision and recall of machine translation. In *Proceedings of HLT-NAACL 2003*, pages 61–63.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the ACL*, pages 440–447.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 295–302.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30:417–449.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2004a. A smorgasbord of features for statistical machine translation. In *Proceedings of HLT-NAACL 2004*.

Franz Josef Och, Ignacio Thayer, Daniel Marcu, Kevin Knight, Dragos Stefan Munteanu, Quamrul Tipu, Michel Galley, and Mark Hopkins. 2004b. Arabic and Chinese MT at USC/ISI. Presentation given at NIST Machine Translation Evaluation Workshop.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the ACL*, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 311–318.

Adwait Ratnaparkhi. 1996. A maximum-entropy model for part-of-speech tagging. In *Proceedings of EMNLP*, pages 133–142.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 901–904.

Stephan Vogel, Ying Zhang, Fei Huang, Alicia Tribble, Ashish Venugopal, Bing Zhao, and Alex Waibel. 2003. The CMU statistical machine translation system. In *Proceedings of MT-Summit IX*, pages 402–409.

Dekai Wu. 1996. A polynomial-time algorithm for statistical machine translation. In *Proceedings of the 34th Annual Meeting of the ACL*, pages 152–158.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23:377–404.

Kenji Yamada and Kevin Knight. 2002. A decoder for syntax-based statistical MT. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 303–310.