Are Multiple Reference Translations Necessary? Investigating the Value of Paraphrased Reference Translations in Parameter Optimization

Nitin Madnani[§], Philip Resnik[§], Bonnie J. Dorr[§] & Richard Schwartz[†]
[§]Laboratory for Computational Linguistics and Information Processing
[§]Institute for Advanced Computer Studies
[§]University of Maryland, College Park
[†]BBN Technologies
{nmadnani,resnik,bonnie}@umiacs.umd.edu schwartz@bbn.com

Abstract

Most state-of-the-art statistical machine translation systems use log-linear models, which are defined in terms of hypothesis features and weights for those features. It is standard to tune the feature weights in order to maximize a translation quality metric, using held-out test sentences and their corresponding reference translations. However, obtaining reference translations is expensive. In our earlier work (Madnani et al., 2007), we introduced a new full-sentence paraphrase technique, based on English-to-English decoding with an MT system, and demonstrated that the resulting paraphrases can be used to cut the number of human reference translations needed in half. In this paper, we take the idea a step further, asking how far it is possible to get with just a single good reference translation for each item in the development set. Our analysis suggests that it is necessary to invest in four or more human translations in order to significantly improve on a single translation augmented by monolingual paraphrases.

1 Introduction

Most state-of-the-art statistical machine translation systems use log-linear models, which are defined in terms of hypothesis features and weights for those features. Such models usually take the form

$$\sum_i \lambda_i h_i(\bar{f}, \bar{e}) \quad (1)$$

where h_i are features of the hypothesis e and λ_i are weights associated with those features.

It is standard practice to tune the feature weights in models of this kind in order to maximize a translation quality metric such as BLEU (Papineni et al., 2002) or TER (Snover et al., 2006), using held-out "development" sentences paired with their corresponding reference translations. Och (2003) showed that system achieves its best performance when the model parameters are tuned using the same objective function being used for evaluating the system. However, this reliance on multiple reference translations creates a problem, because reference translations are labor intensive and expensive to obtain. For example, producing reference translations at the Linguistic Data Consortium, a common source of translated data for MT research, requires undertaking an elaborate process that involves translation agencies, detailed translation guidelines, and quality control processes (Strassel et al., 2006).

In our previous work (Madnani et al., 2007), we introduced an automatic paraphrasing technique based on English-to-English translation of full sentences using a statistical MT system, and demonstrated that, using this technique in the context of parameter tuning, it is possible to cut in half the usual number of reference translations used—when each of two human reference translations is paraphrased automatically, tuning on the resulting four translations yields translation performance that is no worse than that obtained using four human translations. Our method enables the generation of paraphrases for thousands of sentences in a very short amount of time (much shorter than creating other low-cost human references).

In this paper, we take the idea a step further, ask-

ing how far it is possible to get with just a *single* good reference translation for each item in the development set. This question is important for a number of reasons. First, with a few exceptions — notably NIST's annual MT evaluations — most new MT research data sets are provided with only a single reference translation. Second, obtaining multiple reference translations in rapid development, low-density source language scenarios (e.g. (Oard, 2003)) is likely to be severely limited (or made entirely impractical) by limitations of time, cost, and ready availability of qualified translators. Finally, if a single good reference translation turns out to suffice for parameter tuning, this opens the door to future investigations in which we ask *how* good such translations need to be. Ultimately, it may be possible to remove human development-set translations from the statistical MT process altogether, instead simply holding out a subset of sentence pairs that are already part of the training bitext.

The next section lays out the critical research questions that we wish to address in this work. Section 3 describes the paraphrasing model that we used for the experiments in this paper. Section 4 presents experimentation and results, followed by discussion and conclusions in Section 5.

2 Research Questions

There are a number of important research questions that need to be answered in order to determine whether it is feasible to eliminate the need for multiple reference translations, using automatic paraphrases of a single reference translation instead.

1. If only a single reference translation is available for tuning, can adding a paraphrased reference provide significant gains?

2. Can k-best paraphrasing instead of just 1-best lead to better optimization, and how does this compare with using additional human references translations?

3. Does the full-sentence paraphraser always need to be trained on *all* of the training data being used by the MT system (as it was in our previous work) or can it be trained on only a subset of the data? The answer to this question is essential to test the hypothesis that the paraphraser may not actually be producing the claimed n-gram diversity but just performing a form of smoothing over the feature value estimates.

4. To what extent are the gains obtained from this technique contingent on the quality of the human references that are being paraphrased, if at all?

5. How severely does the genre mismatch affect any gains that are to be had? For example, can using paraphrased references still provide large gains if the validation set is of a different genre than the one that the paraphraser is trained on?

6. Given the claim that the paraphraser provides additional n-gram diversity, can it be useful in situations where the tuning criterion does not depend heavily on such overlap?

Answering these questions will make it possible to characterize the utility of paraphrase-based optimization in real-world scenarios, and and how best to leverage it in those scenarios where it does prove useful.

3 Paraphrasing Model

We generate sentence-level paraphrases via English-to-English translation using phrase table pivoting, following (Madnani et al., 2007). The translation system we use (for both paraphrase generation and translation) is based on a state-of-the-art hierarchical phrase-based translation model as described in (Chiang, 2007). English-to-English hierarchical phrases are induced using the pivot-based technique proposed in (Bannard and Callison-Burch, 2005) with primary features similar to those used by (Madnani et al., 2007): the joint probability $p(\bar{e_1}, \bar{e_2})$, the two conditionals $p(\bar{e_1}|\bar{e_2})$ & $p(\bar{e_2}|\bar{e_1})$ and the target length.

To limit noise during pivoting, we only keep the top 20 paraphrase pairs resulting from each pivot, as determined by the induced fractional counts.

Furthermore, we pre-process the source to identify all named entities using BBN IdentiFinder (Bikel et al., 1999) and strongly bias our decoder to leave them unchanged against during the

paraphrasing (translation) process to avoid any erroneous paraphrasing of entities.

4 Experiments

Before presenting paraphrase-based tuning experiments, we outline some general information that is common to all of the experiments described below:

- We choose Chinese-English translation as our test-bed since there are sufficient resources available in this language pair to conduct all of our desired experiments.
- Unless otherwise specified, we use 2 million sentences of newswire text as our training corpus for the Chinese-English MT system for all experiments but train the paraphraser only on a subset—1 million sentences—instead of the full set.
- We use a 1-3 split of the 4 reference translations from the NIST MT02 test set to tune the feature weights for the paraphraser similar to Madnani et al. (2007).
- No changes are made to the number of references in any validation set. Only the tuning sets differed in the number of references across different experiments.
- BLEU and TER are calculated on lowercased translation output. Brevity penalties for BLEU are indicated if not equal to 1.
- For each experiment, BLEU scores shown in bold are significantly better (Koehn, 2004) than the appropriate baselines for that experiment ($p < 0.05$).

4.1 Single Reference Datasets

In this section, we attempt to gauge the utility of the paraphrase approach in a realistic scenario where only a single reference translation is available for the tuning set. We use the NIST MT03 data, which has four references per development item, to simulate a tuning set in which only a single reference translation is available.[1].

[1]The reasons for choosing a set with 4 references will become clear in Section 4.2

One way to create such a simulated set is simply to choose one of the 4 reference sets, i.e., all the translations with the same system identifier for all source documents in the set. However, for the NIST sets, each of the reference sets is typically created by a different human translator. In order to imitate a more realistic scenario where multiple human translators collaborate to produce a single set of reference translations instead of multiple sets, it is essential to normalize over any translator idiosyncrasies so as to avoid any bias. Therefore, we create the simulated single-reference set by choosing, at random, for *each* source document in the set, one of the 4 available reference translations.

As our baseline, we use this simulated single-reference set as the tuning set (1H=1 Human) and evaluate on a held-out validation set consisting of both the NIST MT04 and MT05 data sets (a total of 2870 sentences), hereafter referred to as MT04+05. We then paraphrase the simulated set, extract the 1-best paraphrase as an additional reference, and tune the MT system on this new 2 reference tuning set (1H+1P=1 Human, 1 Paraphrase).

The results, shown in Table 1, confirm that using a paraphrased reference when only a single human reference is available is extremely useful and leads to huge gains in both the BLEU and TER scores on the validation set. In addition, since we see gains despite the fact that the paraphraser is only trained on half of the MT training corpus, we can conclude that these improvements are not the result of fortuitous smoothing, but rather of increased n-gram diversity on the target side of the development set.

Table 1: BLEU and TER scores are shown for MT04+05. 1H=Tuning with 1 human reference, 1H+1P=Tuning with the human reference *and* its paraphrase. Lower TER scores are better.

	BLEU	TER
1H	37.65	56.39
1H+1P	**39.32**	54.39

4.2 Using k-best Paraphrases

Since the paraphraser is an English-to-English SMT system, it can generate n-best hypothesis paraphrases from the chart for each source sentence. An obvious extension to the above experiment then is to

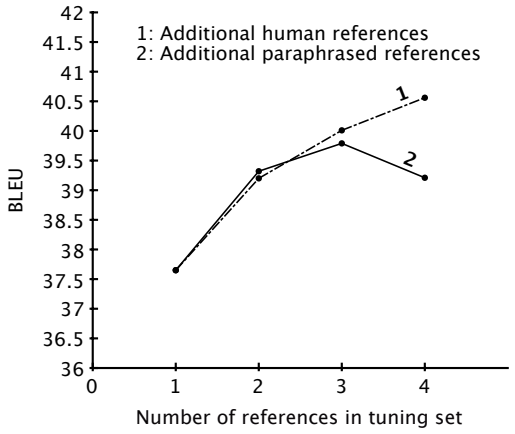

Figure 1: Using the k-best paraphrases are added as references, the graph depicts MT04+05 BLEU scores as additional references—human and paraphrased—are added to the single reference tuning set.

see whether using k-best paraphrase hypotheses as additional reference translations, instead of just the 1-best, can alleviate the reference sparsity to a larger extent during the optimization process. For this experiment, we use the top 1, 2 and 3 paraphrases for the MT03 simulated single reference set as additional references; three tuning sets 1H+1P, 1H+2P and 1H+3P respectively. As points of comparison, we also construct the tuning sets 2H, 3H and 4H from MT03 in the same simulated fashion[2] as the single reference tuning set 1H. The results for this experiment are shown in Figure 1.

Table 2: MT04+05 BLEU and TER scores are shown, as additional references—human and paraphrased—are added to the single reference tuning set.

# tuning refs	Human		Paraphrased	
	BLEU	TER	BLEU	TER
1 (1H+0)	37.65	56.39	37.65	56.39
2 (1H+1)	**39.20**	54.48	**39.32**	54.39
3 (1H+2)	**40.01**	53.50	**39.79**	53.71
4 (1H+3)	**40.56**	53.31	**39.21**	53.46

The graph shows that starting from the simulated single reference set, adding one more human reference translation leads to a significant gain in BLEU score, and adding more human references provides smaller but consistent gains at each step. Table 2 shows the BLEU and TER scores corresponding to Figure 1. With paraphrased references, gains continue up to 3 references, and then drop off; presumably beyond the top two paraphrases or so, n-best paraphrasing adds more noise than genuine diversity (one can observe this drop off in provided diversity in the example shown in Figure 2).[3] Crucially, however, it is important to note that *only* the performance difference with four references—between the human and the paraphrase condition—is statistically significant.

O:	(hong kong, macau and taiwan) macau passed legalization to avoid double tax.
$\mathbf{P_1}$:	macao adopted bills to avoidance of double taxation (hong kong, macao and taiwan).
$\mathbf{P_2}$:	(hong kong, macao and taiwan) macao adopted bills and avoidance of double taxation.
$\mathbf{P_3}$:	(hong kong, macao and taiwan) macao approved bills and avoidance of double taxation.

Figure 2: The 3-best paraphrase hypotheses for the original sentence O with Chinese as the pivot language. The amount of n-gram diversity decreases with each successive hypothesis.

4.3 Effect of Genre Mismatch

It is extremely important to test the utility of optimization with paraphrased references when there is a mismatch between the genre of the data that the paraphraser is trained on and the genre of the actual test set that the system will eventually be scored on. To measure the effect of such mismatch, we conducted two different sets of experiments, each related to a common scenario encountered in MT research.

[2]By randomly choosing the sufficient number of random reference translations from the available 4 for each source document.

[3]This lack of diversity is found in most forms of n-best lists used in language processing systems and has been documented elsewhere in more detail (Langkilde, 2000; Mi et al., 2008).

4.3.1 Mixed-genre Test Set

For this experiment, we use the same paraphraser training data, MT training data and tuning sets as in Section 4.2. However, we now use a mixed-genre test set (MT06-GALE) as our validation set. MT06-GALE is a data set released by NIST in 2006 with 779 sentences, each with only a single reference translation. The composition of this set is as follows: 369 from the newswire genre and 410 sentences from the newsgroup genre. Since we are using MT03 for this experiment as well, we can also test whether using k-best paraphrases instead of just the 1-best helps on this mixed-genre validation set. The results are shown in Figure 3 and Table 3.

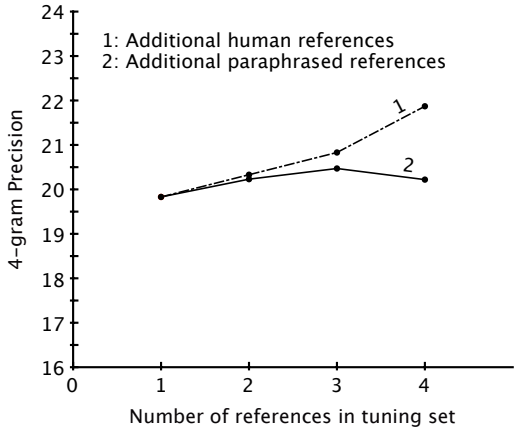

Figure 3: Testing the paraphraser on a mixed-genre validation set MT06-GALE. The graph depicts MT006-GALE 4-gram precision scores as additional references—human and paraphrased—are added to the single reference tuning set.

The first thing to notice about these results is that as we use additional references (human or paraphrased) for tuning the system, the brevity penalty on the validation set increases significantly. This is a well-known weakness of tuning for BLEU with multiple references and testing on a set with a single reference.[4] However, we can focus on the 4-gram precision which is the component that would be directly affected by larger n-gram diversity. The precision increases fairly regularly with additional human references. However, with additional paraphrased references, there are no statistically significant gains to be seen. In fact, as seen in Section 4.2, adding more paraphrases leads to a noisier tuning set. The TER scores, although following a similar trend, seem to provide no statistically significant evidence for either the human or the paraphrase portion of this experiment.

[4] In the NIST formulation of the BLEU metric, the brevity penalty is calculated against the shortest of the available reference translations. With multiple references available, it's very likely that the brevity penalty will be higher than if there was only a single reference.

Table 3: The BLEU scores (Prec.=4-gram precision, BP=brevity penalty) are shown here along with TER scores for MT06-GALE as additional references—human and paraphrased—are added to the single reference tuning set.

		BLEU			
		1H+0	1H+1	1H+2	1H+3
Human	Prec.	19.83	20.33	**20.83**	**21.87**
	BP	0.86	0.79	0.76	0.72
Para	Prec.	19.83	20.23	20.47	20.22
	BP	0.86	0.77	0.76	0.76

	TER			
	1H+0	1H+1	1H+2	1H+3
Human	64.09	64.02	63.99	63.37
Para	64.09	64.78	63.99	63.35

4.3.2 Porting to New Genres

Another important challenge in the MT world arises when systems are used to translate data from genres that are fairly new and for which a large amount of parallel data is not yet available. One such genre that has recently gained in popularity is the weblog genre. In order to test how the paraphrase approach works in that genre, we train both the MT system and the paraphraser on $400,000$ sentences of weblog data. Note that this is less than half the amount of newswire text that we previously used to train the paraphraser. From our experience with this genre, we find that if BLEU is used as the tuning criterion for this genre, the TER scores on held-out validation sets tend to be disproportionately worse and that a better criterion to use is a hybrid TER-BLEU measure given by

$$\text{TERBLEU} = 0.5 * \text{TER} + 0.5 * (1 - \text{BLEU})$$

We used the same measure for tuning our MT system in this experiment because we want to test how

the use of a criterion that's not as heavily dependent on n-gram diversity as BLEU affects the utility of the paraphrasing approach in a real-world scenario. As our tuning set, we use an actual weblog data set with only a single reference translation. As our validation set, we used a different weblog data set (WEB) containing 767 sentences, also with a single reference translation. The results are shown in Table 4.

Table 4: BLEU and TER scores for using paraphrases in tuning the web genre.

	BLEU		TER
	Prec.	BP	
1H	16.85	0.90	68.35
1H+1P	17.25	0.88	68.00

Since our validation set has a single reference translation, we separate out the 4-gram precision and brevity penalty components of BLEU scores so that we can focus on the precision which is directly affected by the increased n-gram diversity supplied by the paraphrase. However, for this experiment, we find that while there seem to be improvements in both the the 4-gram precision and TER scores, they are statistically insignificant. In order to isolate whether the lack of improvement is due to the relatively small size of the training data or the metric mismatch, we re-run the same experiment with BLEU as the tuning criterion instead of TER-BLEU.

Table 5: A significant gain in BLEU is achieved only when the tuning criterion for the MT system can take advantage of the diversity.

	BLEU		TER
	Prec.	BP	
1H	17.05	0.89	70.32
1H+1P	**18.30**	0.87	69.94

The results, shown in Table 5, indicate a significant gain in both the 4-gram precision and the overall BLEU score. They indicate that while a relatively small amount of training data may not hamper the paraphraser's effectiveness for parameter tuning, a tuning criterion that doesn't benefit from added n-gram diversity certainly can.

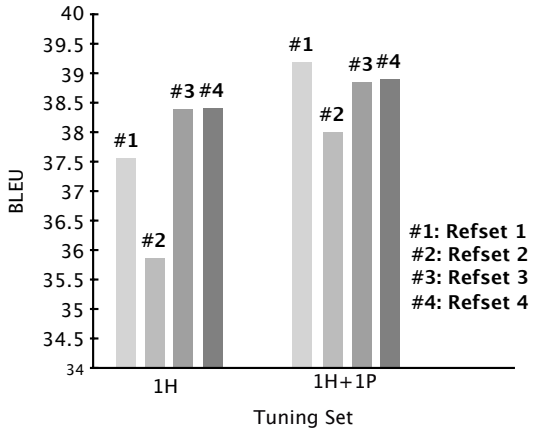

Figure 4: Measuring the impact of reference quality on use of paraphrased references. The graph shows the BLEU and TER scores computed for MT04+05 for cases where tuning utilizes reference translations created by different human translators and their corresponding paraphrases. Tuning usefulness of human translations vary widely (e.g., Refset #2 vs Refset #4) and, in turn, impact the utility of the paraphraser.

4.4 Impact of Human Translation Quality

Each of the 4 sets of references translations in MT03 was created by a different human translator. Since human translators are likely to vary significantly in the quality of translations that they produce, it is important to gauge the impact of the quality of a reference on the effectiveness of using its paraphrase, at least as produced by the paraphraser, as an additional reference. To do this, we choose each of the 4 reference sets from MT03 in turn to create the simulated single-reference set[5] (1H), paraphrased it and used the 1-best paraphrase as an additional reference to create a 2-reference tuning set (1H+1P). We then use each of the 8 tuning sets to tune the SMT system and compute BLEU and TER scores on MT04+05.

Figure 4 and Table 6 show these results in graphical form and tabular form, respectively. These results allow for two very interesting observations:

- The human reference translations do vary significantly in quality. This is clearly seen from the significant differences in the BLEU and TER scores between the 1H conditions, e.g., the third and the fourth human reference trans-

[5]Note that these per-translator simulated sets are different from the bias-free simulated set created in Sections 4.1 and 4.2.

Table 6: MT04+05 BLEU and TER results are shown for cases where tuning utilizes reference translations created by different human translators and their corresponding paraphrases.

	BLEU			
	#1	#2	#3	#4
1H	37.56	35.86	38.39	38.41
1H+1P	**39.19**	**37.94**	38.85	38.90

	TER			
	#1	#2	#3	#4
1H	57.23	60.55	54.50	54.12
1H+1P	54.21	56.42	53.40	53.51

lations seem to be better suited for tuning than, say, the second reference. Note that the term "better" does not necessarily refer to a more fluent translation but to one that is closer to the output of the MT system.

- The quality of the human reference has a significant impact on the effectiveness of its paraphrase as an additional tuning reference. Using paraphrases for references that are not very informative, e.g. the second one, leads to significant gains in both BLEU and TER scores. On the other hand, references that are already well-suited to the tuning process, e.g., the fourth one, show much smaller improvements in both BLEU and TER on MT04+05.

In addition, we also want to see how genre mismatch interacts with reference quality. Therefore, we also measure the BLEU and TER scores of each tuned MT system on MT06-GALE, a mixed-genre validation set with a single reference translation described earlier. These results—shown in Table 7—confirm our observations. The improvements in the TER scores with additional paraphrased references are proportional to how good the original reference was; in fact, for the fourth set of reference translations that seem best suited to tuning, adding a paraphrased reference amounts to adding noise and leads to lower performance on the mixed-genre set. As for the BLEU scores, we see similar trends with its 4-gram precision[6] component: it improves significantly for reference sets that are not as useful for tuning on their own but does not change (or even degrades) for the others.

Table 7: Measuring the impact of reference quality on MT06-GALE, a mixed-genre validation set.

		BLEU			
		#1	#2	#3	#4
1H	Prec.	19.09	19.19	20.34	20.61
	BP	0.88	0.88	0.82	0.84
1H+1P	Prec.	**20.63**	**19.98**	20.60	20.31
	BP	0.79	0.83	0.73	0.74

	TER			
	#1	#2	#3	#4
1H	64.98	66.30	63.42	62.98
1H+1P	63.19	64.01	63.64	63.98

4.5 Effect of Larger Tuning Sets

An obvious question to ask is whether the paraphrased references are equally useful with larger tuning sets. More precisely, would using a larger set of sentences (with a single human reference translation) be as effective as using a paraphraser to produce additional artificial reference translations? Given that creating additional human reference translations is so expensive, the most realistic and cost-effective option of scaling to larger tuning sets is to take the required number of sentences from the training data and add them to the tuning set. The parallel nature of the training corpus facilitates the the use of the same corpus as a tuning set with a single human-authored reference translation.

In order to replicate this scenario, we choose the single reference MT03 bias-free tuning set described previously as our starting point. To add to this tuning set, we remove a block of sentences from the MT training corpus[7] and added sentences from this block to the baseline MT03 tuning set in three steps to create three new tuning sets as shown in Table 8.

Once we create the larger tuning sets, we use each of them to tune the parameters of the MT system (which is trained on bitext excluding this block of sentences) and score the MT04+05 validation set. To see how this compares to the paraphrase-based approach, we paraphrase each of the tunings sets and

[6]Since MT06-GALE is a single reference validation set, brevity penalties are usually higher when scoring a system tuned with multiple references.

[7]We made sure that these sentences did not overlap with the paraphraser training data.

Table 8: Creating larger single reference tuning sets by adding sentences from the training corpus to the single reference base tuning set (MT03).

Tuning Set	# of Sentences
Base (MT03)	919
T1 (Base+600)	1519
T2 (T1+500)	2019
T3 (T2+500)	2519

used the paraphrases as additional reference translations for tuning the MT system. Figure 5 and Table 9 show these results in graphical form and tabular form, respectively.

The most salient observation we can make from the results is that doubling or even tripling the tuning set by adding more sentences from the training data does not lead to statistically significant gains. However, adding the paraphrased of the corresponding human reference translations as additional references for tuning always leads to significant gains, irrespective of the size of the tuning set.

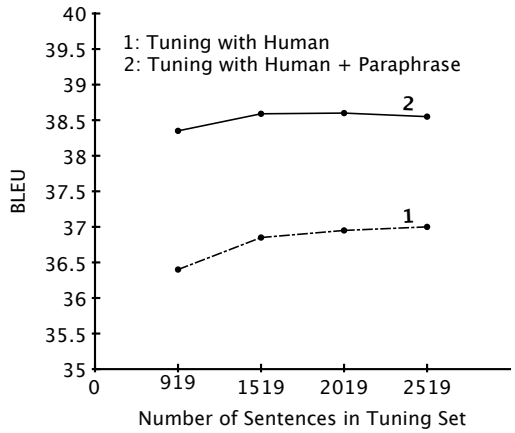

Figure 5: BLEU scores for the MT04+05 validation set as the tuning set is enlarged—by adding sentences from the training data.

5 Conclusion & Future Work

In this paper, we have examined in detail the value of multiple human reference translations, as compared with a single human reference augmented by means of fully automatic paraphrasing obtained via English-to-English statistical translation. We found that for the largest leap in performance, going from a single reference to two references, an automated paraphrase does quite as well as a second human translation, and using n-best paraphrasing we found that the point of diminishing returns is not hit until four human translations are available. In addition, we performed a number of additional analyses in order to understand in more detail how the paraphrase-based approach is affected by a variety of factors, including genre mismatch, human translation quality and tuning criteria that may not find additional n-gram diversity as valuable as BLEU does. The same analyses also validate the hypothesis that the paraphraser indeed works by providing additional n-gram diversity and not by means of accidental smoothing.

Table 9: BLEU and TER scores are shown for the MT04+05 validation set as the tuning set is enlarged by borrowing from the training data.

	BLEU			
	Base	T1	T2	T3
1H	36.40	36.85	36.95	37.00
1H+1P	**38.25**	**38.59**	**38.60**	**38.55**

	TER			
	Base	T1	T2	T3
1H	56.17	58.23	58.60	59.03
1H+1P	54.20	55.43	55.59	55.77

For these analyses, we used only a subset of the data used to train the MT system (2 million sentences). The point of this artificial restriction was to verify that the gains achieved by paraphrasing are not simply due to an inadvertent smoothing of the feature values in the MT system. Of course, a great advantage of the pivot-based full-sentence paraphrase technique is that it does not require any resources beyond those needed for building the MT system: a bitext and an MT decoder. Therefore, the best (and simplest) way to employ this technique is to use the full MT training set for training the paraphraser which, we believe, should provide even larger gains.

Another important issue that must be discussed concerns the brevity penalty component of the BLEU score. One might question whether the success of the paraphrase-based references derives primarily from the potential for generating longer outputs, thereby bypassing the brevity penalty. How-

ever, our TER results offer conclusive evidence that this is, in fact, not the case. If all this method did was to force longer MT outputs without contributing any meaningful content, then we would have observed a large loss in TER scores (due to an increase in the number of errors).

In order to achieve detailed comparisons with multiple human reference translations, our experimentation was done using a carefully translated NIST development set. However, the results here clearly point in a more ambitious direction: doing away entirely with any human translations beyond those already a part of the training material already expected by statistical MT systems. If the quality of the translations in the training set are good enough — or if a high quality subset can be identified — then the paraphrasing techniques we have applied here may suffice to obtain the target-language variation needed to tune statistical MT systems effectively. Experimentation of this kind is clearly a priority for future work.

We also intend to take advantage of one aspect of the paraphraser that radically differentiates it from an MT system: the fact that the source and the target languages are the same. This fact will allow to develop features and incorporate additional knowledge—much more easily than for a bilingual MT system—that can substantially improve the performance of the paraphraser and make it even more useful in scenarios where it may not yet perform up to its potential.

Finally, another avenue of further research is the tuning metric used for the paraphrasers. Currently the feature weights for the paraphraser features are tuned as described in (Madnani et al., 2007), i.e., by iteratively "translating" a set of source paraphrases, comparing the answers to a set of reference paraphrases according to the BLEU metric and updating the feature weights to maximize the BLEU value in the next iteration. While this is not unreasonable, it is not optimal or even close to optimal: in addition to striving for semantic equivalence, an automatic paraphraser should also aim for lexical diversity especially if said diversity is required in a downstream application. However, the BLEU metric is designed to reward larger n-gram overlap with reference translations. Therefore, using BLEU as the metric for the tuning process might actually lead to paraphrases with lower lexical diversity. Metrics recently proposed for the task of detecting paraphrases and entailment (Dolan et al., 2004; João et al., 2007a; João et al., 2007b) might be better suited to this task.

6 Acknowledgments

This work was supported, in part, by BBN under DARPA/IPTO contract HR0011-06-C-0022 and IBM under HR0011-06-2-001. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the view of DARPA. We are grateful to Necip Fazil Ayan, Christof Monz, Adam Lopez, Smaranda Muresan, Chris Dyer and other colleagues for their valuable input. Finally, we would also like to thank the anonymous reviewers for their useful comments and suggestions.

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of ACL*.

Daniel M. Bikel, Richard L. Schwartz, and Ralph M. Weischedel. 1999. An algorithm that learns what's in a name. *Machine Learning*, 34(1-3):211–231.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

William Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of COLING 2004*, Geneva, Switzerland.

Cordeiro João, Dias Gaël, and Brazdil Pavel. 2007a. A metric for paraphrase detection. In *Proceedings of the The Second International Multi-Conference on Computing in the Global Information Technology*.

Cordeiro João, Dias Gaël, and Brazdil Pavel. 2007b. New functions for unsupervised asymmetrical paraphrase detection. *Journal of Software*, 2(4).

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*.

Irene Langkilde. 2000. Forest-based statistical sentence generation. In *Proceedings of NAACL*.

Nitin Madnani, Necip Fazil Ayan, Philip Resnik, and Bonnie J. Dorr. 2007. Using paraphrases for parameter tuning in statistical machine translation. In *Proceedings of the Second ACL Workshop on Statistical Machine Translation*.

Haitao Mi, Liang Huang, and Qun Liu. 2008. Forest-based translation. In *Proceedings of ACL-08: HLT*,

pages 192–199, Columbus, Ohio, June. Association for Computational Linguistics.

D. W. Oard. 2003. The surprise langauge exercises. *ACM Transactions on Asian Language Information Processing*, 2(3).

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL*.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA*.

S. Strassel, C. Cieri, A. Cole, D. DiPersio, M. Liberman, X. Ma, M. Maamouri, and K. Maeda. 2006. Integrated linguistic resources for language exploitation technologies. In *Proceedings of LREC*.