System Combination for Machine Translation through Paraphrasing

Wei-Yun Ma
Institute of Information science
Academia Sinica
Taipei 115, Taiwan
ma@iis.sinica.edu.tw

Kathleen McKeown
Department of Computer Science
Columbia University
New York, NY 10027, USA
kathy@cs.columbia.edu

Abstract

In this paper, we propose a paraphrasing model to address the task of system combination for machine translation. We dynamically learn hierarchical paraphrases from target hypotheses and form a synchronous context-free grammar to guide a series of transformations of target hypotheses into fused translations. The model is able to exploit phrasal and structural system-weighted consensus and also to utilize existing information about word ordering present in the target hypotheses. In addition, to consider a diverse set of plausible fused translations, we develop a hybrid combination architecture, where we paraphrase every target hypothesis using different fusing techniques to obtain fused translations for each target, and then make the final selection among all fused translations. Our experimental results show that our approach can achieve a significant improvement over combination baselines.

1 Introduction

In the past several years, many machine translation (MT) combination approaches have been developed. Word-level combination approaches, such as the confusion network decoding model, have been quite successful (Matusov et al., 2006; Rosti et al., 2007a; He et al. 2008; Karakos et al. 2008; Chen et al. 2009a; Narsale 2010; Leusch 2011; Freitag et al. 2014).

In addition to word-level combination approaches, some phrase-level combination approaches have also recently been developed; the goal is to retain coherence and consistency between the words in a phrase. The most common phrase-level combination approaches are re-decoding methods: by constructing a new phrase table from each MT system's source-to-target phrase alignments, the source sentence can also be re-decoded using the new translation table (Rosti et al., 2007b; Huang and Papineni, 2007; Chen et al., 2007; Chen et al., 2009b). One problem with these approaches is that, just with a new phrase table, existing information about word ordering present in the target hypotheses is not utilized; thus the approaches are likely to make new mistakes of word reordering which do not appear in the target hypotheses of MT engines. Huang and Papineni (2007) attacked this issue through a reordering cost function that encourages search along with decoding paths from all MT engines' decoders.

Another phrase-level combination approach relies on a lattice decoding model to carry out the combination (Feng et al 2009; Du and Way 2010; Ma and McKeown 2012). In a lattice, each edge is associated with a phrase (a single word or a sequence of words) rather than a single word. The construction of the lattice is based on the extraction of phrase pairs from word alignments between a selected best MT system hypothesis (the backbone) and the other translation hypotheses. One challenge of the lattice decoding model is that it is difficult to consider structural consensus among target hypotheses from multiple MT engines, i.e, the consensus among occurrences of discontinuous words.

In this paper, we propose another phrase-level combination approach – a paraphrasing model using hierarchical paraphrases (paraphrases contain subparaphrases), to fuse target hypotheses. We dynamically learn hierarchical paraphrases from target hypotheses without any syntactic annotations and form a synchronous context-free grammar (SCFG) (Aho and Ullman 1969) to

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1053–1058,
Lisbon, Portugal, 17-21 September 2015.

guide a series of transformations of target hypotheses into fused translations. Through these structural transformations, the paraphrasing model is able to exploit phrasal and structural system-weighted consensus and also able to utilize existing information about word ordering present in the target hypotheses. In addition, to consider a diverse set of plausible fused translations, we develop a hybrid combination architecture, where we paraphrase every target hypothesis using different fusing techniques to obtain fused translations for each target, and then make the final selection among all fused translations through a sentence-level selection-based model.

In short, compared with other related work, our approach features the following advantages:

1. It can consider structural system-weighted consensus among target hypotheses from multiple MT engines through its hierarchical paraphrases, which non-hierarchical paraphrases are not able to do.

2. It can utilize existing information about word ordering present in the target hypotheses.

3. It can retain coherence and consistency between the words in a phrase.

4. The hybrid combination architecture enables us to consider a diverse set of plausible fused translations produced by different fusing techniques.

2 Hybrid Combination Architecture

In the context of system combination, discriminative reranking or post editing, MT researchers (Rosti et al., 2007a; Huang and Papineni, 2007; Devlin and Matsoukas, 2012, Matusov et al., 2008; Gimpel et al., 2013) have recently shown many positive results if more diverse translations are considered. Inspired by them, we develop a hybrid combination architecture in order to consider more diverse fused translations. We paraphrase every target hypothesis to obtain the corresponding fused translation, and then make the final selection among all fused translations through a sentence-level selection-based model, shown in Figure 1. In the architecture, different fusing techniques can be used to generate fused translations for the further sentence-level selection, enabling us to exploit more sophisticated information of the whole sentence.

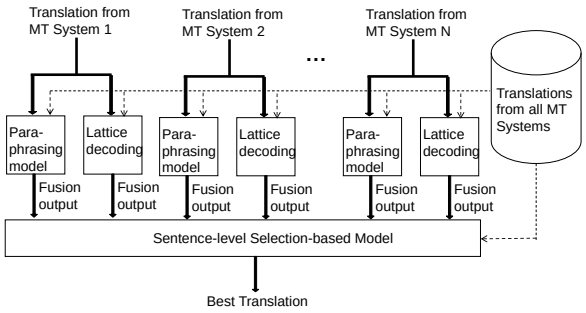

Figure 1. An example of *hybrid combination architecture*

3 Paraphrasing Model

In this section, we introduce our paraphrasing model. For each single target hypothesis, we extract a set of hierarchical paraphrases from monolingual word alignments between the hypothesis and other hypotheses. Each set of hierarchical paraphrases forms a synchronous context-free grammar to guide a series of transformations of that target hypothesis into a fused translation.

Any monolingual word aligner can be used to produce the monolingual word alignments. In our system, we adopt TERp (Snover et al. 2009), one of the state-of-the-art alignment tools, to serve this purpose. TERp is an extension of TER (Snover et al. 2006). Both TERp and TER are automatic evaluation metrics for MT, based on measuring the ratio of the number of edit operations between the reference sentence and the MT system hypothesis. The edit operations of TERp include TER's Matches, Insertions, Deletions, Substitutions and Shifts—as well as three new edit operations: Stem Matches, Synonym Matches and Paraphrases. A valuable side product of TERp is the monolingual word alignment. A constructed example is shown in Figure 2.

3.1 Hierarchical Paraphrase Extraction

We first introduce our notation. For a given sentence *i*, we use E_h^i to denote the target hypothesis from MT system *h*, use EP_h^i to denote E_h^i attached with related word positions, use e_h^i to denote a phrase within E_h^i, and use ep_h^i to denote e_h^i attached with related word positions. For instance, If E_h^i is "you buy the book", then EP_h^i would be "you[1] buy[2] the[3] book[4]". If e_h^i is "the book", then ep_h^i is "the[3] book[4]".

For a given sentence *i*, a MT system *h* and a MT system *k*, we use a SCFG denoted by $Q_{h,k}^i$ to represent the set of hierarchical paraphrases learned from EP_h^i and EP_k^i. Adapting (Chiang

2007), we design the following rules to obtain $Q^i_{h,k}$, based on the monolingual word alignment, obtained by a aligner, such as TERp.

- If $< ep^i_h , ep^i_k >$ is consistent[1] with the monolingual word alignment, then $X \rightarrow < ep^i_h , e^i_k >$ is added to $Q^i_{h,k}$.
- If $X \rightarrow < \gamma , \alpha >$ is in $Q^i_{h,k}$, and $< ep^i_h , ep^i_k >$ is consistent with monolingual word alignment such that $\gamma = \gamma_1 ep^i_h \gamma_2$ and $\alpha = \alpha_1 e^i_k \alpha_2$, then $X \rightarrow < \gamma_1 X_a \gamma_2 , \alpha_1 X_a \alpha_2 >$ is added to $Q^i_{h,k}$, where *a* is an index.

Please note that for each extracted hierarchical paraphrase - $X \rightarrow < \gamma , \alpha >$, γ would include information of word positions while α would not.

For a certain target hypothesis - EP^i_h, our goal is to paraphrase it to get the fusion output by using a set of hierarchical paraphrases, denoted by Q^i_h. Thus we create the union of all related hierarchical paraphrases learned from EP^i_h and other target hypotheses. Two special "glue" rules - $S \rightarrow < S_1 X_2 , S_1 X_2 >$ and $S \rightarrow < X_1 , X_1 >$ are also added to Q^i_h. The process can be represented formally in the following:

$$Q^i_h = \bigcup_{k=1}^{N} Q^i_{h,k} \bigcup \{ S \rightarrow < S_1 X_2 , S_1 X_2 > , S \rightarrow < X_1 , X_1 > \}$$

where *N* is the total number of MT systems.

3.1.1 An Example

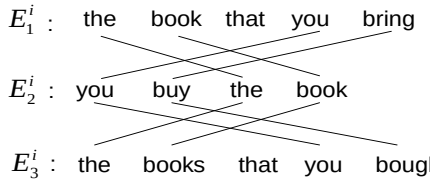

Figure 2. A constructed example of a sentence - "你買的書 (the book that you bought)" and its translations from three MT systems – E^i_1, E^i_2 and E^i_3, and word alignments between E^i_2 and E^i_1, and between E^i_2 and E^i_3, obtained through TERp.

We use a Chinese-to-English example in Figure 2 to illustrate the extraction process. The extracted hierarchical paraphrases to paraphrase EP^i_2 - "you[1] buy[2] the[3] book[4]" are shown in Table 1. Because of limited space, only part of the paraphrases, i.e, part of the rules of Q^i_2, are shown.

part of rules of Q^i_2		in $Q^i_{2,1}$?	in $Q^i_{2,2}$?	in $Q^i_{2,3}$?
X →< you[1] , you>	(a)	✔	✔	✔
X →< you[1] buy[2] , you buy >	(b)		✔	
X →< you[1] buy[2] , you bought >	(c)			✔
X →< you[1] buy[2] , you bring >	(d)	✔		
X →< book[4] , book >	(e)	✔	✔	
X →< book[4] , books >	(f)			✔
X →< the[3] book[4] , the book >	(g)	✔	✔	
X →< the[3] book[4] , the books >	(h)			✔
X →< you[1] buy[2] the[3] book[4] , the books that you bought >	(i)			✔
X →< X_1 the[3] book[4] , the books that X_1 >	(j)			✔
X →< you[1] buy[2] the[3] X_1 , the X_1 that you bought >	(k)			✔
X →< X_1 the[3] X_2 , the X_2 that X_1 >	(l)	✔		✔

Table 1. Part of extracted hierarchical paraphrases to paraphrase EP^i_2, i.e, part of the rules of Q^i_2.

Note that, in Table 1, the rules (j), (k) and (l) can be regarded as structural paraphrases, and they utilize existing information about word ordering present in the target hypotheses. Since rule (l) is included in both $Q^i_{2,1}$ and $Q^i_{2,3}$, we can say that rule (l) has more structural consensus than rule (j) and (k). And rule (l) also models the word reordering through reversing the order of X_1 and X_2. By the example, we can see the reason why our model is able to exploit structural consensus and also to utilize existing information about word ordering present in the target hypotheses.

3.2 Decoding

Given a certain target hypothesis - EP^i_h, and its set of hierarchical paraphrases - Q^i_h, the decoder aims to paraphrase EP^i_h using Q^i_h by performing a search for the single most probable derivation via the CKY algorithm with a Viterbi approximation. The derivation is the paraphrased result, i.e, the fusion result indicated in Figure 1. The single most probable derivation can be represented as

$$\arg\max_{\overline{E}^i_h} \log p(\overline{E}^i_h \mid EP^i_h) = \arg\max_{\overline{E}^i_h} \sum_{j=1}^{J} \left(\sum_{k=1}^{N} \lambda_k * f(q^{i,j}_h , k) \right) + \lambda_p * J + \lambda_l * \log(LM(\overline{E}^i_h)) + \lambda_w * length(\overline{E}^i_h)$$

$$f(q^{i,j}_h , k) = \begin{cases} 1 \text{ if } q^{i,j}_h \in Q^i_{h,k} \\ 0 \text{ otherwise} \end{cases}$$

[1] This means that words in a legal paraphrase are not aligned to words outside of the paraphrase, and should include at least one pair of words aligned with each other.

where $q_h^{i,j}$ is the jth paraphrase in Q_h^i used to generate $\overline{E}_h^i$, J is the number of paraphrases used to generate $\overline{E}_h^i$. N is the total number of MT systems. λ_k is the weight of MT system k, in charge of the system-weighted consensus. λ_p is phrase penalty. λ_l is the LM weight and λ_w is word penalty. All weights are trained discriminatively for Bleu score using Minimum Error Rate Training (MERT) procedure (Och 2004).

The ideal result of paraphrasing EP_2^i is shown in the following, which is supposed to be generated with a higher chance if, regardless of system weights. That is because of the use of the rules with higher degree of structural consensus, such as (l) and (e).

$< S_1 >$
$\Rightarrow < X_2, X_2 >$ using glue rule
$\Rightarrow < X_3 \text{ the}^3 X_4, \text{the } X_4 \text{ that } X_3 >$ using rule (l)
$\Rightarrow < \text{you}^1 \text{ buy}^2 \text{ the}^3 X_4, \text{the } X_4 \text{ that you bought} >$ using rule (c)
$\Rightarrow < \text{you}^1 \text{ buy}^2 \text{ the}^3 \text{ book}^4, \text{the book that you bought} >$ using rule (e)

4 Sentence-Level Selection-based Model

For a given sentence i and its M multiple fusion outputs - $\overline{E}_f^i$, $1 \le f \le M$ generated by the paraphrasing model or the lattice decoding model, the goal here is to select the best one among them, as shown in Figure 1 (For the case shown in the figure, M is $2N$). The idea is to compare system-weighted consensus among all fusion outputs and translations from all MT systems, and then select the one with the highest consensus. We adopt Minimum Bayes Risk (MBR) decoding (Kumar and Byrne, 2004; Sim et al., 2007) to serve our purpose and develop the following TER-based MBR:

$$\arg\max_f \log p(\overline{E}_f^i) = \arg\max_f \omega_l * \log(LM(\overline{E}_f^i)) + \sum_{m=1}^{M}\left(\omega_m * \log(1 - TER(\overline{E}_f^i, \overline{E}_m^i))\right) + \sum_{k=1}^{N}\left(\omega_k * \log(1 - TER(\overline{E}_f^i, E_k^i))\right)$$

where *TER* is Translation Tdit Ratio. ω_m is the fusion weight specific to a certain MT system and a certain fusion model, ω_k is the weight of MT system k and ω_l is the LM weight. All weights are trained discriminatively for Bleu score using MERT.

5 Experiments

Our experiments are conducted and reported on three datasets: The first dataset includes Chinese-English system translations and reference translations from DARPA GALE 2008 (GALE Chi-Eng). The second dataset includes Chinese-English system translations and reference translations and from NIST 2008 (NIST Chi-Eng). And the third dataset includes Arabic-English system translations and reference translations and from NIST 2008 (NIST Ara-Eng).

	MTSystem#	TuneSent#	TestSent#
GALE Chi-Eng	5	422	422
NIST Chi-Eng	5	524	788
NIST Ara-Eng	5	592	717

Table 2. Experimental setting

MT System	Approach	Bleu
nrc	phrase-based SMT	30.95
rwth-pbt-aml	phrase-based SMT + source reordering	31.83
rwth-pbt-jx	phrase-based SMT + word segmentation	31.78
rwth-pbt-sh	phrase-based SMT + source reordering + rescoring	32.63
sri-hpbt	hierarchical phrase-based SMT	32.00

Table 3: Techniques of top five MT of GALE Chi-Eng Dataset

Table 3 lists distinguishing machine translation approaches of top five MT of GALE Chi-Eng Dataset. And "rwth-pbt-sh" performs the best in Bleu score.

Two combination baselines are implemented for comparison: one is an implementation based on confusion network decoding, and the other is Lattice Decoding from (Ma and McKeown 2012), both of which are using TERp to obtain word alignments between a selected backbone hypothesis and other target hypotheses. The former uses these word alignments to construct a confusion network while the latter extracts phrases which are consistent with these word alignments to construct a lattice. For both baselines, backbone hypotheses are selected sentence by sentence based on system-weighted consensus among translation of all MT systems.

5.1 Results

In Table 4, CN represents confusion network; LD represents Lattice Decoding (Ma and McKeown 2012); PARA represents paraphrasing model proposed in this paper; Backbone_* represents that * is carried out on selected backbones, in contrast with the hybrid combination architecture. Arch_LD represents that only lattice decoding is carried out using hybrid combination architecture. Arch_PARA represents that only paraphrasing model is carried out using hybrid combination

architecture. Arch_LD_PARA represents that LD and PARA are both carried out using hybrid combination architecture, which is the example shown in Figure 2.

	GALE Chi-Eng	NIST Chi-Eng	NIST Ara-Eng
Best MT system	32.63	30.16	48.40
Backbone_CN (baseline)	33.04	31.21	48.56
Backbone_LD (baseline)	33.16	32.65	49.33
Backbone_PARA	33.09	32.59	49.46
Arch_LD	33.24	32.66	50.48
Arch_PARA	33.32	32.90	50.20
Arch_LD_PARA	33.72	33.42	50.44

Table 4. Experimental results in Bleu score

From Table 4, we can first observe that, for the three datasets, Backbone_PARA and Backbone_LD outperform Backbone_CN, which shows the advantage of using phrases over words in combination. However, Backbone_PARA does not show improvement over Backbone_LD. The reason could be that selected backbones already have a high level of quality and fewer words need to be replaced or re-ordered in contrast with other target hypotheses.

We find that Arch_PARA performs better than Backbone_PARA, and Arch_LD performs better than Backbone_LD. This observation supports our claim that it is beneficial to consider more diverse sets of plausible fused translations.

Arch_LD_PARA achieves the best performance among all techniques used in this paper. It not only supports our claim, but also brings a conclusion that the paraphrasing model and lattice decoding can compensate for the weaknesses of the other in our architecture.

Since the paraphrasing model uses hierarchical paraphrases to carry out the fusion, it is able to make a bigger degree of word-reordering or structural change on the input hypothesis in comparison with lattice decoding. We suppose that when more word-reordering and structural changes are needed, paraphrasing model can bring more benefits than lattice decoding. Because the quality of a given translation hypothesis is highly related to word reordering and structural change, it can be expected that when a poorly translated hypothesis is paraphrased, paraphrasing model can bring more benefits than lattice decoding. In order to obtain the evidence to support this hypothesis, we carried out the following experiment on NIST Chi-Eng Dataset.

For each MT system from the selected top 5 system A-E, we paraphrase its translations using the paraphrasing model and lattice decoding separately, aiming to compare the performances of the two models on each MT system. In other words, we do not first do backbone selection. Every MT system's translation is regarded as a backbone. The results are shown in Table 5.

	MT	Lattice Decoding	Paraphrasing model
Sys A	30.16	32.17	31.76
Sys B	30.06	31.93	31.72
Sys C	**28.15**	**30.66**	**31.00**
Sys D	29.94	31.86	31.46
Sys E	**29.52**	**31.52**	**31.92**

Table 5. The Bleu score of each MT system, the Bleu score of paraphrasing each MT system using lattice decoding and the Bleu score of paraphrasing each MT system using paraphrasing model.

Among the five MT systems, "Sys C" and "Sys D" perform poorer than the other three MT systems. When we paraphrase the two systems, we find that paraphrasing model outperforms lattice decoding. These results support our hypothesis that when more word-reordering and structural changes are needed, paraphrasing model can bring more benefits than lattice decoding.

6 Conclusion

We view MT combination as a paraphrasing process using a set of hierarchical paraphrases, in which more complicated paraphrasing phenomena are able to be modeled, such as phrasal and structural consensus. Existing information about word ordering present in the target hypotheses are also considered. The experimental results show that our approach can achieve a significant improvement over combination baselines.

There are many possibilities for enriching the simple framework. Many ideas from recent translation developments can be borrowed and modified for combination. Our future work aims to incorporate syntactic or semantic information into our paraphrasing framework.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments and suggestions. This work is supported by the National Science Foundation via Grant No. 0910778 entitled "Richer Representations for Machine Translation".

Reference

A. V. Aho and J. D. Ullman. 1969. Syntax directed translations and the pushdown assembler. Journal of Com-puter and System Sciences, 3:37–56.

Yu Chen, Andreas Eisele, Christian Federmann, Eva Hasler, Michael Jellinghaus, and Silke Theison. 2007. Multi-engine machine translation with an open-source SMT decoder. In Proceedings of WMT07

Boxing Chen, Min Zhang and Aiti Aw. 2009a. A Comparative Study of Hypothesis Alignment and its Improvement for Machine Translation System Combination. In: Proceedings of ACL-IJCNLP. pp. 1067-1074. Singapore. August.

Yu Chen, Michael Jellinghaus, Andreas Eisele, Yi Zhang, Sabine Hunsicker, Silke Theison, Christian Federmann, Hans Uszkoreit. 2009b. Combining Multi-Engine Translations with Moses. In Proceedings of the Fourth Workshop on Statistical Machine Translation

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

J. Devlin and S. Matsoukas. 2012. Trait-based hypothesis selection for machine translation. In Proc. of NAACL

Jinhua Du and Andy Way. 2010. Using TERp to Augment the System Combination for SMT. In Proceedings of the Ninth Conference of the Association for Machine Translation (AMTA2010)

Yang Feng, Yang Liu, Haitao Mi, Qun Liu, and Yajuan Lu. 2009 Lattice-based System Combination for Statistical Machine Translation. In Proceedings of ACL

Markus Freitag, Matthias Huck, and Hermann Ney. 2014. Jane: Open source machine translation system combination. In Conference of the European Chapter of the Association for Computational Linguistics, pages 29–32, Gothenburg, Sweden, April. Association for Computational Linguistics.

Kevin Gimpel, Dhruv Batra, Chris Dyer, and Gregory Shakhnarovich. 2013. A Systematic Exploration of Diversity in Machine Translation. In Proc. of EMNLP

Xiaodong He, Mei Yang, Jianfeng Gao, Patrick Nguyen, and Robert Moore. 2008. Indirect-hmm-based hypothesis alignment for computing outputs from machine translation systems. In Proceedings of EMNLP

Fei Huang and Kishore Papineni. 2007. Hierarchical System Combination for Machine Translation. In Proceed-ings of EMNLP-CoNLL

Damianos Karakos, Jason Eisner, Sanjeev Khudanpur, and Markus Dreyer. 2008. Machine translation system combination using ITG-based alignments. In Proceedings of ACL-HLT

S. Kumar and W. Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In Proceedings of HLT

Wei-Yun Ma and Kathleen McKeown. 2012. Phrase-level System Combination for Machine Translation Based on Target-to-Target Decoding. In Proceedings of the 10th Biennial Conference of the Association for Ma-chine Translation in the Americas (AMTA), San Diego, CA.

Gregor Leusch, Markus Freitag, and Hermann Ney. The RWTH System Combination System for WMT 2011. 2011. In Proceedings of the Sixth Workshop on Statistical Machine Translation

Evgeny Matusov, Nicola Ueffing, and Hermann Ney 2006. Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In Proceedings of EACL

E. Matusov, G. Leusch, R. E. Banchs, N. Bertoldi, D. Dechelotte, M. Federico, M. Kolss, Y. S. Lee, J. B. Marino, M. Paulik, S. Roukos, H. Schwenk, and H. Ney. 2008. System combination for machine translation of spoken and written language. IEEE Transactions on Audio, Speech and Language Processing, 16(7):1222–1237, September.

Sushant Narsale. JHU System Combination Scheme for WMT 2010. 2010. In Proceedings of the Fifth Workshop on Statistical Machine Translation

Franz Josef Och. 2004. Minimum Error Rate Training in Statistical Machine Translation. In Proceedings of ACL

Antti-Veikko I. Rosti, Spyros Matsoukas, and Richard Schwartz. 2007a. Improved word-level system combination for machine translation. In Proceedings of ACL

Antti-Veikko I. Rosti, Necip F. Ayan, Bing Xiang, Spyros Matsoukas, Richard Schwartz, and Bonnie J. Dorr. 2007b. Combining outputs from multiple machine translation systems. In Proceedings of NAACL-HLT

Khe Chai Sim, William J. Byrne, Mark J.F. Gales, Hichem Sahbi, and Phil C. Woodland. 2007. Consensus network decoding for statistical machine translation system combination. In Proceedings of ICASSP

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.

M. Snover, N. Madnani, B. Dorr, and R. Schwartz. 2009. TER-Plus: Paraphrase, Semantic, and Alignment Enhancements to Translation Edit Rate. Machine Translation, 23(2–3):117–127.