Statistical Post-Editing for a Statistical MT System

Hanna Béchara[†] **Yanjun Ma** [‡] **Josef van Genabith**[†]
[†]Centre for Next Generation Localisation
School of Computing, Dublin City University
{hbechara, josef}@computing.dcu.ie
[‡]Baidu Inc., Beijing China
yma@baidu.com

Abstract

Statistical post-editing (SPE) techniques have been successfully applied to the output of Rule Based MT (RBMT) systems. In this paper we investigate the impact of SPE on a standard Phrase-Based Statistical Machine Translation (PB-SMT) system, using PB-SMT both for the first-stage MT and the second stage SPE system. Our results show that, while a naive approach to using SPE in a PB-SMT pipeline produces no or only modest improvements, a novel combination of source context modelling and thresholding can produce statistically significant improvements of 2 BLEU points over baseline using technical translation data for French to English.

1 Introduction

Statistical post-editing (SPE) has been used successfully to improve the output of Rule-Based MT (RBMT) systems: Simard et al. (2007a) train a “mono-lingual” PB-SMT system (the Portage system) on the output of an RBMT system for the source side of the training set of the PB-SMT system and the corresponding human translated reference. A complete translation pipeline consists of a rule-based first-stage system, whose output on some (unseen) test set, in turn, is translated by the second-stage “mono-lingual” SPE system. Simard et al. (2007a) present experiments using Human Resources and Social Development (HRSDC) Job Bank[1] French and English parallel data and found that in combination, the RBMT system post-edited by the PB-SMT system performed significantly better than each of the individual systems on their own. Simard et al. (2007a) also tested the SPE technique with Portage PB-SMT both as first-stage MT and as second stage SPE system (i.e. Portage post-editing its own output) and reported that nothing could be gained. In a number of follow-up experiments, Simard et al. (2007b) used an SPE system to adapt RBMT-systems to a specific domain, once again using Portage in the SPE phase. Adding the SPE system produced BLEU score increases of about 20 points over the original RBMT baseline.

SPE was also applied in an attempt to improve Japanese to English patent translations. Teramusa (2007) uses RBMT to translate patent texts, which tend to be difficult to translate without syntactic analysis. Combining RBMT with SPE in the post-editing phase produced an improved score on the NIST evaluation compared to that of the RBMT system alone. Dugast et al. (2007) report research on combining SYSTRAN with PB-SMT systems Moses and Portage. Comparison between raw SYSTRAN output and SYSTRAN+SPE output shows significant improvements in terms of lexical choice, but almost no improvement in word order or grammaticality. Dugast et al. (2009) trained a similar post-editing system with some additional treatment to prevent the loss of entities such as dates and numbers.

Oflazer and El-Kahlout (2007) explore selective segmentation based models for English to Turkish translation. As part of their experiments they present a short section towards the end of the paper on statistical post-editing of an SMT system, which they

[1] www.jobbank.gc.ca

call model iteration. They train a post-editing SMT model on the training set decoded by the first stage SMT model and iterate the approach, post-editing the output of the post-editing system. BLEU results show positive improvements, with a cumulative 0.46 increase after 2 model iterations. It is not clear whether the result is statistically significant.

Our experiments follow the statistical post-editing design of Simard et al. (2007a), where the output of a first-stage system is used to train a mono-lingual second stage system, that has the potential to correct or otherwise improve on (i.e. post-edit) the output of the first-stage system. In contrast to Simard et al. (2007a), but like Oflazer and El-Kahlout (2007), our experiments use PB-SMT systems throughout both stages. The objective is to investigate in more detail whether and to what extent state-of-the-art PBSMT technology can be used to post-edit itself, i.e. its own output.

2 Methodology

In our experiments we focus on English and French as these are the languages considered in the original research by Simard et al.(Simard et al., 2007a). As SPE on the output of RBMT systems is already a commercial reality,[2] we use industry translation memories (TMs) from the technical computing domain as our data. We use a standard Moses PB-SMT set-up (Koehn et al., 2007).

2.1 Data

The data for our experiments come from an English–French Translation Memory from an IT company (Symantec). The domain of the data is technical software user help information. The translation memory was preprocessed to remove all TMX markup and meta-data information. The translation memory contains 55,322 unique segments. From this we randomly extract a training set of 52,383 English–French segment pairs, between 1 and 98 words in length for English, and 1 to 100 for French. The average segment length in the training set is 13 words for English and 15 words for French. The training set has a vocabulary size of 9,273 for the English side of the data and 12,070 on the French side. We use the remaining 972 and 1967 segments from the TM as a development set and a test set, respectively. As we are working with a translation memory, all translation memory data segments are unique, i.e. there is no repetition in the data (and hence no overlap between the training, development and test sets in our experiments).

2.2 SPE Architecture

All our experiments follow the original statistical post-editing design of Simard et al. (2007a), but as in Oflazer and El-Kahlout (2007) using PB-SMT systems throughout (i.e. both as first and second stage systems in the post-editing pipeline, rather than RBMT as the first-stage followed by PB-SMT as the second stage as in the original work).

The first-stage PB-SMT system is trained in the usual way using the English (E) and French (F) parallel training data, providing us with the output F' (MT output French), which will be the source data for our second-stage system. In order to obtain the "source" training data F' for the second-stage monolingual PB-SMT system, we train another first-stage PB-SMT systems from E to F using a 10-fold cross-validation approach on the E to F training set to avoid translation of already seen data in the creation of the mono-lingual training section part F' of the F'–F second stage PB-SMT system. The intuition here is that, as the eventual test data is unseen data, the training data for the second stage mono-lingual PB-SMT system should not be generated by simply training a first-stage PB-SMT system on all of the E and F training data and then applying it to the (already seen) training data E to generate the source side F' of the training data for the second stage PB-SMT system. Note that this (subtlety) applies to the generation of the (source side of the) second stage training data only, and that in all the experiments, unseen test data are always translated using first and second stage PB-SMT systems trained on the full training sets E and F (for the stage one system) and F' and F (for the stage two post-editing system).

2.3 MT System

We use a standard Moses PB-SMT system (Koehn et al., 2007), 5-gram language models with Kneser-Ney smoothing trained with SRILM (Stolcke, 2002), the GIZA++ implementation of IBM word

[2] http://www.systran.co.uk/translation-products/server/systran-enterprise-server

alignment model 4 (Och and Ney, 2003), and the refinement and phrase-extraction heuristics described in Koehn et al. (2003). All systems are tuned using minimum error rate training (MERT) (Och, 2003) on the development set. During the decoding phase, the stack size was limited to 500 hypotheses. We use approximate randomisation methods (Noreen, 1989) as implemented in FASTMTEVAL (Stroppa and Owczarzak, 2007) to test for statistical significance.

2.4 Contextual SPE

In our basic SPE pipeline (PE) – translating, say, from English to French (E–F) – the second-stage SPE system is trained on the output (F') of the (10-fold cross validation version of the) first-stage MT system, effectively resulting in a "mono-lingual" SPE system (F'-F). In as sense, however, the second-stage SPE system has lost the connection to the original source data: ideally we would like to be able to be in a position to distinguish between situations where f' is a good translation of some source word (or phrase) e, and situations where f' should be post-edited to f. In some of the experiments reported, we model this by recording the source word (or phrase) e that gave rise to f' as f'#e (i.e. concatenating f' with # and e), effectively creating a new intermediate language F'#E as the source language for a context-aware second-stage SPE system (PE-C). In our experiments we do this using GIZA++ word-alignments as illustrated in the following example:

- Source E: `if an original file has been deleted , but backup files are still available ...`
- Target F: `si un fichier original a été supprimé , mais si les fichiers de sauvegarde sont toujours disponibles ...`
- Baseline Output F': `si un fichier initial a été supprimé , mais les fichiers de sauvegarde sont encore disponibles ...`
- Context F'#E: `si#if un#an fichier#file initial#original a#has été#been supprimé#deleted ,#, mais#but les#files fichiers#files de#backup sauvegarde#backup sont#are encore#still disponibles#available ...`
- PE-C Output F": `si un fichier original a été supprimé , mais les fichiers de sauvegarde sont toujours disponibles ...`

Here, the baseline output `initial` and `encores` was changed to `original` and `toujours`, ensuring a better match with the target text.

While this new intermediate language preserves context information, the vocabulary size increases from 9273 in the EN training set to 70780 in the F'#E training set. This increase and the ensuing data sparseness have potentially adverse effects on translation quality. Furthermore, the word alignment data used to create this new language is not always reliable. In order to address the issue of data sparseness and unreliable word alignment data, we experiment with thresholding context information on alignment strength in some of the experiments reported below.

3 Experiments

Below we present results for English to French and French to English translation and post-editing experiments.

3.1 English to French

3.1.1 Experimental Results Using SPE

In order to evaluate our SPE approach, we train two PB-SMT systems for a post-editing pipeline, the first-stage system (baseline) between E and F, producing output F' given input E, and the second stage mono-lingual post-editing system between F' and F, producing output F" given F' as input. We train post-editing pipeline systems without (PE) and with (PE-C) context information.

Table 1 shows that simple PE fails to improve over the baseline and that the drop in the BLEU score for the PE-C (post-editing with context information) compared to the baseline (and PE) is marked. The most likely reason for this drop is the explosion in the size of the vocabulary set between E and F'#E in the post-editing with context information setting (PE-C). This is visible in the output of the second stage post-editing system in the form of untranslated f'#e items. These are effectively OOV (out-of-vocabulary) items that the second stage system has not encountered during training. As the f' part of an f'#e item is already a word in the target language, we simply filter the f'#e items in the output by automatically deleting the source context information suffix #e from such items.[3] This is illustrated in the example below:

- PE-C: `dell recommande de renseigne#populate la baie de disque avec les disques physiques de la même capacité .`
- PE-CF(filtered): `dell recommande de renseigne la baie de disque avec les disques physiques de la même capacité .`

We refer to this output as PE-CF in Table 1 (and elsewhere in the paper), and the BLEU score for PE-CF is much closer to the baseline than that for PE-C. In all our experiments reported in the remainder of this paper we use this simple output filtering prior to evaluating context-informed SPE models (SPE models where context information is included in the post-editing phase). [4]

Score	Baseline	PE	PE-C	PE-CF
BLEU	60.30	60.15	46.89	58.55

Table 1: English–French SPE results

Overall results show that for our data set, a simple second-stage PB-SMT system (with and without context information) is unable to improve on the first-stage PB-SMT system in a pure PB-SMT post-editing pipeline for English to French machine translation.

3.1.2 Thresholding Context Information by Alignment Strength

In order to obtain a more fine-grained view on the effects of OOVs and data-sparseness on context-informed SPE pipelines, we carried out experiments restricting the amount of context information available to PE-C systems. In particular, we used directional Source to Target GIZA++ word alignment strengths to filter context information, using the word alignment levels of $\geq$0.6, $\geq$0.7, $\geq$0.8 and $\geq$0.9 as thresholds: that is for each threshold, source words that are aligned with translation output words with an alignment score greater than or equal to the threshold are used as source context words e in f'#e pairs for the source side of the second stage PB-SMT system.

Table 2 shows the results for the context aware post-editing pipeline PE-CF with alignment strength thresholding on the full test set.

Threshold	0.6	0.7	0.8	0.9
PE-CF	59.80	60.30	60.23	59.73

Table 2: English–French Translation using Contextual SPE with Alignment Thresholding (BLEU scores)

Thresholding shows clear improvements over simple PE-CF in Table 1, however, none of them show improvements over the baseline in Table 1. Clearly, for the English to French translation direction and our data set, all our PB-SMT SPE pipelines (even those that are context aware and use thresholding) fail to improve on the PB-SMT baseline.

3.2 French to English

We ran the same set of experiments for the other translation direction, French to English, which is often considered the easier of the two translation directions (Callison-Burch et al., 2009).

3.2.1 Experimental Results Using SPE

Simple PE results on the test set show that for our data set a simple second-stage PB-SMT system is able to improve on the first-stage PB-SMT system in a pure PB-SMT post-editing pipeline, with a small increase in BLEU of 0.65 absolute over baseline. This result is statistically significant at the p $\leq$ 0.05 level. Compared to baseline and PE, BLEU scores deteriorate for the context-aware post-editing pipeline PE-C, as any beneficial impact of the post-editing pipeline is swamped by data sparseness and OOV items in the output of the second stage PE-C system. The most likely reason for this drop is again the explosion in the size of the vocabulary set between E and E'#F in the post-editing with context information setting: the training set vocabulary size is 9,273 for E compared to 47,730 for E'#F, resulting in both data-sparseness and OOV occurrences for the second stage PB-SMT system in the context informed post-editing pipeline PE-C. Filtering out the

[3] This was suggested to us by Ondrej Bojar following an early presentation of the material.

[4] Note that the source context information suffix #e filtering is crucially different from the alignment strength based context thresholding developed below.

#f tags in the output, leaving only the target word, brings the BLEU score up to 61.36 for PE-CF.

Score	Baseline	PE	PE-C	PE-CF
BLEU	61.60	62.25	57.33	61.36

Table 3: French–English SPE results

3.2.2 Thresholding Context Information by Alignment Strength

Mirroring the English to French Experiments, we carry out experiments restricting the amount of context information available to PE-C systems, filtering context information by thresholding word alignment strengths, using the GIZA++ based word alignment levels of $\geq$0.6, $\geq$0.7, $\geq$0.8 and $\geq$0.9 as thresholds. Results are presented in Table 4.

Threshold	0.6	0.7	0.8	0.9
PE-CF	63.76	63.54	63.89	63.80

Table 4: French–English Translation using Contextual SPE with Alignment Thresholding (BLEU scores)

The results in Table 4, when compared to the baseline (61.60) in Table 3, show that for all alignment thresholds, for this data set and the French to English translation direction, context aware post-editing pipeline PE-CF results outperform the baseline by about 2 BLEU points absolute. All results are statistically significant at the $p \leq$ 0.05 level.

4 Analysis and Discussion

In order to obtain a better understanding of the translation quality gains in French to English translation and respectively the quality loss in English to French translation, we performed both automatic and manual sentence-level evaluation in a bid to reveal the advantages and disadvantages of our statistical post-editing systems under different translation conditions. Firstly, we report edit statistics for each sentence sorted by TER edit types (Snover et al., 2006). As a reasonable approximation to human post-editing efforts, these statistics can be of help in gauging the applicability of our post-editing systems in real world localisation work-flows. Secondly, we perform automatic sentence-level evaluation using sentence-level BLEU (SBLEU) (Liang et al., 2006) to classify the test sets into three subsets, i.e. “better”, “worse” and “equal”, based on the translation quality comparison between our post-editing systems and the baseline systems. This is a straightforward evaluation method to review the relative merits of each system, and can be used to view the overall strength and weakness of our post-editing systems.

4.1 Edit Statistics

Table 5 shows the number of average edits per sentence, based on the TER edit types i.e. insertion (Ins), substitution (Sub), deletion (Del) and shift and the number of errors (N-Err) for our French to English experiments. These numbers, like all of the numbers in this section, have been normalised using sentence length in order to make them comparable. The numbers show that the post-editing system without contextual information (PE) achieves slight gains in the “substitutions” and “shift” categories and as a result a reduction in the number of errors compared to baseline. This reflects that the PE system can improve the baseline translation in terms of better lexical choice (i.e. less substitutions) and better reordering (i.e. less shifts).

Additionally, we observe significant gains over both the baseline and simple PE in terms of “insertion”, “substitution” and “shifts”.

This demonstrates that restricted context-aware approaches are effective in improving reordering and lexical selection. On the other hand, we observe that the restricted context-aware systems also tend to produce longer translations which increases the “deletion” edits.

System	Ins	Del	Sub	Shift	N-Err
Baseline	0.67	0.83	1.52	0.77	3.58
PE	0.67	0.83	1.51	0.75	3.53
PE-CF	0.66	0.89	1.54	0.77	3.65
PE-C06	0.59	0.96	1.42	0.73	3.47
PE-C07	0.58	0.98	1.42	0.73	3.52
PE-C08	0.56	0.99	1.43	0.72	3.48
PE-C09	0.60	0.95	1.43	0.72	3.51

Table 5: TER Edit Statistics - French to English

Table 6 shows the number of TER edits for English to French translation. The numbers show that compared to the baseline all PE systems require more “insertion” and “substitution” edits, indicating that PE systems failed to produce gains resulting in improved translations and better lexical selection.

System	Ins	Del	Subs	Shifts	N-Err
Baseline	0.95	1.05	2.10	0.50	4.53
PE	1.02	0.83	2.19	0.50	4.53
PE-CF	0.98	0.91	2.29	0.52	4.70
PE-C06	1.02	0.89	2.21	0.50	4.62
PE-C07	0.99	0.91	2.19	0.50	4.50
PE-C08	1.00	0.89	2.19	0.50	4.58
PE-C09	1.04	0.88	2.21	0.51	4.64

Table 6: TER Edit Statistics - English to French

4.2 Sentence-level Automatic and Manual Analysis

As mentioned, we also performed a sentence-level BLEU (SBLEU) evaluation in order to identify the number of improved sentences produced by our post-editing systems and manually evaluated these sentences. Tables 7 and 8 show a summary of the number of sentences that got better, worse, or remained unchanged in the post-editing phase.

System	Better	Worse	Equal
PE	137	88	1742
PE-CF	489	511	967
PE-C06	511	451	1005
PE-C07	497	460	1010
PE-C08	528	455	930
PE-C09	496	454	1017

Table 7: Sentence Level Comparisons to Baseline Results - French to English

Table 7 shows that for the French to English translation direction only 225 out of 1967 baseline outputs were changed by the basic PE system. Among the affected sentences, 137 sentences are better than the baseline while 88 sentence are worse, indicating an overall positive effect using the simple PE system. For context-aware PE systems, a total of around a 1000 sentences were changed in the post-editing phase with more improved sentences than disimproved sentences.

A manual analysis of the improved sentences shows that almost half of improvements are lexical improvements, including insertions of missing words, deletions of unnecessarily added words, and translations of previously untranslated items. The analysis of the disimproved sentences also showed a high number of lexical changes, over 55%, mostly in terms of deletions of important items. Many sentences worsened in grammatical quality after post-editing. However, a larger percentage of sentences (41%) improved in terms of word order.

System	Better	Worse	Equal
PE	166	179	1662
PE-CF	259	434	1274
PE-C06	198	318	1451
PE-C07	238	317	1412
PE-C08	181	265	1531
PE-C09	170	306	1491

Table 8: Sentence Level Comparisons to Baseline Results - English to French

In the English-to-French translation, the lexical and grammatical improvements seem to be outweighed by the large amount of sentences that decline in quality. Table 10 shows a summary of the number of better and worse sentences between the baseline system and each of the post-editing systems. Compared to French-to-English translation, we can see fewer sentences changed in the post-editing phase.

Additionally, sentence length analysis shows that post-editing decreases the average sentence length by as much as 2.7% in the French to English direction, while it has very little effect on sentence length ($< 1\%$ increase) in the English to French direction. The effect of these seemingly random changes will be explored further in future investigations.

4.2.1 Examples

The examples in this section are taken from the French to English baseline system and the context-aware post-editing system with thresholding at 0.6 (PE-CF06). The outputs from the context-aware system have been filtered to remove the #f tags.

Example 1 - Paraphrasing

- **Source:** `supprimer un serveur de supports de la liste de publication .`

- **Baseline:** remove a media server from the list to publish to
- **PE:** delete a media servers in the list to publish to .
- **PE-C06:** remove a media server from the publishing list
- **Target:** remove a media server from the publishing list

Example 1 shows that the post-editing system post-edited “the list to publish to” to “the publishing list” with improvements in both ordering and lexical selection.

Example 2 - Recovery of previously deleted word

- **Source:** restauration manuelle utilisant l’ assistant disaster recovery
- **Baseline:** manual restore using the disaster recovery wizard
- **PE:** performing a manual restore by using the disaster recovery wizard
- **PE-C06:** performing a manual restore using the disaster recovery wizard
- **Target:** performing a manual restore using the disaster recovery wizard

Example 2 shows that the post-editing systems successfully recovered a missing translation “performing”.

Example 3 - Lexical Substitution: translation of a previously untranslated sentence

- **Source:** list published updates
- **Baseline:** list publiées updates .
- **PE:** list publiées updates
- **PE-C06:** list published updates.
- **Target:** list published updates.

Example 3 shows that the post-editing system successfully translates out-of-vocabulary words (cf. “publiés”) contained in the baseline translation output.

Example 4 - Worsening Through Deletion

- **Source:** si l’ ordinateur distant n’ a pas été ajouté aux ressources favorites
- **Baseline:** if the remote computer has not been added to favorite resources
- **PE:** if the remote computer has not been added to favorite resources
- **PE-C06:** if the remote computer has been added to favorite resources
- **Target:** if the remote computer has not been added to favorite resources

Example 4 demonstrates how the context-aware post-editing system deleted a word it deemed unnecessary (not) thereby changing the entire meaning of the sentence and dropping the translation quality.

The above examples clearly demonstrate the capabilities of statistical post-editing systems in improving lexical translation and word reordering. Given the degradations incurred together with the improvement, in the future, we plan to use statistical post-editing in a conservative manner, i.e. to automatically identify the segments where post-editing is desirable by using a classifier to predict which sentences profit the most from post-editing (Ma et al., 2011).

5 Conclusion and Further Work

Previous research on SPE has focused on pipelines with RBMT as a first-stage system followed by PB-SMT post-editing systems, trained on output of the RBMT system on the PB-SMT training set. The only references in the literature (as far as we are aware) that have considered PB-SMT systems as both first and second (and further) stage systems in an SPE pipeline is a short note (a single line) in Simard et al. (2007b), and a short section in Oflazer and El-Kahlout (2007), the former reporting that nothing could be gained from such a setup and the latter reporting small improvements < 0.5 BLEU (where it not known whether the result is statistically significant).

As far as we are aware, the research presented in this paper paper is the first attempt to analyse full SPE pipelines using PB-SMT systems throughout as both first and second stage systems more systematically. Our results show that a novel context-aware approach (PE-CF) with context alignment strength thresholding shows clear and statistically significant improvements of about 2 BLEU points absolute for all thresholds for the translation direction of French to English. As far as we are aware, this is the first time such improvements have been demonstrated for a purely PB-SMT based post-editing pipeline. By contrast, the “naive” approach to using statistical post-editing (PE) in a pure PB-SMT pipeline does not improve translation for English to French, and only shows a modest improvement for French to English.

In order to verify whether and to what extent our findings generalise to other languages and data sets. We will test this using English–German translation memory data in the same domain as well as standard JRC-Aquis and Europarl data sets. Moreover, we plan to conduct further investigation into the reasons why our SPE approaches improve French to English translation, but not English to French.

6 Acknowledgments

This work is supported by Science Foundation Ireland (Grant No. 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University. We thank the reviewers for their insightful comments and Symantec for providing us with the data.

References

Chris Callison-Burch, Phillip Koehn, Josh Schroeder, and Christof Monz. 2009. Findings of the 2009 workshop on statistical machine translation. In *Proceedings of the 4th EACL Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece.

Loic Dugast, Jean Senellart, and Phillip Koehn. 2007. Statistical post-editing on systran's rule-based translation system. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 220–223, Prague.

Loic Dugast, Philipp Koehn, and Jean Senellart. 2009. Statistical post-editing and dictionary extraction: Systran/edinburgh submissions for ACL-WMT2009. In *Systran/Edinburgh submissions for ACL-WMT2009*.

Teramusa Ehara. 2007. Rule based machine translation combined with statistical post-editor for Japanese to English patent translation. Tokyo University of Science, Suwas.

Philipp Koehn, Franz Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of the 2003 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics*, pages 48–54, Edmonton, AB, Canada.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Demo and Poster Sessions of the 45th Annual Meeting of the Association for Computational Linguistics (ACL07)*, pages 177–180. Prague, Czech Republic.

Percy Liang, Alexandre Bouchard-Cote, Dan Klein, and Ben Taskar. 2006. An end-to-end discriminative approach to machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th annual meeting of the ACL*, pages 761–768. Sydney Australia.

Yanjun Ma, Yifan He, Andy Way, and Josef van Genabith. 2011. Consistent translation using discriminative learning: A translation memory-inspired approach. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 1239–1248. Portland, OR.

Eric W. Noreen. 1989. Computer intensive methods for testing hypotheses: An introduction.

Franz Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.

Kemal Oflazer and Ilknur Durgar El-Kahlout. 2007. Exploring different representational units in english-to-turkish statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 25–32. Prague.

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007a. Statistical phrase-based post-editing. In *Proceedings of NAACL HLT 2007*, pages 508–515. Rochester, NY.

Michel Simard, Pierre Isabelle, and Cyrill Goutte. 2007b. Domain adaptation of MT systems through automatic post-editing. In *MT Summit XI*, pages 225–261, Copenhagen, Denmark.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate and targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231. Cambridge, MA.

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

Nicolas Stroppa and Karolina Owczarzak. 2007. A cluster-based representation for multi-system MT evaluation.