The trouble with SMT consistency

Carpuat, Marine; Simard, Michel

Publisher's version / Version de l'éditeur:

Proceedings of the 7th Workshop on Statistical Machine Translation, pp. 442-449, 2012-06-08

The Trouble with SMT Consistency

Marine Carpuat and **Michel Simard**
National Research Council Canada
283 Alexandre-Taché Boulevard
Building CRTL, Room F-2007
Gatineau (Québec) J8X 3X7
Firstname.Lastname@nrc.ca

Abstract

SMT typically models translation at the sentence level, ignoring wider document context. Does this hurt the consistency of translated documents? Using a phrase-based SMT system in various data conditions, we show that SMT translates documents remarkably consistently, even without document knowledge. Nevertheless, translation inconsistencies often indicate translation errors. However, unlike in human translation, these errors are rarely due to terminology inconsistency. They are more often symptoms of deeper issues with SMT models instead.

1 Introduction

While Statistical Machine Translation (SMT) models translation at the sentence level (Brown et al., 1993), human translators work on larger translation units. This is partly motivated by the importance of producing consistent translations at the document level. Consistency checking is part of the quality assurance process, and complying with the terminology requirements of each task or client is crucial. In fact, many automatic tools have been proposed to assist humans in this important task (Itagaki et al., 2007; Dagan and Church, 1994, among others).

This suggests that wider document-level context information might benefit SMT models. However, we do not have a clear picture of the impact of sentence-based SMT on the translation of full documents. From a quality standpoint, it seems safe to assume that translation consistency is as desirable for SMT as for human translations. However, consistency needs to be balanced with other quality requirements. For instance, strict consistency might result in awkward repetitions that make translations less fluent. From a translation modeling standpoint, while typical SMT systems do not explicitly enforce translation consistency, they can learn lexical choice preferences from training data in the right domain.

In this paper, we attempt to get a better understanding of SMT consistency. We conduct an empirical analysis using a phrase-based SMT system in a variety of experimental settings, focusing on two simple, yet understudied, questions. Is SMT output consistent at the document level? Do inconsistencies indicate translation errors?

We will see that SMT consistency issues are quite different from consistency issues in human translations. In fact, while inconsistency errors in SMT output might be particularly obvious to the human eye, SMT is globally about as consistent as human translations. Furthermore, high translation consistency does not guarantee quality: weaker SMT systems trained on less data translate more consistently than stronger larger systems. Yet, inconsistent translations often indicate translation errors, possibly because words and phrases that translate inconsistently are the hardest to translate.

After discussing related work on consistency and document modeling for SMT (Section 2), we describe our corpora in Section 3 and our general methodology in Section 4. In Section 5, we discuss the results of an automatic analysis of translation consistency, before turning to manual analysis in Section 6.

Proceedings of the 7th Workshop on Statistical Machine Translation, pages 442–449,
Montréal, Canada, June 7-8, 2012.

2 Related work

While most SMT systems operate at the sentence level, there is increased interest in modeling document context and consistency in translation.

In earlier work (Carpuat, 2009), we investigate whether the "one sense per discourse" heuristic commonly used in word sense disambiguation (Gale et al., 1992) can be useful in translation. We show that "one translation per discourse" largely holds in automatically word-aligned French-English news stories, and that enforcing translation consistency as a simple post-processing constraint can fix some of the translation errors in a phrase-based SMT system. Ture et al. (2012) provide further empirical support by studying the consistency of translation rules used by a hierarchical phrase-based system to force-decode Arabic-English news documents from the NIST evaluation.

Several recent contributions integrate translation consistency models in SMT using a two-pass decoding approach. In phrase-based SMT, Xiao et al. (2011) show that enforcing translation consistency using post-processing and redecoding techniques similar to those introduced in Carpuat (2009) can improve the BLEU score of a Chinese-English system. Ture et al. (2012) also show significant BLEU improvements on Arabic-English and Chinese-English hierarchical SMT systems. During the second decoding pass, Xiao et al. (2011) use only translation frequencies from the first pass to encourage consistency, while Ture et al. (2012) also model word rareness by adapting term weighting techniques from information retrieval.

Another line of work focuses on cache-based adaptive models (Tiedemann, 2010a; Gong et al., 2011), which lets lexical choice in a sentence be informed by translations of previous sentences. However, cache-based models are sensitive to error propagation and can have a negative impact on some data sets (Tiedemann, 2010b). Moreover, this approach blurs the line between consistency and domain modeling. In fact, Gong et al. (2011) reports statistically significant improvements in BLEU only when combining pure consistency caches with topic and similarity caches, which do not enforce consistency but essentially perform domain or topic adaptation.

There is also work that indirectly addresses consistency, by encouraging the re-use of translation memory matches (Ma et al., 2011), or by using a graph-based representation of the test set to promote similar translations for similar sentences (Alexandrescu and Kirchhoff, 2009).

All these results suggest that consistency can be a useful learning bias to improve overall translation quality, as measured by BLEU score. However, they do not yet give a clear picture of the translation consistency issues faced by SMT systems. In this paper, we directly check assumptions on SMT consistency in a systematic analysis of a strong phrase-based system in several large data conditions.

3 Translation Tasks

We use PORTAGE, the NRC's state-of-the-art phrase-based SMT system (Foster et al., 2009), in a number of settings. We consider different language pairs, translation directions, training sets of different nature, domain and sizes. Dataset statistics are summarized in Table 1, and a description follows.

Parliament condition These conditions are designed to illustrate an ideal situation: a SMT system trained on large high-quality in-domain data.

The training set consists of Canadian parliamentary text, approximately 160 million words in each language (Foster et al., 2010). The test set also consists of documents from the Canadian parliament: 807 English and 476 French documents. Each document contains transcript of speech by a single person, typically focusing on a single topic. The source-language documents are relatively short: the largest has 1079 words, the average being 116 words for English documents, 124 for French. For each document, we have two translations in the other language: the first is our SMT output; the second is a postedited version of that output, produced by translators of the Canadian Parliamentary Translation and Interpretation services.

Web condition This condition illustrates a perhaps more realistic situation: a "generic" SMT system, trained on large quantities of heterogeneous data, used to translate slightly out-of-domain text.

The SMT system is trained on a massive corpus of documents harvested from the Canadian federal government's Web domain "gc.ca": close to 40M

lang	train data	# tgt words	test data	#tgt words	#docs	BLEU	WER
en-fr	parl	167M	parl	104k	807	45.2	47.1
fr-en	parl	149M	parl	51k	446	58.0	31.9
en-fr	gov web	641M	gov doc	336k	3419	29.4	60.4
zh-en	small (fbis)	10.5M	nist08	41k	109	23.6	68.9
zh-en	large (nist09)	62.6M	nist08	41k	109	27.2	66.1

Table 1: Experimental data

unique English-French sentence pairs. The test set comes from a different source to guarantee that there is no overlap with the training data. It consists of more than 3000 English documents from a Canadian provincial government organization, totalling 336k words. Reference translations into French were produced by professional translators (not postedited). Documents are quite small, each typically focusing on a specific topic over a varied range of domains: agriculture, environment, finance, human resources, public services, education, social development, health, tourism, etc.

NIST conditions These conditions illustrate the situation with a very different language pair, Chinese-to-English, under two different scenarios: a system built using small in-domain data and one using large more heterogeneous data.

Following Chen et al. (2012), in the *Small* data condition, the SMT system is trained using the FBIS Chinese-English corpus (10.5M target words); the *Large* data condition uses all the allowed bilingual corpora from NIST Open Machine Translation Evaluation 2009 (MT09), except the *UN*, *Hong Kong Laws* and *Hong Kong Hansard* datasets, for a total of 62.6M target words. Each system is then used to translate 109 Chinese documents from the 2008 NIST evaluations (MT08) test set. For this dataset, we have access to four different reference translations. The documents are longer on average than for the previous conditions, with approximately 470 words per document.

4 Consistency Analysis Method

We study *repeated phrases*, which we define as a pair $\langle p, d \rangle$ where d is a document and p a phrase type that occurs more than once in d.

Since this study focuses on SMT lexical choice consistency, we base our analysis on the actual translation lexicon used by our phrase-based translation system (i.e., its phrase-table.) For each document d in a given collection of documents, we identify all source phrases p from the SMT phrase-table that occur more than once. We only consider source phrases that contain at least one content word.

We then collect the set of translations T for each occurrence of the repeated phrase in d. Using the word-alignment between source and translation, for each occurrence of p in d, we check whether p is aligned to one of its translation candidates in the phrase-table. A repeated phrase is translated consistently if all the strings in T are identical — ignoring differences due to punctuation and stopwords.

The word-alignment is given by the SMT decoder in SMT output, and is automatically infered from standard IBM models for the reference[1].

Note that, by design, this approach introduces a bias toward components of the SMT system. A human annotator asked to identify translation inconsistencies in the same data would not tag the exact same set of instances. Our approach might detect translation inconsistencies that a human would not annotate, because of alignment noise or negligible variation in translations for instance. We address these limitations in Section 6. Conversely, a human annotator would be able to identify inconsistencies for phrases that are not in the phrase-table vocabulary. Our approach is not designed to detect these inconsistencies, since we focus on understanding lexical choice inconsistencies based on the knowledge available to our SMT system at translation time.

[1] We considered using forced decoding to align the reference to the source, but lack of coverage led us to use IBM-style word alignment instead.

lang	train	test	translator	# repeated phrases	consistent (%)	avg within doc freq (inconsistent)	avg within doc freq (all)	#docs with repeated phrases	% consistent that match reference	% inconsistent that match reference	% easy fixes
en-fr	parl	parl	SMT	4186	73.03	2.627	2.414	529	70.82	34.37	10.12
en-fr	parl	parl	reference	3250	75.94	2.542	2.427	468			
fr-en	parl	parl	SMT	2048	85.35	2.453	2.351	303	82.72	52.67	3.52
fr-en	parl	parl	reference	1373	82.08	2.455	2.315	283			
en-fr	gov web	gov doc	SMT	79248	88.92	6.262	3.226	2982	60.71	13.05	15.53
en-fr	gov web	gov doc	reference	25300	82.73	4.071	2.889	2166			
zh-en	small	nist08	SMT	2300	63.61	2.983	2.725	109	56.25	18.40	9.81
zh-en	small	nist08	reference	1431	71.49	2.904	2.695	109			
zh-en	large	nist08	SMT	2417	60.20	3.055	2.717	109	60.00	17.88	10.89
zh-en	large	nist08	reference	1919	68.94	2.851	2.675	109			

Table 2: Statistics on the translation consistency of repeated phrases for SMT and references in five translation tasks. See Section 5 for details

5 Automatic Analysis

Table 2 reports various statistics for the translations of repeated phrases in SMT and human references, for all tasks described in Section 3.

5.1 Global SMT consistency

First, we observe that SMT is remarkably consistent. This suggests that consistency in the source-side local context is sufficient to constrain the SMT phrase-table and language model to produce consistent translations for most of the phrases considered in our experiments.

The column "*consistent (%)*" in Table 2 shows that the majority of repeated phrases are translated consistently for all translation tasks considered. For French-English tasks, the percentage of repeated phrases ranges from 73 to 89% . The consistency percentages are lower for Chinese-English, a more distant language pair. The *Parliament* task shows that translating into the morphologically richer language yields slightly lower consistency, all other dimensions being identical. However, morphological variations only explain part of the difference: translating into French under the *Web* condition yields the highest consistency percentage of all tasks, which might be explained by the very short and repetitive nature of the documents. As can be expected, inconsistently translated phrases are repeated in a document more often than average for all tasks (columns "*avg within doc freq*").

Interestingly, the smaller and weaker Chinese-English translation system (23.6 BLEU) is more consistent than its stronger counterpart (27.2 BLEU) according to the consistency percentages.The smaller training condition yields a smaller phrase-table with a lower coverage of the *nist08* source, fewer translation alternatives and therefore more consistent translations. Clearly consistency does not correlate with translation quality, and global consistency rates are not indicators of the translation quality of particular system.

5.2 Consistency of reference translations

Surprisingly, the percentage of consistently translated phrases are very close in SMT output and human references, and even higher in SMT for 2 out of 5 tasks (Table 2).

Note that there are fewer instances of repeated phrases for human references than for SMT, because the phrase-table used as a translation lexicon naturally covers SMT output better than independently produced human translations. Word alignment is also noisier between source and reference.

lang	train	test	translator	# repeated phrases	consistent (%)	avg within doc freq (inconsistent)	avg within doc freq (all)	#docs with repeated phrases	% consistent that match reference	% inconsistent that match reference	% easy fixes
zh-en	small	nist08	human1	1496	71.59	2.974	2.725	109	68.91	34.59	9.71
			human2	1356	69.40	2.913	2.687	109	73.22	36.63	7.60
			human2	1296	71.60	2.870	2.671	109	71.88	36.68	8.15
zh-en	large	nist08	human1	2017	70.25	2.943	2.692	109	66.13	30.83	9.64
			human2	1855	67.17	2.854	2.667	109	69.42	31.86	9.16
			human3	1739	69.70	2.854	2.660	109	68.23	33.78	8.31

Table 3: Statistics on the translation consistency of repeated phrases in the multiple human references available on the Chinese-English NIST08 test set. See Section 5 for details

There is a much wider gap in coherence percentages between references and SMT for Chinese-English than French-English tasks, as can be expected for the harder language pair. In addition, the same *nist08* reference translations are more consistent according to the phrase-table learned in the small training condition than according to the larger phrase-table. This confirms that consistency can signal a lack of coverage for new contexts.

5.3 Consistency and correctness

While translation consistency is generally assumed to be desirable, it does not guarantee correctness: SMT translations of repeated phrases can be consistent and incorrect, or inconsistent and correct. In order to evaluate correctness automatically, we check whether translations of repeated phrases are found in the corresponding reference sentences. This is an approximation since the translation of a source phrase can be correct even if it is not found in the reference, and a target phrase found in the reference sentence is not necessarily a correct translation of the source phrase considered. Post-edited references alleviate some approximation errors for the *Parliament* tasks: if the translated phrase matches the references, it means that it was considered correct by the human post-editor who left it in. However, phrases modified during post-edition are not necessarily incorrect. We will address this approximation in Section 6.

The columns "% consistent that match reference" and "% inconsistent that match reference" in Table 2 show that consistently translated phrases match the references more often than the inconsistent ones. With the post-edited references in the *Parliament* condition, a non-negligible percentage of consistently translated phrases are wrong: 17% when translating into English, and 30% when translating into French. In contrast, inconsistently translated phrases are more likely to be incorrect: more than 65% into French and 47% into English. For all other tasks, fewer translations match the references since the references are not produced by post-edition, but we still observe the same trend as in the *Parliament* condition: inconsistent translations are more likely to be incorrect than consistent translations overall.

Four reference translations are available for the Chinese-English *nist08* test set. We only use the first one as a reference translation (in order to minimize setting variations with French-English conditions.) The three remaining human translations are used differently. We compare them against the reference, exactly as we do for SMT output. The resulting statistics are given in Table 3. Since we know that the human translations are correct, this shows that many correct translations are not identified when using our simple match technique to check correctness. However, it is interesting to note that (1) consistent human translations tend to match the human references more often than the inconsistent ones, and (2) inconsistent MT translations match references much less often than inconsistent human references.

Language	Examples	False Inconsistencies			
	$\langle p, d \rangle$	Same lemma		Misaligned	
en→fr	79	15	(19%)	8	(10%)
fr→en	92	12	(13%)	24	(26%)
Total	171	27	(16%)	32	(19%)

Table 4: False positives in the automatic identification of translation inconsistencies.

What goes wrong when inconsistent translations are incorrect? This question is hard to answer with automatic analysis only. As a first approximation, we check whether we could correct translations by replacing them with machine translations produced elsewhere in the document. In Table 2, we refer to this as "easy fixes" and show that only very few inconsistency errors can be corrected this way. These errors are therefore unlikely to be fixed by post-processing approaches that enforce hard consistency constraints (Carpuat, 2009).

6 Manual Analysis

In order to better understand what goes wrong with inconsistent translations, we conduct a manual analysis of these errors in the *Parliament* test condition (see Table 1). We randomly sample inconsistently translated phrases, and examine a total of 174 repeated phrases ($\langle p, d \rangle$ pairs, as defined in Section 4.)

6.1 Methodological Issues

We first try to quantify the limitations of our approach, and verify whether the inconsistencies detected automatically are indeed real inconsistencies. The results of this analysis are presented in Table 4. Given the set of translations for a repeated phrase, we ask questions relating to morphology and automatic word-level alignment:

Morphology Are some of the alternate translations for phrase p only different inflections of the same lemma? Assuming that inflectional morphology is governed by language-internal considerations more often than translational constraints, it is probably inaccurate to label morphological variations of the same word as inconsistencies. The annotations reveal that this only happens for 16% of our sample (column "*Same lemma*" in Table 4). Work is under way to build an accurate French lemmatizer to automatically abstract away from morphological variations.

Alignment Are some of the alternate translations only a by-product of word alignment errors? This happens for instance when the French word *partis* is identified as being translated in English sometimes as *parties* and sometimes as *political* in the same document: the apparent inconsistency is actually due to an incorrect alignment within the frequent phrase *political parties*. We identify 19% of word alignment issues in our manually annotated sample (column "*Misaligned*" in Table 4). While it is clear that alignment errors should be avoided, it is worth noting that such errors are sometimes indicative of translation problems: this happens, for instance, when a key content word is left untranslated by the SMT system.

Overall, this analysis confirms that, despite the approximations used, a majority of the examples detected by our method are real inconsistencies.

6.2 Analysis of Translation Errors

We then directly evaluate translation accuracy in our sample by checking whether the system translation match the post-edited references. Here we focus our attention on those 112 examples from our sample of inconsistently translated phrases that do not suffer from lemmatization or misalignment problems. For comparison, we also analyze 200 randomly sampled examples of consistently translated phrases. Note that the identification of consistent phrases is not subject to alignment and lemmatization problems, which we therefore ignore in this case. Details of this analysis can be found in Table 5.

We first note that 40% of all inconsistently translated phrase types were not postedited at all: their translation can therefore be considered correct. In the case of consistently translated phrases, the rate of unedited translations rises to 75%.

Focusing now on those phrases whose translation was postedited, we classify each in one of three broad categories of MT errors: *meaning*, *terminology*, and *style/syntax* errors (columns labeled "*Type of Correction*" in Table 5).

Terminology Errors Surprisingly, among the inconsistently translated phrases, we find only 13% of true terminological consistency errors, where

	Language	Examples $\langle p, d \rangle$	Unedited	(%)	Type of Correction (% of edited examples) Meaning		Terminology		Style/Syntax	
Inconsistent	en→fr	56	20	(36%)	8	(22%)	4	(11%)	27	(75%)
Translations	fr→en	56	25	(45%)	10	(32%)	5	(16%)	20	(65%)
	Total	112	45	(40%)	16	(24%)	9	(13%)	47	(70%)
Consistent	en→fr	100	70	(70%)	3	(10%)	0	(0%)	27	(90%)
Translations	fr→en	100	79	(79%)	5	(24%)	0	(0%)	16	(76%)
	Total	200	149	(75%)	8	(16%)	0	(0%)	43	(84%)

Table 5: Manual Classification of Posteditor Corrections on the *Parliament* Task

the SMT output is acceptable but different from standard terminology in the test domain. For instance, the French term *personnes handicapées* can be translated as either *persons with disabilities* or *people with disabilities*, but the former is prefered in the Parliament domain. In the case of consistently translated phrases, no such errors were detected. This contrasts with human translation, where enforcing term consistency is a major concern. In the large-data in-domain condition considered here, SMT mostly translates terminology consistently and correctly. It remains to be seen whether this still holds when translating out-of-domain, or for different genres of documents.

Meaning Errors *Meaning* errors occur when the SMT output fails to convey the meaning of the source phrase. For example, in a medical context, our MT system sometimes translates the French word *examen* into English as *review* instead of the correct *test* or *investigation*. Such errors make up 24% of all corrections on inconsistently translated phrases, 16% in the case of consistent translations.

Style/Syntax Errors By far the most frequent category turns out to be *style/syntax errors* (70% of corrections on inconsistently translated phrases, 84% on consistently translated phrases): these are situations where the SMT output preserves the meaning of the source phrase, but is still post-edited for syntactic or stylistic preference. This category actually covers a wide range of corrections. The more benign cases are more cosmetic in nature, for example when the posteditor changes the MT output "*In terms of the cost associated with...*" into "*With regard to spending related to...*". In the more severe cases, the posteditor completely rewrites a seriously disfluent machine translation. However, errors to which we have assigned this label have a common denominator: the inconsistent phrase that is the focus of our attention is not the source of the error, but rather "collateral damage" in the war against mediocre translations.

Taken together, these results show that translation inconsistencies in SMT tend to be symptoms of generic SMT problems such as meaning and fluency or syntax errors. Only a minority of observed inconsistencies turn out to be the type of terminology inconsistencies that are a concern in human translations.

7 Conclusion

We have presented an in-depth study of machine translation consistency, using state-of-the-art SMT systems trained and evaluated under various realistic conditions. Our analysis highlights a number of important, and perhaps overlooked, issues regarding SMT consistency.

First, SMT systems translate documents remarkably consistently, even without explicit knowledge of extra-sentential context. They even exhibit global consistency levels comparable to that of professional human translators.

Second, high translation consistency does not correlate with better quality: as can be expected in phrase-based SMT, weaker systems trained on less data produce translations that are more consistent than higher-quality systems trained on larger more heterogeneous data sets.

However, this does not imply that inconsistencies are good either: inconsistently translated phrases coincide with translation errors much more often than consistent ones. In practice, translation inconsistency could therefore be used to detect words and phrases that are hard to translate for a given system.

Finally, manual inspection of inconsistent transla-

tions shows that only a small minority of errors are the kind of terminology problems that are the main concern in human translations. Instead, the majority of errors highlighted by inconsistent translations are symptoms of other problems, notably incorrect meaning translation, and syntactic or stylistic issues. These problems are just as prevalent with consistent as with inconsistent translations.

While directly enforcing translation consistency in MT may prove useful in some situations, our analysis suggests that the phrase-based SMT systems considered here would benefit more from directly tackling the underlying —- and admittedly more complex — problems of meaning and syntactic errors.

In future work, we plan to improve our analysis by extending our diagnosis methods, and consider additional data conditions and genres. We also plan to explore the potential of consistency for confidence estimation and error detection.

Acknowledgments

We would like to thank the Canadian Translation Bureau and the Portage team at the National Research Council for providing the post-edited and machine translations used in this study.

References

Andrei Alexandrescu and Katrin Kirchhoff. 2009. Graph-based learning for statistical machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 119–127, Boulder, CO, June.

Peter E. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–312.

Marine Carpuat. 2009. One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 19–27, Boulder, CO, June.

Boxing Chen, Roland Kuhn, and Samuel Larkin. 2012. PORT: a Precision-Order-Recall MT evaluation metric for Tuning. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL-2012)*.

Ido Dagan and Ken Church. 1994. Termight: Identifying and translating technical terminology. In *Proceedings of the Fourth Conference on Applied Natural Language Processing*, pages 34–40, Stuttgart, Germany, October.

George Foster, Boxing Chen, Eric Joanis, Howard Johnson, Roland Kuhn, and Samuel Larkin. 2009. PORTAGE in the NIST 2009 MT Evaluation. Technical report, NRC-CNRC.

George Foster, Pierre Isabelle, and Roland Kuhn. 2010. Translating structured documents. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010)*, Denver, Colorado, November.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. One Sense Per Discourse. In *Proceedings of the workshop on Speech and Natural Language*, Harriman, NY, February.

Zhengxian Gong, Min Zhang, and Guodong Zhou. 2011. Cache-based document-level statistical machine translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 909–919, July.

Masaki Itagaki, Takako Aikawa, and Xiaodong He. 2007. Automatic validation of terminology translation consistency with statistical method. In *Proceedings of Machine Translation Summit XI*, pages 269–274, September.

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: Human Language Technologies*, pages 1239–1248, Portland, Oregon, USA, June.

Jörg Tiedemann. 2010a. Context adaptation in statistical machine translation using models with exponentially decaying cache. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 8–15, Uppsala, Sweden, July.

Jörg Tiedemann. 2010b. To Cache or Not To Cache? Experiments with Adaptive Models in Statistical Machine Translation. In *Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 195–200, Uppsala, Sweden, July.

Ferhan Ture, Douglas W. Oard, and Philip Resnik. 2012. Encouraging consistent translation choices. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2012)*, Montreal, Canada, June.

Tong Xiao, Jingbo Zhu, Shujie Yao, and Hao Zhang. 2011. Document-level Consistency Verification in Machine Translation. In *Machine Translation Summit XIII*, pages 131–138, Xiamen, China, September.