Idiap Research Report

LEVERAGING COMPOUNDS TO IMPROVE NOUN PHRASE TRANSLATION FROM CHINESE AND GERMAN

Xiao Pu Laura Mascarell Andrei Popescu-Belis
Mark Fishel Ngoc-Quang Luong Martin Volk

Idiap-RR-09-2015

MAY 2015

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Leveraging Compounds to Improve Noun Phrase Translation from Chinese and German

First Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Second Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Abstract

This paper presents a method to improve the translation of polysemous nouns, leveraging on their previous occurrence as the head of a compound noun phrase. First, the occurrences are identified through pattern matching rules, which detect occurrences of an XY compound followed closely by a potentially coreferent occurrence of Y, such as "Mooncakes ... cakes ...". Second, two strategies are proposed to improve the translation of the second occurrence of Y: re-using the cached translation of Y from the XY compound, or post-editing the translation of Y using the head of the translation of XY. Experiments are performed on Chinese-to-English and German-to-French statistical machine translation, with about 250 occurrences of XY/Y, from the WIT3 and Text+Berg corpora. The results and their analysis suggest that while the overall BLEU scores increase only slightly, the translations of the targeted polysemous nouns are significantly improved.

1 Introduction

Words tend to be less ambiguous when considered together, which explains the success of phrase-based statistical machine translation (SMT) systems. In this paper, we take advantage of this observation, and extend the disambiguation potential of n-grams to subsequent occurrences of their individual components. To make this idea tractable, we assume that the translation of a noun-noun compound, noted XY, displays fewer ambiguities than the translation of its components X and Y considered individually. Therefore, on a subsequent occurrence of the head of XY, assumed to refer to the same entity, we hypothesize that its previously-found translation offers a better and more coherent translation than the one proposed by an SMT system that is not aware of the compound.

Our claim is supported by results from experiments on Chinese-to-English (ZH/EN) and German-to-French (DE/FR) translation presented in this paper. In both source languages, noun-noun XY compounds are frequent, and will enable us to disambiguate subsequent occurrences of their head, which is Y in both Chinese and German. For instance, in the example below, the Chinese compound 蔬菜 refers to 'vegetables', and the subsequent mention of this referent using only the second character (菜) should also be translated as 'vegetables'. However, the character 菜 by itself could also be translated as 'dish' or 'green', as seen in the example SMT output, which is not aware of the XY/Y coreference.

CHINESE SOURCE: 这是一种中国特有的**蔬菜** — 这种**菜**含有丰富的维他命。

HUMAN TRANSLATION: This is a special kind of **Chinese vegetables** — these **vegetables** are rich in vitamins.

SMT: This is a unique **Chinese vegetables** — this **dish** is rich in vitamins.

The paper is organized as follows. In Section 2 we present the main components of our proposal: first, the rules for identifying XY/Y pairs, and then the two alternative methods for improving the coherence of the translation of a subsequent mention Y, one based on post-editing and the other one based on caching.[1] In Section 3, we present our experimental setting. In Section 4, we evaluate our proposal on ZH/EN and DE/FR transla-

[1] Initial experiments with caching for DE/FR translation of compounds, with subjective evaluations, appeared in (Mascarell et al., 2014).

1. CHINESE SOURCE SENTENCE	她以为自己买了双两英寸的高跟鞋， 但实际上那是一双三英寸高的鞋。
2. SEGMENTATION, POS TAGGING, IDENTIFICATION OF COMPOUNDS AND THEIR CO-REFERENCE	她#PN 以为#VV 自己#AD 买#VV 了#AS 双#CD 两#CD 英寸#NN 的#DEG 高跟鞋#NN ，#PU 但#AD 实际上#AD 那#PN 是#VC 一#CD 双#M 三#CD 英寸#NN 高#VA 的#DEC 鞋#NN 。#PU
3. BASELINE TRANSLATION INTO ENGLISH (STATISTICAL MT)	She thought since bought a pair of two inches high heel, but in fact it was a pair of three inches high shoes.
4. AUTOMATIC POST-EDITING OF THE BASELINE TRANSLATION USING COMPOUNDS	She thought since bought a pair of two inches high heel, but in fact it was a pair of three inches high heel.
5. COMPARISON WITH A HUMAN REFERENCE TRANSLATION	She thought she'd gotten a two-inch heel but she'd actually bought a three-inch heel. ✓

Figure 1: Compound post-editing method illustrated on ZH/EN. The first translation of 高跟鞋 into 'heel' enables the correct translation of the subsequent occurrence of 鞋 as 'heel', by post-editing the baseline output 'shoes'.

tion, demonstrating that the translation of nouns is indeed improved, both through automatic and human evaluation, mainly based on comparison with the reference translation. We conclude with a brief discussion of related studies (Section 5) and with perspectives for future work (Section 6).

2 Description of the Method

2.1 Overview

The translation of a compound XY is used to improve the translation of a subsequent occurrence of Y, the head of the XY noun phrase, in the following way, represented schematically in Figure 1 (details for each stage are given below).

First, the presence of XY/Y patterns is detected either by examining whether a compound XY is followed by an occurrence of Y, or, conversely, by examining for each Y candidate whether it appears as part of a previous compound XY. Distance constraints and additional filtering rules are implemented to increase the likelihood that XY and Y are actually co-referent, or at least refer to entities of the same type.

Second, each sentence is translated by a baseline SMT system, and the translation of the head Y of each identified compound XY is identified using the word alignment from the SMT decoder. This translation is used as the translation of a subsequent occurrence of Y, either by caching the corresponding source/target word pair in the SMT, or by post-editing the baseline SMT output. For instance, if the Chinese pair (蔬菜, 菜) is identified, where the first compound can unambiguously be translated into English by 'vegetable', then the translation of a subsequent occurrence of 菜 is enforced to 'vegetable'. This has the potential to improve over the baseline translation, because when considered individually, 菜 could also be translated as 'dish', 'greens', 'wild herbs', etc.

2.2 Identifying XY/Y Pairs

Chinese and German share a number of similarities regarding compounds. Although Chinese texts are not word-segmented, once this operation is performed, multi-character words in which each character have individual meanings – such as the above-mentioned 蔬菜 ('vegetable') – are frequent. Similarly, in German, noun-noun compounds such as 'Bundesamt' ('Bund' + 'Amt', for Federal Bureau) or Nordwand ('Nord' + 'Wand', for North face) are frequent as well. While the identification of XY noun-noun compounds is straightforward with morpho-syntactic analysis tools, the identification of a subsequent mention of the head noun, Y, and especially the decision whether this Y refers to the same entity as XY, are more challenging issues. In other words, the main difficulty is to separate true XY/Y pairs from false positives. We present here our main strategies for identifying XY/Y pairs in Chinese and German, respectively.

For Chinese as a source language, given the large number of potential false positives XY/Y, we focus first on the identification of Y. In other words, if we searched first for noun phrases with an XY character structure, and then identified possible occurrences of Y, this would generate a large number of actually unrelated XY/Y pairs.

The processing steps are shown in Figure 1. Due to the fact that written Chinese is not word-segmented, we first use a segmentation tool (here the Stanford word segmenter[2]) to split the Chinese source sentence into a sequence of words (second row in Figure 1). Then, we perform POS tagging (with the Stanford log-linear part-of-speech tagger[3]), which enables us to recognize the nouns. Then, we identify possible referring expressions Y in Chinese using two patterns, one for singular forms and one for plural ones. For singular, we select all Y nouns preceded by 这 or 那 (meaning 'this' and 'that') and by a classifier word.[4] For plural, similarly, we select all Y nouns preceded by 这些 or 那些 (meaning 'these' and 'those'), without the need for a classifier word. Moreover, we set a distance constraint, and select only the cases in which the distance between the classifier and 这 or 那 is smaller than three words, with Y being the first or second noun after the classifier, in the singular case. Similarly, in the plural case, we select the Y which are the first or second nouns after 这些 or 那些. These values were determined empirically over the training data. In Figure 1, the selected Y (in orange) is 鞋 in the second line.

After determining Y, we search for the noun phrases in the three previous sentences which have the structure XY. In the example in Figure 1, the noun phrase 高跟鞋 (in blue) satisfies this condition. Therefore, the pair (高跟鞋 , 鞋) is a Chinese compound for which the translation of both occurrences of Y should be consistent.

For German as a source language, the processing steps are described in detail in a preliminary study (Mascarell et al., 2014). Contrary to Chinese, we first identify XY compounds using the Gertwol morphology system (Koskeniemmi and Haapalainen, 1994), which marks the different morphemes in compounds. We then search for the pattern *determiner* + Y, or alternatively *determiner* + *adjective* + Y, to select occurrences of Y which are likely to co-refer with the selected compound. We restrict the part-of-speech of the *determiner* to one of the following Gertwol labels: PDS (substituting demonstrative pronoun), PDAT (attributive demonstrative pronoun), PPOSS (substituting possessive pronoun), PPOSAT (attributive possessive pronoun), or ART (definite article). The maximum distance allowed between a compound XY and its co-reference by Y is four sentences, except for PDAT – a strong indicator of co-reference equivalent to 'this' – for which no maximal distance is required.

2.3 Enforcing the Translation of Y

Two language-independent methods have been designed to ensure that the translations of XY and Y are a consistent: post-editing vs. caching.

In the post-editing method, for each XY/Y pair, the translations of XY and Y by a baseline SMT system (see Section 3) are first identified through alignment with GIZA++. We verify if the translations of Y in both noun phrases are identical or different. both elements comprising the compound structure XY/Y are identified, for the standard cases, which only one possible XY referring to one Y, and the translation of both words are provided completely by the baseline system, while our system subsequently verifies if the translation of Y in both noun phrases are identical or different. We keep them intact in the first case, while in the second case we replace the translation of Y by the translation of XY, or by its head noun only if it contains several words. In the example in Figure1, XY is translated into 'high heel' and Y into 'shoes', which is a wrong translation of 鞋 in this context. Using the consistency constraint, our method post-edits the translation of Y 'heel', which is the correct word.

Several differences from the ideal case presented above must be handled separately. First, it may occur that several XY are likely co-referent with the same Y. In this case, if their translations differ, given that we cannot resolve the co-reference, we do not post-edit Y; if their translations are the same, but consist of one word, we still do not post-edit Y; we only change it if the translations consist of several words, ensuring that XY is a real compound. Second, the compound XY is sometimes out-of-vocabulary, hence it remains untranslated. In this case, we keep the translation of Y and use it also for the translation of XY.[5]

[2] http://nlp.stanford.edu/software/segmenter.shtml

[3] http://nlp.stanford.edu/software/tagger.shtml

[4] Classifier words in Chinese are mainly inserted between a numeral and the noun qualified by it, such as "one person" or "three books".

[5] This helps to improve BLEU scores, but does not affect

Third, sometimes the alignment of Y is empty in the target sentence (alignment error or untranslated word), in which case we apply post-editing as above on the word preceding Y, if it is aligned.

In the caching method, once a XY compound is identified, we obtain the translation of Y (as a part of the compound) through the word alignment given by the SMT decoder. Next, we check whether that translation appears as a translation candidate of Y in the phrase table of the baseline system, and if so, we cache (Tiedemann, 2010a) both Y and the obtained translation. We then enforce the cached translation every time a coreference Y is identified (Mascarell et al., 2014).

3 Experimental Settings

We first extract all pairs of compounds (XY, Y) with our methods from the WIT3 Chinese-English dataset, and from the Text+Berg corpus (Bubenhofer et al., 2013), a collection of documents in German-French from the Alpine domain. We then combined the sentences which included these noun phrases together as test data, while leaving the rest as training data for SMT. The size of the data sets used for the experiments are given in Table 1.

		Lines	**Tokens**
ZH	Training	188’758	19’880’790
	Tuning	2’457	260’770
	Testing	855	12’344
DE	Training	285’877	5’194’622
	Tuning	1’557	32’649
	Testing	505	12’499

Table 1: Size of SMT data sets.

The Chinese-English data comes from WIT (Web Inventory of Transcribed and Translated Talks), which is a ready-to-use version of the multilingual transcriptions of TED talks for research purposes. In the test data, there are 261 pairs of compounds (XY, Y) with different ”Y”s. Our baseline SMT system is Moses phrase-based one with the translation model is trained on...corpus, and the language model is trained by SRILM over...corpus .....

The effectiveness of proposed systems is measured by several metrics. First, BLEU score is used as an overall evaluation, to verify whether these systems provide better translation for the entire source text. Then, we break the assessment down to noun phrases co-referring to compounds. To do that, the number of cases where these NP translations match (mismatch) the reference, given the fact that the correspondent NPs of Baseline match (mismatch) will be computed. Among these values, we pay attention at the total of cases where each proposed system agrees with reference while Baseline does not, and that of the way round. The higher the former value is and the lower the later one is, the more beneficial our method will be.

However, measuring the effectiveness of these two above methods by automatic metrics is not a trivial and feasible task. The improvements yielded by them cannot constitute a significant gain of the overall BLEU score, since their occurrence presents a small percentage over the entire sentence. Furthermore, even if the post-edition is discrepant from the reference, it still can be more valuable than the hypothesis with closer meaning. For instance,

from the example of [Baseline $\neq$ ref. $\wedge$ Post-editing $\neq$ ref], we could find that, even the translation from Baseline (car) and Post-editing (bicycle) are not equal with Reference (bike), but the ”bicycle” looks closer to ”bike” compare with ”car”. So for such cases which both Post-editing and Baseline are not equal to Reference, we couldn’t determine if our system did any improvement or not, only by automatic BLEU score re-calculation.

Therefore, for evaluating the method’s usefulness, apart from automatic metrics, we conduct as well a manual analysis, which is briefly depicted as follows. All NPs translations which differ from references are considered by three human annotators. Each annotator puts the current translation into context (by looking at the previous sentences) to judge its quality over three levels: good (score 2), acceptable (score 1) and bad (score 0). Finally, the consensus of all annotator is computed to evaluate the system’s performance.

4 Analysis of Results

The BLEU scores given by baseline SMT and our method are as following (Table 2):

Of the total 261 pairs of relevant cases, we observe a total of 39 pairs in which XY couldn’t be translated and we modified the translation of XY by Y. Among the remaining 222 pairs, where XY and Y were completely translated by **BL**, there

the specific scoring of Y in Section 4.1.

	ZH/EN	DE/FR
BASELINE	11.18	27.65
CACHING	11.23	27.26
POST-EDITING	11.27	27.48

Table 2: BLEU scores of our methods.

were 79 cases in which Y was modified during post-editing, and 143 pairs it which the translation of the pairs were kept intact. The analysis of the results is as follows:

Table 3 represents the complete result from both PE and Baseline systems, compared to the reference. Among 222 extracted pairs, there were 45 cases in which the Post-editng system were equal to the reference while results from Baseline were not, which corresponds to a positive result for our method that will improve the BLEU score compared to the baseline. On the other hand, there were 10 cases in which the Baseline translation was equal to the reference but our Post-editing system altered them into another, which would drive the BLEU score down; there were 94 unmodified cases in which both Post-editing and Baseline produced equal results compared to reference; the last case in which both Post-editing and Baseline did not produce results equal to the reference amounts to 73 cases. Such cases could be further divided into 2 sub-categories: Post-editing and Baseline produced the same results (49 pairs) and where they did not (24 pairs).

The Caching method modified the translation of 123 Y cases in Chinese and 184 in German of the total 261 pairs. The coverage of Caching is lower than the Postedited, since the first only enforces a translation when it appears as one of the translation candidates of Y in the phrase table (Mascarell et al., 2014). Among the enforced translations, 17 of them from the Chinese-English test set and 8 from the German-French were improved, matching the reference. However, 5 cases in English and 19 in French worsen the BLEU score, since the baseline matches the reference, but not the enforced translation. The Caching method does not enforce a new translation and both Caching and Baseline match the reference in 73 (EN) and 129 (FR) cases. Among the cases that neither Caching nor Baseline match the reference, 21 cases in English and 17 in French share the same translation, and 7 (EN) and 11 (FR) do not. Overall, Caching improves the BLEU score for the Chinese-English language pair, but drives the score down for the German-French. This is due to the fact that most of the Baseline translations in French match already the reference, and Caching does not match the reference in more cases in French than in English.

It also can be seen from the results that both methods perform significantly better over ZH/EN than DE/FR language pair. While Caching system reaches on ZH/EN a total of 13.8% reference-matched cases when Baseline fails to do that, this percentage on DE/FR is only 4.3%. The situation is similar with Post-editing, when the gains difference is sharp: 20.3% vs. 3.5%, respectively for two above language pairs. This can be originated from the compounds and NPs alignment information, as well as the XY/Y identification in ZH/EN is more accurate than these in DE/FR.

4.1 Manual Evaluation of Undecided Cases

In the case where both the baseline and the post-editing systems generate translations of Y which differ from the reference (73 cases out of 222 for ZH/EN), it is not possible to compare the translations without having them examined by human subjects. Three of the authors, working independently, considered each translation along with the reference one, and rated it on a 3-point scale: 2 (good), 1 (acceptable) or 0 (wrong). To estimate their agreement, we computed the average absolute deviation (i.e. the average over all the ratings of $(\sum_{i=1}^{3} |\text{score}_i - \text{mean}|)/3$) and found a value of 0.15, thus denoting very good agreement. Below, we group '2' and '1' answers into one category, "acceptable", and compare them to '0' answers, i.e. wrong translations.

When both the baseline and the post-edited translations of Y differ from the reference, they can either be identical (49 cases) or different (24). In the former case, of course, neither of the systems outperforms the other. The interesting observation is that the relatively high number of such cases (49) is due to situations where the reference translation of noun Y is by a pronoun (40), which the systems have currently no possibility to generate from a noun in the source sentence. Manual evaluation shows that the systems' translations are correct in 36 out of 40 cases. This large number shows that the "quality" of the systems is actually higher that what can be inferred from Table 3 only. Conversely, in the 9 cases when the refer-

			CACHING		POST-EDITING	
			= ref	$\neq$ ref	= ref	$\neq$ ref
ZH/EN	BASELINE	= ref	59.3	*4.1*	42.3	*4.5*
		$\neq$ ref	**13.8**	22.8	**20.3**	32.9
DE/FR	BASELINE	= ref	70.1	*10.3*	73.9	*5.0*
		$\neq$ ref	**4.3**	15.2	**3.5**	17.5

Table 3: Comparison of each approach with the baseline, for the two language pairs, in terms of Y nouns which are identical or different from a reference translation ('ref'). All scores are percentages of the totals. Numbers in **bold** are improvements over the baseline, while those in *italics* are degradations.

ence translation of Y is not a pronoun, only about half of the translations are correct.

In the latter case, when baseline and post-edited translations differ from the reference *and* among themselves (24 cases), it is legitimate to ask which of the two systems is better. Overall, 10 baseline translations are correct and 14 are wrong, whereas 23 post-edited translations are correct (or at least acceptable) and only one is wrong. The post-edited system thus clearly outperforms the baseline one in this case. Similarly to the observation above, we note that among the 24 cases considered here, almost all (20) involve a reference translation of Y by a pronoun. In these cases, the baseline system translates only about half of them with a correct noun (9 out of 20), while the post-edited system translates correctly 19 out of 20.

5 Related Work

We briefly review in this section several previous studies from which the present one has benefited. Our idea is built upon the one-sense-per-discourse hypothesis (Gale et al., 1992) and its application to machine translation is based on the premise that consistency in discourse (Carpuat, 2009) is desirable. The initial compound idea was first published by Mascarell et al. (2014), in which the coreference of noun phrases (XY, Y) in German (e.g. Nordwand, Wand) is studied and used to improve translation from German by assuming that the last constituent of the compound Y should share the same translation as that of Y in XY. This is then used in into German-French translation.

Several other approaches focused on enforcing consistent lexical choice. (Tiedemann, 2010b) proposes a cache-model to enforce consistent translation of phrases across the document. However, caching is sensitive to error propagation, that is, when a phrase is incorrectly translated and cached, the model propagates the error to the following sentences. (Gong et al., 2011) extend later Tiedemann's proposal by initializing the cache with phrase pairs from similar documents at the beginning of the translation and by applying a topic cache. The latest is introduced to deal with the error propagation issue. (Xiao et al., 2011) describe a three steps procedure that enforces a consistent translation of ambiguous words, achieving improvements for English-Chinese language pair. (Ture et al., 2012) encourage consistency to the Arabic-English translation task by introducing cross-sentence consistency features to the translation model and (Alexandrescu and Kirchhoff, 2009) enforce similar translations to sentences having a similar graph representation.

6 Conclusion and Perspectives

We presented a method to enforce the consistent translation of coreferences to a compound, when the coreference matches with the final constituent of the compound coreferenced. We tried this method on results from Baseline on WIT Chinese-English data and compare with it. We then performed post-editing and returned the result back to the system in order to recalculate the BLEU score.

For the identification of coreference pairs, we detect the Chinese 这/那 (this/that) or 这些/那些 (these/those) which are strong indicators of coreferencing, and with some other constraints like distance and classifier in order to extract possible "Y"s in the compounds. If found, we return back to preceeding text to find the possible noun phrase XY which ends with Y, with additional distance constraints as well.

Moreover, we manually analyse other cases such as Postedited = Baseline $\neq$ reference, and Postedited $\neq$ Baseline $\neq$ reference. All results show that the consistency in both the reference and Baseline systems are reasonable.

Experimental results show that SMT systems often translate consistently coreferences to compounds. However, for some cases in which the noun phrase Y has multiple meanings, the original translation from Moses Baseline shows inconsistent results. Our system reduces the frequency of such mistranslations.

In the future, this work can be extended in various ways. In our work, we only considered conreference cases which matches ”this/that” or ”these/those” in Chinese. However, we are in the process of considering if we could extend the study to include other conditions, such as ”it.../they...”. For example:

Source: 这就是贝加尔湖。它是世界上最大的湖。全场800千米。

Reference: This is **Baikal lake**, it’s the biggest **lake** in the world.

However, due to the difference in construction of such cases compared to our previous methodology, the identification method for it should be considered separately.

Moreover, in our present work we performed post-editing after the original translation. We are considering the addition of features for the various cases before the original translation, so that it would affect other translations in the test in order to improve the overall translation result.

References

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

Noah Bubenhofer, Martin Volk, David Klaper, Manuela Weibel, and Daniel Wüest. 2013. Text+Berg-korpus (release 147_v03). Digitale Edition des Jahrbuch des SAC 1864-1923, Echo des Alpes 1872-1924 und Die Alpen 1925-2011.

Marine Carpuat. 2009. One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 19–27. Association for Computational Linguistics.

William A Gale, Kenneth W Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language*, pages 233–237. Association for Computational Linguistics.

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

Kimmo Koskeniemmi and Mariikka Haapalainen. 1994. Gertwol–lingsoft oy. *Linguistische Verifikation: Dokumentation zur Ersten Morpholympics*, pages 121–140.

Laura Mascarell, Mark Fishel, Natalia Korchagina, and Martin Volk. 2014. Enforcing consistent translation of german compound coreferences. In *Proceedings of the 12th Konvens Conference*.

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. Association for Computational Linguistics.

Jörg Tiedemann. 2010b. 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.

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*, pages 417–426, Montréal, Canada, June.

Tong Xiao, Jingbo Zhu, Shujie Yao, and Hao Zhang. 2011. Document-level consistency verification in machine translation. In *Proceedings of the 13th Machine Translation Summit*, volume 13, pages 131–138, Xiamen, China.