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.