Consistent Translation of Repeated Nouns using Syntactic and Semantic Cues

Xiao Pu[1,2], **Laura Mascarell**[3] **and Andrei Popescu-Belis**[1]

[1]Idiap Research Institute, 1920 Martigny, CH
[2]École Polytechnique Fédérale de Lausanne, 1015 Lausanne, CH
[3]Institute of Computational Linguistics, University of Zürich, 8050 Zürich, CH
{xiao.pu, andrei.popescu-belis}@idiap.ch
{mascarell}@cl.uzh.ch

Abstract

We propose a method to decide whether two occurrences of the same noun in a source text should be translated consistently, i.e. using the same noun in the target text as well. We train and test classifiers that predict consistent translations based on lexical, syntactic, and semantic features. We first evaluate the accuracy of our classifiers intrinsically, in terms of the accuracy of consistency predictions, over a subset of the UN Corpus. Then, we also evaluate them in combination with phrase-based statistical MT systems for Chinese-to-English and German-to-English. We compare the automatic post-editing of noun translations with the re-ranking of the translation hypotheses based on the classifiers' output, and also use these methods in combination. This improves over the baseline and closes up to 50% of the gap in BLEU scores between the baseline and an oracle classifier.

1 Introduction

The repetition of a noun in a text may be due to co-reference, i.e. repeated mentions of the same entity, or to mentions of two entities of the same type. But in other cases, two occurrences of the same noun may simply convey different meanings. The translation of repeated nouns depends, among other things, on the conveyed meanings: in case of co-reference or identical senses, they should likely be translated with the same word, while otherwise they should be translated with different words, if the target language distinguishes the two meanings. State-of-the-art machine translation systems do not address this challenge systematically, and translate two occurrences of the same noun independently, thus potentially introducing unwanted variations in translation.

We exemplify this issue in Figure 1 for Chinese-to-English and German-to-English translations, with examples of inconsistent translations of a repeated source noun by a baseline SMT system, as opposed to consistent translations in the reference. In Example 1, the system's translation of the second occurrence of *politik* is mistaken and should be replaced by the first one (*policy*, not *politics*). In Example 2, although the first translation differs from the reference, it could be acceptable as a legitimate variation, although the second one (*identity documents*) is more idiomatic and more frequent. Of course, in addition to these two examples, there are other configurations of the six nouns involved in a consistency relation across source, candidate and reference translations, but they will be discussed below when designing the training and test data for our problem.

In this paper, we aim to improve the translation of repeated nouns by designing a classifier which predicts, for every pair of repeated nouns in a source text, whether they should be translated by the same noun, i.e. consistently, and if that is the case, which of the two candidate translations generated by an MT system should replace the other one. We thus address one kind of long-range dependencies between words in texts; such dependencies have been the target of an increasing number of studies, presented briefly in Section 2.

To learn a consistency classifier from the data, we consider a corpus with source texts and reference translations, from the parallel UN Corpora in Chinese, German and English. As we explain in Section 3, we mine the corpus for pairs of repeated nouns in the source texts, and examine human and machine translations in order to learn to predict whether the machine translation of the first noun must replace the second one, or vice-versa, or no

Example 1

Source: nach einführung dieser **politik** [...] die **politik** auf dem gebiet der informationstechnik [...]

Reference: once the **policy** is implemented [...] the information technology **policy** [...]

MT: after introduction of **policy** [...] the **politics** in the area of information technology [...]

Example 2

Source: 欺诈性旅行或身份**证件**系指有下列情形之一的任何旅行或身份**证件**

Reference: Fraudulent travel or identity **document**; shall mean any travel or identity **document**

MT: 欺诈性 travel or identity **papers**. 系指 have under one condition; any travel, or identity **document**

Figure 1: Inconsistent translations of repeated nouns, in blue, from German (Example 1) and Chinese (Example 2) into English. While in both examples one noun is different from the reference, only Example 1 is truly mistaken: the second occurrence of the noun should be replaced with the first one.

change should be made. In Section 4, we present the lexical, syntactic and semantic features used by the classifiers. When presented with previously unseen source texts and baseline MT output, the decisions of the classifiers serve to post-edit or re-rank the repeated nouns of the MT baseline.

As shown in Section 5, the new end-to-end MT system generates improved Chinese-English and German-English translations, with larger improvements on the latter pair. Syntactic features appear to be more useful than semantic ones, for reasons that will be discussed. The case of more than two consecutive occurrences of the same noun will be briefly examined. Finally, a combined re-ranking and post-editing approach appears to be the most effective, covering about 50% of the gap in BLEU scores between the baseline MT and the use of an oracle classifier.

2 Related Work

This study is related to several research topics in MT: lexical consistency, caching, co-reference, and long-range dependencies between words in general. Our proposal aims to improve the consistency of noun translation, and thus has a narrower scope than the "one translation per discourse" hypothesis (Carpuat, 2009; Carpuat and Simard, 2012), which aimed to implement for MT the broader hypothesis of "one sense per discourse" (Gale et al., 1992).

We focus on nouns because of their referential properties, which are a strong requirement for consistency in case of co-reference, although in many cases consistency should not be blindly enforced, in order to avoid the "trouble with MT consistency" (Carpuat and Simard, 2012) which may induce translation errors. As indicated in that study, MT systems trained on small datasets are often more consistent but of lower quality than systems trained on larger and more diverse data sets. In any case, in our study, we never alter consistent translations, but we address inconsistencies, which are often translation errors (Carpuat and Simard, 2012), and attempt to find those that can be corrected simply by enforcing consistency.

Similarly, our scope is narrower than the caching approach (Tiedemann, 2010; Gong et al., 2011), which encourages *a priori* consistent translations of any word, with the risk on propagating cached incorrect translations. In our study, the first and second translation in a pair have equal status.

Noun phrase consistency is often due to co-reference. Several recent studies consider co-reference to improve pronoun resolution, but none of them exploits noun phrase co-reference, likely due to an insufficient accuracy of co-reference resolution systems (**?**; **?**). The improvement of pronoun translation was only marginal with respect to a baseline SMT system in a 2015 shared task (Hardmeier et al., 2015), while the 2016 shared task (Guillou et al., 2016) somewhat shifted its focus to pronoun prediction in a lemmatized reference translation.

This study builds upon and extends our previous work on the translation of compounds (Mascarell et al., 2014; Pu et al., 2015), which constrained the translation of the head of a compound when it was repeated separately after it. The present study is considerably more general, as it makes no assumption on either of the repeated nouns, i.e. it does not require them to be part of compounds.

Our study contributes to a growing corpus of research on modeling longer-range dependencies than those modeled in phrase-based SMT or neural MT, often across different sentences of a document. Ture et al. (2012) used cross-sentence consistency features in a translation model, while Hardmeier (2012) designed the Docent decoder, which can use document-level features to improve the coherence across sentences of a translated document. Our classifier for repeated nouns outputs decisions that can serve as features in Docent, but as the frequency of repeated nouns in documents is quite low, we use here post-editing and/or reranking rather than Docent.

3 Datasets for Noun Consistency in MT

3.1 Overview of the Method

Our method flexibly enforces noun consistency in discourse to improve noun phrase translation. We first detect two neighboring occurrences of the same noun in the source text, i.e. closer than a fixed distance, and which satisfy some basic conditions. Then, we consider their baseline translations by a phrase-based statistical MT system, which are identified from word-level alignments. If the two baseline translations of the repeated noun differ, then our classifier uses the source and target nouns and a large set of features (presented in Section 4) to decide whether one of the translations should be edited, and how. This decision will serve to post-edit and/or re-rank the baseline MT's output (Section 4.4). To design the classifier, we train machine-learning classifiers over examples that are extracted from parallel data and from a baseline MT system, as described in Section 3.3. A separate subset of unseen examples will be used to test classifiers, first intrinsically and then in combination with MT.

3.2 Corpora and Pre-processing

Our data comes from WIT^3 Corpus[1] (Cettolo et al., 2012), a collection of transcripts of TED talks, and the UN Corpora,[2] a collection of documents from the United Nations. The experiments are on Chinese-to-English and German-to-English.

We first build a phrase-based SMT system for each language pair with Moses (Koehn et al., 2007), with its default settings. Both MT systems are trained on the WIT^3 data, and are used to generate candidate translations of the UN Corpora. Then, the ML classifiers are trained on noun pairs extracted from the UN Corpora, using semantic and syntactic features extracted from both source and target sides. The test sets also come from the UN Corpora, with the same features on the source side. Table 1 presents statistics about the data.

3.3 Extraction of Training/Testing Instances

At this stage, the goal is to automatically extract for training the pairs of repeated nouns in the source texts, noted $N \dots N$, which are translated differently by the SMT baseline, noted $T_1 \dots T_2$, with $T_1 \neq T_2$. Indeed, when the translations are identical, we have no element in the 1-best translation to post-edit them, therefore we do not consider such pairs. We examine the reference translations of T_1 and T_2, noted RT_1 and RT_2, from which we derive the answer we expect from the classifiers (as specified below), and which will be used for supervised learning. We obtain the T_i and RT_i values using word-alignment with GIZA++.

Prior to the identification of repeated nouns in the source text, we tokenize the texts and identify parts-of-speech (POS) using the Stanford NLP tools[3]. In particular, as Chinese texts are not word-segmented, we first perform this operation and then identify multi-character nouns. We then consider each noun in turn, and look for a second occurrence of the same noun in what follows, limiting the search to the same sentence for Chinese, and to the same and next three sentences for German. The difference in the distance settings is based on observations of the Chinese vs. German datasets: average length of sentences, average distance of repeated nouns, and sentence segmentation issues.

Once the pairs of repeated nouns have been identified, we check the SMT translations of each pair, and if the two translations are different, we include the pair in our dataset. For instance, in Figure 1, the noun 证件 appears twice in the sentence, and the baseline translations of the two occurrences are *papers* and *document*; therefore, this pair is included in our dataset. We extracted from the UN Corpora 3,301 pairs for training and 647 pairs for testing on ZH/EN, and 11,289 pairs for training and 695 pairs for testing on DE/EN. We selected a smaller amount of noun pairs for ZH/EN than DE/EN for reasons of availability, because DE/EN dataset is more than 10 times larger than the ZH/EN one. We kept similar test set sizes to enable comparison.

The word-aligned reference translations are

[1] http://wit3.fbk.eu

[2] http://www.uncorpora.org

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

WIT^3	MT training		MT tuning		Language modeling	
	Sentences	Words	Sentences	Words	Sentences	Words
DE-EN	193,152	3.6M	2,052	40K	217K	4.4M
ZH-EN	185,443	3.4M	2,457	54K	4.8M	800M

UN Data	Classifier training			Classifier testing		
	Sentences	Words	Nouns	Sentences	Words	Noun
DE-EN	150K	4.5M	11,289	7,771	225K	695
ZH-EN	10K	368K	3,301	3,000	121K	647

Table 1: WIT^3 data for building the SMT systems and UN data to train/test the classifiers.

used to set the ground-truth class (or decision) for training the classifiers, as follows. With the notations above (baseline translations of N noted T_1 and T_2, with $T_1 \neq T_2$), if the reference translations differ ($RT_1 \neq RT_2$), then we label the pair as 'none', i.e. none of T_1 and T_2 should be post-edited and changed into the other, because this would not help to reach the reference translation anyway (recall that the only possible actions knowing the SMT baseline are replacing T_1 by T_2 or vice-versa).

If the reference translations are the same ($RT_1 = RT_2$), then we examine this word, noted RT. If this word is equal to one of the baseline translations ($T_1 = RT$ or $T_2 = RT$), then this value should be given to other baseline (e.g., if $T_1 = RT \neq T_2$, then $T_2 := T_1$). For classification, we simply label these examples with the index of the word that must be used, 1 or 2. However, if the reference differs from both baseline translations, then the label is again 'none', because we cannot infer which of them is a better translation.

After labeling all the pairs, we extract the features in an attribute/value format to be used for machine learning.

4 Classifiers for Translation Consistency

4.1 Role and Nature of the Classifiers

We describe here the machine learning classifiers that are trained to predict one of the three classes – '1', '2' or 'none' – for each pair of identical source nouns with different baseline SMT translations. The sense of the predicted classes is the following: '1' means that T_1 should replace T_2, '2' means the opposite, and 'none' means translations should be left unchanged. For instance, if Example 2 in Figure 1 was classified as '2', we would replace the translation of the first occurrence (*papers*) with the second one (*documents*).

We use the WEKA environment[4] to train and test several different learning algorithms: SVMs (Cortes and Vapnik, 1995), C4.5 Decision Trees (noted J48 in Weka) (Quinlan, 1993), and Random Forests (Breiman, 2001). We use 10-fold cross validation on the training set, and then test them once on the test set, and later on in combination with MT. For performance reasons, we used the Maximum Entropy classifier (Manning and Klein, 2003) from Stanford[5] instead of WEKA's Logistic Regression.

The hyper-parameters of the above classifiers were set as follows, mostly following the default settings from WEKA, and setting others on the cross-validation sets (not the unseen test sets). For SVMs, the round-off error is $\epsilon = 10^{-12}$. For Decision Trees, we set the minimal number of instances per leaf ('minNumObj') at 2 and the confidence factor used for pruning to 0.25. For Random Forests, we defined the number of trees to be generated ('numTree') as 100 and set their maximal depth ('maxDepth') as unlimited. Finally, we set the MaxEnt smoothing (σ) to 1.0, and the tolerance used for convergence in parameter optimization to 10^{-5}.

We evaluate our proposal in two ways. First, we measure the classification accuracy in terms of *accuracy* and *kappa* (κ) agreement (Cohen, 1960) with the correct class, either in 10-fold cross-validation experiments, or on the test set. Second, we compare the updated translations with the reference, to check if we obtain a result that is closer to it, using the popular BLEU measure (Papineni et al., 2002).

4.2 Syntactic Features

We defined 19 syntactic features, mainly with the assumption that out of a pair of repeated source nouns $N \ldots N$, the occurrence which is embed-

[4] http://www.cs.waikato.ac.nz/ml/weka/

[5] http://nlp.stanford.edu/software/classifier.shtml

ded in a more complex local parse tree, i.e. has more information syntactically bound to it, is more “determined” and has a higher probability of been translated correctly by the baseline MT system, since this information can help the system to disambiguate it. The results tend to confirm this assumption.

The features are listed in Figure 2, left side, with an explicit description of each feature and its value on a Chinese text (top of the figure). In the last line of the table we show the ground-truth class of this example.

The sentences are parsed using the Stanford parser,[6], and the values of the features are obtained from the parse trees, using the sizes (in nodes or words) of the siblings and ancestor sub-trees for each analyzed noun. In the sample parse trees on the right side of Figure 2, the first NP ancestor is marked with a red rectangle, and the values of the features are computed

We can distinguish three subsets of features. The first subset includes lexical and positional features: the original noun, automatic baseline translations of both occurrences from the baseline MT system, and the distance between the sentences that contain the two nouns. The second subset includes features that capture the size of the siblings in the parse trees of each of the two nouns. The third subset includes the size of sub-tree for the latest noun phrase ancestor for each analyzed noun, and also the depth distances to the next noun phrase ancestor.

4.3 Semantic Features

The semantic features, to be used independently or in combination with the syntactic ones, are divided into two groups: discourse vs. local context features, which differ by the amount of context they take into account. On the one hand, local context features represent the immediate context of each of the nouns in the pair and their translations, i.e. three words to their left and three words to their right in both source and MT output, always within the same sentence.

On the other hand, discourse features capture those cases where the inconsistent translations of a noun might be due to a disambiguation problem of the source noun, and semantic similarity can be leveraged to decide which of the two translations best matches the context. To compute the discourse features, we use the word2vec word vector representations generated from a large corpus (Mikolov et al., 2013), which have been successfully used in the recent past to compute similarity between words (Schnabel et al., 2015). Specifically, we employ the model trained on the English Google News corpus [7] with about 100 billion words.

For each pair of inconsistent translations (T_1, T_2) of a source noun N, we compute the cosine similarities c_1 and c_2 between the vector representation of each translation and the mean vector of their contexts. These mean vectors, noted $\vec{v}_1$ and $\vec{v}_2$, are computed by averaging all vectors of the words in the respective contexts of T_1 and T_2. Here, the contexts consist of 20 words to the left and 20 words to the right of each T_i, possibly crossing sentence boundaries. The cosine similarities c_1 and c_2 are thus:

$$c_1 = \cos(\vec{T}_1, \vec{v}_1) = \frac{\vec{T}_1 \cdot \vec{v}_1}{\|\vec{T}_1\| \|\vec{v}_1\|} \qquad (1)$$

$$c_2 = \cos(\vec{T}_2, \vec{v}_2) = \frac{\vec{T}_2 \cdot \vec{v}_2}{\|\vec{T}_2\| \|\vec{v}_2\|} \qquad (2)$$

The two values c_1 and c_2 are used as features, allowing classifiers to learn that, in principle, higher values indicate a better translation in the sense of its semantic similarity with the context.

In the Example 1 from Figure 1, the German word *Politik* is translated into the English words *policy* and then *politics*. The semantic similarity between the word *politics* and its context (c_2) is lower than the similarity between *policy* and its context (c_1), which we consider to be an indication that the first occurrence, namely *policy*, has better chances to be the correct translation – which is actually the case in this example.

4.4 Integration with the MT System

The classifier outputs a post-editing decision for each pair of repeated nouns: replace T_1 with T_2, replace T_2 with T_1, or do nothing. This decision can be directly executed, or it can be combined in a more nuanced fashion with the MT system. Therefore, to modify translations using this decision, we propose and test three approaches for using in in an MT system:

Post-editing: directly edit the translations T_1 or T_2 depending on the classifier’s decision.

[6] http://nlp.stanford.edu/software/lex-parser.html

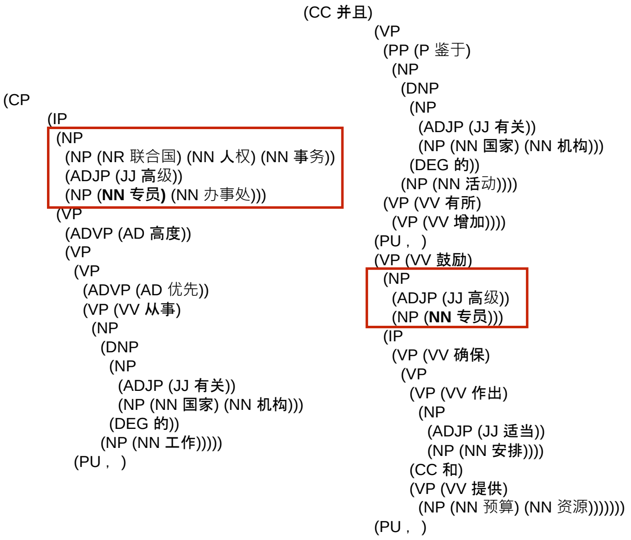
[7] https://code.google.com/p/word2vec/

Source: 赞扬 联合国 人权 事务 高级 专员 办事处 高度 优先 从事 有关 国家 机构 的 工作 ，[. . .] ， 鼓励 高级 专员 确保 作出 适当 安排 和 提供 预算 资源

Reference: commends the high priority given by the office of the united nations high **commissioner** for human rights to work on national institutions , [. . .] , encourages the high **commissioner** to ensure that appropriate arrangements are made and budgetary resources provided

MT: praise the human rights high **commissioner** was the high priority to offices in the country , [. . .] , to encourage senior **specialists** to make sure that make appropriate and provide budget resources

Features	**Values**
Source noun (Chinese)	专员
Distance in sentences between the two source occurrences	0
Translation of the first occurrence (labeled NN)	commissioner
Translation of the second occurrence (labeled NN)	specialists
Number of sibling nodes of the 1st occurrence	4
Number of sibling nodes of the 2nd occurrence	2
Sign of the difference between the above (+1, 0, −1)	1
Number of words of the 1st occurrence and its siblings	2
Number of words of the 2nd occurrence and its siblings	1
Sign of the difference between the above (+1, 0, −1)	1
Number of nodes in the first NP ancestor of 1st occ.	15
Number of nodes in the first NP ancestor of 2nd occ.	7
Sign of the difference between the above (+1, 0, −1)	1
Number of words in the first NP ancestor of the 1st occ.	6
Number of words in the first NP ancestor of the 2nd occ.	2
Sign of the difference between the above (+1, 0, −1)	1
Distance between the first NP ancestor and the 1st occ.	3
Distance between the first NP ancestor and the 2nd occ.	3
Sign of the difference between the above (+1, 0, −1)	0
Class (1, 2, 0)	1



```
(CP
  (IP
    (NP
      (NP (NR 联合国) (NN 人权) (NN 事务))
      (ADJP (JJ 高级))
      (NP (NN 专员) (NN 办事处)))
    (VP
      (ADVP (AD 高度))
      (VP
        (VP
          (ADVP (AD 优先))
          (VP (VV 从事)
            (NP
              (DNP
                (NP
                  (ADJP (JJ 有关))
                  (NP (NN 国家) (NN 机构)))
                (DEG 的))
              (NP (NN 工作)))))
        (PU ， )
```

```
(CC 并且)
  (VP
    (PP (P 鉴于)
      (NP
        (DNP
          (NP
            (ADJP (JJ 有关))
            (NP (NN 国家) (NN 机构)))
          (DEG 的))
        (NP (NN 活动))))
    (VP (VV 有所)
      (VP (VV 增加))))
  (PU ， )
  (VP (VV 鼓励)
    (NP
      (ADJP (JJ 高级))
      (NP (NN 专员)))
    (IP
      (VP (VV 确保)
        (VP
          (VP (VV 作出)
            (NP
              (ADJP (JJ 适当))
              (NP (NN 安排))))
          (CC 和)
          (VP (VV 提供)
            (NP (NN 预算) (NN 资源)))))))
  (PU ， )
```

Figure 2: Definition of syntactic features (left) and illustration of their values on a Chinese text (top). The red boxes in the parse trees (right) show the first NP ancestors of the examined nouns.

Re-ranking: search among the translation hypotheses provided by the SMT system (in practice, the first 10,000 ones) for those where T_1 and T_2 are translated as predicted by the classifier, and select the highest ranking one as the new translation. If none is found, the baseline 1-best hypothesis is kept.

Re-ranking + Post-editing: after applying re-ranking, if no hypothesis conforms to the prediction of the classifier, instead of keeping the baseline translation we post-edit it as in the first approach.

	Syntactic features		Semantic features		All features	
	Acc. (%)	κ	Acc. (%)	κ	Acc. (%)	κ
J48	72.1	0.48	60.2	0.00	60.2	0.00
SVM	74.5	0.54	60.2	0.00	73.9	0.51
RF	75.3	0.54	68.4	0.29	70.7	0.35
MaxEnt	76.7	0.65	69.5	0.32	**83.3**	**0.75**

Table 2: Prediction of the correct translation (1^{st} / 2^{nd} / None) for repeated nouns in *Chinese*, in terms of accuracy (%) and kappa scores, *on the development set* with 10-fold cross-validation. Methods are sorted by average accuracy over the three feature sets. When using semantic or all features, no decision tree outperformed the majority class baseline, hence $\kappa = 0$.

5 Results and Analysis

We first present the results of the classification task, i.e. the prediction of the correct translation variant (1^{st} / 2^{nd} / None), for Chinese-English and German-English translation respectively in Tables 2 and 3, with 10-fold cross-validation on the training sets. Then, we present the scores on the test sets for both the classification task and its combination with MT, for ZH/EN and DE/EN respectively in Tables 4 and 5. We compare the results obtained with several ML methods: Decision Trees (J48), SVMs, Random Forests and MaxEnt, ordered in the tables by average increasing scores. Moreover, we compare the merits of syntactic vs. semantic features, as well as post-editing vs. re-ranking the MT output.

	Syntactic features					Semantic features					All features				
	Acc.	κ	BLEU			Acc.	κ	BLEU			Acc.	κ	BLEU		
			PE	RR	RR+PE			PE	RR	RR+PE			PE	RR	RR+PE
Baseline	-	-	11.07	11.07	11.07	-	-	11.07	11.07	11.07	-	-	11.07	11.07	11.07
J48	66.3	0.42	11.17	11.20	11.30	33.1	0.00	11.07	11.07	11.07	33.1	0.00	11.07	11.07	11.07
SVM	71.9	0.53	11.23	11.27	11.33	33.1	0.00	11.07	11.07	11.07	62.1	0.43	11.18	11.26	11.26
RF	71.7	0.53	11.22	11.24	11.27	55.2	0.33	11.04	11.07	11.12	54.9	0.32	11.16	11.20	11.24
MaxEnt	73.7	0.60	11.27	11.33	**11.35**	56.1	0.34	10.87	11.11	11.18	72.5	0.56	11.21	11.33	**11.36**
Oracle	100	1.00	11.40	11.52	11.64	100	1.00	11.40	11.52	11.64	100	1.00	11.40	11.52	11.64

Table 4: Prediction of the correct translation (accuracy (%) and *kappa*) and translation quality (BLEU) for repeated nouns on the *Chinese test set*. Maximum Entropy was the best method found on the dev set.

	Syntactic features					Semantic features					All features				
	Acc.	κ	BLEU			Acc.	κ	BLEU			Acc.	κ	BLEU		
			PE	RR	RR+PE			PE	RR	RR+PE			PE	RR	RR+PE
Baseline	-	-	17.10	17.10	17.10	-	-	17.10	17.10	17.10	-	-	17.10	17.10	17.10
SVM	71.4	0.57	17.59	17.65	17.72	32.8	0.00	17.10	17.10	17.10	32.8	0.00	17.10	17.10	17.10
J48	70.5	0.56	17.59	17.61	17.70	48.2	0.23	17.13	17.27	17.33	69.4	0.54	17.56	17.60	17.66
RF	70.2	0.55	17.55	17.62	17.68	54.4	0.32	17.21	17.34	17.37	67.6	0.52	17.53	17.57	17.63
MaxEnt	78.3	0.67	17.63	17.66	**17.75**	63.5	0.49	17.39	17.47	17,49	68.7	0.53	17.58	17.59	**17.67**
Oracle	100	1.00	17.78	17.83	17.99	100	1.00	17.78	17.83	17.99	100	1.00	17.78	17.83	17.99

Table 5: Prediction of the correct translation (accuracy (%) and *kappa*) and translation quality (BLEU) for repeated nouns on the *German test set*. Maximum Entropy was the best method found on the dev set.

	Syntactic features		Semantic features		All features	
	Acc. (%)	κ	Acc. (%)	κ	Acc. (%)	κ
SVM	77.8	0.67	38.1	0.00	38.1	0.00
J48	77.0	0.66	64.8	0.45	79.7	0.69
RF	82.0	0.73	73.5	0.60	**84.5**	**0.77**
MaxEnt	80.8	0.71	76.8	0.65	83.4	0.75

Table 3: Prediction of the correct translation (1st / 2nd / None) for repeated nouns in *German*, in terms of accuracy (%) and kappa scores, *on the development set* with 10-fold c.-v. Methods are sorted by average accuracy over the three feature sets. The best scores are in bold.

5.1 Best Scores of Classification and MT

The classification accuracy is above 80% when applying 10-fold cross-validation, for both language pairs, and reaches 74–78% on the test sets. As the classes are quite balanced, a random baseline would reach around 33% only. Kappa values reach 0.75 on the dev sets and 0.60–0.67 on the test sets. The performances of the classifiers appear thus to be well above chance, and the comparable performances achieved on the unseen test sets indicate that over-fitting is unlikely.

The ordering of methods by performance is remarkably stable: Decision Trees (J48) and SVMs get the lowest scores, followed by Random Forests, and then by the MaxEnt classifier. The ordering $\{$J48, SVM$\} <$ RF $<$ MaxEnt is observed over both language pairs, over the three types of features, and the four datasets, with 1-2 exceptions only. Overall, the best configuration of our method found on the training sets is, for both languages, the MaxEnt classifier with all features.

There is a visible rank correlation between the increase in classification accuracy and the increase in BLEU score, for all languages, features, classifiers, and combination methods with MT. The best configurations found on the training sets bring the following BLEU improvements: for ZH/EN, from 11.07 to 11.36, and for DE/EN, from 17.10 to 17.67. In fact, syntactic features turn out to reach an even higher value on the test set, at 17.75. To interpret these improvements, they should be compared to the oracle BLEU scores obtained by using a "perfect" classifier, which are 11.64 for ZH/EN and 17.99 for DE/EN. Our method thus bridges 51% of the BLEU gap between baseline and oracle on ZH/EN and 64% on DE/EN – a significant improvement.

The BLEU scores of the three different methods for using classification for MT (Tables 4 and 5) clearly show that the combined method outperforms both post-editing and re-ranking alone, for all languages and features. Post-editing, the easiest one to implement, has little consideration for the words surrounding the nouns, while re-ranking works on MT hypotheses and thus ensures that a better global translation is found that is also consistent. However, in some cases, no hypothesis conforms to the consistency decision, and in this case post-editing the best hypothesis appears to be beneficial.

5.2 Feature Analysis: Syntax vs. Semantics

On the training sets, syntactic features always outperform the semantic ones when using the Max-

ZH/EN		DE/EN	
Translation of the 2$^{\text{nd}}$ occurrence	0.165	Translation of the 1$^{\text{st}}$ occurrence	0.162
Translation of the 1$^{\text{st}}$ occurrence	0.163	Translation of the 2$^{\text{nd}}$ occurrence	0.162
Source noun	0.110	Source noun	0.099
#words in the first NP ancestor of the 2$^{\text{nd}}$ occ.	0.060	#words in the first NP ancestor of the 2$^{\text{nd}}$ occ.	0.062
#words in the first NP ancestor of the 1$^{\text{st}}$ occ.	0.050	#words in the first NP ancestor of the 1$^{\text{st}}$ occ.	0.057
#nodes in the first NP ancestor of the 2$^{\text{nd}}$ occ.	0.036	#nodes of the 2$^{\text{nd}}$ occ.	0.054
#nodes in the first NP ancestor of the 1$^{\text{st}}$ occ.	0.033	#sibling nodes of the 1$^{\text{st}}$ occ.	0.052
Sign of difference between #words and its siblings	0.031	#nodes in the first NP ancestor of the 1$^{\text{st}}$ occ.	0.042
Dist. between the first NP ancestor and the 2$^{\text{nd}}$ occ.	0.025	#nodes in the first NP ancestor of the 2$^{\text{nd}}$ occ.	0.037
#words of the 1$^{\text{st}}$ occ. and its siblings	0.023	#words of the 2$^{\text{nd}}$ occ. and its siblings	0.037

Table 6: Top ten syntactic features ranked by information gain for each language pair.

Ent classifier, and their joint use outperforms their separate uses. For the other classifiers (not the best ones on the training sets), on ZH/EN, adding semantic features to syntactic ones decreases the performance. Indeed, semantic features (specifically the discourse ones) are intended to disambiguate nouns based on contexts, but here, manual inspection of the data showed that these are similar for T_1 and T_2, which makes prediction difficult.

Semantic features appear to be more useful in German compared to Chinese. We hypothesize that this is because translation ambiguities of Chinese nouns, i.e. cases when the same noun can be translated into English with two very different words, are less frequent and less semantically divergent than in German. In other words, semantic features are less useful in Chinese because cases of strong polysemy or homonymy seem to be less frequent than in German. Such a characteristic is suggested for English vs. Chinese by Huang (1995), and we believe it extends to German.

These facts might also explain the results obtained when using all features, for German and Chinese. As in Chinese semantic features are less helpful, given also the limited amount of data, combining them with syntactic ones actually decreases the performance of the syntactic ones used independently. In contrast, semantic features are more helpful on German dataset, and also improve results when we considered along with the syntactic ones together.

Table 6 shows the top ten syntactic features for ZH/EN and for DE/EN, ranked by information gain computed using Weka. These features include both lexical information and properties of the parse tree. The analysis shows that lexical features are significantly more important than purely syntactic ones, for both languages. However, the syntactic ones are not negligible.

		Local Context		Discourse		Both	
cosSim.	Inst.	Acc.	κ	Acc.	κ	Acc.	κ
0.0–0.1	141	63.8	0.27	73.8	0.47	66.0	0.31
0.1–0.2	341	70.1	0.40	75.4	0.51	71.0	0.42
0.2–0.3	350	73.1	0.43	68.0	0.35	72.3	0.41
0.3–0.4	350	72.6	0.45	66.0	0.32	68.6	0.37

Table 7: Effects of semantic similarity (cosSim) on classification (10-fold c.-v.). The scores with discourse features increase as similarity between T_1 and T_2 decreases.

Table 7 shows an analysis of the effect of the semantic features on different training sets in terms of accuracy and kappa scores. These training sets are built according to the cosine similarity between T_1 and T_2, as follows: for each training instance (pair of nouns), we compute the cosine similarity between the vector representation of T_1 and T_2; then, we group instances by intervals and carry out 10-fold c.-v. classification experiments for each subset. The lower the range values, the more dissimilar the translation pairs T_1 and T_2, and the better the scores of discourse features. Specifically, when the translations are dissimilar, the classifier makes better predictions with the discourse features, i.e. considering a larger context. However, the more similar the words are, the better the local context features, i.e. the surrounding words.

5.3 Extension to Triples of Repeated Nouns

Finally, we consider briefly the case of nouns that appear more than twice. Using our dataset, we identified them as noun pairs that share the same word, i.e. triples of repeated nouns, to which we limit our investigation. There are 129 ZH/EN triples and 138 DE/EN ones.

We defined the following method to determine the translation of such nouns when their baseline translations are different across the two pairs. If T_1, T_2 and T_3 are the translation candidates, we aim to find the consistent translation T_c as follows. If two of the T_i are identical, we use this value as T_c, but if they all differ, then we compare the syntactic features of the three source occurrences, and select the one with the highest number of features with highest values, and use its value as T_c. Going back to our classifier, if the decision for a particular instance pair is not 'none', then we replace the translations of the instance pairs with T_c.

We tested the method with the three feature types and the four classifiers, i.e. 12 cases per language. On ZH/EN, a small increase of BLEU is observed in 5 cases (0.01), a decrease in two cases (0.02), and no variation in 5 cases. On DE/EN, half of the cases show a small improvement (up to 0.03) and the rest stay the same. The method appears to work better on DE/EN, possibly because the initial accuracy on pairs is lower, but improvements are overall very small. The main conclusion from experimenting with triples, and considering also longer lexical chains of consistent nouns, is that the pairwise method should be replaced by a different type of consistency predictor, which remains to be found.

6 Conclusion and Perspectives

We presented a method for flexibly enforcing consistent translations of repeated nouns, by using a machine learning approach with syntactic and semantic features to decide when it should be enforced. We experimented with Chinese-English and German-English data. To build our datasets, we detected source-side nouns which appeared twice within a fixed distance and were translated differently by MT. Syntactic features were defined based on the complexity of the parse trees containing the nouns, thus capturing which of the two occurrences of a noun is more syntactically bound, while semantic features focused on the similarity between each translated noun and its context. The trained classifiers have shown that they can predict consistent translations above chance, and that, when combined to MT, bridge 50–60% of the gap between the baseline and an oracle classifier.

In future work, we will consider whether neural MT is prone to similar consistency problems, and whether they can be addressed by a similar method. The answer is likely positive, because both PBSMT and NMT assume that consistency simply results from correct individual translations, whereas human translators often take consistency into account for lexical choice. Moreover, a better consideration of legitimate lexical variation, e.g. using multiple references or human evaluators, should improve the assessment of consistency enforcement strategies.

Acknowledgments

We are grateful for their support to the Swiss National Science Foundation (SNSF) under the Sinergia MODERN project (grant n. 147653, see www.idiap.ch/project/modern/) and to the European Union under the Horizon 2020 SUMMA project (grant n. 688139, see www.summa-project.eu). We thank the three anonymous reviewers for their helpful suggestions.

References

Leo Breiman. 2001. Random Forests. *Machine Learning*, 45(1):5–32.

Marine Carpuat and Michel Simard. 2012. The trouble with SMT consistency. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, WMT '12, pages 442–449.

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

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20:37–46.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

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.

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 (EMNLP)*, pages 909–919, Edinburgh.

Liane Guillou, Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, Mauro Cettolo, Bonnie Webber, and Andrei Popescu-Belis. 2016. Findings of the 2016 WMT shared task on cross-lingual pronoun prediction. In

Proceedings of the First Conference on Machine Translation (WMT16), Berlin, Germany.

Christian Hardmeier, Joakim Nivre, and Jörg Tiedemann. 2012. Document-Wide Decoding for Phrase-Based Statistical Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Natural Language Learning (EMNLP-CoNLL)*, Jeju, Korea.

Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. 2015. Pronoun-focused MT and cross-lingual pronoun prediction: Findings of the 2015 DiscoMT shared task on pronoun translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16, Lisbon, Portugal.

Shuan Fan Huang. 1995. Chinese as a metonymic language. In *In Honor of William Wang: Interdisciplinary studies on Language and Language Change*, pages 223–252, Taipei, Taiwan. Pyramid Press.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbs. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of 45th Annual Meeting of the Association for Computational Linguistics (ACL), Demonstration Session*, pages 177–180, Prague, Czech Republic.

Christopher Manning and Dan Klein. 2003. Optimization, MaxEnt Models, and Conditional Estimation without Magic. In *Tutorial at HLT-NAACL and 41st ACL conferences*, Edmonton, Canada and Sapporo, Japan.

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

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at the International Conference on Learning Representations*, Scottsdale, Arizona, USA.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA.

Xiao Pu, Laura Mascarell, Andrei Popescu-Belis, Mark Fishel, Ngoc Quang Luong, and Martin Volk. 2015. Leveraging compounds to improve noun phrase translation from Chinese and German. In *Proceedings of the ACL-IJCNLP 2015 Student Research Workshop*, pages 8–15, Beijing, China.

J. R. Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, Lisbon, Portugal, September. Association for Computational Linguistics.

Jörg Tiedemann. 2010. 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.

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)*, pages 417–426, Montréal, Canada.