Encouraging Consistent Translation Choices

Ferhan Ture,[1] **Douglas W. Oard,**[2,4] **Philip Resnik**[3,4]

[1]Department of Computer Science
[2]College of Information Studies
[3]Department of Linguistics
[4]Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20740 USA
fture@cs.umd.edu, oard@umd.edu, resnik@umd.edu

Abstract

It has long been observed that monolingual text exhibits a tendency toward "one sense per discourse," and it has been argued that a related "one translation per discourse" constraint is operative in bilingual contexts as well. In this paper, we introduce a novel method using forced decoding to confirm the validity of this constraint, and we demonstrate that it can be exploited in order to improve machine translation quality. Three ways of incorporating such a preference into a hierarchical phrase-based MT model are proposed, and the approach where all three are combined yields the greatest improvements for both Arabic-English and Chinese-English translation experiments.

1 Introduction

In statistical Machine Translation (MT), the state-of-the-art approach is to translate phrases in the context of a sentence and to re-order those phrases appropriately. Intuitively, it seems as if it should also be possible to draw on information outside of a single sentence to further improve translation quality. In this paper, we challenge the conventional approach of translating each sentence independently, and argue that it can indeed also be beneficial to consider document-scale context when translating text. Motivated by the success of a "one sense per discourse" heuristic in Word Sense Disambiguation (WSD), we explore the potential benefit of leveraging a "one translation per discourse" heuristic in MT.

The paper is organized as follows. We begin with related work in Section 2. Next, we provide new confirmation that the hypothesized one-translation-per-discourse condition does indeed often hold, based on a novel analysis using forced decoding (Section 3). We incorporate this idea into a hierarchical MT framework by adding three new document-scale features to the translation model (Section 4). We then present experimental results demonstrating solid improvements in translation quality obtained by leveraging these features, both for Arabic-English (Ar-En) and Chinese-English (Zh-En) translation (Section 5). Conclusions and future work are presented in Section 6.

2 Related work

Exploiting discourse-level context has to date received only limited attention in MT research (e.g., (Giménez and Màrquez, 2007; Liu et al., 2010; Carpuat, 2009; Brown, 2008; Xiao et al., 2011)). Exploratory analysis of reference translations by Carpuat (2009) motivates a hypothesis that MT systems might benefit from the "one sense per discourse" heuristic, first introduced by Gale et al. (1992), which has proven to be effective in the context of WSD (Yarowsky, 1995). Carpuat's approach was to do post-processing on the translation output to impose a "one translation per discourse" constraint where the system would otherwise have made a different choice. A manual evaluation on a sample of sentences suggested promise from the technique, which Carpuat suggested in favor of exploring more integrated approaches.

Xiao et al. (2011) took this one step further and implement an approach where they identified ambiguous translations within each document, and at-

tempt to fix them by replacing each ambiguity with the most frequent translation choice. Based on their error analysis, the authors indicate two shortcomings when trying to find the correct translation of a given phrase. First, frequency may not provide sufficient information to distinguish between translation candidates, which is why we take rareness into account when scoring translation candidates. Another problem is, like any other heuristic, that there may be cases where the heuristic fails and there are multiple senses per discourse. Guaranteeing consistency hurts performance in such situations, which is why we implement the heuristic as a model feature, and let the model score decide for each case.

We are aware of a few other analyses that have shown promising results based on a similar motivation. For instance, Wasser and Dorr (2008)'s approach biases the MT system based on term statistics from relevant documents in comparable corpora. Ma et al. (2011) show that a translation memory can be used to find similar source sentences, and consecutively adapt translation choices towards consistency. Domain adaptation for MT has has also been shown to be useful in some cases (Bertoldi and Federico, 2009; Hildebrand et al., 2005; Sanchis-Trilles and Casacuberta, 2010; Tiedemann, 2010; Zhao et al., 2004), so to the extent we consider documents to be micro-domains we might expect similar approaches to be useful at document scale. Indeed, hints that such ideas may work have been available for some time. For example, there is clear evidence that the behavior of human translators can provide evidence that is often useful for automating WSD (Diab and Resnik, 2002; Ng et al., 2003). When coupled with the one-sense-per-discourse heuristic, this suggests that the reverse may also be true.

3 Exploratory analysis

It is well known that writing styles vary by genre, and in particular that the amount of vocabulary variation within a document depends to some extent on the genre (e.g., higher in poetry than in engineering writing). The degree to which authors tend to make consistent word choices in any particular genre is, therefore, an empirical question. In order to gain insight into the extent to which human translators make consistent vocabulary choices in the types of materials that we wish to translate (in this work, news stories), we first explore the degree of support for our one-translation-per-discourse hypothesis in the reference translations of a standard MT test collection.

We used the Ar-En MT08 data set, which contains 74 newswire documents with a total of 813 sentences, each of which has four reference translations. Throughout this paper we consistently use the document (i.e., one news story) as a convenient discourse unit, although of course finer-scale or broader-scale discourse units might also be explored in future work. Moreover, throughout this paper we use the hierarchical phrase-based translation system (Hiero), which is based on a synchronous context-free grammar (SCFG) model (Chiang, 2005). In a SCFG, the rule [X] ||| α ||| β indicates that context free expansion $X \rightarrow \alpha$ in the source language can occur synchronously with $X \rightarrow \beta$ in the target language. In this case, we call α the left hand side (LHS) of the rule, and β the right hand side (RHS) of the rule.

To determine the extent and nature of translation consistency choices made by human translators, we randomly selected one of the four sets of reference translations (first set, with id 0) and we used forced decoding to find all possible sequences of rules that could transform the source sentence into the target sentence. In forced decoding, given a pair of source and target sentences, and a grammar consisting of learned translation rules with associated probabilities, the decoder searches all possible derivations for the one sequence of rules that is most likely (under the learned translation model) to synchronously produce the source sentence on the LHS and the target sentence on the RHS. For instance, consider the following Arabic sentence as input:

الثلاثة الاعتداءات بين رابط .

and its uncased reference translation:

there is a link between the three attacks .

The following four rules, which are part of the SCFG learned from the the same translation pairs, allows the decoder to find a sequence of derivations that "translates" the source-side Arabic sentence into the

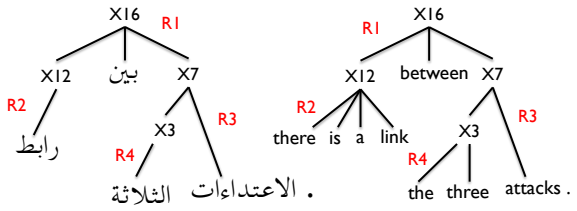

Figure 1: Illustration of forced decoding.

target-side reference translation.[1]

R_1. $[X_{12}]$ ||| رابط ||| there is a link
R_2. $[X_{16}]$ ||| [2] بين [1] ||| $[X_{12}, 1]$ between $[X_7, 2]$
R_3. $[X_7]$ ||| [1] الاعتداءات . ||| $[X_3, 1]$ attacks .
R_4. $[X_3]$ ||| الثلاثة ||| the three

Figure 1 illustrates how the decoder uses these rules to produce the source and target sides synchronously.

As we repeated this procedure for all sentence pairs, we kept track of all rules that were actually used by the decoder to generate a reference English translation from the corresponding Arabic sentences.

Our next step was to identify cases in which the SCFG could reasonably have produced a substantially different translation. Whenever an Arabic phrase f occurs multiple times in a document, and f appears on the LHS of two or more different grammar rules in the SCFG, we count this as a single "case".[2] These cases correspond to unique (source phrase f, document d) pairs in which a translation process using that SCFG *could have* chosen to produce two or more *different* translations of f in d. Since the multiple appearances of f are distributed among sentences of d, each counted case may correspond to a number of sentences ranging from 1 to the number of sentences in that document.

Table 1 shows a small sample of the cases (i.e., (source phrase f, document d) pairs) identified as a result of forced decoding. There were 321 such cases in our dataset and there were 672 sentences in which at least one case occurred. This is not an uncommon phenomenon; these 672 sentences comprise 83% of the test set. However, many of these cases represent either unlikely choices or inconsequential differences, so some post-processing is called for.

Since grammar rules are typically more fine-grained than is necessary for our purposes (e.g., to capture various punctuation and determiner differences that do not affect the "sense" of the translation), we applied a few simple heuristics to edit the source and target sides and group all such minor variations into a single "mega-rule" (e.g., "how"∼", how", "third"∼"a third", "want"∼"we want"). For this, we removed nonterminal symbols and punctuation, and considered two target phrases e and e' to be *different* only if $edit\_distance(e, e') > max(length(e), length(e'))/2$, where the edit distance is based on character removal and insertion. For instance, the third example in Table 1 would have been considered to be translated consistently as a result of this heuristic, as opposed to the first example. We also eliminated cases in which no reasonable alternatives were available in the translation grammar (i.e., cases where the second most probable rule with the same LHS was assigned a probability below 0.1 in the grammar). Cases 4 and 5 would have been removed by this heuristic.

After this filtering and aggregation we were left with 176 (f, d) pairs in which the translation model could reasonably have selected between rules that would have produced substantially different English translations of f in d (such as cases 1–3 and 6–9). It was these 176 cases, affecting a total of 512 sentences (63% of test set) for which we then examined what forced decoding could tell us about translation consistency.

So now that we know what the human who produced the reference translations actually did (according to forced decoding), and in which cases they might reasonably have chosen to do something substantially different (according to the SCFG), we can ask in which cases the human (effectively) made a consistent choice of translation rules when encountering the same Arabic phrase in the same document. In 128 of the 176 cases, that is what they did (i.e., when the same phrase occurred multiple times in a single document and more than one translation was reasonably possible, forced decoding indicated that the human translator translated that phrase in essentially the same way). These cases affected the trans-

[1] Since our goal was an exploratory analysis, the MT08 test set was combined with the training set in order to ensure reachability of the reference translations using the learned grammar. Proper train/dev/test splits were, of course, used for the evaluation results reported in Section 5.

[2] We define a phrase as any text that constitutes the entire LHS of a grammar rule.

Case		Translation counts
Source phrase	Doc #	
مقتل	566	that killed = 1 killing of = 1
الرهائن	782	hostages = 2
الرهائن	138	hostage = 1 hostages = 2
كوريا	466	korea = 2
كوريا	763	korea = 2
من	30	from = 2
من	7	of = 1 from = 1
الراهن من	717	of the current = 2
التي	30	the = 1 which =1

Table 1: A sample of cases (i.e., (source phrase f, document d) pairs) identified as a result of forced decoding.

lation of 455 sentences (56% of the test set), suggesting that if we can replicate this human behavior in a system, it might affect a nontrivial number of translation choices.

These statistics also suggest, however, that there may be some risk incurred in such a process, since in 48 of the 176 cases, the human translator opted for a substantially different translation. When we closely examined these 48 instances, we found that 19 (40%) involved changing a content-bearing word (sometimes to a word with similar meaning). The remaining 29 (60%) involved function words or similar constructions. See Figures 2 and 3 for examples.

1a. [X] ||| سمحت ||| had allowed
1b. [X] ||| سمحت ||| has permitted
2a. [X] ||| [X,1] تدرس ||| examining [X,1]
2b. [X] ||| [X,1] تدرس ||| is considering [X,1]
3a. [X] ||| [X,1] الجوار ||| neighbors
3b. [X] ||| [X,1] الجوار ||| neighboring countries

Figure 2: Examples of differences in lexical choice for content-bearing words within the same document.

We can make several observations based on this analysis. First, there does indeed seem to be evidence to support the one-translation-per-discourse heuristic, and to suggest that respecting that heuristic could improve translation outcomes for a substantial number of sentences. Second, even when a reference translation contains different translations of the same phrase, this may sometimes be the result of stylistic choices rather than an intent by the translator to affect the expressed meaning. If a system were try to "fix" such cases by enforcing consistent translation, the resulting translation might be somewhat more stilted, but perhaps not less accurate or less intelligible. Finally, sentence structure conventions or other language-specific phenomena may sometimes require the same phrase to be translated differently, so some way of encouraging consistency while still allowing the model to consider other contextual factors might be better than always imposing a hard consistency constraint.

4a. [X] ||| في ||| on
4b. [X] ||| في ||| in
4c. [X] ||| في ||| 's
5a. [X] ||| قد ||| had
5b. [X] ||| قد ||| was

Figure 3: Examples of differences in lexical choice for other types of lexical units within the same document.

4 Approach

To incorporate document-level features into an MT system that would otherwise operate with only sentence-level evidence, we added three supersentential "consistency features" to the translation model. The decoder computes scores for these features in two passes over each document; in each pass, each sentence in the document is decoded. In the first pass, the decoder keeps track of the number of occurrences of some aspects of each grammar rule and stores that information. The consistency features are disabled during this pass, and do not affect decoder scoring. In the second pass, each grammar rule is assigned as many as three consistency feature scores, each of which is based on some frozen counts from the first pass. These features are designed to introduce a bias towards translation consistency, but to leave the final decision to the decoder, which of course also has access to other features from the translation and language model. At this point we are more interested in effectiveness than efficiency, so

we simply note that this approach doubles the running time of the decoder and that future work on a more elegant implementation might be productive.

We explore three ways to compute features in this section. The essential idea behind all of them is to define some feature function that increases monotonically with an increase in some count that we believe to be informative, and in which the rate of increase is damped more strongly as that count increases. Several feature functions could satisfy those broad requirements; in this section, we describe three variants, C_1, C_2 and C_3, and discuss the potential benefits and drawbacks of each.

C_1: Counting rules In this variant, we count instances of the same entire grammar rule, where a rule r contains both the source phrase f and the target phrase e. During the first pass, whenever a grammar rule is chosen by the decoder for the one-best output, the count for that rule is incremented. Given a grammar rule r and the number of times r was counted in the first pass (given by $N\{r\}$), the consistency feature score is computed as follows:

$$C_1(r) = \frac{2.2N\{r\}}{1.2 + N\{r\}} \quad (1)$$

Equation 1 is the term frequency component of the well known Okapi BM25 term weighting function, when parameters are set to the conventional values $k = 1.2, b = 0$. This is an increasing and concave function in which the count has a diminishing marginal effect on the feature score. It has proven to be useful in information retrieval applications, in which the goal is to model "aboutness" based on term counts (Robertson et al., 1994). Because our goal is to demonstrate the potential of consistency features, it seemed reasonable to work with some simple function that has a shape like the one we desired. We leave exploration of optimal damping functions for future work.

A drawback of this C_1 approach is that as we saw in Section 3, grammar rules in phrase-based MT systems tend to be somewhat more fine-grained than seems optimal for constructing a consistency feature. For instance, consider the following rules that all translate the same Arabic term:

R_1. [X] ||| [X,1] اجهزة ||| [X,1] the bodies
R_2. [X] ||| [X,1] اجهزة ||| [X,1] the organs
R_3. [X] ||| [X,1] اجهزة ||| [X,1] organs
R_4. [X] ||| اجهزة [X,1] ||| the organs of [X,1]
R_5. [X] ||| اجهزة [X,1] ||| [X,1] bodies

Based on these grammar rules, we as human readers infer that this Arabic phrase can be translated in two different ways: as *organs* or as *bodies*. An optimal application of the one-translation-per-discourse heuristic would thus group the rules based on the presence of one of those words. However, in the C_1 variant, each of these rules would be counted separately because of differences that in some cases do not directly affect the choice of content words. For instance, on the source side, the Arabic token appears to the right of the nonterminal symbol in R_1, R_2 and R_3, while it is to the left of the nonterminal in R_4 and R_5. On the target side, differences are due to both nonterminal symbol position and the existence of determiners. Motivated by many examples like this, we came up with an alternative way of counting rules.

C_2: Counting target tokens To partially address this sparseness issue, variant C_2 focuses only on the target side. We extract all target tokens whenever a grammar rule is used by the decoder in a one-best derivation and increment a counter for each. Since we are mainly interested in content words (e.g. *bodies*, *organs*), we use simple pattern matching to discard nonterminal symbols and punctuation, and we ignore terms that appear in more than 50% of all documents (a convenient way of discarding common tokens such as *the*, *or*, *and*). This approach separates the rules in the example above into two groups: rules with *bodies* on the target side and rules with *organs* on the target side. Upon completion of the first pass, the consistency feature score for rule r is then determined by first computing a score for each unique target-side token w using:

$$bm25(w) = \frac{2.2N\{w\}}{1.2 + N\{w\}} \log \frac{D + 1}{DF(w) + 0.5} \quad (2)$$

where in this case $N\{w\}$ maps tokens to their respective counts in the document, D is the total number of documents in the collection, and DF (document frequency) is the number of documents in which the token occurs. This is a fuller version of the BM25 function in which (in the information retrieval application) both high term frequencies and rare terms

are rewarded. We then set the feature score for each rule r to the maximum score of any of its target-side terminal tokens:

$$C_2(r) = \max_{e \in RHS(r)} bm25(e) \quad (3)$$

Our motivation for choosing the maximum is that when there is more than one content word that survives the pruning of common terms, we want the score to be influenced most strongly by the most important of those terms. Since BM25 term weights can be thought of as a measure of term importance, taking the maximum is a simple expedient.

Although counting only target-side tokens yields coarser granularity than counting rules, ignoring the source side of the rule risks combining target side statistics from translations of unrelated source language terms. Consider the following grammar rule:

R_6. [X] ||| <s> [X,1] اجهزة ||| <s> [X,1] life support

Since the counter for *life* and *support* will both be incremented whenever rule R_6 fires in the one-best decoding during the first pass, problems could arise if a rule with a different LHS that also contains *support* on the RHS were to fire in the same document, for example:

R_7. [X] ||| الارها ||| support

If we don't take the source side into account, both occurrences of *support* will be grouped together when counting and R_7 will receive extra score from the consistency feature whenever R_6 is used by the decoder. Of course, this problem will only arise when the LHS of R_6 and R_7 are present in the same document, and how often that happens (and thus how large the risk from this factor is) is an empirical question. We therefore developed a third alternative as a middle ground between the fine-grained C_1 and the coarse-grained C_2.

C_3: Counting token translation pairs In this variant, we count each terminal (source token, target token) pair that survives pruning. Specifically, if grammar rule $[X]|||f_1 f_2 ... f_m|||e_1 e_2 ... e_n$ fires, we increment the count of every pair $\langle f_i, e_j \rangle$, where f_i is aligned to e_j. After the first pass, we compute the feature value of each observed pair, based on this count and the DF of the target-side of the pair. We chose to use only the target token in the DF computation (i.e., aggregating over all source tokens) to reduce sparsity effects. Similar to C_2, the feature of a rule r is defined by the maximum of scores of all pairs extracted from r.

$$C_3(r) = \max_{\substack{f \in LHS(r) \\ e \in RHS(r) \\ \langle f,e \rangle \text{ aligned}}} bm25(\langle f, e \rangle) \quad (4)$$

Since each variant has its benefits and drawbacks, we can include all three in the system and let the tuning process decide on how each should be weighted.

5 Evaluation and Discussion

We have evaluated the one-translation-per-discourse feature using the cdec MT system (Dyer et al., 2010). We started by building a baseline system using standard features in cdec: lexical and phrase translation probabilities in both directions, word and arity penalty features, and a 5-gram language model. We then added each of the three consistency feature variants, along with all two-way and the one three-way combinations of them, thus yielding a total of eight systems for comparison, including the baseline.

For training the Ar-En system, we used the dataset from the DARPA GALE evaluation (Olive et al., 2011), which consists of NIST and LDC releases. The corpus was filtered to remove sentence pairs with anomalous length ratios and subsampled to yield a training set containing 3.4 million parallel sentence pairs. The Arabic text was preprocessed to produce two different segmentations (simple punctuation tokenization with orthographic normalization, and LDC's ATBv3 representation (Maamouri et al., 2008)), represented together using cdec's lattice input format (Dyer et al., 2008).

The Zh-En system was trained on parallel training text consisting of the non-UN portions and non-HK Hansards portions of the NIST training corpora. Chinese was automatically segmented by the Stanford segmenter (Tseng et al., 2005), and traditional characters were simplified. After subsampling and filtering, we obtain a training corpus of 1.6 million parallel sentences.

Both training sets were word-aligned with GIZA++ (Och and Ney, 2003), using 5 Model 1 and 5 HMM iterations. A SCFG was then extracted from these alignments using a suffix array extractor (Chiang, 2007). Evaluation was done with multi-reference BLEU (Papineni et al., 2002) on test

sets with four references for each language pair, and MIRA was used for tuning (Crammer et al., 2006). In our experiments, we run the first decoding phase using feature weights that are guessed heuristically based on weights from previously tuned systems. All feature weights, including the discourse feature, were then tuned together, based on the output of the second decoding phase. For Ar-En parameter tuning, we used the MT06 newswire dataset, which contains 104 documents and a total of 1,797 sentences. For testing, we used the MT08 dataset described above (74 documents, 813 sentences). For Zh-En experiments, the MT02 newswire dataset (100 documents, 878 sentences) was used for tuning, and evaluation was done on the MT06 test set (79 documents, 1,664 sentences). For both language pairs, DF values were computed from the tuning set for both tuning and evaluation experiments.

When we used NIST's official metric (BLEU-4) to compare our results to the official NIST evaluation (NIST, 2006; NIST, 2008), our baseline system achieved 54.70 for Ar-En and 31.69 for Zh-En. Based on reported NIST results, our baseline would have ranked 4^{th} in the Zh-En MT06 evaluation, and would have outperformed all Ar-En MT08 systems. We used a slightly different IBM-BLEU metric for the rest of our evaluation. In this case, the baseline system achieved 53.07 BLEU points for Ar-En and 30.43 points for Zh-En. Among more recent papers, the best reported results were 56.87 for Ar-En MT08 (Zhao et al., 2011a) and 35.87 for Zh-En MT06 (Zhao et al., 2011b), although many papers report BLEU scores below 53 points for Arabic (Carpuat et al., 2011) and 32 points for Chinese (Monz, 2011). The systems that outperformed our baseline applied novel techniques, and used larger language models, as well as many non-standard features. We argue that these novelties are complementary to our approach, and therefore do not damage the credibility of our baseline.

Among the single-feature runs, C_3 had the best performance in Ar-En experiments, with 53.84 BLEU points, whereas C_2 yielded the best results for Zh-En with a BLEU score of 30.96. In any case, all three variants outperformed the baseline (see Table 2). When multiple features were combined, we generally observed an increase in BLEU, suggesting that our features have usefully different error characteristics. The combination of all three variants, C_{123}, yielded the best results, nearly 1.0 BLEU point higher than the baseline for both language pairs. Evaluation results are summarized in Table 2.

Method	**BLEU**	
	Ar-En	Zh-En
Baseline	53.07	30.43
C_1	53.82	30.59
C_2	53.70	30.96
C_3	53.84	30.54
C_{12}	53.82	30.79
C_{13}	53.82	30.76
C_{23}	53.88	30.63
C_{123}	**53.98**	**31.42**

Table 2: Evaluation results: BLEU scores with four references for Ar-En and Zh-En experiments.

Method	**# documents**	
	Ar-En	Zh-En
	74	79
C_1	37	30
C_2	37	35
C_3	42	36
C_{123}	**43**	**41**

Table 3: Document-level analysis: Number of documents where each variant outperforms baseline.

Given our focus on documents, it is natural to ask what fraction of the documents were helped or harmed by consistency features. Document-level BLEU scores for Arabic-to-English translations show that C_3 outperformed the baseline on a larger number of documents than any other single feature (42/74=57%), compared with 37/74 (50%) for both C_1 and C_2. C_{123} did better by this measure as well, with BLEU increasing for 43 of the documents. There were no documents where the BLEU score was exactly the same, therefore the BLEU score declined for the remaining documents. As Table 3 indicates, document-level BLEU for the Zh-En experiments shows similar results.

We can also look at our results in a more fine-grained way, focusing on differences in how each system translated the same source-language phrase. For this analysis, we defined English phrases e and e' to be *different* if $edit\_distance(e, e') >$

Method	Ar-En		Zh-En	
	Cases	Test set	Cases	Test set
C_1	77	24%	401	48%
C_2	127	35%	686	60%
C_3	101	33%	491	53%
Any	197	68%	968	94%
C_{123}	141	41%	651	59%

Table 4: Effect of applying variants of the consistency feature (Any=C_1 or C_2 or C_3).

$max(length(e), length(e'))/2$. By this way of counting, there are 197 unique (Arabic phrase, document) pairs for which at least one single-feature system produced translations differently from the baseline system. Together, these cases affect 553 sentences (68%) in 67 of the 74 documents, with as many as 12 differences observed in a single document. The number of such differences is even higher for Chinese-to-English translation, probably due to lower confidence from the translation model and longer documents. Table 4 shows the number of changes by each system, and the percentage of the test set affected by these changes.

In order to gain greater insight into the effect of the consistency features, we randomly sampled 60 of the 197 cases and analyzed the influence of the change to the document BLEU score. In 26 of the sampled cases, at least one of the three systems made a change that improved the BLEU score, whereas the score was adversely affected for at least one system in 14 cases. BLEU remained unchanged in 21 cases,[3] mostly due to the use of multiple reference translations. When we analyze the effect of each system separately, we see that C_2 was the most aggressive, making 25 changes that influenced BLEU (16 positive, 9 negative). C_1 was the most conservative, with only 13 such changes (8 positive, 5 negative). Consistent with the overall BLEU scores, C_3 evidenced the best ratio between benefit and harm, making 20 changes that affected the score (16 positive, 4 negative).

Looking at specific cases can yield some insight into how the consistency features achieve improvements. For example, results improved when translating the phrase تنظيمية, (Eng. *organizational*, *regulatory*), which appears in the context of organizational groups that support terrorist ideology. The baseline system translated this as *organizational* in one sentence, and *regulatory* in another. Variants C_1 and C_2 changed this behavior, so that the translation was *organizational* in both cases. One of the reference translations used *organizational* in one case and dropped the phrase in the other, and the other three translators provided consistent translations (using *organized* and *organizational*). As a result, applying the one-translation-per-discourse heuristic improved the multi-reference BLEU score.

On the other hand, here is one of the cases where our feature hurt performance. The phrase 边防部队 (Eng. *border/frontier troops/guards*) appears in two sentences of a Chinese news story about violence along the India - Nepal border. All reference translations consistently used the word *border* in the translation, as it is a better choice in this context. The baseline system translated the phrase as *frontier guards* and *border troops* in the two sentences. All system variants replaced *border* with *frontier* to maintain consistency, and therefore produced worse translations, causing a decrease in BLEU score.

Examples can, however, also point up limitations in our ability to measure improvements. In one of the test documents, the Arabic phrase الي التسلل (Eng. *sneak*, *infiltrate*, *enter without approval*) appears in the context of Turkey trying to enter the European Union. This was translated by the baseline system as *sneak into* in one occurrence and *infiltrate into* in another. C_1 didn't change the output, but C_2 and C_3 translated the phrase as *infiltrate into* in both cases. Although all of the four reference translators were consistent within their choices, each of them chose different translations, namely *worm its way*, *enter*, *sneak* and *sneak into*. This resulted in a decrease in BLEU score for the two systems that chose *infiltrate into*. This case illustrates a limitation to fine-grained use of BLEU alone as a basis for analysis, since we might argue that *infiltrate into* is no less appropriate than *sneak into* in this context. In other words, some of the reductions we see in BLEU may not be actual errors but rather simply changes that take us outside of the coverage of the test set. We did not find any cases in our sample

[3]There was one case for which one system improved overall BLEU and another reduced it.

in which improvements in BLEU seemed to reward changes that adversely affected meaning. From this, we conclude that BLEU is a somewhat conservative measure when used in this way, and that the actual overall improvement in translation quality over our baseline may be somewhat more than our roughly 1.0 measured BLEU improvement would suggest.

6 Conclusions and Future Work

In this paper, we started with a new way of looking at, and largely supporting, the "one translation per discourse" hypothesis using forced decoding of human reference translations. We then leveraged insights from that analysis to design the translation model consistency features, obtaining solid improvements for both Ar-En and Zh-En translation. In future work, we plan to explore additional variants. For example, we can further address sparsity by incorporating monolingual paraphrase detection on the source side, the target side or both. We can and should explore other monotonically increasing concave feature functions in addition to the Okapi BM25 function that we have found to be useful in this work, we should explore alternatives to our use of the maximum function in C_2 and C_3, and we should consider optimizing to measures other than BLEU (e.g., METEOR) that extend the range of rewarded lexical choices by leveraging monolingual paraphrase evidence.

In designing our features we were guided by our intuition about which kinds of consistency should be rewarded. Data can be superior to intuition, however, and our forced decoding technique might also be helpful in generating new insights that could help to guide the design of even more useful features. For example, our forced decoding clearly points to cases in which translators have chosen different structural variants when translating the same phrase, and closer examination of these cases might help us to automatically detect which kinds of structural variation can most profitably be moderated using a consistency feature. We should also note that we have only done forced decoding to date in one language pair (Ar-En), and there might be more to be learned about language-specific issues from doing the same analysis for additional language pairs.

Finally, the time seems propitious to reconsider our choice of document-scale as our discourse context. Documents have much to recommend them, but much of the content that we might wish to translate (conversational speech, text chat, email threads, ...) doesn't present the kinds of obvious and unambiguous document boundaries that we find in MT test collections that are built from news stories. Moreover, some documents (e.g., textbooks) may be too diverse for an entire document to be the right scale for consistency. We might also be able to productively group similar documents into clusters in which the vocabulary choices are (or should be) mutually reinforcing.

We therefore end where we began, with many questions to be answered. Now, however, we have somewhat different questions – not *whether* to encourage consistency at a super-sentential scale, but rather *when* and *how best* to do that.

Acknowledgements

This research was supported in part by the BOLT program of the Defense Advanced Research Projects Agency, Contract No. HR0011-12-C-0015. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the view of DARPA.

References

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation (StatMT '09)*, pages 182–189.

Ralf D. Brown. 2008. Exploiting document-level context for data-driven machine translation. In *Proceedings of the the Eighth Conference of the Association for Machine Translation in the Americas (AMTA '08)*.

Marine Carpuat, Yuval Marton, and Nizar Habash. 2011. Improved Arabic-to-English statistical machine translation by reordering post-verbal subjects for word alignment. *Machine Translation*, pages 1–16.

Marine Carpuat. 2009. One translation per discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, DEW '09, pages 19–27.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL '05*.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33:201–228.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of ACL '02*.

Christopher Dyer, Smaranda Muresan, and Philip Resnik. 2008. Generalizing Word Lattice Translation. In *Proceedings of ACL-HLT'08*, pages 1012–1020, June.

Chris Dyer, Jonathan Weese, Hendra Setiawan, Adam Lopez, Ferhan Ture, Vladimir Eidelman, Juri Ganitkevitch, Phil Blunsom, and Philip Resnik. 2010. cdec: a decoder, alignment, and learning framework for finite-state and context-free translation models. In *ACLDemos '10*, pages 7–12.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language*, HLT '91, pages 233–237.

Jesús Giménez and Lluís Màrquez. 2007. Context-aware discriminative phrase selection for statistical machine translation. In *Proceedings of StatMT '07*, pages 159–166.

AS Hildebrand, M Eck, S Vogel, and Alex Waibel. 2005. Adaptation of the translation model for statistical machine translation based on information retrieval. In *Proceedings of The European Association for Machine Translation (EAMT '05)*.

Zhanyi Liu, Haifeng Wang, Hua Wu, and Sheng Li. 2010. Improving statistical machine translation with monolingual collocation. In *ACL '10*.

Yanjun Ma, Yifan He, Andy Way, and Josef van Genabith. 2011. Consistent translation using discriminative learning: a translation memory-inspired approach. In *Proceedings of ACL-HLT'11*, pages 1239–1248.

Mohamed Maamouri, Ann Bies, and Seth Kulick. 2008. Enhancing the Arabic Treebank: A Collaborative Effort toward New Annotation Guidelines. In *LREC '08*.

Christof Monz. 2011. Statistical Machine Translation with Local Language Models. In *EMNLP '11*.

Hwee Tou Ng, Bin Wang, and Yee Seng Chan. 2003. Exploiting parallel texts for word sense disambiguation: an empirical study. In *ACL '03*.

NIST. 2006. http://www.itl.nist.gov/iad/mig/tests/mt/2006/.

NIST. 2008. http://www.itl.nist.gov/iad/mig/tests/mt/2008/.

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

Joseph Olive, Caitlin Christianson, and John McCary. 2011. *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. Springer Publishing Company, Inc., 1st edition.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL '02*.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *TREC*.

Germán Sanchis-Trilles and Francisco Casacuberta. 2010. Bayesian adaptation for statistical machine translation. In *Proceedings of the workshop on Structural and Syntactic Pattern Recognition (SSPR '10)*, pages 620–629.

Jörg Tiedemann. 2010. Context adaptation in statistical machine translation using models with exponentially decaying cache. In *Proceedings of the workshop on Domain Adaptation for Natural Language Processing (DANLP '10)*, pages 8–15.

Huihsin Tseng, Pi-Chuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter. In *Fourth SIGHAN Workshop on Chinese Language Processing*.

Michael M. Wasser and Bonnie Dorr. 2008. Machine translation with cross-lingual information retrieval based document relevance scores. Unpublished.

Tong Xiao, Jingbo Zhu, Shujie Yao, and Hao Zhang. 2011. Document-level consistency verification in machine translation. In *Machine Translation Summit XIII (MTS'11)*.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *ACL '95*.

Bing Zhao, Matthias Eck, and Stephan Vogel. 2004. Language model adaptation for statistical machine translation with structured query models. In *COLING '04*.

Bing Zhao, Young-Suk Lee, Xiaoqiang Luo, and Liu Li. 2011a. Learning to transform and select elementary trees for improved syntax-based machine translations. In *ACL-HLT '11*, pages 846–855.

Yinggong Zhao, Yangsheng Ji, Ning Xi, Shujian Huang, and Jiajun Chen. 2011b. Language model weight adaptation based on cross-entropy for statistical machine translation. In *Pacific Asia Conference on Language, Information and Computation (PACLIC '11)*.