PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 108 JUNE 2017 85–96

Using Word Embeddings to Enforce Document-Level Lexical Consistency in Machine Translation

Eva Martínez Garcia,[a] Carles Creus,[b] Cristina España-Bonet,[c] Lluís Màrquez[d]

[a] TALP Research Center, Universitat Politècnica de Catalunya
[b] Universitat Politècnica de Catalunya
[c] University of Saarland; DFKI, German Research Center for Artificial Intelligence
[d] ALT Group, Qatar Computing Research Institute, HBKU, Qatar Foundation

Abstract

We integrate new mechanisms in a document-level machine translation decoder to improve the lexical consistency of document translations. First, we develop a document-level feature designed to score the lexical consistency of a translation. This feature, which applies to words that have been translated into different forms within the document, uses word embeddings to measure the adequacy of each word translation given its context. Second, we extend the decoder with a new stochastic mechanism that, at translation time, allows to introduce changes in the translation oriented to improve its lexical consistency. We evaluate our system on English–Spanish document translation, and we conduct automatic and manual assessments of its quality. The automatic evaluation metrics, applied mainly at sentence level, do not reflect significant variations. On the contrary, the manual evaluation shows that the system dealing with lexical consistency is preferred over both a standard sentence-level and a standard document-level phrase-based MT systems.

1. Introduction

Statistical Machine Translation (SMT) systems have been traditionally designed at sentence level, without paying special attention to document-level information. However, taking into account some linguistic phenomena that go beyond the sentence boundaries, such as coreference or discourse markers, can be useful to improve the quality of the translation. Lexical coherence and consistency are also expected in a

Corresponding author: emartinez@cs.upc.edu

Cite as: Eva Martínez Garcia, Carles Creus, Cristina España-Bonet, Lluís Màrquez. Using Word Embeddings to Enforce Document-Level Lexical Consistency in Machine Translation. The Prague Bulletin of Mathematical Linguistics No. 108, 2017, pp. 85–96. doi: 10.1515/pralin-2017-0011.

document, but they are difficult to attain if the document is translated in a sentence by sentence basis.

In this paper we focus on improving the quality of the translations by handling lexical selection consistency across sentences in the document. The hypothesis is that making translations more consistent will lead to more coherent documents, perceived as globally better translations by humans (Carpuat, 2009; Carpuat and Simard, 2012). We tackle this problem by integrating new mechanisms inside a document-level decoder based on Docent (Hardmeier et al., 2013), which evaluate lexical consistency at the document level, and which provide translation changes oriented to improve it.

First, we design and implement a new document-level feature. Our feature scores the document lexical consistency by measuring how suitable the translation of a word is according to its context and to the other possible translations for that word found within the document. The feature uses word embeddings to make these adequacy assessments.

Second, we design a new change operation affecting how the translation search space is explored by the document-level decoder. This operation guides the translation process to improve lexical consistency. In particular, our operation detects those words that present translation inconsistencies within a document and proposes alternative, consistent translations for them.

Finally, we evaluate our system on benchmark datasets for English to Spanish translation, comparing its results to a phrase-based MT system. Although the usual automatic MT evaluation metrics are mostly insensitive to the changes introduced by our document-based MT system, a manual evaluation conducted on the output shows that those changes are important and noticeable by humans when assessing the quality of the document translations.

2. Related Work

In recent years, several efforts have been devoted to deal with document-level translation. Usually, authors focus on a particular phenomenon, such as pronominal anaphora (Hardmeier and Federico, 2010; Nagard and Koehn, 2010), topic cohesion (Gong et al., 2011), or topic coherence (Xiong et al., 2015). Lexical consistency has also been addressed before. For instance, Xiao et al. (2011) and Martínez Garcia et al. (2014a) used a post-process to re-translate source words that have been translated in different ways in a document. This is similar to our work in the sense that they consider inconsistent terms to be those words translated in different ways throughout a document, but differs from ours in that we want to consider the consistency information at decoding time and not as a post-process. The way we measure the consistency also differs: we use (bilingual) distributed word representations for this purpose.

Distributed word representation or word embeddings (WE) models have been successfully applied to several different NLP tasks. An efficient implementation of the Context Bag of Words (CBOW) and the Skipgram algorithms is presented in Mikolov et al. (2013a) and implemented in the word2vec toolkit. These models proved to be robust and powerful to predict semantic relations between words even across languages. However, they are unable to handle lexical ambiguity as they conflate word senses of polysemous words into one common representation. This limitation is discussed in Mikolov et al. (2013b) and Wolf et al. (2014), where bilingual extensions of the standard architecture are also proposed. Another bilingual approach is presented in Martínez Garcia et al. (2014b), where the resulting models are also evaluated in a cross-lingual lexical substitution task. Recently, WEs have been used in Pu et al. (2017) to improve the consistency of noun translations by means of a post-editing/re-ranking procedure with a phrase-based SMT system.

Closely related to our work, Hardmeier et al. (2012) used distributional vector models to implement semantic space language models (SSLM) within a document-oriented MT decoder. When working with SSLMs, the decoder uses the information of the word vector model to evaluate the adequacy of a word inside a translation by calculating the distance between the current word and its preceding context. In a similar way, Martínez Garcia et al. (2015) used, as SSLMs, bilingual and monolingual embedding models obtained with word2vec. Both studies used Docent (Hardmeier et al., 2013), a document-oriented SMT decoder that implements the algorithms in Hardmeier et al. (2012) and offers the possibility of using word embeddings as SSLMs. For our work, we use an in-house implementation of Hardmeier et al. (2012), named Lehrer as a homage to Docent.[1] These decoders work by performing hill climbing in a translation search space. This space can be seen as a graph where nodes are full-document translations and an edge connects two nodes when one translation can be transformed into the other. This transformation depends on the change operations provided by the decoders, which in general are simple operations such as changing the translation of a phrase, swapping phrase-pairs, or resegmenting the data. At each step of the search a full-document translation is available to the decoder. Thus, it is possible to develop features that capture properties of document-level phenomena. This makes these decoders flexible frameworks to develop and test different document-level strategies at translation time.

3. Lexical Consistency Feature

We strive to obtain translations where the same word appears translated into similar forms and with similar or related meanings throughout a document. In other terms, we want to avoid inconsistent translations for the same word. Thus, we are tackling a lexical-choice problem. Inspired by the SSLMs and with these aims, we develop a new lexical consistency feature that uses a Semantic Space to measure the Lexical Consistency of a document translation (SSLC).

[1] "*Lehrer*" means "teacher" in German. Source code at: `http://www.cs.upc.edu/~emartinez/lehrer.tgz`

Intuitively, SSLC scores each occurrence of an inconsistently translated source word with a value in $[-\infty, 0]$. For each such occurrence, this value is intended to measure how worse (in terms of adequacy) the current translation option is when compared to the other translation options seen in the document. More precisely, this value is computed as a subtraction between two numbers: the first one represents the adequacy of the current translation option, and the second one represents the best adequacy that could be obtained if another translation option (among the ones used somewhere in the document) had been used there instead. We consider a translation option to have better adequacy the more semantically similar it is to the context surrounding the occurrence being scored, and we compute it with WEs as the cosine similarity between the translation option and the context. Overall, note that SSLC does not try to enforce a strict lexically consistent translation, as long as lexical inconsistencies are semantically similar to their surrounding context.

To formalize SSLC we require some preliminary artillery. Let the source and target documents be the sequences of words $s_1, s_2, \ldots, s_N$ and $t_1, t_2, \ldots, t_M$, respectively, for some $N, M > 0$. Let $\tau : \{1, \ldots, N\} \to \{1, \ldots, M\}$ be a partial, injective mapping that associates to a source word index its corresponding target word index according to the current translation, if any.[2] In order to detect inconsistencies we need a way to identify whether two source or two target words must be considered to be the same word or not. To this end, we introduce the normalization functions $norm_{src}$ and $norm_{tgt}$ that take as input a source or target word, respectively, and return a normalized version of it. Then, two source or two target words are considered the same if they have the same normalized form through $norm_{src}$ or $norm_{tgt}$, respectively. In our experiments, $norm_{src}$ and $norm_{tgt}$ are implemented by, first, lower-casing the word and, second, by stemming it with the Snowball library.[3] Let $occ : \{1, \ldots, N\} \to 2^{\{1,\ldots,N\}}$ be the function that associates to each source word index i the set of indexes of the source words that have the same normalized form as s_i, i.e., $occ(i) = \{j \in \{1, \ldots, N\} \mid norm_{src}(s_j) = norm_{src}(s_i)\}$. Let $\tau occ(i)$ be a shorthand for $\tau(occ(i) \cap dom(\tau))$, where the intersection with $dom(\tau)$ is only necessary since τ is partial. We say that the ith source word is *inconsistent* in the current translation, denoted $incons(i)$, if the source words s_j that have the same normalized form as s_i have been translated into more than two distinct normalized targets. Formally:

$$incons(i) = \left(|\{norm_{tgt}(t_j) \mid j \in \tau occ(i)\}| > 2\right)$$

Let μ be the mapping defined by the word vector model in use by the decoder, i.e., a function that maps a word to a vector in a certain space $\mathbb{R}^n$ for some $n > 0$. Let $C > 0$ be the size of the context to either side of the target word, possibly crossing sentence

[2] Recall that phrase-based decoders perform translations by, in particular, using arbitrary alignments from source words to target words. For the SSLC feature we consider only the one-to-one word alignments.

[3] `http://snowballstem.org/`

boundaries. We tried several values for C and decided to fix $C = 15$ in the experiments as a good trade-off between performance and results. For each target word index $j \in \tau occ(i)$, where the source word index i satisfies that $incons(i)$ is true, we define its associated *score*, denoted $score(j)$, depending on the cosine similarity between the context and the current used translation option, and the cosine similarity between the context and the other translation options in the document. More precisely:

$$score(j) = sim(ctxt(j), \mu(t_j)) - \max_{k \in \tau occ(i)} sim(ctxt(j), \mu(t_k))$$

where $ctxt(j)$ is the sum of the vector representations of the words in the context of the jth target word, i.e., $ctxt(j) = \sum_{k \in \{max(1,j-C),\ldots,min(j+C,M)\} \setminus \{j\}} \mu(t_k)$, and sim of two vectors is the natural logarithm of their cosine similarity linearly scaled to the range $[0, 1]$, i.e., $sim(\vec{a}, \vec{b}) = \ln\left((\vec{a} \cdot \vec{b}/(\|\vec{a}\|\|\vec{b}\|) + 1)/2\right)$. Note that sim ranges in $[-\infty, 0]$, with $-\infty$ corresponding to the case where the vectors are diametrically opposed (semantically distant) and 0 to the case where they have the same orientation (semantically close). The final SSLC score for the whole document simply adds together the individual scores: $\sum_{i \in dom(\tau),\ incons(i)} score(\tau(i))$.

As a final remark, note that for ease of presentation we have assumed that the word vector model is monolingual. If it were bilingual, the expressions like $\mu(t_j)$ would be $\mu(t_j, s_{\tau^{-1}(j)})$ instead. Also, unknown words for the vector model, i.e., words w such that $\mu(w)$ is undefined, are ignored when computing the scores, and not taken into account when considering the C-sized context of the target word.

4. Lexical Consistency Change Operation

Recall that the decoding process of Lehrer performs a hill climbing in a translation search space. At each step, the decoder explores the neighbourhood of the current translation by randomly applying to it one of the available change operations. The default operations perform simple modifications such as changing the translation of a phrase, swapping phrase-pairs, or resegmenting the data. Unfortunately, these simple operations do not aid in our goal of reaching more lexically consistent translations. The reason for this fact is twofold. On the one hand, to increase the consistency it is in general necessary to perform multiple changes within the document and, since the default change operations only perform one change at a time, it would take several steps to fix one of the lexical choice inconsistencies. On the other hand, since hill climbing only performs a step when it strictly increases the score, each of the intermediate steps that try to fix an inconsistency would need to increase the score. To ameliorate this limitation on the hill climbing, we introduce the Lexical Consistency Change Operation (LCCO) that shortcuts the process by, at a single step, performing simultaneous changes that fix inconsistent translations of the same source word.

Intuitively, LCCO randomly selects an inconsistently translated source word, randomly chooses one of its translation options used in the document, and re-translates its occurrences throughout the document to match the chosen translation option. Both random decisions follow uniform distributions (the first one is uniform on all the distinct source words that appear inconsistently translated in the document, and the second one is uniform on all the distinct translation options seen in the document for the selected source word) in order to allow the hill climbing to fully explore the neighbourhood (given enough time) while minimizing the repetition of failed steps.

To formalize LCCO we need a more refined view of the source and target documents than in Section 3. Nevertheless, we reuse some of the previous definitions. Since the decoder works at phrase level, the documents are processed as sequences of phrases. Hence, we now consider that all the s_i and t_j are phrases instead of words. The definition of τ is still the same (although we can now guarantee that it is a total bijection since the decoder works with phrase-pairs) and $norm_{src}, norm_{tgt}$ are similar to before but have phrases as input and output instead of single words. The goal of LCCO is to change the translation of inconsistently translated words but, since the decoder works at phrase level, it can only change them safely when the inconsistent word appears alone in a phrase (otherwise LCCO would need to resegment the data too). For this reason, let us consider a more restricted definition of occ that only deals with indexes of source phrases having a single word. That is, for any $i \in \{1, \ldots, N\}$:

$$occ(i) = \{j \in \{1, \ldots, N\} \mid norm_{src}(s_j) = norm_{src}(s_i) \ \wedge \ |s_j| = 1\}$$

where $|s_j|$ is the number of words in the source phrase s_j. Using this redefined occ, we can keep the same definition for τocc and $incons$ as before.

LCCO works as follows. First, it selects a source phrase index $i \in \{1, \ldots, N\}$ such that $incons(i)$ is true. This is done by uniformly drawing that i from $\{\min(occ(k)) \mid k \in \{1, \ldots, N\} \ \wedge \ incons(k)\}$, where min is used to pick a representative from $occ(k)$. Second, it selects an occurrence $j \in occ(i)$ of that source phrase and considers $t_{\tau(j)}$ as the translation to use in the other occurrences. This is done by uniformly drawing that j from $\{k \in occ(i) \mid \nexists k' \in occ(i) : (k' < k \wedge norm_{tgt}(t_{\tau(k')}) = norm_{tgt}(t_{\tau(k)}))\}$. The new document translation $t'_1, t'_2, \ldots, t'_M$ is obtained by setting for each $k \in \{1, \ldots, M\}$:

$$t'_k := \begin{cases} t_k & \text{if } k \notin \tau occ(i) \\ t_k & \text{if } k \in \tau occ(i) \wedge norm_{tgt}(t_k) = norm_{tgt}(t_{\tau(j)}) \\ t_k & \text{if } k \in \tau occ(i) \wedge \nexists t \in \rho(s_{\tau^{-1}(k)}) : norm_{tgt}(t) = norm_{tgt}(t_{\tau(j)}) \\ t & \text{otherwise, with random } t \in \rho(s_{\tau^{-1}(k)}), norm_{tgt}(t) = norm_{tgt}(t_{\tau(j)}) \end{cases}$$

where ρ maps a source phrase to the set of target phrases that are its possible translations according to the phrase table in use by the decoder. Note that we do not change all the target phrases in $\tau occ(i)$, as in some of them we might already have a phrase with the same normal form as $t_{\tau(j)}$ (second case of the definition) and in some others the phrase table might not contain any entry with the same normal form as $t_{\tau(j)}$ (third case). The third case would never arise if $norm_{src}$ had been defined as the identity.

5. Experiments

We use as baselines a standard sentence-level SMT system based on Moses (Koehn et al., 2007) and our document-level Lehrer system implementing the algorithms in Hardmeier et al. (2012). We use the Europarl corpus (Koehn, 2005) for training an English to Spanish translation system, GIZA++ (Och and Ney, 2003) for word alignments, and the 5-gram language model described in Specia et al. (2013). We build monolingual and bilingual WEs as described in Martínez Garcia et al. (2014b, 2015) using the CBOW implementation in word2vec. We use newsCommentary2009 as development set and newsCommentary2010 as test set.

Weight optimization for the baseline Moses system is done with MERT (Och, 2003) against the BLEU metric (Papineni et al., 2002). The same weights are used for the baseline Lehrer system. Since automatic weight optimization for document-level features is not straightforward (Smith, 2015), we optimize the weights for the document-level features of non-baseline Lehrer system variants with manual grid searches.

We analyze the performance of 17 systems: the standard baseline Moses, 8 variants of Lehrer, and another 8 analogous variants of Lehrer+LCCO. More precisely, the first mentioned 8 system variants are: a baseline Lehrer system, three systems that implement the SSLMs within Lehrer using either the bilingual (+SSLMbi), the monolingual (+SSLMmo), or both (+SSLMbi&mo) embeddings, two systems implementing our SSLC feature within Lehrer using the bilingual embeddings (+SSLCbi) and its combination with the SSLM features (+SSLMbi&mo+SSLCbi), and finally, two more systems with the monolingual embeddings in SSLC (+SSLCmo) and its combination with the SSLMs (+SSLMbi&mo+SSLCmo). For Lehrer+LCCO, its 8 system variants are analogous and we denote them with equivalent names.

5.1. Automatic Evaluation

We use the Asiya toolkit (González et al., 2012) for automatic evaluation and include several lexical metrics (TER, BLEU, NIST, METEOR).

In Table 1 we show the performance of the systems. On the development set, results without LCCO show that bilingual information in SSLM appears to be more helpful than monolingual, but it also seems that both kinds of models can work together to improve the final system output. Looking at the scores of both SSLC systems, there are almost no noticeable improvements with respect to baseline Lehrer. The best results have been obtained combining all the information: bilingual and monolingual SSLMs with either of the SSLCs. When introducing LCCO, we observe roughly the same trends as before, except that combining SSLC and SSLM does not seem to provide the same benefit. On the test set we observe a similar behaviour, although differences among system scores are smaller. In this occasion both SSLC appear to improve the baseline Lehrer. Note that, in contrast with the trend observed on the development set, now both SSLC seem to work better alone than combined with SSLM.

	Development set				**Test set**			
System	**TER↓**	**BLEU↑**	**NIST↑**	**METEOR↑**	**TER↓**	**BLEU↑**	**NIST↑**	**METEOR↑**
Moses	58.28	24.27	6.826	46.84	53.70	27.52	7.323	50.02
Lehrer	58.34	24.28	6.820	46.92	53.78	27.58	7.313	50.08
+SSLMbi	58.08	24.35	6.845	46.93	**53.49**	27.60	**7.349**	**50.13**
+SSLMmo	58.28	24.27	6.827	46.89	53.70	27.57	7.319	50.07
+SSLMbi&mo	58.01	24.36	6.854	46.91	**53.49**	27.48	7.344	50.10
+SSLCbi	58.38	24.26	6.817	46.90	53.77	**27.61**	7.315	50.07
+SSLCmo	58.37	24.24	6.818	46.91	53.78	27.59	7.313	50.07
+SSLMbi&mo+SSLCbi	**57.99**	**24.39**	6.861	**46.95**	53.50	27.50	7.344	50.07
+SSLMbi&mo+SSLCmo	**57.99**	24.37	**6.863**	**46.95**	53.51	27.51	7.347	50.08
Lehrer+LCCO	58.36	24.27	6.819	46.92	53.77	27.57	7.308	50.07
+SSLMbi	58.04	**24.38**	6.849	**46.94**	53.45	**27.61**	7.352	50.14
+SSLMmo	58.29	24.27	6.825	46.91	53.71	27.58	7.320	50.09
+SSLMbi&mo	58.04	24.35	6.848	46.92	**53.43**	27.60	**7.355**	**50.15**
+SSLCbi	58.36	24.25	6.819	46.89	53.81	27.59	7.310	50.07
+SSLCmo	58.35	24.27	6.819	46.91	53.77	27.59	7.311	50.07
+SSLMbi&mo+SSLCbi	58.06	24.34	6.846	46.93	53.46	27.57	7.351	50.12
+SSLMbi&mo+SSLCmo	**58.03**	24.36	**6.851**	46.92	53.47	27.57	7.348	50.12

Table 1. Scores of the automatic evaluation of the systems.

As a general remark, the differences between most of the systems are not statistically significant.[4] Several causes contribute to this effect. On the one hand, a pairwise comparison of all the system outputs shows that the amount of different sentences is only between 8% and 42%. On the other hand, SSLC and LCCO deal with very sparse phenomena, and thus, they cannot have a huge impact on the automatic metrics. For instance, in average, LCCO is applied on 8% of the documents on the development and test sets, and in those cases it comprises between 4% and 9% of the total amount of change operation applications. Nevertheless, this does not necessarily hinder our goals, as consistent lexical selection improvements can also be introduced by the default change operations (although taking more search steps in decoding than LCCO, as the latter performs several modifications at once), which are boosted by SSLC.

These results make necessary a human evaluation of the translations, since we expect that the few changes induced by SSLC and LCCO will be appreciated by humans.

5.2. Human Evaluation

We carry out two distinct evaluation tasks. The first one tries to assess the quality of the different systems, working with and without LCCO. The second one is a small document-level evaluation task that compares the adequacy of the lexical choices between pairs of system variants that differ on whether they use LCCO or not.

For the first evaluation task, we select a common subset of sentences from the test set translated by the Moses system and by the 8 variants of the Lehrer system. More

[4] According to bootstrap resampling (Koehn, 2004) over BLEU and NIST metrics with a p-value of 0.05.

ID	**System**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	Moses	-	39 / 39	**44 / 43**	35 / 45	38 / 48	37 / 41	**43 / 39**	36 / 47	40 / 46
2	Lehrer	39 / 39	-	28 / 32	24 / 28	37 / 40	11 / 14	**14 / 11**	35 / 45	34 / 44
3	+SSLMbi	43 / 44	**32 / 28**	-	**36 / 33**	34 / 34	33 / 34	**37 / 29**	23 / 34	23 / 34
4	+SSLMmo	**45 / 35**	**28 / 24**	33 / 36	-	31 / 35	**31 / 30**	**32 / 26**	27 / 38	26 / 39
5	+SSLMbi&mo	**48 / 38**	**40 / 37**	34 / 34	**35 / 31**	-	**42 / 36**	**44 / 36**	18 / 27	20 / 25
6	+SSLCbi	**41 / 37**	**14 / 11**	**34 / 33**	30 / 31	36 / 42	-	**13 / 8**	34 / 43	36 / 45
7	+SSLCmo	39 / 43	11 / 14	29 / 37	26 / 32	36 / 44	8 / 13	-	31 / 47	33 / 47
8	+SSLMbi&mo+SSLCbi	**47 / 36**	**45 / 35**	**34 / 23**	**38 / 27**	**27 / 18**	**43 / 34**	**47 / 31**	-	**21 / 18**
9	+SSLMbi&mo+SSLCmo	**46 / 40**	**44 / 34**	**34 / 23**	**39 / 26**	**25 / 20**	**45 / 36**	**47 / 33**	18 / 21	-

ID	**System**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	Moses	-	**40 / 38**	44 / 45	39 / 43	41 / 49	36 / 40	39 / 40	40 / 46	**44 / 42**
2	Lehrer+LCCO	38 / 40	-	32 / 40	23 / 32	28 / 38	14 / 19	13 / 19	31 / 41	35 / 38
3	+SSLMbi	**45 / 44**	**40 / 32**	-	38 / 39	21 / 26	**40 / 36**	36 / 36	21 / 28	24 / 26
4	+SSLMmo	**43 / 39**	**32 / 23**	**39 / 38**	-	36 / 37	**31 / 27**	**32 / 26**	34 / 36	**37 / 36**
5	+SSLMbi&mo	**49 / 41**	**38 / 28**	**26 / 21**	**37 / 36**	-	**39 / 34**	**40 / 35**	18 / 24	22 / 23
6	+SSLCbi	**40 / 36**	**19 / 14**	36 / 40	27 / 31	34 / 39	-	**16 / 13**	35 / 40	**36 / 35**
7	+SSLCmo	**40 / 39**	**19 / 13**	36 / 36	26 / 32	35 / 40	13 / 16	-	37 / 44	37 / 37
8	+SSLMbi&mo+SSLCbi	**46 / 40**	**41 / 31**	**28 / 21**	**36 / 34**	**24 / 18**	**40 / 35**	**44 / 37**	-	**21 / 19**
9	+SSLMbi&mo+SSLCmo	42 / 44	**38 / 35**	**26 / 24**	36 / 37	**23 / 22**	35 / 36	37 / 37	19 / 21	-

Table 2. The two pairwise system comparisons done in the human evaluation. Each entry is the mean % of times a row system is evaluated better/worse than the column system.

precisely, we randomly choose 100 sentences with at least 5 and at most 30 words, and with at least 3 different translations among all the considered system outputs. We set up an evaluation environment where 3 native-Spanish annotators (including two of the authors) with a high English level have been asked to rank the output of all the systems for each of the 100 selected sentences, from best to worst general translation quality and with possible ties. System outputs were presented in random order to avoid system identification. The same evaluation procedure is also carried out with the 8 variants of Lehrer+LCCO. Table 2 shows the results obtained, where each entry of the table contains the mean number of times that the row system is better/worse than the column system according to the annotators, the remainder being ties. For the ranking with Lehrer variants, (pairs of) annotators agreed 70% of the time when ranking (pairs of) distinct outputs, and with Lehrer+LCCO variants, 72% of the time.

From the results in Table 2, we can say that Lehrer and Lehrer+LCCO are equivalent to Moses: they have a few ties, and either system is considered better than the other in roughly the same amount of cases. On the other hand, most non-baseline variants of Lehrer and Lehrer+LCCO seem to surpass Moses on wins. Translations from the systems including the combination of several features seem to be preferred in general; for instance, annotators prefer the combination SSLMbi&mo over SSLMbi or SSLMmo alone. Another interesting detail is that the SSLC systems seem analogous to the corresponding Lehrer and Lehrer+LCCO baselines, as they have many ties (although the SSLC systems have a slight advantage on wins). Also, SSLCbi and SSLCmo seem analogous, with SSLCbi having a slight win advantage over SSLCmo. This fact

source	[...] Due to the choice of the camera and the equipment, these **portraits** remember the classic photos. [...] The passion for the *portrait* led Bauer to repeat the idea [...]
reference	[...] Son **retratos** que, debido a la selección de la cámara y del material recuerdan la fotografía clásica. [...] La pasión por los *retratos* de Bauer le llevó a repetir la idea [...]
Moses	[...] Debido a la elección de la cámara y el equipo, estos **retratos** recordar el clásico fotos. [...] la pasión por el *cuadro* conducido Bauer a repetir la idea [...]
Lehrer+LCCO	[...] Debido a la elección de la cámara y el equipo, estos **retratos** recordar el clásico fotos. [...] la pasión por el *retrato* conducido Bauer a repetir la idea [...]

Table 3. Systems translation example with (in)consistent lexical choices.

shows that bilingual information has helped SSLC more than monolingual information. Both combinations of SSLMbi&mo with either of the SSLCs also seem analogous. As final remarks, the SSLMbi&mo+SSLCbi variants of Lehrer and Lehrer+LCCO systematically beat the other systems, and the non-baseline Lehrer and Lehrer+LCCO variants beat their respective baseline variant (except for Lehrer+SSLCmo).

The second, small evaluation task is a comparison between three system pairs with and without LCCO: the baseline, +SSLCbi, and +SSLMbi&mo+SSLCbi variants of Lehrer against the analogous variants of Lehrer+LCCO. We selected 10 documents with lexical changes introduced by LCCO, and asked an annotator to choose the translation with best lexical consistency and adequacy, given the source and two translated documents obtained by a system pair. The annotator preferred the translations of the variants with LCCO 60% of the time, and 20% of the time considered the translations of either system to have the same quality. So, systems with LCCO provided better translations according to the annotator regarding lexical consistency and adequacy.

To conclude, we provide in Table 3 a translation example from a news-piece about a photographer and his portraits work. Moses has not translated consistently an occurrence of the word *portrait* (the one in italics) which wrongly becomes *cuadro* (painting) instead of the correct choice *retrato*. Without LCCO, only the baseline, +SSLMbi, and both SSLC variants of Lehrer correctly produce *retrato* instead of *cuadro*. With LCCO, on the contrary, all the system variants are able to produce the consistent translation.

6. Conclusions

We have presented two new document-level strategies that aid MT systems in producing more coherent translations by improving the lexical consistency of the translations during the decoding process. In particular, we have developed a new document-level feature and change operation. The feature scores the lexical selection consistency of a translation document. To this end, it uses word embeddings to measure the adequacy of word translations given their context, computed on words that have been translated in several different forms within a document. The change operation helps the decoder explore the translation search space by performing simultaneous lexical changes in a translation step. Since it is able to modify several words at a time, even across sentences, it boosts the process of correcting the lexical inconsistencies. Both the feature and the change operation are implemented within our Lehrer decoder.

Results show that, although differences among systems are not statistically significant for the automatic evaluation metrics, they are noticeable for human evaluators that prefer the outputs from the enhanced systems.

As future work, we plan to study the impact of applying SSLC at lemma and seme levels, and conduct thorougher evaluations. Additionally, we are interested in tackling the same phenomena when using neural machine translation systems (Cho et al., 2014). These systems have recently achieved state-of-the-art results; however most are designed at sentence-level, and thus far, only a handful of works have studied the impact of using context information (Wang et al., 2017; Jean et al., 2017).

Bibliography

Carpuat, M. One Translation Per Discourse. In *Proc. of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, DEW '09, pages 19–27, 2009.

Carpuat, M. and M. Simard. The Trouble with SMT Consistency. In *Proc. of the 7th Workshop on Statistical Machine Translation, WMT@NAACL-HLT 2012*, pages 442–449, 2012.

Cho, K., B. van Merrienboer, D. Bahdanau, and Y. Bengio. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. In *Proc. of SSST-8*, pages 103–111, 2014.

Gong, Z., M. Zhang, and G. Zhou. Cache-based document-level statistical machine translation. In *Proc. of the 2011 Conference on Empirical Methods in NLP*, pages 909–919, 2011.

González, M., J. Giménez, and L. Màrquez. A Graphical Interface for MT Evaluation and Error Analysis. In *Proc. of the 50th ACL, System Demonstrations*, pages 139–144, 2012.

Hardmeier, C. and M. Federico. Modelling pronominal anaphora in statistical machine translation. In *Proc. of the 7th IWSLT*, pages 283–289, 2010.

Hardmeier, C., J. Nivre, and J. Tiedemann. Document-Wide Decoding for Phrase-Based Statistical Machine Translation. In *Proc. of EMNLP-CoNLL*, pages 1179–1190, 2012.

Hardmeier, C., S. Stymne, J. Tiedemann, and J. Nivre. Docent: A Document-Level Decoder for Phrase-Based Statistical Machine Translation. In *Proc. of the 51st ACL Conference*, pages 193–198, 2013.

Jean, S., S. Lauly, O. Firat, and K. Cho. Does Neural Machine Translation Benefit from Larger Context? *CoRR*, abs/1704.05135, 2017.

Koehn, P. Statistical Significance Tests for Machine Translation Evaluation. In *Proc. of EMNLP*, pages 388–395, 2004.

Koehn, P. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proc. of the MT Summit X*, pages 79–86, 2005.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: open source toolkit for statistical machine translation. In *Proc. of the 45th ACL*, pages 177–180, 2007.

Martínez Garcia, E., C. España-Bonet, and L. Màrquez. Document-level machine translation as a re-translation process. In *Procesamiento del Lenguaje Natural, Vol. 53*, pages 103–110, 2014a.

Martínez Garcia, E., C. España-Bonet, J. Tiedemann, and L. Màrquez. Word's Vector Representations meet Machine Translation. In *Proc. of SSST-8*, pages 132–134, 2014b.

Martínez Garcia, E., C. España-Bonet, and L. Màrquez. Document-Level Machine Translation with Word Vector Models. In *Proc. of the 18th EAMT*, pages 59–66, 2015.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. In *Proc. of Workshop at ICLR*, 2013a.

Mikolov, T., I. Sutskever, G. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Proc. of NIPS*, 2013b.

Nagard, R. Le and P. Koehn. Aiding pronouns translation with co-reference resolution. In *Proc. of Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261, 2010.

Och, F. Josef. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of the ACL 2003*, pages 160–167, 2003.

Och, F. Josef and H. Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, 2003.

Papineni, K., S. Roukos, T. Ward, and W.J. Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proc. of the 40th ACL*, pages 311–318, 2002.

Pu, X., L. Mascarell, and A. Popescu-Belis. Consistent Translation of Repeated Nouns using Syntactic and Semantic Cues. In *Proc. of the 15th EACL*, 2017.

Smith, A. BLEU Decoding and Feature Weight Tuning in Docent (Master Thesis). Uppsala Universitet, 2015.

Specia, L., K. Shah, J. G. C. de Souza, and T. Cohn. QuEst - A translation quality estimation framework. In *Proc. of ACL Demo Session*, 2013.

Wang, L., Z. Tu, A. Way, and Q. Liu. Exploiting Cross-Sentence Context for Neural Machine Translation. *CoRR*, abs/1704.04347, 2017.

Wolf, L., Y. Hanani, K. Bar, and N. Dershowitz. Joint word2vec Networks for Bilingual Semantic Representations. In *Poster sessions at CICLING*, 2014.

Xiao, T., J. Zhu, S. Yao, and H. Zhang. Document-level Consistency Verification in Machine Translation. In *Proc. of MT Summit XIII*, pages 131–138, 2011.

Xiong, D., M. Zhang, and X. Wang. Topic-Based Coherence Modeling for Statistical Machine Translation. In *IEEE/ACM Trans. on audio, speech & language processing*, pages 483–493, 2015.

Address for correspondence:
Eva Martínez Garcia
emartinez@cs.upc.edu
TALP Research Center – Universitat Politècnica de Catalunya
Jordi Girona, 1-3, 08034 Barcelona, Spain