Context-Aware Neural Machine Translation Decoding

Eva Martínez Garcia
TALP Research Center,
Universitat Politècnica de Catalunya
Barcelona
emartinez@cs.upc.edu

Carles Creus
Vicomtech
Donostia
ccreus@vicomtech.org

Cristina España-Bonet
Saarland University
DFKI GmbH
Saarbrücken
cristinae@dfki.de

Abstract

This work presents a decoding architecture that fuses the information from a neural translation model and the context semantics enclosed in a semantic space language model based on word embeddings. The method extends the beam search decoding process and therefore can be applied to any neural machine translation framework. With this, we sidestep two drawbacks of current document-level systems: (*i*) we do not modify the training process so there is no increment in training time, and (*ii*) we do not require document-level annotated data. We analyze the impact of the fusion system approach and its parameters on the final translation quality for English–Spanish. We obtain consistent and statistically significant improvements in terms of BLEU and METEOR and observe how the fused systems are able to handle synonyms to propose more adequate translations as well as help the system to disambiguate among several translation candidates for a word.

1 Introduction

Neural Machine Translation (NMT) systems represent the current state-of-the-art for machine translation technologies and even some evaluations claim that they have reached human performance (Hassan et al., 2018). These systems typically translate documents sentence by sentence, ignoring in the process inter-sentence context and document-level information, and this fact limits the maximum quality that they can achieve. Läubli et al. (2018) show how human translations are preferred over machine translations when they are evaluated at document level, even if the opposite happens at sentence level.

Although there exist several approaches that successfully enhance state-of-the-art neural machine translation systems to take into account document-level information, these systems usually propose modifications to the neural architecture (Wang et al., 2017a; Jean et al., 2017; Voita et al., 2018; Tu et al., 2018; Maruf and Haffari, 2018; Miculicich Werlen et al., 2018; Jean and Cho, 2019) making the training process slower, or require the training data to be annotated with document-level information, such as the document boundaries (Tiedemann and Scherrer, 2017; Junczys-Dowmunt, 2019; Talman et al., 2019; España-Bonet et al., 2019). The main benefit of these approaches is that the neural translation models they obtain are better tuned and able to handle document-level information. However, the training data with the document-level annotations that they require is still scarce, and also, their design increments the training times since the complexity of their neural architectures increase the model parameters to learn.

We propose an alternative to introducing inter-sentence information in an NMT system that follows the encoder–decoder architecture with attention of Bahdanau et al. (2015) without changing the neural translation model architecture. Furthermore, our approach does not need a costly training process with scarce document-level tagged data. Roughly, we modify the beam search algorithm to allow the introduction of a Semantic Space Language Model (SSLM) (Hardmeier et al., 2012) working in shallow fusion (Gülçehre et al., 2017) with a pre-trained NMT model. When evaluated on English–Spanish translations, we observe promising improvements in the automatic evaluation metrics used for the analysis.

The paper is organized as follows. Section 2 revisits the related work. We present the particularities of our approach in Section 3. We describe the experiments and results in Section 4, including an evaluation with oracles to assess the potential impact of our techniques, and conclude in Section 5.

Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019), pages 13–23
Hong Kong, China, November 3, 2019.

2 Related Work

The interest in making NMT systems able to include wider context information in the translation process has increased in recent years (Jean et al., 2017; Popescu-Belis, 2019), and even in some cases the *necessity* for exploring new approaches of document-level machine translation has been argued (Läubli et al., 2018).

On the one hand, several approaches tried to extend the context beyond the sentence information by modifying the system's input. Tiedemann and Scherrer (2017) concatenate the previous source sentence to the current one, whereas Bawden et al. (2018) also concatenate the previous predicted target sentence.

On the other hand, more sophisticated context-aware approaches propose to modify the NMT architecture. Jean et al. (2017) propose a variation of an attentional recurrent NMT system (Bahdanau et al., 2015) by including an additional encoder and attentional model to encode as context sentence the previous source sentence, showing how NMT systems can also benefit from larger contexts. Wang et al. (2017a) propose a cross-sentence context-aware approach that integrates the historical contextual information within the NMT system. However, these approaches only extend the source context but ignore the target side context. In contrast, Tu et al. (2018) take into account the target side context by using a lightweight cache-like memory network which stores bilingual hidden representations as translation history. More recent approaches implement system extensions that handle both source and target side contexts. Maruf and Haffari (2018) use memory networks to capture global source and target document context. Also, Maruf et al. (2019) present an approach to selectively focus on relevant sentences in the document context and not only consider a few previous sentences as context.

There are other approaches that study the effect of introducing context information within Transformer-based translation systems. Voita et al. (2018) present a variation of the Transformer (Vaswani et al., 2017) that extends the handled context by taking in the input both the current and previous sentences. Miculicich Werlen et al. (2018) extend it by integrating a hierarchical attention model to capture inter-sentence connections, Jean and Cho (2019) by including a context-aware regularization, and Zhang et al. (2018) propose to use a new context encoder to represent document-level context in combination with the original Transformer encoder-decoder architecture.

The importance of document-level translation is also seen in the recent WMT2019[1] news translation shared task, where for the first time a specific track for document-level MT was included. The systems presented at the shared task follow the previously explained strategies: introducing the inter-sentence context information into the NMT system by augmenting the training data including document-level information, i.e., including coreference information (España-Bonet et al., 2019), or just by increasing the training-sequence length in order to capture a larger data context (Popel et al., 2019; Talman et al., 2019; Junczys-Dowmunt, 2019), or introducing variations in the NMT architecture to take into account document-level information (Stahlberg et al., 2019; Talman et al., 2019).

Also related to our work, but far from machine translation, is the work by Wang and Cho (2016). They present an approach to include document-level context into language modeling by implementing fusion approaches that help the LSTM maintain separated the inter- and the intra-sentence context dependencies. Their conclusions show how using a wider context helps neural language models. We borrow the idea of (shallow) fusion and apply it to neural machine translation. In this line, Ji et al. (2015) presented new language models able to capture contextual information within and beyond the sentence level.

3 Context-Aware Decoding

Our document-level extension of the NMT decoding process benefits from the shallow fusion technique. In particular, it exploits the flexibility of being able to combine a general NMT model with a more domain specific Language Model (LM) to guide the NMT system towards a more adequate translation. In our approach, this other model is an SSLM used to introduce inter-sentence context information into the NMT decoding process. In the remaining of this section we briefly describe the SSLM (Section 3.1), the shallow fusion technique (Section 3.2), and finalize detailing our proposed combination (Section 3.3).

[1] http://www.statmt.org/wmt19/translation-task.html

3.1 Semantic Space Language Model (SSLM)

A semantic space language model is a probabilistic model able to predict the following word on a sequence taking into account the semantics, and so able to score the semantic relationship among a bunch of words in a sequence. In particular, we follow the SSLM definition presented by Hardmeier et al. (2012), who describe an SSLM based on a word dense vector model built with latent semantic analysis (Foltz et al., 1998; Bellegarda, 2000) and the cosine similarity, which is converted into a probability by a histogram lookup, as proposed by Bellegarda (2000). However, we substitute their LSA model for a word vector model built on the CBOW implementation of the WORD2VEC toolkit (Mikolov et al., 2013).

Intuitively, an SSLM mimics a traditional n-gram language model, but it is computed over semantic information and its expected effect is to promote translation choices that are semantically similar to the target context. To this end, for each candidate word w to append to the target translation, a score is computed based on the cosine similarity between the vector representation of w and the sum of the vector representations of the n target words that precede w in the document translation. In our system, the non-content words and the words unknown to the model are handled specially, both when computing their associated score and when considering them as part of the context of any later word. More precisely, given an already generated word sequence $y_{k-1} = w_1 w_2 \dots w_{k-1}$, the score associated by the SSLM to a translation candidate w_k proposed to extend the translation sequence, denoted $p_{SSLM}(w_k|y_{k-1})$, is:

$$\begin{cases} p_{uni}(w_k) & \text{if } w_k \text{ is a SW} \\ \alpha\, sim(\vec{c}_{y_{k-1}}, \mu(w_k)) & \text{if } w_k \in \mathrm{dom}(\mu) \text{ is not SW} \\ \varepsilon & \text{otherwise} \end{cases}$$

where p_{uni} maps each stop word (SW) to its relative frequency in the training corpus, α is the proportion of content words in the training corpus, $\vec{c}_{y_{k-1}}$ is the vector representing the preceding context of w_k (i.e., the sum of the vector representations of the last n non-stop known words of y_{k-1}), sim of two vectors is their cosine similarity linearly scaled to the range $[0, 1]$, i.e.,

$$sim(\vec{a}, \vec{b}) = \frac{1}{2} \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \|\vec{b}\|} + \frac{1}{2},$$

the word vector model is represented by μ and maps words to their associated vector representations, with $\mathrm{dom}(\mu)$ being its domain, and ε is a small fixed probability. Note that the lower bound 0 of sim corresponds to the case where the vectors are diametrically opposed (semantically distant) whereas the upper bound 1 to the case where they have the same orientation (semantically close). We use the value $n = 30$ chosen by Hardmeier et al. (2012) to make it possible that the context $\vec{c}_{y_{k-1}}$ used in the computations crosses sentence boundaries. Note that although our system does not need any document-level annotation, it will understand that any set of sentences in its input can be understood as a document. Thus, we need to translate document per document.

3.2 Shallow Fusion

Fusion techniques (Gülçehre et al., 2015, 2017) have shown to be successful in several natural language tasks to merge information from two different neural models. In our context, they are motivated by how Statistical Machine Translation (SMT) systems integrate the information from different feature functions that represent different probabilistic models. There are four main fusion techniques: deep, shallow, cold, and simple fusion. All of them extend the conditional probability learned by one model introducing the information from a second one, where the specific method that is used to combine both models is the main differentiator between the approaches.

Deep, cold, and simple fusion are techniques that need to train the resulting fused network. Deep fusion (Gülçehre et al., 2015, 2017; Stahlberg et al., 2018; Sriram et al., 2018) proposes a method to merge a translation model and a language model by introducing a gating mechanism that learns to balance the weight of the additional language model. Cold fusion (Sriram et al., 2018) goes a step beyond and proposes to implement a deep fusion where the NMT model is trained from scratch including the LM as a fixed part of the network. This allows the NMT to better model the conditioning on the source sequence while the target language modeling is covered by the LM. Simple fusion (Stahlberg et al., 2018) is the latest approach. It arises as an alternative simple method to use monolingual data for NMT training. Roughly, it integrates the shallow fusion technique in training time.

Shallow fusion is a simpler approach that fol-

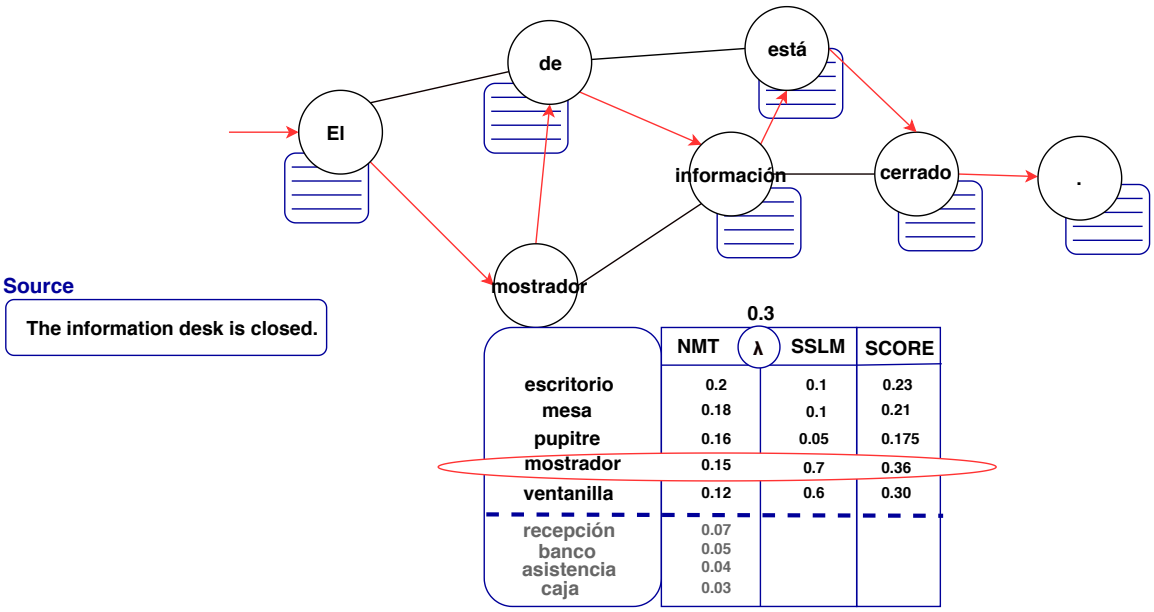

Figure 1: Sketch of the shallow fusion of an SSLM and an NMT inside the beam search algorithm. In this example, the process re-scores the $N = 5$ best candidates from the NMT model using the scores from the SSLM. Directed edges in the graph mark the path found by the beam search that maximizes the translation probability, whereas undirected edges mark possible steps considered by the beam search algorithm.

lows the same idea as deep fusion but, in contrast, proposes the combination of the probabilities from the two models at inference time. To this end, it changes the decoding objective function to integrate an LM prediction. The usual decoding objective function for an MT system with input x can be written as:

$$\hat{y} = \arg\max_{y} \log p(y|x)$$

whereas the shallow fusion variation introduces the LM in a manner inspired by the SMT log-linear model:

$$\hat{y} = \arg\max_{y} \big(\log p(y|x) + \lambda \log p_{LM}(y) \big) \quad (1)$$

where p_{LM} is a language model trained on monolingual target data and λ is its weight. The LM used by Gülçehre et al. (2017) is an LSTM-based RNN language model, but could be any model that generates as output a probability distribution on the discrete space of the target vocabulary shared with the translation model.

An advantage of shallow fusion over the other fusion techniques is that it only needs to adjust the weight λ for the language model by a grid-search on development data, avoiding a long training on large corpora. Furthermore, this technique can be easily applied to any NMT model, either RNN-based or purely attention-based neural models. In the same way as deep fusion, it uses independently pre-trained LM and NMT models. Although this can hinder the system performance, it can also be seen as an advantage due to the flexibility it confers.

3.3 Shallow Fusion between NMT and SSLM

In our model, we substitute the language model probability p_{LM} in the shallow fusion decoding function (Eq. 1) by the SSLM associated probability:

$$\hat{y} = \arg\max_{y} \big(\log p(y|x) + \lambda \log p_{SSLM}(y) \big)$$

Since the computation of p_{SSLM} for each generated word takes into account the preceding context of that word, it is necessary to modify the beam search of the NMT decoder. We implement a cache mechanism to keep track of the context information from the previously generated words, extending beyond sentence boundaries. The cache allows to add together the word embeddings from the previously generated words to obtain $\vec{c}_{y_{k-1}}$.

Additionally, the NMT model requires not only an estimate for a given target word, but a distribution probability over the entire target vocabulary space. Thus, $p_{SSLM}(w_k|y_{k-1})$ must be computed for each word w_k in the target vocabulary. Unfortunately, such an approach would

have a high computational cost. Following the ranking/filtering approaches of Jean et al. (2015) and Wang et al. (2017b), we speed up this computation by filtering the words to score by the SSLM. In particular, p_{SSLM} is only computed on the N target words with the highest probabilities from the NMT model, that is, only the N best candidates from the NMT model are considered by the SSLM. Figure 1 depicts how the filtering process works in combination with the shallow fusion of the NMT and the SSLM models during the beam search. Recall that although our system does not need any document-level annotation, it will understand any set of sentences in its input as a document, and thus we translate document per document.

4 Experiments

4.1 Settings

Our baseline NMT model follows the encoder–decoder architecture with attention of Bahdanau et al. (2015) and it is built using the OpenNMT-lua toolkit (Klein et al., 2017). We use a 4-layered bidirectional RNN encoder and a 4-layered RNN-based decoder with 800-dimensional hidden layers. Word embeddings are set to 500 dimensions for both source and target vocabularies. Stochastic gradient descent is used as optimizer algorithm for training, setting an initial learning rate of 1 and a learning decay of 0.7 after epoch 10 or if there is no loss improvement over the validation set. Training data is distributed on batches of 64 sentences and we use a 0.3 dropout probability between recurrent layers. Finally, a maximum sentence length of 50 tokens is used for both source and target sides and the vocabulary size is 50,000 for both target and source languages. The system is trained on the Europarl-v7 parallel corpus, using the newscommentary2009 corpus as validation set. The system at epoch 20 is to be shallow fused with the SSLM.

We implement the shallow fusion of the SSLM and an NMT as an extension of the attentional encoder–decoder NMT baseline. The Word Vector Models (WVM) used as SSLMs are built using word2vec with the CBOW algorithm (Mikolov et al., 2013), using a context window size of 5 and 600-dimensional vectors. The training data set for this model is the Spanish side of a set of parallel English–Spanish corpora available in Opus[2] (Tiedemann, 2012, 2009). We select the Europarl-v7, United Nations, Multilingual United Nations, and Subtitles-2012 corpora, which total 759 million words for Spanish. We use newscommentary2011 as test set. We take advantage of the document annotations from the newscommentary corpus to translate the test set document by document to avoid addition of random noise.

We evaluate the quality of the outputs with two automatic metrics, BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005).

4.2 Oracle Analysis

We implement three oracles to assess the potential impact of our techniques. The oracles behave as our fused approach, but leverage the reference translation to bias the decoding towards the word choices that are present in the reference. The goal of Oracle1 and Oracle2 is to assess the utility of the information enclosed in the WVM used by the SSLM, i.e., to check whether the semantic information of SSLM can help in producing better translations. Oracle3 mimics our fused decoding approach and its goal is to evaluate the potential gain of using an SSLM in combination with an NMT. In other words, with Oracle3 we check how much the SSLM can help the NMT disambiguate between its best translation candidates, thus obtaining an upper bound for the improvements that can be achieved by shallow fusing an SSLM and an NMT system.

Oracle1 proceeds offline as follows: once a sentence has been translated, for each target word t (*i*) it uses the attention information to map that t to its corresponding source word s and, in turn, maps that s to its corresponding target word r found in the reference, and (*ii*) it replaces the target word t by r whenever $t \neq r$ and r is among the M words that are closest to t (w.r.t. cosine similarity) according to our WVM. Note that the use of attention in step (*i*) to map between target and source words is not as straightforward as the alignment information in an SMT system. In particular, we consider that a target word t and a source word s are one-to-one mapped, denoted $t \overset{1}{\leftrightarrow} s$, when the following holds: the attention from t to s is maximal among the attentions from that t to any source word s' and also among the attentions from any

[2] http://opus.lingfil.uu.se/

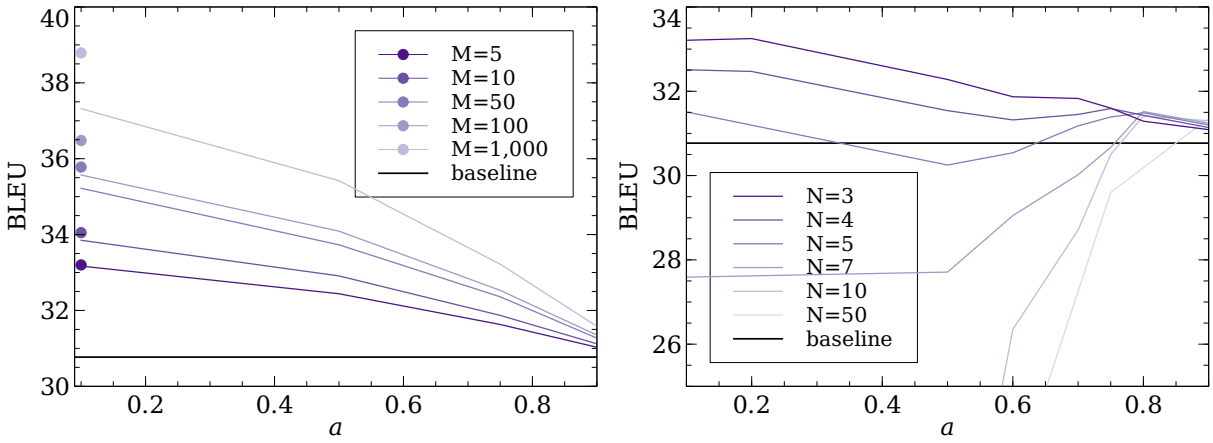

Figure 2: BLEU score of ORACLE1 (left, bullets), ORACLE2 (left, line plots), and ORACLE3 (right, line plots), as a function of the threshold a (ORACLE2 and ORACLE3) and for several values of the parameters M (ORACLE1 and ORACLE2) and N (ORACLE3). For ORACLE1 and ORACLE2, increasing the value of M beyond 1,000 does not affect the obtained scores noticeably.

target word t' to that s, i.e., $t \overset{1}{\leftrightarrow} s$ if and only if $att(t,s) = \max\{att(t',s') : t' = t \vee s' = s\}$, where $att(.,.)$ denotes the attention value between two words. We use an analogous definition for the one-to-one mapping $s \overset{1}{\leftrightarrow} r$ between the source and reference words.[3] Thus, for the target word t in consideration, step (*i*) tries to find the word r of the reference satisfying $t \overset{1}{\leftrightarrow} s \overset{1}{\leftrightarrow} r$, for some source word s. Table 1 and Figure 2 show the results for ORACLE1. We observe that the WVM encodes semantically-valid candidates close together, as there is a noticeable improvement in the BLEU score even when considering just the $M = 5$ closest candidates. Also, the accuracy of the oracle's translations increases with the number M of considered closest words. This is expected since augmenting the number M also increases the coverage of the target vocabulary. In the limit, when M allows to encompass the whole $50K$-word vocabulary, ORACLE1 simply rewrites the translation into the reference as far as the attention information allows, reaching an increase of $+8.02$ in BLEU score.

ORACLE2 works as ORACLE1 but proceeds online with the beam search. That is, when a hypothesis of the beam is to be extended with a new target word t, the oracle (*i*) analyzes the attention information to identify the actual word r used in the reference to translate the source word s that t corresponds to and (*ii*) replaces t with r under the same circumstances as before (i.e., when $t \neq r$ and r appears in the list of M words closest to t according to our WVM). In this occasion, however, the attention information needed in step (*i*) to deduce the one-to-one mappings between the target and source is not fully available, as the target sentence is still being generated. For this reason, we need to add a minimal threshold a for the attention and refine our criterion as $t \overset{1,a}{\leftrightarrow} s$ if and only if $t \overset{1}{\leftrightarrow} s \wedge att(t,s) \geq a$. Thus, for the target word t in consideration, step (*i*) tries to find the word r of the reference satisfying $t \overset{1,a}{\leftrightarrow} s \overset{1}{\leftrightarrow} r$, for some source word s. Table 1 and Figure 2 present also the results for ORACLE2. The results are analogous to those of ORACLE1, but with lower scores. This difference of score between both oracles is almost negligible for the smallest values of M and a, but the distance widens as either M or a increases. This shows that our definition of $\overset{1,a}{\leftrightarrow}$ is a proper approximation to obtain the mappings when not having the full attention information, as the permissive value $a = 0.1$ does not seem to be affected by noisy alignments for low values of M. This is because the oracle only replaces words by other

System	BLEU↑	MTR↑	N	M	a
baseline	30.77	49.86	-	-	-
ORACLE1	38.79	57.85	-	1,000	-
ORACLE2	37.32	54.35	-	1,000	0.1
ORACLE3	33.25	51.74	3	-	0.2

Table 1: BLEU and METEOR (MTR) scores obtained with the oracles defined in Section 4.2.

[3] The mapping from source to reference is done through attention by using the OpenNMT option of passing the target gold standard in the input.

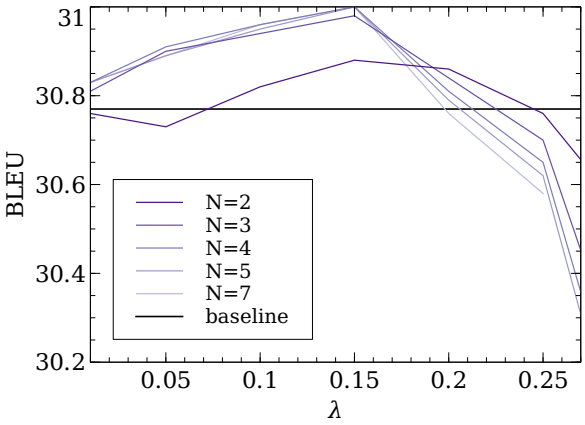

Figure 3: BLEU score of the fused system as a function of the weight λ, for several values of the parameter N.

N	**BLEU↑**	**METEOR↑**	**#unknown**
-	30.77	49.86	5901
2	30.88	50.17	4632
3	† 30.98	50.14	4501
4	† 31.00	50.15	4475
5	† 31.00	50.14	4459
7	† 31.00	50.14	4463
10	† 31.00	50.14	4463

Table 2: BLEU and METEOR scores obtained with the fused systems with $\lambda = 0.15$, together with the amount of unknown words in their output, where the first row corresponds to the baseline. † marks systems that are significantly different to the baseline with a p-value of 0.05, according to bootstrap resampling (Koehn, 2004).

semantically-close words (e.g., by synonyms), and thus, each of the substitutions preserves the meaning of the replaced word even if in some occasions the computed alignment is not adequate. Conversely, by increasing M the oracle handles lists of candidates that are more semantically distant, and thus, in combination with the uncertainty of the alignments, the system introduces more errors.

ORACLE3 proceeds online with the beam search like ORACLE2, just differing on the criterion used to replace the target word t by the corresponding reference word r: the replacement is done when $t \neq r$ and, moreover, r appears among the N best candidates proposed by the NMT model. Note that this oracle does not use in any way the WVM underlying the SSLM: it simply assumes that such model will properly promote the correct word (i.e., the reference word) whenever it is present among the N top candidates of the NMT. Table 1 and Figure 2 present also the results for ORACLE3, which show that there is some margin for improvement for the fused system with respect to the NMT working in isolation. In contrast with ORACLE2, ORACLE3 produces more errors the more candidates that it considers, i.e., the greater the value of N is. Also, considering alignments with lower probabilities only helps when the value of N is small. In particular, considering more candidates by increasing N needs a stronger (i.e., higher) attention threshold a in order to filter out noisy substitutions. Nevertheless, in that more restrictive configuration of a, the results for the various values of N tend to converge.

In summary, ORACLE1 shows that the WVM of the SSLM properly clusters semantically-valid candidates close together, ORACLE2 that incomplete attention information does not hinder the oracle's ability to approximate the alignments, and ORACLE3 that there is a wide enough margin for improvement when fusing the systems.

4.3 System Results and Analysis

Our system has two main hyperparameters: the number N of NMT translation options that are used in the fusion, and the weight of the semantic language model λ. Table 2 and Figure 3 show the results of the automatic evaluation of the different variations of the presented fused system. The figure shows how the maximum quality is achieved around $\lambda = 0.15$, independently of the number N of re-scored candidates. All of our systems are able to improve the baseline for every value of N that we explored, achieving a statistically significant improvement of $+0.23$ in BLEU score and $+0.31$ in METEOR. Nevertheless, there is still room for further gains since, as seen in Table 1, ORACLE3 is able to increase $+2.48$ BLEU and $+1.88$ METEOR points.

We observe in Table 2 that the scores improve as long as we increase the value of N until it seems to stabilize for $N \geq 4$. Furthermore, comparing the outputs for $\lambda = 0.15$, the translations that the system produces with $N = 4$ only differ in 95 sentences with respect to those for $N = 5$ and in 107 for $N = 7$, while having 1,407 sentences out of 3,003 that differ with respect to the baseline. Also, the translations for $N = 5$ are almost exactly the same as with $N = 7$, differing only in 30 sentences, whereas the translations for $N = 7$ and $N = 10$ coincide. These facts support that the systems with $N \geq 4$ are converging towards

an equivalent output. Looking into these differences, we realize that they manage different synonyms that may or not be in the reference. Like translating "*I have to*" as "*Tengo*" or "*Voy a tener*" which can be equivalent depending on the context.

We also observe that with larger values of N, the translations tend to be noisier or less adequate with respect to the source. For instance, "*Offices need a kindergarten nearby, architects have understood.*" is translated as:

"*las Oficinas necesitan una guardería cercana, los arquitectos han comprendido*" (N=4)

"*las oficinas de las oficinas de asistencia necesitan una guardería cercana.*" (N=7)

Notice in the second one the useless repetition of the translation for "*Offices*" and the appearance of the extra concept of assistance ("*asistencia*") that does not appear in the source sentence. Also, the information regarding the architects is missing in the second translation. Two important error types in NMT systems, word omission and new word creation, are exacerbated with large values for N.

Another example of more accurate translation occurs when translating "*According to Meteo France*". The best system using $N \geq 5$ translates this as "*Según Francia*" losing the reference to the meteorological company. In contrast, using $N = 4$, the system is able to generate a more accurate translation "*Según Meteo Francia*". This analysis reflects the noise introduced by increasing the number of re-scored translation candidates by the system. In other words, it is important to have enough candidates to see more adequate translations, but there is a trade-off that the system needs to maintain between the number of new options and the noise introduced by these re-scored options.

Finally, we observe that the increase in the translation quality is also related to the decrease in the number of unknown words generated by the system. Since we use complete tokens without BPE (Sennrich et al., 2016) or SENTENCEPIECE (Kudo and Richardson, 2018) as translation units, several tokens are unknowns to the system. In general, the number of generated unknown words with the shallow fusion approach drops almost a 25% with respect to the unknown words generated by the baseline. For instance, the worst case-scenario sentence "*I'm rather a novice in Prague politics responded Lukas Kaucky.*" is translated by the baseline as:

"*Más bien soy un ⟨unk⟩ en la política de Praga, ⟨unk⟩ a Lucas ⟨unk⟩.*"

whereas our fused system is able to produce:

"*Más bien soy un **novato** en la política de Praga, **respondió** a Lucas ⟨unk⟩.*"

generating good translations for "*novice*" and "*responded*". These examples illustrate how fusing the SSLM with the NMT model helps the latter to disambiguate between the considered translation candidates for a word.

Finally, we pursue a little manual evaluation with 3 native-Spanish speakers with fluent English. We select a common subset of sentences from the test set translated by the baseline NMT and by the fused system with $N = 4$ and $\lambda = 0.15$. We randomly choose 100 sentences with at least 5 and at most 30 words with different translations. The annotators were asked for each of the 100 selected sentences to rank the output of both systems according to their general translation quality, allowing to rank them as tying. System outputs were presented in random order to avoid system identification. The annotators find 49% of the time that the translation from the fused system is better than the baseline, and they consider the quality of both translations to tie 19% of the time. They agreed 67.33% of the time, reaching a $\kappa = 0.4733$ (Fleiss, 1971) showing a "moderate" inter-annotator agreement (Landis and Koch, 1977). These results support that fused systems are able to improve the translations' quality.

5 Summary and Conclusions

We presented a new approach that extends NMT decoding by introducing information from the preceding context on the target side. It fuses an attentional RNN with an SSLM by modifying the computation of the final score for an element of the target vocabulary inside the beam search algorithm. It is a flexible approach since it is compatible with any NMT architecture, and it allows to combine pre-trained models.

We reach improvements in the BLEU and METEOR scores of up to $+0.23$ and $+0.31$ respectively for English-to-Spanish translations. We analyze the impact of the different parameters of the system on these scores, observing that it is important to maintain a trade-off between the number of re-scored candidates, the SSLM weight, and the noise that will be introduced in the final translations. It is remarkable that our systems are able to

propose valid translations where the baseline fails to choose one, making the number of unknown words drop while the translation quality increases. Also, a small manual evaluation shows that humans tend to prefer fused system outputs.

As future work, we find interesting to pursue an in-depth manual evaluation to analyze how end users perceive the variations produced by our systems. The next step will be to test this implementation within Transformer-based NMT systems (Vaswani et al., 2017) to analyze how the inter-sentence information can affect the quality of attention-based translation systems and also to use BPEed input to compare the positive effect on unknown words that we observed. These two studies will improve the quality of the systems as a whole (both baseline and fused). In order to better capture the improvements reachable by our oracles, we want to analyse the validity of the cosine similarity as a measure and use other alternatives such as CSLS (cross-domain similarity local scaling) (Lample et al., 2018), or other margin-based scores instead (Artetxe and Schwenk, 2019).

Finally, we are interested in making a thorougher evaluation of the domain adaptation power of this technique by carrying out experiments designed to show how an embedding model trained on several specific domain data can guide a general-oriented NMT system towards more specific and adequate translations.

Acknowledgments

The project on which this paper is based was partially funded by the German Federal Ministry of Education and Research under the funding code 01IW17001 (Deeplee). Responsibility for the content of this publication is with the authors.

References

Mikel Artetxe and Holger Schwenk. 2019. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL) – Volume 1*, pages 3197–3203.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization (IEEvaluation@ACL)*, pages 65–72.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) – Volume 1*, pages 1304–1313.

Jerome R. Bellegarda. 2000. Exploiting latent semantic information in statistical language modeling. *Proceedings of the IEEE*, 88(8):1279–1296.

Cristina España-Bonet, Dana Ruiter, and Josef van Genabith. 2019. UdS-DFKI participation at WMT 2019: Low-resource (en–gu) and coreference-aware (en–de) systems. In *Proceedings of the Fourth Conference on Machine Translation (WMT) – Volume 2: Shared Task Papers*, pages 382–389.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Peter W. Foltz, Walter Kintsch, and Thomas K. Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25(2–3):285–307.

Çağlar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *CoRR*, abs/1503.03535.

Çağlar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, and Yoshua Bengio. 2017. On integrating a language model into neural machine translation. *Computer Speech & Language*, 45:137–148.

Christian Hardmeier, Joakim Nivre, and Jörg Tiedemann. 2012. Document-wide decoding for phrase-based statistical machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1179–1190.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic Chinese to English news translation. *CoRR*, abs/1803.05567.

Sébastien Jean and Kyunghyun Cho. 2019. Context-aware learning for neural machine translation. *CoRR*, abs/1903.04715.

Sébastien Jean, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. Montreal neural machine translation systems for WMT'15. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT)*, pages 134–140.

Sébastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context? *CoRR*, abs/1704.05135.

Yangfeng Ji, Trevor Cohn, Lingpeng Kong, Chris Dyer, and Jacob Eisenstein. 2015. Document context language models. *CoRR*, abs/1511.03962.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (WMT) – Volume 2: Shared Task Papers*, pages 424–432.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL)*, pages 67–72.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 388–395.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 66–71.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

John Richard Landis and Gary Grove Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? A case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4791–4796.

Sameen Maruf and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL) – Volume 1*, pages 1275–1284.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. *CoRR*, abs/1903.08788.

Lesly Miculicich Werlen, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2947–2954.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, pages 311–318.

Martin Popel, Dominik Macháček, Michal Auersperger, Ondřej Bojar, and Pavel Pecina. 2019. English-Czech systems in WMT19: Document-level transformer. In *Proceedings of the Fourth Conference on Machine Translation (WMT) – Volume 2: Shared Task Papers*, pages 541–547.

Andrei Popescu-Belis. 2019. Context in neural machine translation: A review of models and evaluations. *CoRR*, abs/1901.09115.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL) – Volume 1*, pages 1715–1725.

Anuroop Sriram, Heewoo Jun, Sanjeev Satheesh, and Adam Coates. 2018. Cold fusion: Training Seq2Seq models together with language models. In *Proceedings of the 19th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 387–391.

Felix Stahlberg, James Cross, and Veselin Stoyanov. 2018. Simple fusion: Return of the language model. In *Proceedings of the Third Conference on Machine Translation: Research Papers (WMT)*, pages 204–211.

Felix Stahlberg, Danielle Saunders, Adrià de Gispert, and Bill Byrne. 2019. CUED@WMT19:EWC&LMs. In *Proceedings of the Fourth Conference on Machine Translation (WMT) – Volume 2: Shared Task Papers*, pages 563–572.

Aarne Talman, Umut Sulubacak, Raúl Vázquez, Yves Scherrer, Sami Virpioja, Alessandro Raganato, Arvi Hurskainen, and Jörg Tiedemann. 2019. The University of Helsinki submissions to the WMT19 news translation task. In *Proceedings of the Fourth Conference on Machine Translation (WMT) – Volume 2: Shared Task Papers*, pages 611–622.

Jörg Tiedemann. 2009. News from OPUS – A collection of multilingual parallel corpora with tools and interfaces. *Recent Advances in Natural Language Processing (RANLP)*, 5:237–248.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pages 2214–2218.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation (DiscoMT@EMNLP)*, pages 82–92.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics (TACL)*, 6:407–420.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS)*, pages 6000–6010.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL) – Volume 1*, pages 1264–1274.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017a. Exploiting cross-sentence context for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2826–2831.

Tian Wang and Kyunghyun Cho. 2016. Larger-context language modelling with recurrent neural network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL) – Volume 1*, pages 1319–1329.

Yuguang Wang, Shanbo Cheng, Liyang Jiang, Jiajun Yang, Wei Chen, Muze Li, Lin Shi, Yanfeng Wang, and Hongtao Yang. 2017b. Sogou neural machine translation systems for WMT17. In *Proceedings of the Second Conference on Machine Translation (WMT)*, pages 410–415.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the Transformer translation model with document-level context. *CoRR*, abs/1810.03581.