Evaluating Discourse Phenomena in Neural Machine Translation

Rachel Bawden[1] **Rico Sennrich**[2,3] **Alexandra Birch**[2] **Barry Haddow**[2]

[1]LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, F-91405 Orsay, France
[2]School of Informatics, University of Edinburgh, Scotland
[3]Institute of Computational Linguistics, University of Zurich, Switzerland

`rachel.bawden@limsi.fr`
`{rico.sennrich, a.birch}@ed.ac.uk`
`bhaddow@inf.ed.ac.uk`

Abstract

For machine translation to tackle discourse phenomena, models must have access to extra-sentential linguistic context. There has been recent interest in modelling context in neural machine translation (NMT), but models have been principally evaluated with standard automatic metrics, poorly adapted to evaluating discourse phenomena. In this article, we present hand-crafted, discourse test sets, designed to test the models' ability to exploit previous source and target sentences. We investigate the performance of recently proposed multi-encoder NMT models trained on subtitles for English to French. We also explore a novel way of exploiting context from the previous sentence. Despite gains using BLEU, multi-encoder models give limited improvement in the handling of discourse phenomena: 50% accuracy on our coreference test set and 53.5% for coherence/cohesion (compared to a non-contextual baseline of 50%). A simple strategy of decoding the concatenation of the previous and current sentence leads to good performance, and our novel strategy of multi-encoding and decoding of two sentences leads to the best performance (72.5% for coreference and 57% for coherence/cohesion), highlighting the importance of target-side context.

1 Introduction

Machine translation (MT) systems typically translate sentences independently of each other. However, certain textual elements cannot be correctly translated without linguistic context, which may appear outside the current sentence. The most obvious examples of context-dependent phenomena problematic for MT are coreference (Guillou, 2016), lexical cohesion (Carpuat, 2009) and lexical disambiguation (Rios Gonzales et al., 2017), an example for each of which is given in (1-3). In each case, the English element in italic is ambiguous in terms of its French translation. The correct translation choice (in bold) is determined by linguistic context (underlined), which can be outside the current sentence. This disambiguating context can be source or target-side; the correct translation of anaphoric pronouns *it* and *they* depends on the gender of the translated antecedent (1). In lexical cohesion, a translation may depend on target factors, but may also be triggered by source effects and linguistic mechanisms such as repetition or alignment (2). In lexical disambiguation, source or target information may provide the appropriate context (3).

(1) The bee is busy. // *It* is making honey.
L'abeille$_{[f]}$ est occupée. // **Elle**$_{[f]}$/#il$_{[m]}$ fait du miel.

(2) Do you fancy some soup? // *Some soup?*
Tu veux de la soupe? // **De la soupe**/#du potage?

(3) And the code? // Still some *bugs*...
Et le code ? // Encore quelques **bugs**/#insectes...

Recent work on multi-encoder neural machine translation (NMT) appears promising for the integration of linguistic context (Zoph and Knight, 2016; Libovický and Helcl, 2017; Jean et al., 2017a; Wang et al., 2017). However models have almost only been evaluated using standard automatic metrics, which are poorly adapted to evaluating discourse phenomena. Targeted evaluation, in particular of coreference in MT, has proved to be time-consuming and laborious (Guillou, 2016).

In this article, we address the evaluation of discourse phenomena for MT and propose a novel contextual model. We present two hand-crafted, discourse test sets designed to test models' capacity to exploit linguistic context for coreference and coherence/cohesion for English to French translation. Using these sets, we review contextual NMT strategies trained on subtitles in a high-resource

Proceedings of NAACL-HLT 2018, pages 1304–1313
New Orleans, Louisiana, June 1 - 6, 2018. ©2018 Association for Computational Linguistics

setting. Our new combination of strategies outperforms previous methods according to our targeted evaluation and the standard metric BLEU.

2 Evaluating contextual phenomena

Traditional automatic metrics are notoriously problematic for the evaluation of discourse in MT (Hardmeier, 2014); discursive phenomena may have an impact on relatively few word forms with respect to their importance, meaning that improvements are overlooked, and a correct translation may depend on target-side coherence rather than similarity to a reference translation.

Coreference has been a major focus of discourse translation, spurred on by shared tasks on cross-lingual pronoun prediction (Guillou et al., 2016; Loáiciga et al., 2017). Participants were provided with lemmatised versions of reference translations,[1] in which pronoun forms were to be predicted. Evaluation in this setting (with the use of reference translations) was possible with traditional metrics, because the antecedents were fixed in advance. However there are at least two disadvantages to the approach: (i) models must be trained on lemmatised data and cannot be used in a real translation setting, and (ii) many of the pronouns did not need extra-sentential context; easier gains were seen for the pronouns with intra-sentential antecedents and therefore the leaderboard was dominated by sentence-level systems.

Guillou and Hardmeier's (2016) pronoun translation test suite succeeds in overcoming some of these problems by creating an automatic evaluation method, with a back-off manual evaluation. Manual evaluation has always been an essential part of evaluating MT quality, and targeted translation allows us to isolate a model's performance on specific linguistic phenomena; recent work using in-depth, qualitative manual evaluation (Isabelle et al., 2017; Scarton and Specia, 2015) is very informative. Isabelle et al. (2017) focus on specially constructed challenging examples in order to analyse differences between systems. They cover a wide range of linguistic phenomena, but since manual evaluation is costly and time-consuming, only a few examples per phenomenon are analysed, and it is difficult to obtain quick, quantitative feedback.

An alternative method, which overcomes the problem of costly, one-off analysis, is to evaluate models' capacity to correctly rank contrastive pairs of pre-existing translations, of which one is correct and the other incorrect. This method was used by Sennrich (2017) to assess the grammaticality of character-level NMT and again by Rios Gonzales et al. (2017) in a large-scale setting for lexical disambiguation for English-German. The method allows automatic quantitative evaluation of specific phenomena at large scale, at the cost of only testing for very specific translation errors. It is also the strategy that we will use here to evaluate translation of discourse phenomena.

2.1 Our contrastive discursive test sets

We created two contrastive test sets to help compare how well different contextual MT models handle (i) anaphoric pronoun translation and (ii) coherence and cohesion.[2] For each test set, models are assessed on their ability to rank the correct translation of an ambiguous sentence higher than the incorrect translation, using the disambiguating context provided in the previous source and/or target sentence.[3] All examples in the test sets are hand-crafted but inspired by real examples from OpenSubtitles2016 (Lison and Tiedemann, 2016) to ensure that they are credible and that vocabulary and syntactic structures are varied. The method can be used to evaluate any NMT model, by making it produce a score for a given source sentence and reference translation.

Our test sets differ from previous ones in that examples necessarily need the previous context (source and/or target-side) for the translations to be correctly ranked. Unlike the shared task test sets, the ambiguous pronouns' antecedents are guaranteed not to appear within the current sentence, meaning that, for MT systems to score highly, they must use discourse-level context. Compared to other test sets suites, ours differs in that evaluation is performed completely automatically and concentrates specifically on the model's ability to use context. Each of the test sets contains

[1] This was to avoid pronoun forms being trivial to predict from the morphological inflections of other forms within the sentence, an unrealistic setting for MT output.

[2] The test sets are freely available at https://diamt.limsi.fr/eval.html.

[3] We acknowledge that in reality, the disambiguating context is not guaranteed to be in the previous sentence (cf. Guillou (2016, p. 161), for the distribution of intra- and inter-sentential anaphoric pronouns). However it is important to first judge in a controlled way whether or not models are actually capable of using extra-sentential linguistic context at all, before investigating longer distance context.

200 contrastive pairs and is designed such that a non-contexual baseline system would achieve 50% accuracy.

Source:

context:	Oh, I hate **flies**. Look, there's another one!
current sent.:	Don’t worry, I'll kill **it** for you.

Target:

1	context:	Ô je déteste les **mouches**. Regarde, il y en a une autre !
	correct:	T'inquiète, je **la** tuerai pour toi.
	incorrect:	T'inquiète, je **le** tuerai pour toi.
2	context:	Ô je déteste les **moucherons**. Regarde, il y en a un autre !
	correct:	T'inquiète, je **le** tuerai pour toi.
	incorrect:	T'inquiète, je **la** tuerai pour toi.
3	context:	Ô je déteste les **araignées**. Regarde, il y en a une autre !
	semi-correct:	T'inquiète, je **la** tuerai pour toi.
	incorrect:	T'inquiète, je **le** tuerai pour toi.
4	context:	Ô je déteste les **papillons**. Regarde, il y en a un autre !
	semi-correct:	T'inquiète, je **le** tuerai pour toi.
	incorrect:	T'inquiète, je **la** tuerai pour toi.

Figure 1: Example block from the coreference set.

Coreference test set This set contains 50 example blocks, each containing four contrastive translation pairs (see the four examples in Fig. 1). The test set’s aim is to test the integration of target-side linguistic context. Each block is defined by a source sentence containing an occurrence of the anaphoric pronoun *it* or *they* and its preceding context, containing the pronoun’s nominal antecedent.[4] Four contrastive translation pairs of the previous and current source sentence are given, each with a different translation of the nominal antecedent, of which two are feminine and two are masculine per block. Each pair contains a correct translation of the current sentence, in which the pronoun’s gender is coherent with the antecedent’s translation, and a contrastive (incorrect) translation, in which the pronoun’s gender is inversed (along with agreement linked to the pronoun choice). Two of the pairs contain what we refer to as a “semi-correct” translation of the current sentence instead of a “correct” one, for which the antecedent in the previous sentence is strangely or wrongly translated (e.g. *flies* translated as *araignées* “spiders” and *papillons* “butterflies” in Fig. 1). In the “semi-correct” translation, the pronoun, whose translation is wholly dependent on the translated antecedent, is coherent with this translation choice. These semi-correct examples assess the use of target-side context, taking into account previous translation choices.

Target pronouns are evenly distributed according to number and gender with 50 examples (25 correct and 25 semi-correct) for each of the pronoun types (m.sg, f.sg, m.pl and f.pl). Since there are only two possible translations of the current sentence per example block, an MT system can only score all examples within a block correctly if it correctly disambiguates, and a non-contextual baseline system is guaranteed to score 50%.

Source:

context:	What's **crazy** about me?
current sent.:	Is this **crazy**?

Target:

context:	Qu'est-ce qu'il y a de **dingue** chez moi ?
correct:	Est-ce que ça c'est **dingue** ?
incorrect:	Est-ce que ça c'est fou ?

Source:

context:	What's **crazy** about me?
current sent.:	Is this **crazy**?

Target:

context:	Qu'est-ce qu'il y a de **fou** chez moi ?
correct:	Est-ce que ça c'est **fou** ?
incorrect:	Est-ce que ça c'est dingue ?

Figure 2: Example block from the coherence/cohesion test: alignment.

Source:

context:	So what do you say to £50?
current sent.:	It's a little **steeper** than I was expecting.

Target:

context:	Qu'est-ce que vous en pensez de 50£ ?
correct:	C'est un peu plus **cher** que ce que je pensais.
incorrect:	C'est un peu plus **raide** que ce que je pensais.

Source:

context:	How are your feet holding up?
current sent.:	It's a little **steeper** than I was expecting.

Target:

context:	Comment vont tes pieds ?
correct:	C'est un peu plus **raide** que ce que je pensais.
incorrect:	C'est un peu plus **cher** que ce que je pensais.

Figure 3: Example block from the coherence/cohesion test: lexical disambiguation.

[4] The choice to use only nominal antecedents and only two anaphoric pronouns *it* and *they* is intentional in order to provide a controlled environment in which there are two contrasting alternatives for each example. This ensures that a non-contextual baseline necessarily gives a score of 50%, and also enables us to explore this simpler case before expanding the study to explore more difficult anaphoric phenomena.

Coherence and cohesion test set Coherence and cohesion concern the interpretation of a text in the context of discourse (i.e. beyond sentence level). De Beaugrande and Dressler (1981) define the dichotomous pair as representing two separate aspects: coherence relating to the consistency of the text to concepts and world knowledge, and cohesion relating to the surface formulation of the text, as expressed through linguistic mechanisms.

This set contains 100 example blocks, each containing two contrastive pairs (see Figs. 2 and 3). Each of the blocks is constructed such that there is a single ambiguous source sentence, with two possible translations provided. The use of one translation over the other is determined by disambiguation context found in the previous sentence. The context may be found on the source side, the target side or both. In each contrastive pair, the incorrect translation of the current sentence corresponds to the correct translation of the other pair, such that the block can only be entirely correct if the disambiguating context is correctly used.

All test set examples have in common that the current English sentence is ambiguous and that its correct translation into French relies on context in the previous sentence. In some cases, the correct translation is determined more by cohesion, for example the necessity to respect alignment or repetition (Fig. 2). This means that despite two translations of an English source word being synonyms (e.g. *dingue* and *fou*, "crazy"), they are not interchangeable in a discourse context, given that the chosen formulation (alignment) requires repetition of the word of the previous sentence. In other cases, lexical choice is determined more by cohesion, for example by a general semantic context provided by the previous sentence, in a more classic disambiguation setting as in Fig. 3, where the English *steeper* is ambiguous between French *cher* "more expensive" and *raide* "sharply sloped". However, these types are not mutually exclusive and the distinction is not always so clear.

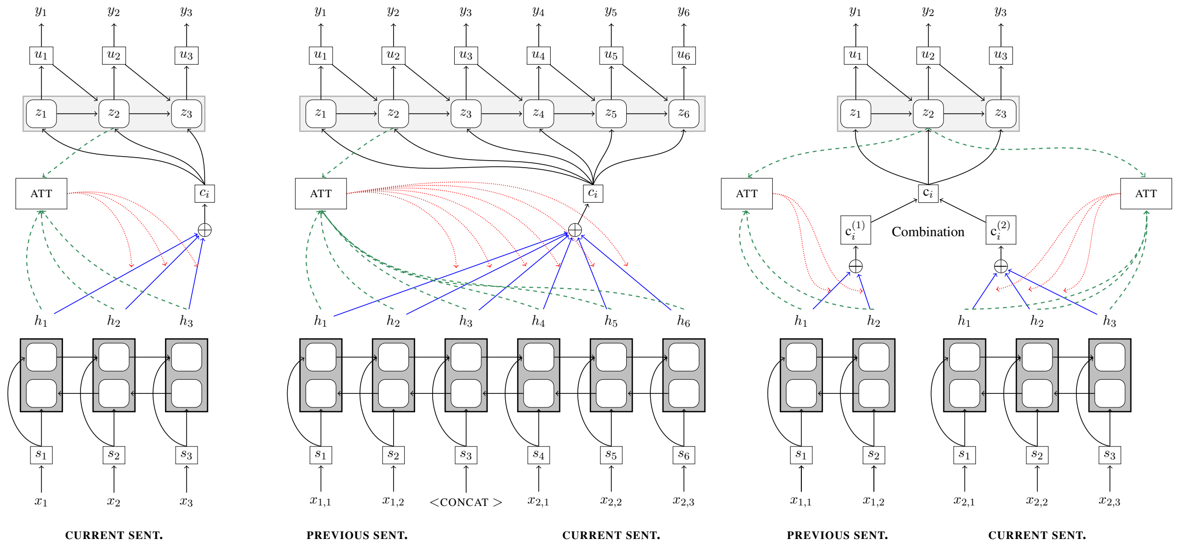
3 Contextual NMT Models

In order to correctly translate the type of phenomena mentioned in Sec. 1, translation models need to look beyond the sentence. Much of the previous work, mainly in statistical machine translation (SMT), focused on post-edition, particularly for anaphoric pronoun translation (Guillou et al., 2016; Loáiciga et al., 2017). However, coreference resolution is not yet sufficient for high quality post- or pre-edition (Bawden, 2016), and for other discourse phenomena such as lexical cohesion and lexical disambiguation, detecting the disambiguating context is far from trivial.

Recent work in NMT has explored multi-input models, which integrate the previous sentence as an auxiliary input. A simple strategy of concatenating the previous sentence to the current sentence and using a basic NMT architecture was explored by Tiedemann and Scherrer (2017), but with mixed results. A variety of multi-encoder strategies have also been tested, including using a representation of the previous sentence to initialise the main encoder and/or decoder (Wang et al., 2017) and using multiple attention mechanisms, with different strategies to combine the resulting context vectors, such as concatenation (Zoph and Knight, 2016), hierarchical attention (Libovický and Helcl, 2017) and gating (Jean et al., 2017a).

Although some of the models were evaluated in a contextual setting, for example on the cross-lingal pronoun prediction task at DiscoMT17 (Jean et al., 2017b), certain strategies only appear to give gains in a low-resource setting (Jean et al., 2017a), and, more importantly, there has yet to be an in-depth study into which strategies work best specifically for context-dependent discursive phenomena. Here we provide such a study, using the targeted test sets described in Sec. 2 to isolate and evaluate the different contextual models' capacity to exploit extra-sentential context. We test several contextual variants, using both a single encoder (Sec. 3.1) and multiple encoders (Sec. 3.2).

NMT notation All models presented are based on the widely used encoder-decoder NMT framework with attention (Bahdanau et al., 2015). At each decoder step i, the context (or summary) vector c_i of the input sequence is a weighted average of the recurrent encoder states at each input position depending on the attention weights. We refer to the recurrent state of the decoder as z_i. When multiple inputs are concerned, inputs are noted $x_j^{(k)}$, where k is the input number and j the input position. Likewise, when multiple encoders are used, $c_i^{(k)}$ refers to the k^{th} context vector where k is the encoder number. In the following section, all Ws, Us and bs are learned parameters.



(a) S2S with attention (BASELINE).

(b) Concatenate input (2-TO-2, 2-TO-1).

(c) Multi-source S2S with attention. The three combination methods tested are CONCAT, HIER and GATE.

Figure 4: The baseline model and the two contextual strategies tested (single and multi-encoder).

3.1 Single-encoder models

We train three single-source models: a baseline model and two contextual models. The baseline model translates sentences independently of each other (Fig. 4a). The two contextual models, described in (Tiedemann and Scherrer, 2017), are designed to incorporate the preceding sentence by prepending it to the current one, separated by a <CONCAT> token (Fig. 4b). The first method, which we refer to as 2-TO-2, is trained on concatenated source and target sentences, such that the previous and current sentence are translated together. The translation of the current sentence is obtained by extracting the tokens following the translated concatenation token and discarding preceding tokens.[5] The second method, 2-TO-1, follows the same principle, except that only source (and not target) sentences undergo concatenation; the model directly produces the translation of the current sentence. The comparison of these two methods allows us to assess the impact of the decoder in producing contextual translations.

3.2 Multi-encoder models

Inspired by work on multi-modal translation (Caglayan et al., 2016; Huang et al., 2016), multi-encoder translation models have recently been used to incorporate extra-sentential linguistic context in purely textual NMT (Zoph and Knight, 2016; Libovický and Helcl, 2017; Wang et al., 2017). Unlike multi-modal translation, which typically uses two complementary representations of the main input, for example a textual description and an image, linguistically contextual NMT has focused on exploiting the previous linguistic context as auxiliary input alongside the current sentence to be translated. Within this framework, we encode the previous sentence using a separate encoder (with separate parameters) to produce a context vector of the auxiliary input in a parallel fashion to the current source sentence. The two resulting context vectors $c_i^{(1)}$ and $c_i^{(2)}$ are then combined to form a single context vector c_i to be used for decoding (see Fig. 4c). We study three combination strategies here: concatenation, an attention gate and hierarchical attention. We also tested using the auxiliary context to initialise the decoder, similar to Wang et al. (2017), which was ineffective in our experiments and which we therefore do not report in this paper.

Attention concatenation The two context vectors $c_i^{(1)}$ and $c_i^{(2)}$ are concatenated and the resulting vector undergoes a linear transformation in order to return it to its original dimension to produce c_i (similar to work by Zoph and Knight (2016)).

$$c_i = W_c[c_i^{(1)}; c_i^{(2)}] + b_c \tag{1}$$

[5] Although the non-translation of the concatenation symbol is possible, in practice this was rare (<0.02%). If this occurs, the whole translation is kept.

Attention gate A gate r_i is learnt between the two vectors in order to give differing importance to the elements of each context vector, similar to the strategy of Wang et al. (2017).

$$r_i = \tanh\left(W_r c_i^{(1)} + W_s c_i^{(2)}\right) + b_r \tag{2}$$

$$c_i = r_i \odot \left(W_t c_i^{(1)}\right) + (1 - r_i) \odot \left(W_u c_i^{(2)}\right) \tag{3}$$

Hierarchical attention An additional (hierarchical) attention mechanism (Libovický and Helcl, 2017) is introduced to assign a weight to each encoder's context vector (designed for an arbitrary number of encoders).

$$e_i^{(k)} = v_b^\top \tanh\left(W_b z_{(i-1)} + U_b^{(k)} c_i^{(k)}\right) + b_e \tag{4}$$

$$\beta_i^{(k)} = \frac{\exp\left(e_i^{(k)}\right)}{\sum_{k'=1}^{K} \exp\left(e_i^{(k')}\right)} \tag{5}$$

$$c_i = \sum_{k=1}^{K} \beta_i^{(k)} U_c^{(k)} c_i^{(k)} \tag{6}$$

3.3 Novel strategy of hierarchical attention and context decoding

We also test a novel strategy of combining multiple encoders and decoding of both the previous and current sentence. We use separate, multiple encoders to encode the previous and current sentence and combine the context vectors using hierarchical attention. We train the model to produce the concatenation of the previous and current target sentences, of which the second part is kept, as in the contextual single encoder models.

4 Experiments

Each of the multi-encoder strategies is tested using the previous source and target sentences as an additional input (prefixed as S- and T- respectively) in order to test which is the most useful disambiguating context. Two additional models tested are triple-encoder models, which use both the previous source and target (prefixed as S-T-).

4.1 Data

Models are trained and tested on fan-produced parallel subtitles from OpenSubtitles2016[6] (Lison and Tiedemann, 2016). The data is first corrected using heuristics (e.g. minor corrections of OCR and encoding errors). It is then tokenised, further cleaned (keeping subtitles ≤80 tokens) and truecased using the Moses toolkit (Koehn et al., 2007) and finally split into subword units using BPE (Sennrich et al., 2016).[7] We run all experiments in a high-resource setting, with a training set of ≈29M parallel sentences, with vocabulary sizes of ≈55k for English and ≈60k for French.

4.2 Experimental setup

All models are sequence-to-sequence models with attention (Bahdanau et al., 2015), implemented in Nematus (Sennrich et al., 2017). Training is performed using the Adam optimiser with a learning rate of 0.0001 until convergence. We use embedding layers of dimension 512 and hidden layers of dimension 1024. For training, the maximum sentence length is 50.[8] We use batch sizes of 80, tied decoder embeddings and layer normalisation. The hyper-parameters are the same for all models and are the same as those used for the University of Edinburgh submissions to the news translation shared task at WMT16 and WMT17. Final models are ensembled using the last three checkpointed models.

Models that use the previous target sentence are trained using the previous reference translation. During translation, baseline translations are used. For the targeted evaluation, the problem does not apply since the translations that are being scored are given.

5 Results and Analysis

Overall translation quality is evaluated using the traditional automatic metric BLEU (Papineni et al., 2002) (Tab. 1) to ensure that the models do not degrade overall performance. We test the models' ability to handle discursive phenomena using the test sets described in Sec. 2 (Tab. 2). The models are described in the first half of Table 1: *#In* is the number of input sentences, the type of auxiliary input of which (previous source or target) is indicated by *Aux.*, *#Out* is the number of sentences translated, and *#Enc* is the number of encoders used to encode the input sentences. When there is a single encoder and more than one input, the input sentences are concatenated to form a single input to the encoder.

[6] http://www.opensubtitles.org

[7] 90,000 merge operations with a minimum theshold of 50.

[8] 76 when source sentences are concatenated to the previous sentence in order to keep the same percentage of training sentences as for other models.

	System Description				BLEU ↑			
	Aux.	#In	#Out	#Enc.	Comedy	Crime	Fantasy	Horror
Single-encoder, non-contexual model								
BASELINE	✗	1	1	1	19.52	22.07	26.30	33.05
Single-encoder with concatenated input								
2-TO-2	src	2	2	1	**20.09**	**22.93**	26.60	33.59
2-TO-1	src	2	1	1	19.51	21.81	26.78	**34.37**
Multi-encoder models (+previous target sentence)								
T-CONCAT	trg	2	1	2	18.33	20.90	24.36	32.90
T-HIER	trg	2	1	2	17.89	20.77	25.42	31.93
T-GATE	trg	2	1	2	18.25	20.76	25.55	32.64
Multi-encoder models (+previous source sentence)								
S-CONCAT	src	2	1	2	19.35	22.41	26.50	33.67
S-HIER	src	2	1	2	**20.22**	21.90	26.81	**34.04**
S-GATE	src	2	1	2	19.89	**22.80**	**26.87**	33.81
S-T-HIER	src, trg	3	1	3	19.53	22.53	**26.87**	33.24
Multi-encoder with concatenated output								
S-HIER-TO-2	src	2	2	2	**20.85**	**22.81**	**27.17**	**34.62**
S-T-HIER-TO-2	src, trg	3	2	3	18.80	21.18	**27.68**	33.33

Table 1: Results (de-tokenised, cased BLEU) of the ensembled models on four different test sets, each containing three films from each film genre. The best, second- and third-best results are highlighted by decreasingly dark shades of green.

5.1 Overall performance

Results using the automatic metric BLEU are given in Tab. 1. The models are tested on four different genres of film: comedy, crime, fantasy and horror.[9] Scores vary dramatically depending on the genre and the best model is not always the same for each of the genres.

Contrary to intuition, using the previous target sentence as an auxiliary input (prefix T-) degrades the overall performance considerably. Testing at decoding time with the reference translations did not significantly improve this result, suggesting that it is unlikely to be a case of overfitting during training. The highest performing model is our novel S-HIER-TO-2 model with more than +1 over the baseline BLEU on almost all test sets. There is no clear second best model, since performance depends strongly on the test set used.

5.2 Targeted evaluation

Tab. 2 shows the results on the discourse test sets.

Coreference The multi-encoder models do not perform well on the coreference test set; all multi-encoder models giving at best random accuracy, as with the baseline. This set is designed to test the model's capacity to exploit previous target context. It is therefore unsurprising that multi-encoder models using just the previous source sentence perform poorly. It is possible that certain pronouns could be correctly predicted from the source antecedents, if the antecedent only has one possible translation. However, this non-robust way of translating pronouns is not tested by the test set. More surprisingly, the multi-encoder models using the previous target sentence also perform poorly on the test set. An explanation could be that the target sentence is not being encoded sufficiently well in this framework, resulting in poor learning. This hypothesis is supported by the low overall translation performance shown in Tab. 1.

Two models perform well on the test set: 2-TO-2 and our S-HIER-TO-2. The high scores, particularly on the less common feminine pronouns, which can only be achieved through using contextual linguistic information, show that these models are capable of using previous linguistic context to disambiguate pronouns. The progressively high performance of these models can be seen in Fig. 5, which illustrates the training progress of these models. The S-T-HIER-TO-2 model (which uses the previous target sentence as a third auxiliary input) performs much worse than S-HIER-TO-2, showing that the addition of the previous target sentence is detrimental to performance. Whilst the

[9] Each of the test sets contains three films from that genre, with varying sizes and difficulty. The number of sentences in each test set is as follows: comedy: 4,490, crime: 4,227, fantasy: 2,790 and horror: 2,158.

	Coreference (%)							Coherence/cohesion (%)
	ALL	M.SG.	F.SG.	M.PL.	F.PL	CORR.	SEMI	ALL
BASELINE	50.0	80.0	20.0	80.0	20.0	53.0	47.0	50.0
2-TO-2	**63.5**	**92.0**	**50.0**	84.0	**28.0**	**68.0**	**59.0**	52.0
2-TO-1	52.0	72.0	28.0	84.0	**24.0**	54.0	50.0	**53.0**
T-CONCAT	49.0	88.0	8.0	**96.0**	4.0	50.0	48.0	51.5
T-HIER	47.0	78.0	10.0	**90.0**	10.0	47.0	47.0	50.5
T-GATE	47.0	80.0	6.0	82.0	20.0	45.0	49.0	49.0
S-CONCAT	50.0	68.0	32.0	88.0	12.0	**55.0**	45.0	**53.5**
S-HIER	50.0	64.0	**36.0**	80.0	20.0	**55.0**	45.0	**53.0**
S-GATE	50.0	68.0	32.0	84.0	16.0	**55.0**	45.0	51.5
S-T-HIER	49.5	**94.0**	4.0	88.0	12.0	53.0	46.0	**53.0**
S-HIER-TO-2	**72.5**	**100.0**	**40.0**	**90.0**	**36.0**	**77.0**	**68.0**	**57.0**
S-T-HIER-TO-2	**56.5**	84.0	**36.0**	86.0	20.0	**55.0**	**58.0**	51.5

Table 2: Results on the discourse test sets (% correct). Results on the coreference set are also given for each pronoun class. CORR. and SEMI correspond respectively to the "correct" and "semi-correct" examples. The best, second- and third-best results are highlighted by decreasingly dark shades of green.

results for the "correct" examples (CORR.) are almost always higher than the "semi-correct" examples (SEMI), for which the antecedent is strangely translated, the TO-2 models also give improved results on these examples, showing that the target context is necessarily being exploited during decoding.

These results show that the translation of the previous sentence is the most important factor in the efficient use of linguistic context. Combining the S-HIER model with decoding of the previous target sentence (S-HIER-TO-2) produces some of the best results across all pronoun types, and the 2-TO-2 model performs almost always second best.

Coherence and cohesion Much less variation in scores can be seen here, suggesting that these examples are more challenging and that there is room for improvement. Unlike the coreference examples, the multi-encoder strategies exploiting the previous source sentences perform better than the baseline (up to 53.5% for S-CONCAT). Yet again, using the previous target sentence achieves near random accuracy. 2-TO-2 and 2-TO-1 achieve similarly low scores (52% and 53%), suggesting that if concatenated input is used, decoding the previous sentence does not add more information.

However, combining multi-encoding with the decoding of the previous and the current sentences (S-HIER-TO-2) greatly improves the handling of the ambiguous translations, improving the accuracy to 57%. Extending this same model to also exploit the previous target sentence (S-T-HIER-TO-2) degrades this result, giving very similar scores to T-HIER and is therefore not illustrated in FIgure 5. This provides further support for the idea that the target sentence is not encoded efficiently as an auxiliary input and adds noise to the model, whereas exploiting the target context as a bias in the recurrent decoder is more effective.

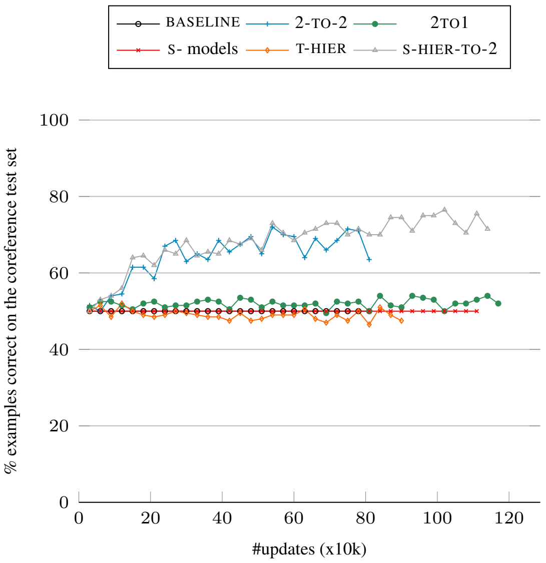


Figure 5: Progression of % correctly ranked examples (from the coreference test set) during training.

5.3 How much is the context being used?

Looking at the attention weights can sometimes offer insights into which input elements are being attended to at each step. For coreference resolution, we would expect the decoder to attend to the pronoun's antecedent. The effect is most expected when the previous target sentence is used, but it could also apply for the previous source sentence when the antecedent has only one possible translation. Unlike Tiedemann and Scherrer (2017), we do not observe increased attention between a translated pronoun and its source antecedent. Given the discourse test set results, which can only give high scores when target-side context is used, the contextual information of the type studied in this paper seems to be best exploited when channelled through the recurrent decoder node rather than when encoded through the input. This could explain why coreference is not easily seen via attention weights; the crucial information is encoded on the decoder-side rather than in the encoder.

6 Conclusion

We have presented an evaluation of discourse-level NMT models through the use of two discourse test sets targeted at coreference and lexical coherence/cohesion. We have shown that multi-encoder architectures alone have a limited capacity to exploit discourse-level context; poor results are found for coreference and more promising results for coherence/cohesion, although there is room for improvement. Our novel combination of contextual strategies greatly outperfoms existing models. This strategy uses the previous source sentence as an auxiliary input and decodes both the current and previous sentence. The observation that the decoding strategy is very effective for the handling of previous context suggests that techniques such as stream decoding, keeping a constant flow of contextual information in the recurrent node of the decoder, could be very promising for future research.

Acknowledgments

Rico Sennrich has received funding from the Swiss National Science Foundation (SNF) in the project CoNTra (grant number 105212_169888). This project has also received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 644333 (SUMMA) and 644402 (HimL).

References

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'15. ArXiv: 1409.0473.

Rachel Bawden. 2016. Cross-lingual Pronoun Prediction with Linguistically Informed Features. In *Proceedings of the 1st Conference on Machine Translation*. Berlin, Germany, WMT'16, pages 564–570.

Ozan Caglayan, Loïc Barrault, and Fethi Bougares. 2016. Multimodal Attention for Neural Machine Translation. In *arXiv:1609.03976*.

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

Robert De Beaugrande and Wolfgang Dressler. 1981. *Introduction to Text Linguistics*. Longman, London.

Liane Guillou. 2016. *Incorporating Pronoun Function into Statistical Machine Translation*. Ph.D. thesis, School of Informatics. University of Edinburgh.

Liane Guillou and Christian Hardmeier. 2016. PROTEST: A Test Suite for Evaluating Pronouns in Machine Translation. In *Proceedings of the 10th Language Resources and Evaluation Conference*. Portorož, Slovenia, LREC'16, pages 636–643.

Liane Guillou, Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, Mauro Cettolo, Bonnie Webber, and Andrei Popescu-Belis. 2016. Findings of the 2016 WMT Shared Task on Cross-lingual Pronoun Prediction. In *Proceedings of the 1st Conference on Machine Translation*. Berlin, Germany, WMT'16, pages 525–542.

Christian Hardmeier. 2014. *Discourse in Statistical Machine Translation*. Ph.D. thesis, Uppsala University, Department of Linguistics and Philology, Uppsala, Sweden.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based Multimodal Neural Machine Translation. In *Proceedings of the 1st Conference on Machine Translation*. Berlin, Germany, volume 2: of *WMT'16*, pages 639–645.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A Challenge Set Approach to Evaluating Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark, EMNLP'17, pages 2476–2486.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017a. Does Neural Machine Translation Benefit from Larger Context? In *arXiv:1704.05135*. ArXiv: 1704.05135.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017b. Neural Machine Translation for Cross-Lingual Pronoun Prediction. In *Proceedings of the 3rd Workshop on Discourse in Machine Translation*. Copenhagen, Denmark, DISCOMT'17, pages 54–57.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. Prague, Czech Republic, ACL'07, pages 177–180.

Jindřich Libovický and Jindřich Helcl. 2017. Attention Strategies for Multi-Source Sequence-to-Sequence Learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Vancouver, Canada, ACL'17, pages 196–202.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the 10th Language Resources and Evaluation Conference*. Portorož, Slovenia, LREC'16, pages 923–929.

Sharid Loáiciga, Sara Stymne, Preslav Nakov, Christian Hardmeier, Jörg Tiedemann, Mauro Cettolo, and Yannick Versley. 2017. Findings of the 2017 DiscoMT Shared Task on Cross-lingual Pronoun Prediction. In *Proceedings of the 3rd Workshop on Discourse in Machine Translation*. Copenhagen, Denmark, DISCOMT'17, pages 1–16.

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*. Philadelphia, USA, ACL'02, pages 311–318.

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. Improving Word Sense Disambiguation in Neural Machine Translation with Sense Embeddings. In *Proceedings of the 2nd Conference on Machine Translation*. Copenhagen, Denmark, WMT'17, pages 11–19.

Carolina Scarton and Lucia Specia. 2015. A Quantitative Analysis of Discourse Phenomena in Machine Translation. *Discours [online]* (16). `https://doi.org/10.4000/discours.9047`.

Rico Sennrich. 2017. How Grammatical is Character-level Neural Machine Translation? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia, Spain, EACL'17, pages 376–382.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nădejde. 2017. Nematus: a Toolkit for Neural Machine Translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia, Spain, EACL'17, pages 65–68.

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*. Berlin, Germany, ACL'16, pages 1715–1725.

Jörg Tiedemann and Yves Scherrer. 2017. Neural Machine Translation with Extended Context. In *Proceedings of the 3rd Workshop on Discourse in Machine Translation*. Copenhagen, Denmark, DISCOMT'17, pages 82–92.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting Cross-Sentence Context for Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Denmark, Copenhagen, EMNLP'17, pages 2816–2821.

Barret Zoph and Kevin Knight. 2016. Multi-source Neural Translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*. San Diego, California, USA, NAACL'16, pages 30–34.