

Context-Aware Neural Machine Translation Learns Anaphora Resolution

Elena Voita

Yandex, Russia

University of Amsterdam, Netherlands

lena-voita@yandex-team.ru

Pavel Serdyukov

Yandex, Russia

pavser@yandex-team.ru

Rico Sennrich

University of Edinburgh, Scotland

University of Zurich, Switzerland

rico.sennrich@ed.ac.uk

Ivan Titov

University of Edinburgh, Scotland

University of Amsterdam, Netherlands

ititov@inf.ed.ac.uk

Abstract

Standard machine translation systems process sentences in isolation and hence ignore extra-sentential information, even though extended context can both prevent mistakes in ambiguous cases and improve translation coherence. We introduce a context-aware neural machine translation model designed in such way that the flow of information from the extended context to the translation model can be controlled and analyzed. We experiment with an English-Russian subtitles dataset, and observe that much of what is captured by our model deals with improving pronoun translation. We measure correspondences between induced attention distributions and coreference relations and observe that the model implicitly captures anaphora. It is consistent with gains for sentences where pronouns need to be gendered in translation. Beside improvements in anaphoric cases, the model also improves in overall BLEU, both over its context-agnostic version (+0.7) and over simple concatenation of the context and source sentences (+0.6).

1 Introduction

It has long been argued that handling discourse phenomena is important in translation (Mitkov, 1999; Hardmeier, 2012). Using extended context, beyond the single source sentence, should in principle be beneficial in ambiguous cases and also ensure that generated translations are coherent. Nevertheless, machine translation systems typically ignore discourse phenomena and translate sentences in isolation.

Earlier research on this topic focused on handling specific phenomena, such as translating pronouns (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010; Hardmeier et al., 2015), discourse connectives (Meyer et al., 2012), verb tense (Gong et al., 2012), increasing lexical consistency (Carpuat, 2009; Tiedemann, 2010; Gong et al., 2011), or topic adaptation (Su et al., 2012; Hasler et al., 2014), with special-purpose features engineered to model these phenomena. However, with traditional statistical machine translation being largely supplanted with neural machine translation (NMT) models trained in an end-to-end fashion, an alternative is to directly provide additional context to an NMT system at training time and hope that it will succeed in inducing relevant predictive features (Jean et al., 2017; Wang et al., 2017; Tiedemann and Scherrer, 2017; Bawden et al., 2018).

While the latter approach, using context-aware NMT models, has demonstrated to yield performance improvements, it is still not clear what kinds of discourse phenomena are successfully handled by the NMT systems and, importantly, how they are modeled. Understanding this would inform development of future discourse-aware NMT models, as it will suggest what kind of inductive biases need to be encoded in the architecture or which linguistic features need to be exploited.

In our work we aim to enhance our understanding of the modelling of selected discourse phenomena in NMT. To this end, we construct a simple discourse-aware model, demonstrate that it achieves improvements over the discourse-agnostic baseline on an English-Russian subtitles dataset (Lison et al., 2018) and study which context information is being captured in the model. Specifically, we start with the Trans-

former (Vaswani et al., 2017), a state-of-the-art model for context-agnostic NMT, and modify it in such way that it can handle additional context. In our model, a source sentence and a context sentence are first encoded independently, and then a single attention layer, in a combination with a gating function, is used to produce a context-aware representation of the source sentence. The information from context can only flow through this attention layer. When compared to simply concatenating input sentences, as proposed by Tiedemann and Scherrer (2017), our architecture appears both more accurate (+0.6 BLEU) and also guarantees that the contextual information cannot bypass the attention layer and hence remain undetected in our analysis.

We analyze what types of contextual information are exploited by the translation model. While studying the attention weights, we observe that much of the information captured by the model has to do with pronoun translation. It is not entirely surprising, as we consider translation from a language without grammatical gender (English) to a language with grammatical gender (Russian). For Russian, translated pronouns need to agree in gender with their antecedents. Moreover, since in Russian verbs agree with subjects in gender and adjectives also agree in gender with pronouns in certain frequent constructions, mistakes in translating pronouns have a major effect on the words in the produced sentences. Consequently, the standard cross-entropy training objective sufficiently rewards the model for improving pronoun translation and extracting relevant information from the context.

We use automatic co-reference systems and human annotation to isolate anaphoric cases. We observe even more substantial improvements in performance on these subsets. By comparing attention distributions induced by our model against co-reference links, we conclude that the model implicitly captures coreference phenomena, even without having any kind of specialized features which could help it in this subtask. These observations also suggest potential directions for future work. For example, effective co-reference systems go beyond relying simply on embeddings of contexts. One option would be to integrate ‘global’ features summarizing properties of groups of mentions predicted as linked in a document (Wiseman et al., 2016), or to use latent relations to trace en-

ties across documents (Ji et al., 2017). Our key contributions can be summarized as follows:

- we introduce a context-aware neural model, which is effective and has a sufficiently simple and interpretable interface between the context and the rest of the translation model;
- we analyze the flow of information from the context and identify pronoun translation as the key phenomenon captured by the model;
- by comparing to automatically predicted or human-annotated coreference relations, we observe that the model implicitly captures anaphora.

2 Neural Machine Translation

Given a source sentence $\mathbf{x} = (x_1, x_2, \dots, x_S)$ and a target sentence $\mathbf{y} = (y_1, y_2, \dots, y_T)$, NMT models predict words in the target sentence, word by word.

Current NMT models mainly have an encoder-decoder structure. The encoder maps an input sequence of symbol representations \mathbf{x} to a sequence of distributed representations $\mathbf{z} = (z_1, z_2, \dots, z_S)$. Given \mathbf{z} , a neural decoder generates the corresponding target sequence of symbols \mathbf{y} one element at a time.

Attention-based NMT The encoder-decoder framework with attention has been proposed by Bahdanau et al. (2015) and has become the de-facto standard in NMT. The model consists of encoder and decoder recurrent networks and an attention mechanism. The attention mechanism selectively focuses on parts of the source sentence during translation, and the attention weights specify the proportions with which information from different positions is combined.

Transformer Vaswani et al. (2017) proposed an architecture that avoids recurrence completely. The Transformer follows an encoder-decoder architecture using stacked self-attention and fully connected layers for both the encoder and decoder. An important advantage of the Transformer is that it is more parallelizable and faster to train than recurrent encoder-decoder models.

From the source tokens, learned embeddings are generated and then modified using positional encodings. The encoded word embeddings are then used as input to the encoder which consists of N

layers each containing two sub-layers: (a) a multi-head attention mechanism, and (b) a feed-forward network.

The self-attention mechanism first computes attention weights: i.e., for each word, it computes a distribution over all words (including itself). This distribution is then used to compute a new representation of that word: this new representation is set to an expectation (under the attention distribution specific to the word) of word representations from the layer below. In multi-head attention, this process is repeated h times with different representations and the result is concatenated.

The second component of each layer of the Transformer network is a feed-forward network. The authors propose using a two-layered network with the ReLU activations.

Analogously, each layer of the decoder contains the two sub-layers mentioned above as well as an additional multi-head attention sub-layer that receives input from the corresponding encoding layer.

In the decoder, the attention is masked to prevent future positions from being attended to, or in other words, to prevent illegal leftward information flow. See Vaswani et al. (2017) for additional details.

The proposed architecture reportedly improves over the previous best results on the WMT 2014 English-to-German and English-to-French translation tasks, and we verified its strong performance on our data set in preliminary experiments. Thus, we consider it a strong state-of-the-art baseline for our experiments. Moreover, as the Transformer is attractive in practical NMT applications because of its parallelizability and training efficiency, integrating extra-sentential information in Transformer is important from the engineering perspective. As we will see in Section 4, previous techniques developed for recurrent encoder-decoders do not appear effective for the Transformer.

3 Context-aware model architecture

Our model is based on Transformer architecture (Vaswani et al., 2017). We leave Transformer’s decoder intact while incorporating context information on the encoder side (Figure 1).

Source encoder: The encoder is composed of a stack of N layers. The first $N - 1$ layers are identical and represent the original layers of Trans-

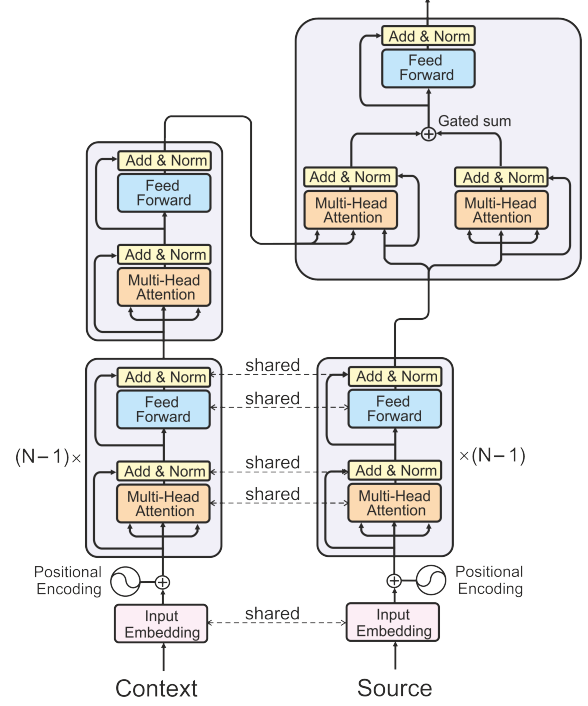


Figure 1: Encoder of the discourse-aware model

former’s encoder. The last layer incorporates contextual information as shown in Figure 1. In addition to multi-head self-attention it has a block which performs multi-head attention over the output of the context encoder stack. The outputs of the two attention mechanisms are combined via a gated sum. More precisely, let $c_i^{(s-attn)}$ be the output of the multi-head self-attention, $c_i^{(c-attn)}$ the output of the multi-head attention to context, c_i their gated sum, and σ the logistic sigmoid function, then

$$g_i = \sigma \left(W_g \left[c_i^{(s-attn)}, c_i^{(c-attn)} \right] + b_g \right) \quad (1)$$

$$c_i = g_i \odot c_i^{(s-attn)} + (1 - g_i) \odot c_i^{(c-attn)} \quad (2)$$

Context encoder: The context encoder is composed of a stack of N identical layers and replicates the original Transformer encoder. In contrast to related work (Jean et al., 2017; Wang et al., 2017), we found in preliminary experiments that using separate encoders does not yield an accurate model. Instead we share the parameters of the first $N - 1$ layers with the source encoder.

Since major proportion of the context encoder’s parameters are shared with the source encoder, we add a special token (let us denote it $\langle \text{bos} \rangle$) to the beginning of context sentences, but not source

sentences, to let the shared layers know whether it is encoding a source or a context sentence.

4 Experiments

4.1 Data and setting

We use the publicly available OpenSubtitles2018 corpus (Lison et al., 2018) for English and Russian.¹ As described in the appendix, we apply data cleaning and randomly choose 2 million training instances from the resulting data. For development and testing, we randomly select two subsets of 10000 instances from movies not encountered in training.² Sentences were encoded using byte-pair encoding (Sennrich et al., 2016), with source and target vocabularies of about 32000 tokens. We generally used the same parameters and optimizer as in the original Transformer (Vaswani et al., 2017). The hyperparameters, preprocessing and training details are provided in the supplementary material.

5 Results and analysis

We start by experiments motivating the setting and verifying that the improvements are indeed genuine, i.e. they come from inducing predictive features of the context. In subsequent section 5.2, we analyze the features induced by the context encoder and perform error analysis.

5.1 Overall performance

We use the traditional automatic metric BLEU on a general test set to get an estimate of the overall performance of the discourse-aware model, before turning to more targeted evaluation in the next section. We provide results in Table 1.³ The ‘baseline’ is the discourse-agnostic version of the Transformer. As another baseline we use the standard Transformer applied to the concatenation of the previous and source sentences, as proposed by Tiedemann and Scherrer (2017). Tiedemann and Scherrer (2017) only used a special symbol to mark where the context sentence ends and the source sentence begins. This technique performed badly with the non-recurrent Transformer architecture in preliminary experiments, resulting in

model	BLEU
baseline	29.46
concatenation (previous sentence)	29.53
context encoder (previous sentence)	30.14
context encoder (next sentence)	29.31
context encoder (random context)	29.69

Table 1: Automatic evaluation: BLEU. Significant differences at $p < 0.01$ are in bold.

a substantial degradation of performance (over 1 BLEU). Instead, we use a binary flag at every word position in our concatenation baseline telling the encoder whether the word belongs to the context sentence or to the source sentence.

We consider two versions of our discourse-aware model: one using the previous sentence as the context, another one relying on the next sentence. We hypothesize that both the previous and the next sentence provide a similar amount of additional clues about the topic of the text, whereas for discourse phenomena such as anaphora, discourse relations and elliptical structures, **the previous sentence is more important.**

First, we observe that our best model is the one using a context encoder for the previous sentence: it achieves 0.7 BLEU improvement over the discourse-agnostic model. We also notice that, unlike the previous sentence, the next sentence does not appear beneficial. This is a first indicator that discourse phenomena are the main reason for the observed improvement, rather than topic effects. Consequently, we focus solely on using the previous sentence in all subsequent experiments.

Second, we observe that the concatenation baseline appears less accurate than the introduced context-aware model. This result suggests that our model is not only more amendable to analysis but also potentially more effective than using concatenation.

In order to verify that our improvements are genuine, we also evaluate our model (trained with the previous sentence as context) on the same test set with shuffled context sentences. It can be seen that the performance drops significantly when a real context sentence is replaced with a random one. This confirms that the model does rely on context information to achieve the improvement in translation quality, and is not merely better regularized. However, the model is robust towards being shown a random context and obtains a performance similar to the context-agnostic baseline.

¹<http://opus.nlpl.eu/OpenSubtitles2018.php>

²The resulting data sets are freely available at http://data.statmt.org/acl18_contextnmt_data/

³We use bootstrap resampling (Riezler and Maxwell, 2005) for significance testing

5.2 Analysis

In this section we investigate what types of contextual information are exploited by the model. We study the **distribution of attention** to context and perform analysis on specific subsets of the test data. Specifically the research questions we seek to answer are as follows:

- For the translation of **which words does the model rely on** contextual history most?
- Are there any non-lexical patterns affecting attention to context, such as **sentence length and word position**?
- Can the context-aware NMT system **implicitly learn coreference phenomena** without any feature engineering?

Since all the attentions in our model are multi-head, by *attention weights* we refer to an average over heads of per-head attention weights.

First, we would like to identify a *useful* attention mass coming to context. We analyze the attention maps between source and context, and find that the model mostly attends to `<bos>` and `<eos>` context tokens, and much less often attends to words. Our hypothesis is that the model has found a way to take no information from context by looking at uninformative tokens, and it attends to words only when it wants to pass some contextual information to the source sentence encoder. Thus we define *useful* contextual attention mass as sum of attention weights to context words, excluding `<bos>` and `<eos>` tokens and punctuation.

5.2.1 Top words depending on context

We analyze the distribution of attention to context for individual source words to see for which words the model depends most on contextual history. We compute the overall average attention to context words for each source word in our test set. We do the same for source words at positions higher than first. We filter out words that occurred less than 10 times in a test set. The top 10 words with the highest average attention to context words are provided in Table 2.

An interesting finding is that contextual attention is high for the translation of “it”, “yours”, “ones”, “you” and “I”, which are indeed very ambiguous out-of-context when translating into Russian. For example, “it” will be translated as third person singular masculine, feminine or neuter, or third person plural depending on its antecedent.

word	attn	pos	word	attn	pos
it	0.376	5.5	it	0.342	6.8
yours	0.338	8.4	yours	0.341	8.3
yes	0.332	2.5	ones	0.318	7.5
i	0.328	3.3	'm	0.301	4.8
yeah	0.314	1.4	you	0.287	5.6
you	0.311	4.8	am	0.274	4.4
ones	0.309	8.3	i	0.262	5.2
'm	0.298	5.1	's	0.260	5.6
wait	0.281	3.8	one	0.259	6.5
well	0.273	2.1	won	0.258	4.6

Table 2: Top-10 words with the highest average attention to context words. *attn* gives an average attention to context words, *pos* gives an average position of the source word. Left part is for words on all positions, right — for words on positions higher than first.

“You” can be second person singular impolite or polite, or plural. Also, verbs must agree in gender and number with the translation of “you”.

It might be not obvious why “I” has high contextual attention, as it is not ambiguous itself. However, in past tense, verbs must agree with “I” in gender, so to translate past tense sentences properly, the source encoder must predict speaker gender, and the context may provide useful indicators.

Most surprising is the appearance of “yes”, “yeah”, and “well” in the list of context-dependent words, similar to the finding by [Tiedemann and Scherrer \(2017\)](#). We note that these words mostly appear in sentence-initial position, and in relatively short sentences. If only words after the first are considered, they disappear from the top-10 list. We hypothesize that the amount of attention to context not only depends on the words themselves, but also on factors such as sentence length and position, and we test this hypothesis in the next section.

5.2.2 Dependence on sentence length and position

We compute useful attention mass coming to context by averaging over source words. Figure 2 illustrates the dependence of this average attention mass on sentence length. We observe a disproportionally high attention on context for short sentences, and a positive correlation between the average contextual attention and context length.

It is also interesting to see the importance given to the context at different positions in the source

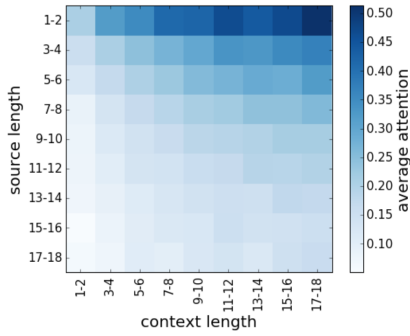


Figure 2: Average attention to context words vs. both source and context length

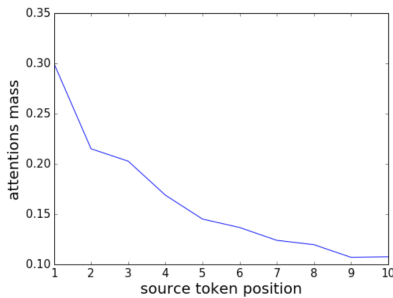


Figure 3: Average attention to context vs. source token position

sentence. We compute an average attention mass to context for a set of 1500 sentences of the same length. As can be seen in Figure 3, words at the beginning of a source sentence tend to attend to context more than words at the end of a sentence. This correlates with standard view that English sentences present hearer-old material before hearer-new.

There is a clear (negative) correlation between sentence length and the amount of attention placed on contextual history, and between token position and the amount of attention to context, which suggests that context is especially helpful at the beginning of a sentence, and for shorter sentences. However, Figure 4 shows that there is no straightforward dependence of BLEU improvement on source length. This means that while attention on context is disproportionally high for short sentences, context does not seem disproportionally more useful for these sentences.

5.3 Analysis of pronoun translation

The analysis of the attention model indicates that the model attends heavily to the contextual history for the translation of some **pronouns**. Here, we investigate whether this context-aware modelling results in empirical improvements in translation

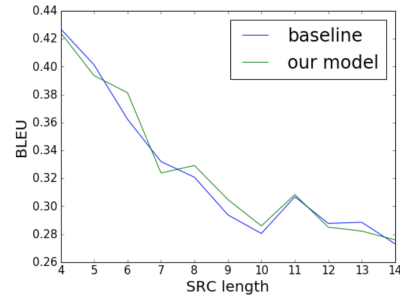


Figure 4: BLEU score vs. source sentence length

quality, and whether the model learns structures related to anaphora resolution.

5.3.1 Ambiguous pronouns and translation quality

Ambiguous pronouns are relatively sparse in a general-purpose test set, and previous work has designed targeted evaluation of pronoun translation (Hardmeier et al., 2015; Miculicich Werlen and Popescu-Belis, 2017; Bawden et al., 2018). However, we note that in Russian, grammatical gender is not only marked on pronouns, but also on adjectives and verbs. Rather than using a pronoun-specific evaluation, we present results with BLEU on test sets where we hypothesize context to be relevant, specifically sentences containing co-referential pronouns. We feed Stanford CoreNLP open-source coreference resolution system (Manning et al., 2014a) with pairs of sentences to find examples where there is a link between one of the pronouns under consideration and the context. We focus on anaphoric instances of “it” (this excludes, among others, pleonastic uses of “it”), and instances of the pronouns “I”, “you”, and “yours” that are coreferent with an expression in the previous sentence. All these pronouns express ambiguity in the translation into Russian, and the model has learned to attend to context for their translation (Table 2). To combat data sparsity, the test sets are extracted from large amounts of held-out data of OpenSubtitles2018. Table 3 shows BLEU scores for the resulting sub-sets.

First of all, we see that most of the antecedents in these test sets are also pronouns. Antecedent pronouns should not be particularly informative for translating the source pronoun. Nevertheless, even with such contexts, improvements are generally larger than on the overall test set.

When we focus on sentences where the antecedent for pronoun under consideration contains

pronoun	N	#pronominal antecedent	baseline	our model	difference
it	11128	6604	25.4	26.6	+1.2
you	6398	5795	29.7	30.8	+1.1
yours	2181	2092	24.1	25.2	+1.1
I	8205	7496	30.1	30.0	-0.1

Table 3: BLEU for test sets with coreference between pronoun and a word in context sentence. We show both N , the total number of instances in a particular test set, and number of instances with *pronominal antecedent*. Significant BLEU differences are in bold.

word	N	baseline	our model	diff.
it	4524	23.9	26.1	+2.2
you	693	29.9	31.7	+1.8
I	709	29.1	29.7	+0.6

Table 4: BLEU for test sets of pronouns having a nominal antecedent in context sentence. N : number of examples in the test set.

type	N	baseline	our model	diff.
masc.	2509	26.9	27.2	+0.3
fem.	2403	21.8	26.6	+4.8
neuter	862	22.1	24.0	+1.9
plural	1141	18.2	22.5	+4.3

Table 5: BLEU for test sets of pronoun “it” having a nominal antecedent in context sentence. N : number of examples in the test set.

a noun, we observe even larger improvements (Table 4). Improvement is smaller for “I”, but we note that verbs with first person singular subjects mark gender only in the past tense, which limits the impact of correctly predicting gender. In contrast, different types of “you” (polite/impolite, singular/plural) lead to different translations of the pronoun itself plus related verbs and adjectives, leading to a larger jump in performance. Examples of nouns co-referent with “I” and “you” include names, titles (“Mr.”, “Mrs.”, “officer”), terms denoting family relationships (“Mom”, “Dad”), and terms of endearment (“honey”, “sweetie”). Such nouns can serve to disambiguate number and gender of the speaker or addressee, and mark the level of familiarity between them.

The most interesting case is translation of “it”, as “it” can have many different translations into Russian, depending on the grammatical gender of the antecedent. In order to disentangle these cases, we train the Berkeley aligner on 10m sentences and use the trained model to divide the test set with “it” referring to a noun into test sets specific to each gender and number. Results are in Table 5.

pronoun	agreement (in %)			
	random	first	last	attention
it	69	66	72	69
you	76	85	71	80
I	74	81	73	78

Table 6: Agreement with CoreNLP for test sets of pronouns having a nominal antecedent in context sentence (%).

We see an improvement of 4-5 BLEU for sentences where “it” is translated into a feminine or plural pronoun by the reference. For cases where “it” is translated into a masculine pronoun, the improvement is smaller because the masculine gender is more frequent, and the context-agnostic baseline tends to translate the pronoun “it” as masculine.

5.3.2 Latent anaphora resolution

The results in Tables 4 and 5 suggest that the context-aware model exploits information about the antecedent of an ambiguous pronoun. We hypothesize that we can interpret the model’s attention mechanism as a latent anaphora resolution, and perform experiments to test this hypothesis.

For test sets from Table 4, we find an antecedent noun phrase (usually a determiner or a possessive pronoun followed by a noun) using Stanford CoreNLP (Manning et al., 2014b). We select only examples where a noun phrase contains a single noun to simplify our analysis. Then we identify which token receives the highest attention weight (excluding <bos> and <eos> tokens and punctuation). If this token falls within the antecedent span, then we treat it as agreement (see Table 6).

One natural question might be: does the attention component in our model genuinely learn to perform anaphora resolution, or does it capture some simple heuristic (e.g., pointing to the last noun)? To answer this question, we consider several baselines: choosing a random, last or first

pronoun	agreement (in %)			
	random	first	last	attention
it	40	36	52	58
you	42	63	29	67
I	39	56	35	62

Table 7: Agreement with CoreNLP for test sets of pronouns having a nominal antecedent in context sentence (%). Examples with ≥ 1 noun in context sentence.

noun from the context sentence as an antecedent.

Note that an agreement of the last noun for “it” or the first noun for “you” and “I” is very high. This is partially due to the fact that most context sentences have only one noun. For these examples a random and last predictions are always correct, meanwhile attention does not always pick a noun as the most relevant word in the context. To get a more clear picture let us now concentrate only on examples where there is more than one noun in the context (Table 7). We can now see that the attention weights are in much better agreement with the coreference system than any of the heuristics. This indicates that the model is indeed performing anaphora resolution.

While agreement with CoreNLP is encouraging, we are aware that coreference resolution by CoreNLP is imperfect and partial agreement with it may not necessarily indicate that the attention is particularly accurate. In order to control for this, we asked human annotators to manually evaluate 500 examples from the test sets where CoreNLP predicted that “it” refers to a noun in the context sentence. More precisely, we picked random 500 examples from the test set with “it” from Table 7. We marked the pronoun in a source which CoreNLP found anaphoric. Assessors were given the source and context sentences and were asked to mark an antecedent noun phrase for a marked pronoun in a source sentence or say that there is no antecedent at all. We then picked those examples where assessors found a link from “it” to some noun in context (79% of all examples). Then we evaluated agreement of CoreNLP and our model with the ground truth links. We also report the performance of the best heuristic for “it” from our previous analysis (i.e. last noun in context). The results are provided in Table 8.

The agreement between our model and the ground truth is 72%. Though 5% below the coreference system, this is a lot higher than the best

	agreement (in %)
CoreNLP	77
attention	72
last noun	54

Table 8: Performance of CoreNLP and our model’s attention mechanism compared to human assessment. Examples with ≥ 1 noun in context sentence.

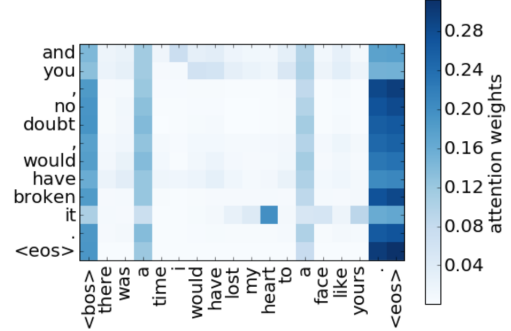


Figure 5: An example of an attention map between source and context. On the y -axis are the source tokens, on the x -axis the context tokens. Note the high attention between “it” and its antecedent “heart”.

	CoreNLP	
	right	wrong
attn right	53	19
attn wrong	24	4

Table 9: Performance of CoreNLP and our model’s attention mechanism compared to human assessment (%). Examples with ≥ 1 noun in context sentence.

heuristic (+18%). This confirms our conclusion that our model performs latent anaphora resolution. Interestingly, the patterns of mistakes are quite different for CoreNLP and our model (Table 9). We also present one example (Figure 5) where the attention correctly predicts anaphora while CoreNLP fails. Nevertheless, there is room for improvement, and improving the attention component is likely to boost translation performance.

6 Related work

Our analysis focuses on how our context-aware neural model implicitly captures anaphora. Early work on anaphora phenomena in statistical machine translation has relied on external systems for coreference resolution (Le Nagard and Koehn, 2010; Hardmeier and Federico, 2010). Results

were mixed, and the low performance of coreference resolution systems was identified as a problem for this type of system. Later work by [Hardmeier et al. \(2013\)](#) has shown that cross-lingual pronoun prediction systems can implicitly learn to resolve coreference, but this work still relied on external feature extraction to identify anaphora candidates. Our experiments show that a context-aware neural machine translation system can implicitly learn coreference phenomena without any feature engineering.

[Tiedemann and Scherrer \(2017\)](#) and [Bawden et al. \(2018\)](#) analyze the attention weights of context-aware NMT models. [Tiedemann and Scherrer \(2017\)](#) find some evidence for above-average attention on contextual history for the translation of pronouns, and our analysis goes further in that we are the first to demonstrate that our context-aware model learns latent anaphora resolution through the attention mechanism. This is contrary to [Bawden et al. \(2018\)](#), who do not observe increased attention between a pronoun and its antecedent in their recurrent model. We deem our model more suitable for analysis, since it has no recurrent connections and fully relies on the attention mechanism within a single attention layer.

7 Conclusions

We introduced a context-aware NMT system which is based on the Transformer architecture. When evaluated on an En-Ru parallel corpus, it outperforms both the context-agnostic baselines and a simple context-aware baseline. We observe that improvements are especially prominent for sentences containing ambiguous pronouns. We also show that the model induces anaphora relations. We believe that further improvements in handling anaphora, and by proxy translation, can be achieved by incorporating specialized features in the attention model. Our analysis has focused on the effect of context information on pronoun translation. Future work could also investigate whether context-aware NMT systems learn other discourse phenomena, for example whether they improve the translation of elliptical constructions, and markers of discourse relations and information structure.

Acknowledgments

We would like to thank Bonnie Webber for helpful discussions and anonymous reviewers for their

comments. The authors also thank David Talbot and Yandex Machine Translation team for helpful discussions and inspiration. Ivan Titov acknowledges support of the European Research Council (ERC StG BroadSem 678254) and the Dutch National Science Foundation (NWO VIDI 639.022.518). Rico Sennrich has received funding from the Swiss National Science Foundation (105212_169888).

References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the Third International Conference on Learning Representations (ICLR 2015)*. San Diego.
- Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating Discourse Phenomena in Neural Machine Translation. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. New Orleans, USA.
- Marine Carpuat. 2009. [One Translation Per Discourse](#). In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*. Association for Computational Linguistics, Boulder, Colorado, pages 19–27. <http://www.aclweb.org/anthology/W09-2404>.
- Zhengxian Gong, Min Zhang, Chew Lim Tan, and Guodong Zhou. 2012. [N-gram-based tense models for statistical machine translation](#). In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, Jeju Island, Korea, pages 276–285. <http://www.aclweb.org/anthology/D12-1026>.
- Zhengxian Gong, Min Zhang, and Guodong Zhou. 2011. [Cache-based Document-level Statistical Machine Translation](#). In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Edinburgh, Scotland, UK., pages 909–919. <http://www.aclweb.org/anthology/D11-1084>.
- Christian Hardmeier. 2012. Discourse in statistical machine translation: A survey and a case study. *Discours* 11.
- Christian Hardmeier and Marcello Federico. 2010. Modelling Pronominal Anaphora in Statistical Machine Translation. In *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*. pages 283–289.

- Christian Hardmeier, Preslav Nakov, Sara Stymne, Jörg Tiedemann, Yannick Versley, and Mauro Cettolo. 2015. [Pronoun-Focused MT and Cross-Lingual Pronoun Prediction: Findings of the 2015 DiscoMT Shared Task on Pronoun Translation](#). In *Proceedings of the Second Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal, pages 1–16. <https://doi.org/10.18653/v1/W15-2501>.
- Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre. 2013. [Latent anaphora resolution for cross-lingual pronoun prediction](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, pages 380–391. <http://www.aclweb.org/anthology/D13-1037>.
- Eva Hasler, Phil Blunsom, Philipp Koehn, and Barry Haddow. 2014. [Dynamic topic adaptation for phrase-based mt](#). In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Gothenburg, Sweden, pages 328–337. <https://doi.org/10.3115/v1/E14-1035>.
- Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does Neural Machine Translation Benefit from Larger Context? In *arXiv:1704.05135*. ArXiv: 1704.05135.
- Yangfeng Ji, Chenhao Tan, Sebastian Martschat, Yejin Choi, and Noah A Smith. 2017. [Dynamic entity representations in neural language models](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 1830–1839. <https://doi.org/10.18653/v1/D17-1195>.
- Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representation (ICLR 2015)*.
- Ronan Le Nagard and Philipp Koehn. 2010. [Aiding pronoun translation with co-reference resolution](#). In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*. Association for Computational Linguistics, Uppsala, Sweden, pages 252–261. <http://www.aclweb.org/anthology/W10-1737>.
- Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.
- Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan.
- Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014a. [The stanford corenlp natural language processing toolkit](#). In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Baltimore, Maryland, pages 55–60. <https://doi.org/10.3115/v1/P14-5010>.
- Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014b. [The Stanford CoreNLP natural language processing toolkit](#). In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Baltimore, Maryland, pages 55–60. <https://doi.org/10.3115/v1/P14-5010>.
- Thomas Meyer, Andrei Popescu-Belis, Najeh Hajlaoui, and Andrea Gesmundo. 2012. [Machine Translation of Labeled Discourse Connectives](#). In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA)*. <http://www.mt-archive.info/AMTA-2012-Meyer.pdf>.
- Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. [Validation of an automatic metric for the accuracy of pronoun translation \(apt\)](#). In *Proceedings of the Third Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, Copenhagen, Denmark, pages 17–25. <https://doi.org/10.18653/v1/W17-4802>.
- Ruslan Mitkov. 1999. Introduction: Special issue on anaphora resolution in machine translation and multilingual nlp. *Machine Translation* 14(3/4):159–161. <http://www.jstor.org/stable/40006919>.
- Stefan Riezler and John T. Maxwell. 2005. [On some pitfalls in automatic evaluation and significance testing for mt](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Association for Computational Linguistics, Ann Arbor, Michigan, pages 57–64. <https://www.aclweb.org/anthology/W05-0908>.
- 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 (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 1715–1725. <https://doi.org/10.18653/v1/P16-1162>.
- Jinsong Su, Hua Wu, Haifeng Wang, Yidong Chen, Xiaodong Shi, Huailin Dong, and Qun Liu. 2012. [Translation model adaptation for statistical machine translation with monolingual topic information](#). In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Jeju Island, Korea, pages 459–468. <http://www.aclweb.org/anthology/P12-1048>.

- Jörg Tiedemann. 2010. [Context Adaptation in Statistical Machine Translation Using Models with Exponentially Decaying Cache](#). In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*. Association for Computational Linguistics, Uppsala, Sweden, pages 8–15. <http://www.aclweb.org/anthology/W10-2602>.
- Jörg Tiedemann and Yves Scherrer. 2017. [Neural Machine Translation with Extended Context](#). In *Proceedings of the Third Workshop on Discourse in Machine Translation*. Association for Computational Linguistics, Copenhagen, Denmark, DISCOMT’17, pages 82–92. <https://doi.org/10.18653/v1/W17-4811>.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *NIPS*. Los Angeles. <http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf>.
- 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*. Association for Computational Linguistics, Denmark, Copenhagen, EMNLP’17, pages 2816–2821. <https://doi.org/10.18653/v1/D17-1301>.
- Sam Wiseman, Alexander M Rush, and Stuart M Shieber. 2016. [Learning global features for coreference resolution](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 994–1004. <https://doi.org/10.18653/v1/N16-1114>.

A Experimental setup

A.1 Data preprocessing

We use the publicly available OpenSubtitles2018 corpus (Lison and Tiedemann, 2016) for English and Russian.⁴ We pick sentence pairs with an overlap of at least 0.9 to reduce noise in the data. For context, we take the previous sentence if its timestamp differs from the current one by no more than 7 seconds.

We use the tokenization provided by the corpus.

Sentences were encoded using byte-pair encoding (Sennrich et al., 2016), with source and target vocabularies of about 32000 tokens. Translation pairs were batched together by approximate sequence length. Each training batch contained a set of translation pairs containing approximately 5000 source tokens.

A.2 Model parameters

We follow the setup of Transformer base model (Vaswani et al., 2017). More precisely, the number of layers in the encoder and decoder is $N = 6$. We employ $h = 8$ parallel attention layers, or heads. The dimensionality of input and output is $d_{model} = 512$, and the inner-layer of a feed-forward networks has dimensionality $d_{ff} = 2048$.

We use regularization as described in (Vaswani et al., 2017).

A.3 Optimizer

The optimizer we use is the same as in (Vaswani et al., 2017). We use the Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\varepsilon = 10^9$. We vary the learning rate over the course of training, according to the formula:

$$l_{rate} = d_{model}^{0.5} \cdot \min(step_num^{0.5}, step_num \cdot warmup_steps^{1.5})$$

We use $warmup_steps = 4000$.

⁴<http://opus.nlpl.eu/OpenSubtitles2018.php>