Document-Level Neural Machine Translation with Hierarchical Attention Networks

Lesly Miculicich[†◇] **Dhananjay Ram**[†◇] **Nikolaos Pappas**[†] **James Henderson**[†]
[†] Idiap Research Institute, Switzerland
[◇]École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{lmiculicich, dram, npappas, jhenderson}@idiap.ch

Abstract

Neural Machine Translation (NMT) can be improved by including document-level contextual information. For this purpose, we propose a hierarchical attention model to capture the context in a structured and dynamic manner. The model is integrated in the original NMT architecture as another level of abstraction, conditioning on the NMT model's own previous hidden states. Experiments show that hierarchical attention significantly improves the BLEU score over a strong NMT baseline with the state-of-the-art in context-aware methods, and that both the encoder and decoder benefit from context in complementary ways.

1 Introduction

Neural machine translation (NMT) (Bahdanau et al., 2015; Wu et al., 2016; Vaswani et al., 2017) trains an encoder-decoder network on sentence pairs to maximize the likelihood of predicting a target-language sentence given the corresponding source-language sentence, without considering the document context. By ignoring discourse connections between sentences and other valuable contextual information, this simplification potentially degrades the coherence and cohesion of a translated document (Hardmeier, 2012; Meyer and Webber, 2013; Sim Smith, 2017). Recent studies (Tiedemann and Scherrer, 2017; Jean et al., 2017; Wang et al., 2017; Tu et al., 2018) have demonstrated that adding contextual information to the NMT model improves the general translation performance, and more importantly, improves the coherence and cohesion of the translated text (Bawden et al., 2018; Lapshinova-Koltunski and Hardmeier, 2017). Most of these methods use an additional encoder (Jean et al., 2017; Wang et al., 2017) to extract contextual information from previous source-side sentences. However, this requires additional parameters and it does not exploit the representations already learned by the NMT encoder. More recently, Tu et al. (2018) have shown that a cache-based memory network performs better than the above encoder-based methods. The cache-based memory keeps past context as a set of words, where each cell corresponds to one unique word keeping the hidden representations learned by the NMT while translating it. However, in this method, the word representations are stored irrespective of the sentences where they occur, and those vector representations are disconnected from the original NMT network.

We propose to use a hierarchical attention network (HAN) (Yang et al., 2016) to model the contextual information in a structured manner using word-level and sentence-level abstractions. In contrast to the hierarchical recurrent neural network (HRNN) used by (Wang et al., 2017), here the attention allows dynamic access to the context by selectively focusing on different sentences and words for each predicted word. In addition, we integrate two HANs in the NMT model to account for target and source context. The HAN encoder helps in the disambiguation of source-word representations, while the HAN decoder improves the target-side lexical cohesion and coherence. The integration is done by (i) re-using the hidden representations from both the encoder and decoder of previous sentence translations and (ii) providing input to both the encoder and decoder for the current translation. This integration method enables it to jointly optimize for multiple-sentences. Furthermore, we extend the original HAN with a multi-head attention (Vaswani et al., 2017) to capture different types of discourse phenomena.

Our main contributions are the following: (i) We propose a HAN framework for translation to capture context and inter-sentence connections in a structured and dynamic manner. (ii) We integrate the HAN in a very competitive NMT ar-

chitecture (Vaswani et al., 2017) and show significant improvement over two strong baselines on multiple data sets. (iii) We perform an ablation study of the contribution of each HAN configuration, showing that contextual information obtained from source and target sides are complementary.

2 The Proposed Approach

The goal of NMT is to maximize the likelihood of a set of sentences in a target language represented as sequences of words $\mathbf{y} = (y_1, ..., y_t)$ given a set of input sentences in a source language $\mathbf{x} = (x_1, ..., x_m)$ as:

$$\max_{\Theta} \frac{1}{N} \sum_{n=1}^{N} log(P_\Theta(\mathbf{y^n}|\mathbf{x^n})) \quad (1)$$

so, the translation of a document $\mathbf{D}$ is made by translating each of its sentences independently. In this study, we introduce dependencies on the previous sentences from the source and target sides:

$$\max_{\Theta} \frac{1}{N} \sum_{n=1}^{N} log(P_\Theta(\mathbf{y^n}|\mathbf{x^n}, \mathbf{D_{x^n}}, \mathbf{D_{y^n}})) \quad (2)$$

where $\mathbf{D_{x^n}} = (\mathbf{x^{n-k}}, ..., \mathbf{x^{n-1}})$ and $\mathbf{D_{y^n}} = (\mathbf{y^{n-k}}, ..., \mathbf{y^{n-1}})$ denote the previous k sentences from source and target sides respectively. The contexts $\mathbf{D_{x^n}}$ and $\mathbf{D_{y^n}}$ are modeled with HANs.

2.1 Hierarchical Attention Network

The proposed HAN has two levels of abstraction. The word-level abstraction summarizes information from each previous sentence j into a vector s^j as:

$$q_w = f_w(h_t) \quad (3)$$

$$s^j = \underset{i}{\text{MultiHead}}(q_w, h_i^j) \quad (4)$$

where h denotes a hidden state of the NMT network. In particular, h_t is the last hidden state of the word to be encoded, or decoded at time step t, and h_i^j is the last hidden state of the i-th word of the j-th sentence of the context. The function f_w is a linear transformation to obtain the *query* q_w. We used the MultiHead attention function proposed by (Vaswani et al., 2017) to capture different types of relations among words. It matches the *query* against each of the hidden representations h_i^j (used as *value* and *key* for the attention).

The sentence-level abstraction summarizes the contextual information required at time t in d_t as:

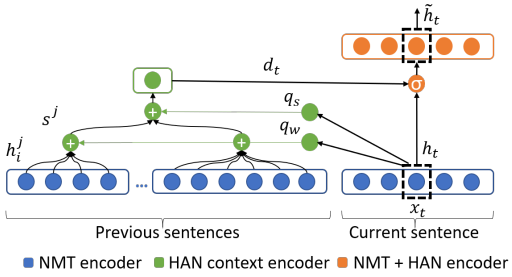


Figure 1: Integration of HAN during encoding at time step t, $\tilde{h}_t$ is the context-aware hidden state of the word x_t. Similar architecture is used during decoding.

$$q_s = f_s(h_t) \quad (5)$$

$$d_t = \text{FFN}(\underset{j}{\text{MultiHead}}(q_s, s^j)) \quad (6)$$

where f_s is a linear transformation, q_s is the query for the attention function, FFN is a position-wise feed-forward layer (Vaswani et al., 2017). Each layer is followed by a normalization layer (Lei Ba et al., 2016).

2.2 Context Gating

We use a gate (Tu et al., 2018, 2017) to regulate the information at sentence-level h_t and the contextual information at document-level d_t. The intuition is that different words require different amount of context for translation:

$$\lambda_t = \sigma(W_h h_t + W_d d_t) \quad (7)$$

$$\widetilde{h}_t = \lambda_t h_t + (1 - \lambda_t) d_t \quad (8)$$

where W_h, W_p are parameter matrices, and $\widetilde{h}_t$ is the final hidden representation for a word x_t or y_t.

2.3 Integrated Model

The context can be used during encoding or decoding a word, and it can be taken from previously encoded source sentences, previously decoded target sentences, or from previous alignment vectors (i.e. context vectors (Bahdanau et al., 2015)). The different configurations will define the input *query* and *values* of the attention function. In this work we experiment with five of them: one at encoding time, three at decoding time, and one combining both. At encoding time the *query* is a function of the hidden state h_{x_t} of the current word to be encoded x_t, and the *values* are the encoded states of previous sentences $h_{x_i}^j$ (HAN encoder). At decoding time, the *query* is a function of the hidden state h_{y_t} of the current word to be decoded y_t, and the *values* can be (a) the encoded states of previous sentences $h_{x_i}^j$ (HAN decoder *source*),

(b) the decoded states of previous sentences $h^j_{y_i}$ (HAN decoder), and (c) the alignment vectors c^j_i (HAN decoder *alignment*). Finally, we combine complementary target-source sides of the context by joining HAN encoder and HAN decoder. Figure 1 shows the integration of the HAN encoder with the NMT model; a similar architecture is applied to the decoder. The output $\tilde{h}_t$ is used by the NMT model as replacement of h_t during the final classification layer.

3 Experimental Setup

3.1 Datasets and Evaluation Metrics

We carry out experiments with Chinese-to-English (Zh-En) and Spanish-to-English (Es-En) sets on three different domains: talks, subtitles, and news.

TED Talks is part of the IWSLT 2014 and 2015 (Cettolo et al., 2012, 2015) evaluation campaigns[1]. We use *dev2010* for development; and *tst2010-2012* (Es-En), *tst2010-2013* (Zh-En) for testing. The Zh-En subtitles corpus is a compilation of TV subtitles designed for research on context (Wang et al., 2018). In contrast to the other sets, it has three references to compare. The Es-En corpus is a subset of OpenSubtitles2018 (Lison and Tiedemann, 2016)[2]. We randomly select two episodes for development and testing each. Finally, we use the Es-En News-Commentaries11[3] corpus which has document-level delimitation. We evaluate on WMT sets (Bojar et al., 2013): *newstest2008* for development, and *newstest2009-2013* for testing. A similar corpus for Zh-En is too small to be comparable. Table 2 shows the corpus statistics.

For evaluation, we use BLEU score (Papineni et al., 2002) (*multi-blue*) on *tokenized* text, and we measure significance with the paired bootstrap resampling method proposed by Koehn (2004) (implementations by Koehn et al. (2007)).

3.2 Model Configuration and Training

As baselines, we use a NMT transformer, and a context-aware NMT transformer with cache memory which we implemented for comparison following the best model described by Tu et al. (2018), with memory size of 25 words. We used the OpenNMT (Klein et al., 2017) implementation of the transformer network. The configuration is the same as the model called "base model" in the original paper (Vaswani et al., 2017). The encoder and decoder are composed of 6 hidden layers each. All hidden states have dimension of 512, dropout of 0.1, and 8 heads for the multi-head attention. The target and source vocabulary size is 30K. The optimization and regularization methods were the same as proposed by Vaswani et al. (2017). Inspired by Tu et al. (2018) we trained the models in two stages. First we optimize the parameters for the NMT without the HAN, then we proceed to optimize the parameters of the whole network. We use $k = 3$ previous sentences, which gave the best performance on the development set.

[1] https://wit3.fbk.eu
[2] http://www.opensubtitles.org
[3] http://opus.nlpl.eu/News-Commentary11.php

4 Experimental Results

4.1 Translation Performance

Table 1 shows the BLEU scores for different models. The baseline NMT transformer already has better performance than previously published results on these datasets, and we replicate previous previous improvements from the cache method over the this stronger baseline. All of our proposed HAN models perform at least as well as the cache method. The best scores are obtained by the combined encoder and decoder HAN model, which is significantly better than the cache method on all datasets without compromising training speed (2.3K vs 2.6K tok/sec). An important portion of the improvement comes from the HAN encoder, which can be attributed to the fact that the source-side always contains correct information, while the target-side may contain erroneous predictions at testing time. But combining HAN decoder with HAN encoder further improves translation performance, showing that they contribute complementary information. The three ways of incorporating information into the decoder all perform similarly.

Table 3 shows the performance of our best HAN model with a varying number k of previous sentences in the test-set. We can see that the best performance for TED talks and news is archived with 3, while for subtitles it is similar between 3 and 7.

4.2 Accuracy of Pronoun/Noun Translations

We evaluate coreference and anaphora using the reference-based metric: accuracy of pronoun translation (Miculicich Werlen and Popescu-Belis, 2017b), which can be extended for nouns. The list of evaluated pronouns is predefined in the metric, while the list of nouns was extracted using NLTK POS tagging (Bird, 2006). The upper part

Models	TED Talks Zh–En BLEU	Δ	TED Talks Es–En BLEU	Δ	Subtitles Zh–En [4] BLEU	Δ	Subtitles Es–En BLEU	Δ	News Es–En BLEU	Δ
NMT transformer	16.87		35.44		28.60		35.20		21.36	
+ cache (Tu et al., 2018)	17.32	$(+0.45)^{***}$	36.46	$(+1.02)^{***}$	28.86	$(+0.26)$	35.49	$(+0.29)$	22.36	$(+1.00)^{***}$
+ HAN encoder	17.61	$(+0.74)^{***}_{\dagger\dagger}$	36.91	$(+1.47)^{***}_{\dagger\dagger}$	29.35	$(+0.75)^{*}_{\dagger}$	35.96	$(+0.76)^{*}_{\dagger}$	22.36	$(+1.00)^{***}$
+ HAN decoder	17.39	$(+0.52)^{***}$	37.01	$(+1.57)^{***}_{\dagger\dagger\dagger}$	29.21	$(+0.61)^{*}$	35.50	$(+0.30)$	22.62	$(+1.26)^{***}_{\dagger\dagger\dagger}$
+ HAN decoder *source*	17.56	$(+0.69)^{***}_{\dagger\dagger}$	36.94	$(+1.50)^{***}_{\dagger\dagger}$	28.92	$(+0.32)$	35.71	$(+0.51)^{*}$	22.68	$(+1.32)^{***}_{\dagger\dagger\dagger}$
+ HAN decoder *alignment*	17.48	$(+0.61)^{***}_{\dagger}$	37.03	$(+1.60)^{***}_{\dagger\dagger\dagger}$	28.87	$(+0.27)$	35.63	$(+0.43)$	22.59	$(+1.23)^{***}_{\dagger\dagger\dagger}$
+ HAN encoder + HAN decoder	**17.79**	$(+0.92)^{***}_{\dagger\dagger\dagger}$	**37.24**	$(+1.80)^{***}_{\dagger\dagger\dagger}$	**29.67**	$(+1.07)^{**}_{\dagger}$	**36.23**	$(+1.03)^{**}_{\dagger\dagger}$	**22.76**	$(+1.40)^{***}_{\dagger\dagger\dagger}$

Table 1: BLEU score for the different configurations of the HAN model, and two baselines. The highest score per dataset is marked in bold. Δ denotes the difference in BLEU score with respect of the NMT transformer. The significance values with respect to the NMT and the cache method are denoted by $*$, and † respectively. The repetitions correspond to the p-values: $^{*}_{\dagger} < .05, ^{**}_{\dagger\dagger} < .01, ^{***}_{\dagger\dagger\dagger} < .001$.

	TED Talks Zh–En	TED Talks Es–En	Subtitles Zh–En	Subtitles Es–En	News Es–En
Training	0.2M	0.2M	2.2M	4.0M	0.2M
Development	0.8K	0.8K	1.1K	1.0K	1.9K
Test	5.5K	4.7K	1.2K	1.0K	13.5K

Table 2: Dataset statistics in # sentence pairs.

k	TED Talks Zh–En	TED Talks Es–En	Subtitles Zh–En	Subtitles Es–En	News Es–En
1	17.70	37.20	29.35	36.20	22.46
3	**17.79**	**37.24**	29.67	**36.23**	**22.76**
5	17.49	37.11	**29.69**	36.22	22.54
7	17.00	37.22	29.64	36.21	22.64

Table 3: Performance for variable context sizes k with the HAN encoder + HAN decoder.

of Table 4 shows the results. For nouns, the joint HAN achieves the best accuracy with a significant improvement compared to other models, showing that target and source contextual information are complementary. Similarity for pronouns, the joint model has the best result for TED talks and news. However, HAN encoder alone is better in the case of subtitles. Here HAN decoder produces mistakes by repeating past translated personal pronouns. Subtitles is a challenging corpus for personal pronoun disambiguation because it usually involves dialogue between multiple speakers.

4.3 Cohesion and Coherence Evaluation

We use the metric proposed by Wong and Kit (2012) to evaluate lexical cohesion. It is defined as the ratio between the number of repeated and lexically similar content words over the total number of content words in a target document. The lexical similarity is obtained using WordNet. Table 4 (bottom-left) displays the average ratio per tested document. In some cases, HAN decoder achieves the best score because it produces a larger quantity of repetitions than other models. However, as previously demonstrated in 4.2, repetitions do not always make the translation better. Although HAN boosts lexical cohesion, the scores are still far from the human reference, so there is room for improvement in this aspect.

For coherence, we use a metric based on Latent Semantic Analysis (LSA) (Foltz et al., 1998). LSA is used to obtain sentence representations, then cosine similarity is calculated from one sentence to the next, and the results are averaged to get a document score. We employed the pre-trained LSA model *Wiki-6* from (Stefanescu et al., 2014). Table 4 (bottom-right) shows the average coherence score of documents. The joint HAN model consistently obtains the best coherence score, but close to other HAN models. Most of the improvement comes from the HAN decoder.

4.4 Qualitative Analysis

Table 5 shows an example where HAN helped to generate the correct translation. The first box shows the current sentence with the analyzed word in bold; and the second, the past context at source and target. For the context visualization we use the toolkit provided by Pappas and Popescu-Belis (2017). Red corresponds to sentences, and blue to words. The intensity of color is proportional to the weight. We see that HAN correctly translates the ambiguous Spanish pronoun "*su*" into the English "*his*". The HAN decoder highlighted a previous mention of "*his*", and the HAN encoder highlighted the antecedent "*Nathaniel*". This shows that HAN can capture interpretable inter-sentence connections. More samples with different attention heads are shown in the Appendix A.

5 Related Work

Statistical Machine Translation (SMT) Initial studies were based on cache memories (Tiede-

[4] *NIST BLEU*: NMT transformer 35.99, cache 36.52, and HAN 37.15.

Model	Noun Translation					Pronoun Translation				
	TED Talks		Subtitles		News	TED Talks		Subtitles		News
	Zh–En	Es–En	Zh–En	Es–En	Zh–En	Es–En	Zh–En	Es–En	Zh–En	Es–En
NMT Transformer	40.16	65.97	46.65	61.79	47.94	63.44	68.00	69.71	65.83	47.22
+ cache	40.87	66.75	46.00	61.87	49.91	63.53	68.66	69.97	66.27	49.34
+ HAN encoder	41.93	67.75	46.78	61.52	50.06	64.05	69.17	**71.04**	**68.56**	49.57
+ HAN decoder	41.61	67.35	46.78	61.99	50.03	64.02	69.36	70.50	67.03	49.33
+ HAN encoder + HAN decoder	**42.99**	**67.81**	**47.43**	**62.30**	**50.40**	**64.35**	**69.60**	70.60	67.47	**49.59**
	Lexical cohesion					Coherence				
NMT Transformer	54.26	51.98	51.87	51.77	30.06	0.298	0.299	0.283	0.262	0.279
+ HAN encoder	54.87	52.35	51.89	52.33	30.34	0.304	0.299	0.285	0.262	0.280
+ HAN decoder	54.95	**52.43**	**52.33**	52.43	30.41	0.302	0.301	0.287	0.265	0.282
+ HAN enc. + HAN dec.	**55.40**	52.36	51.94	**52.75**	**30.58**	**0.305**	**0.302**	**0.287**	**0.265**	**0.282**
Human reference	56.08	57.02	54.81	58.19	35.12	0.310	0.314	0.296	0.270	0.298

Table 4: Evaluation on discourse phenomena. Noun and pronoun translation: Accuracy with respect to a human reference. Lexical cohesion: Ratio of repeated and lexically similar words over the number of content words. Coherence: Average cosine similarity of consecutive sentences (i.e. average of LSA word-vectors)

Currently Translated Sentence

Src.:	y esto es un escape de **su** estado atormentado .
Ref.:	and that is an escape from **his** tormented state .
Base:	and this is an escape from ***its*** $< unk >$ state .
Cache:	and this is an escape from ***their*** state .
HAN:	and this is an escape from **his** $< unk >$ state .

Context from Previous Sentences

HAN decoder context with target. *Query*: **his** (En)

s^{t-3} music is medicine . music changes us .

s^{t-2} and for Nathaniel , music is mine .

s^{t-1} because music allows him to take his thoughts and his delusions and turn through his imagination and his creativity actually .

HAN encoder context with source. *Query*: **su** (Es)

s^{t-3} la música es medicina . la música nos cambia .

s^{t-2} y para Nathaniel la música es cordura .

s^{t-1} porque la música le permite tomar sus pensamientos y sus delirios y transformarlos a través de su imaginación y su creatividad en realidad .

Table 5: Example of pronoun disambiguation using HAN (TED Talks Es-En).

mann, 2010; Gong et al., 2011). However, most of the work explicitly models discourse phenomena (Sim Smith, 2017) such as lexical cohesion (Meyer and Popescu-Belis, 2012; Xiong et al., 2013; Loáiciga and Grisot, 2016; Pu et al., 2017; Mascarell, 2017), coherence (Born et al., 2017), and coreference (Rios Gonzales and Tuggener, 2017; Miculicich Werlen and Popescu-Belis, 2017a). Hardmeier et al. (2013) introduced the document-level SMT paradigm.

Sentence-level NMT Initial studies on NMT enhanced the sentence-level context by using memory networks (Wang et al., 2016), self-attention (Miculicich Werlen et al., 2018; Zhang et al., 2016), and latent variables (Yang et al., 2017).

Document-level NMT Tiedemann and Scherrer (2017) use the concatenation of multiple sentences as NMT's input/output, Jean et al. (2017) add a context encoder for the previous source sentence, Wang et al. (2017) includes a HRNN to summarize source-side context, and Tu et al. (2018) use a dynamic cache memory to store representations of previously translated words. Recently, Bawden et al. (2018) proposed test-sets for evaluating discourse in NMT, Voita et al. (2018) shows that context-aware NMT improves the of anaphoric pronouns, and Maruf and Haffari (2018) proposed a document-level NMT using memory-networks.

6 Conclusion

We proposed a hierarchical multi-head HAN NMT model[5] to capture inter-sentence connections. We integrated context from source and target sides by directly connecting representations from previous sentence translations into the current sentence translation. The model significantly outperforms two competitive baselines, and the ablation study shows that target and source context is complementary. It also improves lexical cohesion and coherence, and the translation of nouns and pronouns. The qualitative analysis shows that the model is able to identify important previous sentences and words for the correct prediction. In future work, we plan to explicitly model discourse connections with the help of annotated data, which may further improve translation quality.

Acknowledgments

We are grateful for the support of the European Union under the Horizon 2020 SUMMA project n. 688139, see www.summa-project.eu.

[5]Code available at https://github.com/idiap/HAN_NMT. Project Towards Document-Level NMT (Miculicich Werlen, 2017)

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*, San Diego, USA.

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 North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, USA. Association for Computational Linguistics.

Steven Bird. 2006. Nltk: The natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Leo Born, Mohsen Mesgar, and Michael Strube. 2017. Using a graph-based coherence model in document-level machine translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 26–35, Copenhagen, Denmark. Association for Computational Linguistics.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2015. The IWSLT 2015 evaluation campaign. In *In proceedins of the International Workshop on Spoken Language Translation*.

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.

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*, pages 909–919, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Christian Hardmeier. 2012. Discourse in statistical machine translation. a survey and a case study. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (11).

Christian Hardmeier, Sara Stymne, Jörg Tiedemann, and Joakim Nivre. 2013. Docent: A document-level decoder for phrase-based statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 193–198, Sofia, Bulgaria. Association for Computational Linguistics.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context? *arXiv preprint arXiv:1704.05135*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej 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 Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Ekaterina Lapshinova-Koltunski and Christian Hardmeier. 2017. Discovery of discourse-related language contrasts through alignment discrepancies in english-german translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 73–81, Copenhagen, Denmark. Association for Computational Linguistics.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. In *NISP 2016 - Deep Learning Symposium paper*.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association.

Sharid Loáiciga and Cristina Grisot. 2016. Predicting and using a pragmatic component of lexical aspect. *LiLT (Linguistic Issues in Language Technology)*, 13.

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 (Volume 1: Long Papers)*, pages 1275–1284. Association for Computational Linguistics.

Laura Mascarell. 2017. Lexical chains meet word embeddings in document-level statistical machine translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 99–109, Copenhagen, Denmark. Association for Computational Linguistics.

Thomas Meyer and Andrei Popescu-Belis. 2012. Using sense-labeled discourse connectives for statistical machine translation. In *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, pages 129–138. Association for Computational Linguistics.

Thomas Meyer and Bonnie Webber. 2013. Implicitation of discourse connectives in (machine) translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 19–26, Sofia, Bulgaria. Association for Computational Linguistics.

Lesly Miculicich Werlen. 2017. Towards document-level neural machine translation. Technical report, Idiap.

Lesly Miculicich Werlen, Nikolaos Pappas, Dhananjay Ram, and Andrei Popescu-Belis. 2018. Self-attentive residual decoder for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1366–1379. Association for Computational Linguistics.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017a. Using coreference links to improve spanish-to-english machine translation. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 30–40, Valencia, Spain. Association for Computational Linguistics.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017b. Validation of an automatic metric for the accuracy of pronoun translation (APT). In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 17–25, Copenhagen, Denmark. Association for Computational Linguistics.

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

Nikolaos Pappas and Andrei Popescu-Belis. 2017. Multilingual hierarchical attention networks for document classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1015–1025, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Xiao Pu, Laura Mascarell, and Andrei Popescu-Belis. 2017. Consistent translation of repeated nouns using syntactic and semantic cues. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 948–957. Association for Computational Linguistics.

Annette Rios Gonzales and Don Tuggener. 2017. Coreference resolution of elided subjects and possessive pronouns in spanish-english statistical machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 657–662. Association for Computational Linguistics.

Karin Sim Smith. 2017. On integrating discourse in machine translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 110–121, Copenhagen, Denmark. Association for Computational Linguistics.

Dan Stefanescu, Rajendra Banjade, and Vasile Rus. 2014. Latent semantic analysis models on wikipedia and tasa. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

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*, pages 8–15, Uppsala, Sweden. Association for Computational Linguistics.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Zhaopeng Tu, Yang Liu, Zhengdong Lu, Xiaohua Liu, and Hang Li. 2017. Context gates for neural machine translation. *Transactions of the Association for Computational Linguistics*, 5:87–99.

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*, 6.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

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 (Volume 1: Long Papers), pages 1264–1274. Association for Computational Linguistics.

Longyue Wang, Zhaopeng Tu, Shuming Shi, Tong Zhang, Yvette Graham, and Qun Liu. 2018. Translating pro-drop languages with reconstruction models. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 1–9, New Orleans, Louisiana, USA. AAAI Press.

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*, pages 2826–2831, Copenhagen, Denmark. Association for Computational Linguistics.

Mingxuan Wang, Zhengdong Lu, Hang Li, and Qun Liu. 2016. Memory-enhanced decoder for neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 278–286, Austin, Texas. Association for Computational Linguistics.

Billy T. M. Wong and Chunyu Kit. 2012. Extending machine translation evaluation metrics with lexical cohesion to document level. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1060–1068, Jeju Island, Korea. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Deyi Xiong, Yang Ding, Min Zhang, and Chew Lim Tan. 2013. Lexical chain based cohesion models for document-level statistical machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1563–1573, Seattle, Washington, USA. Association for Computational Linguistics.

Zichao Yang, Zhiting Hu, Yuntian Deng, Chris Dyer, and Alex Smola. 2017. Neural machine translation with recurrent attention modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 383–387, Valencia, Spain. Association for Computational Linguistics.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Biao Zhang, Deyi Xiong, jinsong su, Hong Duan, and Min Zhang. 2016. Variational neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 521–530, Austin, Texas. Association for Computational Linguistics.

A Examples of HAN Attention

The examples were taken from the Spanish-English TED talks corpus. We show the behavior of the attention function of HAN. First, we show the attention to context for HAN encoder and HAN decoder respectively. Second, we show the multi-head attention only for HAN decoder (English) for better understanding.

A.1 Encoder and Decoder Attention

Currently Translated Sentence

Src.:	y **toqué** el primer movimiento del concierto para violín de Beethoven .
Ref.:	and I **played** the first movement of the Beethoven Violin Concerto .
Base:	and I ***touched*** the first move from the concert to Beethoven .
Cache:	and I ***touched*** the first move of Beethoven 's violin .
HAN:	and I **played** the first move of Beethoven 's violin .

Context from Previous Sentences

HAN decoder context with target. *Query*: **played** (En)

s^{t-3} and he was talking about invisible demons and smoke , and how someone was sleeping with him .

s^{t-2} and I felt fear , not for me , but fear that I was going to lose it ...

s^{t-1} so I just started playing .

HAN encoder context with source. *Query*: **toqué** (Es)

s^{t-3} y hablaba de demonios invisibles y humo , y de cómo alguien lo estaba envenenando mientras dormía .

s^{t-2} y yo sentí miedo , no por mí , sino miedo de que iba a perderlo ...

s^{t-1} por ello sólo empecé a tocar .

Table 6: In this example, the HAN model disambiguates correctly the word "*toqué*", which can be translated as "*touched*" or "*played*". We can see that the HAN decoder uses the semantically close word "*playing*" from the previous sentence. In similar manner, the HAN encoder focused on "*tocar*" which is coherent with "*toqué*".

A.2 Multi-Head Attention

Currently Translated Sentence

Src.:	y $< ellos >$ estarían tan compenetrados en la partida de dados porque los juegos son tan atractivos ...
Ref:	and **they** would be so immersed in playing the dice games because games are so engaging ..
Base:	and ***you*** would be so $< unk >$ in the start of it because games are so attractive ...
HAN:	and **they** would be so $< unk >$ in the start of dice because games are so attractive ...

Context from Previous Sentences. *Query*: **they**

Head 2: Attention to the antecedent "*people*" in s^{t-3}.

s^{t-3} people suffered . people suffered .

s^{t-2} it was an extreme situation . they needed an extreme solution .

s^{t-1} so , according to Indyk , the games of dice and a polic y was established throughout the kingdom : one day , e verybody would eat , and the next day , everybody woul d eat .

Head 4: Attention to the same pronoun "*they*" in s^{t-2}

s^{t-3} people suffered . people suffered .

s^{t-2} it was an extreme situation . they needed an extreme solution .

s^{t-1} so , according to Indyk , the games of dice and a polic y was established throughout the kingdom : one day , e verybody would eat , and the next day , everybody woul d eat .

Head 7: Attention to verbs that conjugate with "*they*"

s^{t-3} people suffered . people suffered .

s^{t-2} it was an extreme situation . they needed an extreme solution .

s^{t-1} so , according to Indyk , the games of dice and a polic y was established throughout the kingdom : one day , e verybody would eat , and the next day , everybody woul d eat .

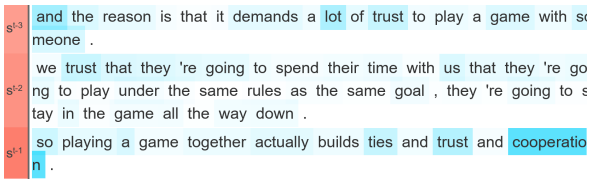
Table 7: This example displays the translation of Spanish pronoun "*ellos*", which is a dropped-pronoun which is implicit in the verb conjugation of "*estarían*". As we can observe, HAN translates correctly the dropped-pronoun into the English "*they*". Each head focuses on a different aspect during translation, for example *head 2* seems to attend to the antecedent of the pronoun "*people*" in the third previous sentence, *head 4* attends to the same pronoun on the second previous sentence, and *head 7* attends to different verbs on all previous sentences.

Currently Translated Sentence

Src:	y como resultado **construimos** relaciones sociales más fuertes .
Ref:	and we actually **build** stronger social relationships as a result .
Base:	and as a result , we ***construct*** stronger social relationships .
HAN:	and as a result , we **build** stronger social relationships .

Context from Previous Sentences. *Query*: **build**

Head 1: Attention to related words "*construimos*", "*trust*"...



Head 4: Attention to a similar translation "*builds*" in s^{t-1}

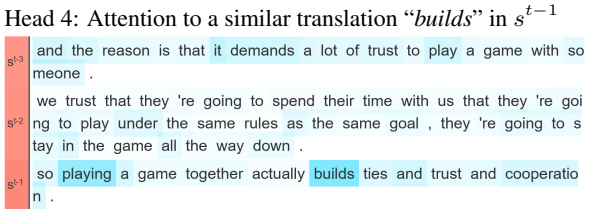


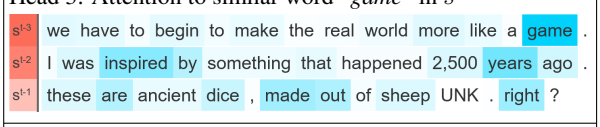
Table 8: This example displays the translation of the ambiguous Spanish word "*construimos*", which can be translated as "*construct*" or "*build*". HAN translates this word correctly according to the context using for example related words "*trust*", "*ties*", and "*cooperation*" on previous sentences with *head 1*, and a previous translation "*builds*" in the previous sentence with *head 4*.

Currently Translated Sentence

Src:	antes de los fantásticos controladores de **juegos** teníamos tabas de oveja .
Ref:	before we had awesome **game** controllers , we had sheep 's knuckles .
Base:	before the fantastic ***TV*** controllers , we had $< unk >$.
HAN:	before the fantastic **game** controllers , we had $< unk >$.

Context from Previous Sentences. *Query*: **game**

Head 3: Attention to similar word "*game*" in s^{t-3}



Head 5: Attention to a related word "*dice*" in s^{t-1}

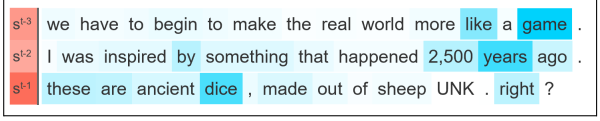


Table 9: This example shows the translation of the Spanish word "*juegos*". The baseline translates it incorrectly, while HAN translates it correctly by spotting a similar translation "*game*" in the third previous sentence with *head 3*, and a related word "*dice*" on previous sentence with *head 5*.