Document-Level Neural Machine Translation with Hierarchical Attention Networks

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, James Henderson

Motivation

Why document-level NMT?

- Not considering the document context and discourse connections affects coherence and cohesion of a text.

Why hierarchical attention networks?

- Different abstraction levels: word-level and sentence-level.
- Allows dynamic access to the context for each predicted word.

Other advantages in our approach

- Joint optimization of multiple sentences.
- Shared hidden representations across sentence translations.
- Exploiting source and target context.
- Multi-head attention to capture different discourse phenomena.

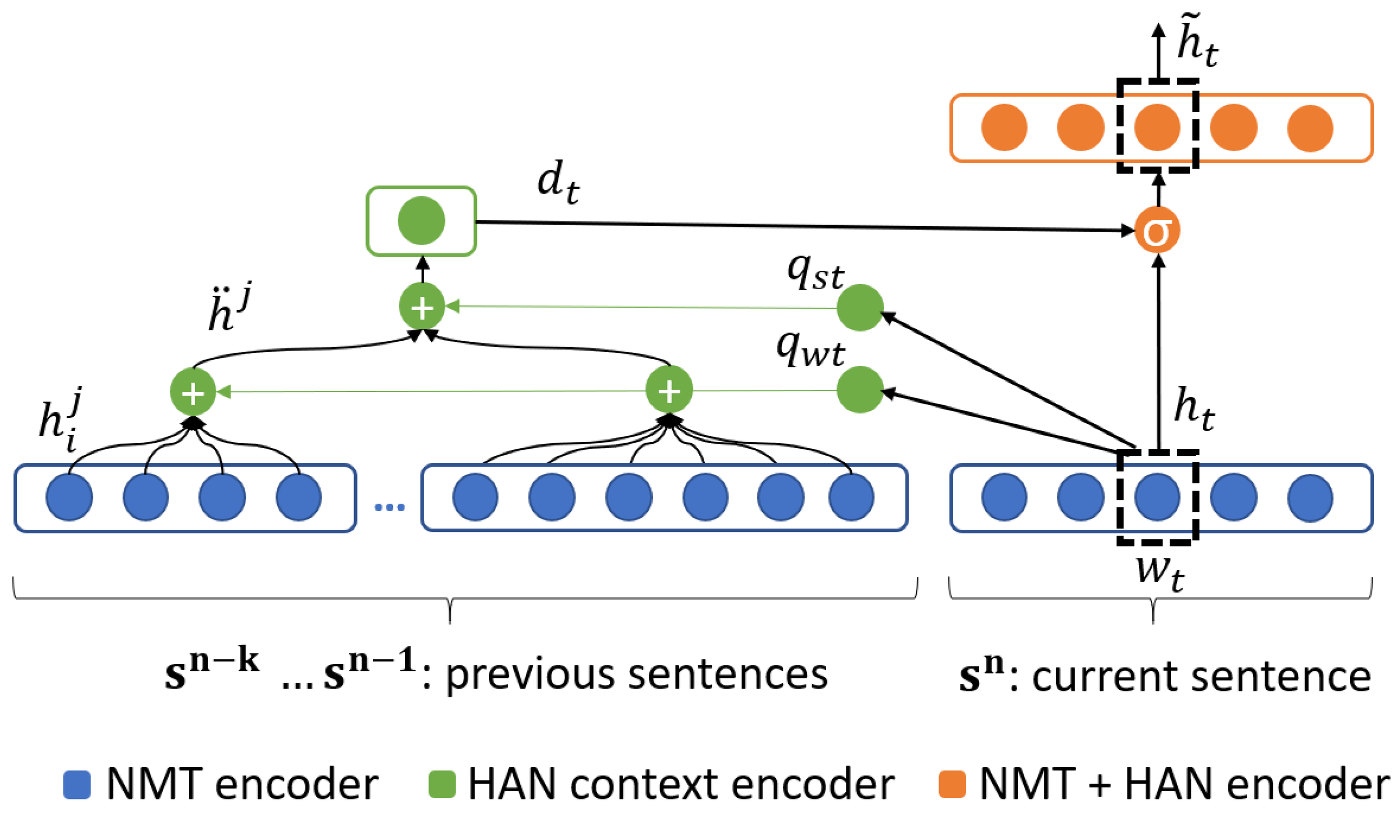
Document-level NMT

$$\max_{\Theta} \sum_{n=1}^{N} log(P_{\Theta}(\mathbf{t^n}|\mathbf{s^n})) \quad \rightarrow \quad \max_{\Theta} \sum_{n=1}^{N} log(P_{\Theta}(\mathbf{t^n}|\mathbf{s^n}, \mathbf{D_{s^n}}, \mathbf{D_{t^n}}))$$

Baseline NMT — **Document-level NMT**

- $\mathbf{s^n}$: source sentence, and $\mathbf{D_{s^n}} = (\mathbf{s^{n-k}}, ..., \mathbf{s^{n-1}})$: source context
- $\mathbf{t^n}$: target sentence, and $\mathbf{D_{t^n}} = (\mathbf{t^{n-k}}, ..., \mathbf{t^{n-1}})$: target context.
- Context (k previous sentences) is modeled by HANs:

Hierarchical Attention Network (HAN)



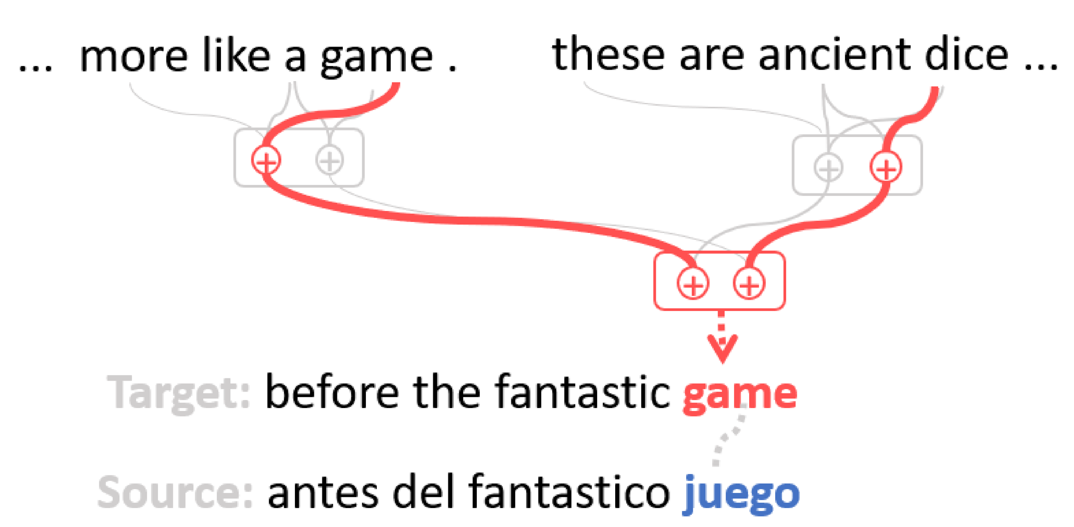
Word-level attention:

$$\ddot{h}^j = \mathrm{MultiHead}_{i}(q_{wt}, h_i^j) \quad q_{wt} = f_w(h_t)$$

Sentence-level attention:

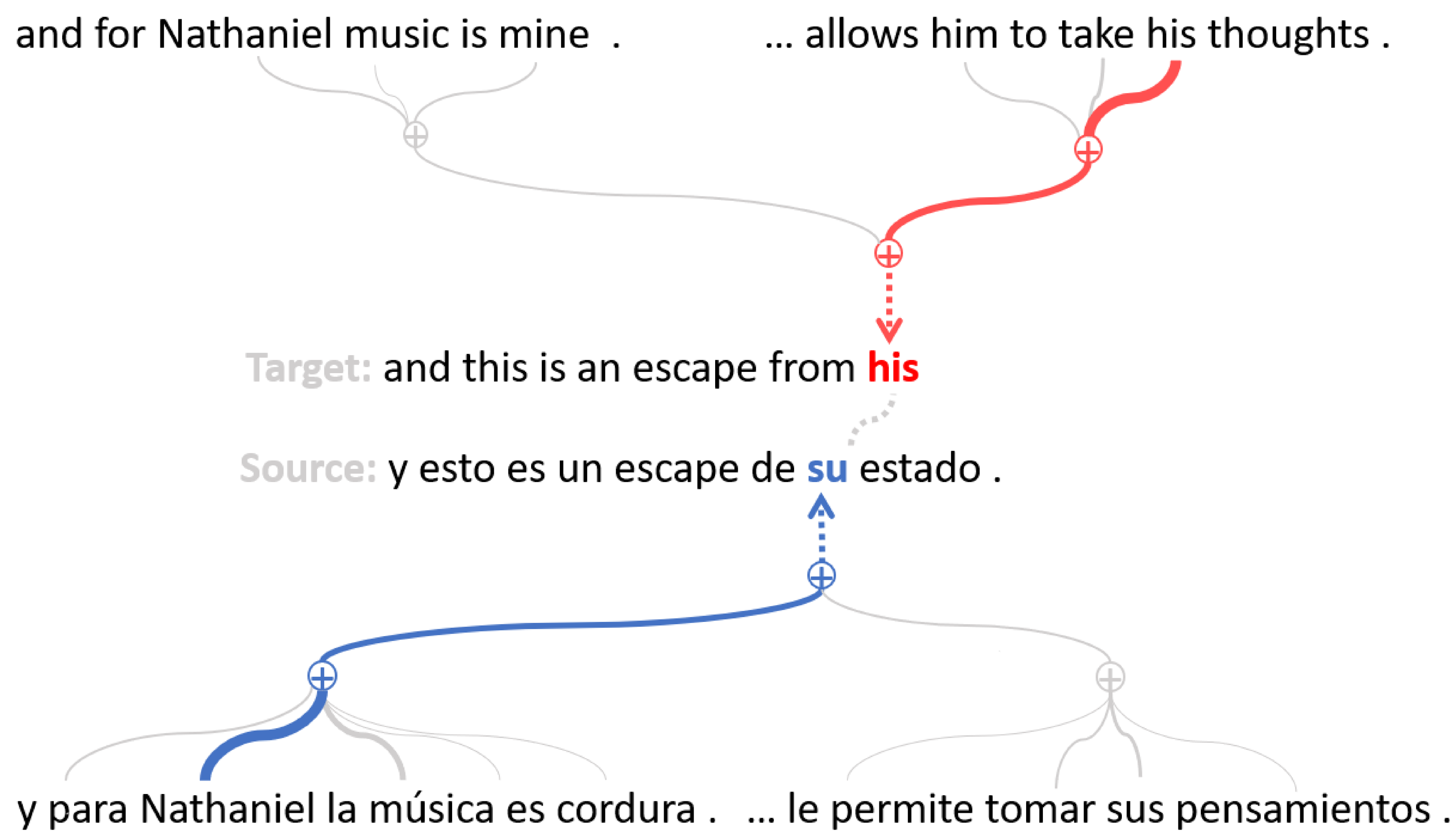
$$d_t = \mathrm{FFN}(\mathrm{MultiHead}_{j \in [n-k,...,n-1]}(q_{st}, \ddot{h}^j)) \quad q_{st} = f_s(h_t)$$

Multi-head attention



- "**juego**" can be translated as "game" or "set"

Source and Target Sides Context



- "**su**" can be translated as "his", "her", or "its"

Experimental Results

	TED Talks		Subtitles		News
	Zh–En	Es–En	Zh–En	Es–En	Es–En
NMT transformer	16.87	35.44	28.60	35.20	21.36
+ cache	17.32 ***	36.46 ***	28.86	35.49	22.36 ***
+ HAN encoder	17.61 $^{***}_{\dagger\dagger}$	36.91 $^{***}_{\dagger\dagger}$	29.35 $^{*}_{\dagger}$	35.96 $^{*}_{\dagger}$	22.36 ***
+ HAN decoder	17.39 ***	37.01 $^{***}_{\dagger\dagger\dagger}$	29.21 *	35.50	22.62 $^{***}_{\dagger\dagger\dagger}$
+ HAN joint	**17.79** $^{***}_{\dagger\dagger\dagger}$	**37.24** $^{***}_{\dagger\dagger\dagger}$	**29.67** $^{**}_{\dagger}$	**36.23** $^{**}_{\dagger\dagger}$	**22.76** $^{***}_{\dagger\dagger\dagger}$

BLEU scores. Significance with respect to NMT $*$, and to cache model $\dagger$.
P-values: $^{*}_{\dagger} < .05, ^{**}_{\dagger\dagger} < .01, ^{***}_{\dagger\dagger\dagger} < .001$.

- Significant improvement over strong baselines on multiple data sets.
- Context from source and target sides are complementary.

Discourse Evaluation

	Coherence	Lexical Cohesion	Pronouns	Nouns
NMT transformer	28.42	47.98	62.84	52.50
+ HAN encoder	28.60	48.35	64.48	53.61
+ HAN decoder	28.78	48.51	64.04	53.55
+ HAN joint	28.82	48.61	64.32	54.19
Human reference	29.79	52.94	100.0	100.0

- HAN decoder helps in lexical cohesion and coherence.
- HAN encoder helps in pronoun and noun disambiguation.

Conclusion

- We proposed a hierarchical multi-head attention model for document-level context.
- It directly connects representations from previous sentence translations into the current sentence translation.
- It significantly outperforms two competitive baselines.
- It improves cohesion and coherence, and noun/pronoun translation.
- We show that target and source context is complementary.

Our multi-head HAN could be used to model context in other NLP tasks. Code available at `https://github.com/idiap/HAN_NMT`

Acknowledgements Supported by the European Union Horizon 2020 SUMMA project (grant 688139, www.summa-project.eu).