Integrating and Evaluating Extra-linguistic Context in Neural Machine Translation

First Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Second Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Abstract

In Machine Translation (MT), taking into account information related to the setting in which a text is produced can be crucial. We investigate the impact of different extra-linguistic factors (speaker gender, speaker age, film genre and film year) on the MT of subtitles. Our starting point is the pseudo-token approach (Sennrich et al., 2016a). We explore the simultaneous addition of multiple factors in various orders in order to assess the limits of treating these factors as a sequence of words. We compare this approach to the encoding of the same factors using an additional, separate encoder. We evaluate both using BLEU and a targeted evaluation of how well the context is used. Our results show that both strategies are well adapted to exploiting such context. Contrarily to our intuitions, the pseudo-token approach appears unperturbed by the use of multiple values in various orders, and also results in significant improvements in BLEU score ($p < 0.01$). The multi-encoder approach proves more effective at integrating context but results in lower overall BLEU scores.

1 Introduction

When translating dialogues, extra-linguistic contextual information (e.g. speaker gender, topic) can be useful, if not necessary, to correctly translate certain elements of text. For example, correctly translating from English into French can be dependent on the gender of the speaker as in (1).

(1) EN: I am **surprised** and **shocked**.
FR_{male}: Je suis **surpris** et **choqué**.
FR_{female}: Je suis **surprise** et **choquée**.

Previous work has sought to adapt Machine Translation (MT) models to such contextual values, using strategies borrowed from domain adaptation (Foster and Kuhn, 2007), including values as side constraints in neural MT (NMT) (Sennrich et al., 2016a) and more recently by learning the traits implicitly (Michel and Neubig, 2018).

A simple and effective ways of integrating extra-linguistic context into NMT is the pseudo-token approach, whereby a contextual value is either prepended or appended to the sentence being translated (Sennrich et al., 2016a). It has proved effective for adapting NMT to politeness (Sennrich et al., 2016a), and has since been used for other types of context: sentence length (Takeno et al., 2017), topic (Jehl and Riezler, 2017) and speaker gender (Elaraby et al., 2018; Vanmassenhove et al., 2018). However little is known about the effectiveness of the approach when including *multiple* contextual values. Since the strategy relies on treating contextual values as linguistic tokens (part of the input sequence), adding several such values is not intuitive: extra-linguistic values are not linguistic tokens, multiple values have no inherent order between them, and it is unclear whether individual values would be well exploited when several values are integrated.[1]

In this article we test the effectiveness of the pseudo-token approach for four types of extra-linguistic context: *speaker gender*, *speaker age*, *film genre* and *film year*. We test whether the addition of multiple tokens and the order in which they are presented have an impact on how well context is exploited. We compare this strategy to a multi-encoder approach (Libovický and Helcl, 2017; Bawden et al., 2018). Within such models, extra-linguistic contextual tokens are encoded separately from the source sentence, and we propose a novel version of the model in which the sequential component of the additional encoder is removed

[1] Although the use of an attention mechanism may somewhat alleviate worries about token order, adding extra tokens will have an impact on the sentence length and the influence of the tokens on the encoded input words (within the recurrent encoder)

(to accommdate the fact that the tokens are not sequential). We report results for English-to-French translation of film subtitles, evaluating using both BLEU (Papineni et al., 2002) and a targeted evaluation of adaptation to speaker gender.

2 Contextual strategies

Pseudo-token approach As mentioned above, this approach consists in prepending contextual values as tokens to the sentence to be translated, as in (2), with the aim of biasing the translation. We test two versions of this strategy, where the token is either added only to the source sentence (*src*) as in (2) or where the model is also trained to produce the token in the translation (*src+trg*) as in (3). This translated token is then removed in post-processing. This has previously shown to give gains for integrating linguistic context (Bawden et al., 2018).

(2) FEM I am happy. → Je suis contente.
MASC I am happy. → Je suis content.

(3) FEM I am happy. → FEM Je suis contente.
MASC I am happy. → MASC Je suis content.

We test the approach with multiple contextual values, as in (4). Note that the order of the tokens is not intrinsic and must be decided in advance.

(4) FEM DRAMA ADULT I am happy. → Je suis contente.
MASC HORROR TEEN I am happy. → Je suis content.

Multi-encoder approach The alternative strategy we test is to encode the extra-linguistic tokens using an additional encoder and attention mechanism and then to combine the two representations using a hierarchical attention mechanism (Libovický and Helcl, 2017). The potential advantage of this approach is that the model may learn to encode the contextual factors differently from the source sentence, accounting for the different nature of the information. As with the pseudo-token approach, we test this strategy in the *src+trg* scenario, also learning to decode the context given as input, as shown in (5).

(5) Input 1: FEM DRAMA ADULT 90S
Input 2: I am happy.
Output: FEM DRAMA ADULT 90S Je suis contente.

We test two types of multi-encoder model: (i) the additional encoder is an RNN encoder analogue to the original encoder, (ii) the recurrent unit is removed from the additional encoder and the attention mechanism applies directly to projected word embeddings. In this second model, the order in which tokens are input has no effect.

3 Evaluation of MT adaptation to speaker gender

Objectively evaluating the impact of film genre/year and speaker age is tricky; they have no clearly identifiable and systematic impact. We therefore choose to provide a targeted evaluation of the models' capacity to ensure suitable agreement with speaker gender, which does have an explicit impact on English-to-French translation (as in (1)). We evaluate models on their capacity to translate using gender agreement that matches the speaker gender. The influence of the other features is evaluated indirectly by evaluating how speaker gender adaptation is influenced by their presence and the order in which they are input.

Automatic detection of gender markings We developed a script to automatically detect French sentences containing the following agreement types, detecting for each sentence the supposed gender of the speaker: male, female or underspecified gender:[2]

1. **Adjectival agreement**
   Je suis content(e) "I am happy"
2. **Nominal agreement**
   Je suis votre voisin(e) "I am your neighbour"
3. **Past participle with auxiliary *être***
   Je suis allé(e) "I went"
4. **Past participle with auxiliary *avoir* and preceding direct object**
   Il m'a grondé(e) "He told me off"

We use manually identified patterns and a morphological lexicon, the Le*fff* (Sagot, 2010), to detect gender-marked sentences and the associated speaker gender. A manual evaluation on 500 randomly selected sentences showed that 98% of automatically detected genders were correct (i.e. the correct gender was identified).

Evaluation of gender-marked sentences We use this detection method to identify the subset of an MT model's translations that are gender-marked and to evaluate the model according to the percentage of this subset that match a specified speaker gender. For each model we translate the same test set (described in Sec. 4) twice, specifying that all speakers are (i) male and (ii) female, and calculate the percentage of gender-marked sentences that match the assigned gender.

[2] *Underspecified* refers to patterns such as the following for which an invariable form is used. E.g. *Je suis confortable* 'I am comfortable' is correct for both male and female speakers.

4 Data

Training context-adapted models requires having data annotated for contextual values. We train and test our models on the English-French portion of OpenSubtitles2016[3] (Lison and Tiedemann, 2016), for which English film transcripts exist online in the IMSDB database (hereafter referred to as *context-enriched* data). The pre-trained model is trained on those films that are not included in this subset.

Film genre and film year OpenSubtitles2016 includes possible film genres for many films and all films' release years. A film can be associated with several genres, in which case we use them all. We bucket year values into decades (e.g. 80s, 90s).

Film speaker and film age The subtitles do not contain speaker information, necessary to determine speaker gender and age. As in (Lison and Meena, 2016; Wang et al., 2016; van der Wees et al., 2016), we automatically align IMSDB film transcripts (which contain character names for each turn) to the English side of the subtitle corpus and transfer the annotation across to the subtitles. We use the Champollion sentence aligner (Ma, 2006) to align subtitles and transcripts. We then automatically determine the actor playing each character in the film, their gender and their age[4] from film information in the IMDB database. We add more gender values by applying our gender detection method (Sec. 3) to the French side of the context-enriched parallel data. We also correct any speaker genders based on the French gender markings where necessary.[5]

Train/dev/test split We only use films that are at least 70% aligned with the film transcripts. We choose the best aligned films for the dev (3000 sentences) and test (50 000 sentences) sets and the remainder of films for the train set (1,696,040 sentences). Data is pre-processed using the Moses tokeniser (Koehn et al., 2007) and split into subword units using BPE (Sennrich et al., 2016b). Full details and statistics are provided in the Appendix.

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

[4] We approximate the age of the character by taking the age of the speaker at the time the film was released and group ages into six categories: *infant*, *child*, *teen*, *20-something*, *adult* and *older adult*.

[5] There could be multiple reasons for the speaker gender not to match the French gender markings: alignment error or (and quite commonly) an agreement mistake in the French translation.

5 Experimental setup

We train all models using an encoder-decoder (and a multi-encoder-decoder) model implemented in Nematus (Sennrich et al., 2017), using the hyperparameters indicated in the Appendix.

Pre-training and fine-tuning We first pre-train a state-of-the-art non-contextual model for each of the approaches used: (i) a single-encoder baseline for the pseudo-token approach and (ii) a multi-encoder baseline in which the additional encoder is initialised with four dummy values (tokens indicating an empty value for each feature).These pre-trained models are trained on OpenSubtitles2016 data that was not included in the enriched data (24M training sentences). The training of the models is then continued with the above-described context-enriched data (Sec. 4) to fine-tune it to exploit the contextual values. We test using an ensemble of three models produced during training.

Baseline model To produce a fair baseline model, we also continue the training of the pre-trained model on the same fine-tuning data, but without any contextual values included.

6 Results

We evaluate models using both BLEU and our above-described targeted evaluation. Tab. 1 shows the impact of individual features on BLEU and Tab. 2 gives results for multiple features. The order in which the features appear correspond to the order in which they were input to the model. The first-, second- and third-best results in each table are indicated in increasingly light shades of green. BLEU scores that are statistically better than that of the baseline model are indicated (* for $p < 0.05$ and ** for $p < 0.01$).

Impact of individual features We begin by looking at the impact of each features individually on the BLEU score as calculated on the test set using gold values. We recall that *src* means that the context token is added only to the source sentence and *src+trg* that the model is also trained to reproduce the token on the target side.

As shown in Tab. 1, learning to translate the token gives systematic gains in the BLEU score, confirming results seen for linguistic context by Bawden et al. (2018). The highest score is seen for speaker gender, following by film year. Just including the token in the source sentence does not

yield significantly better scores than the baseline model (continuing model training on the unannotated data), suggesting that these gains are not necessarily only due to better exploitation of context. This is confirmed by control experiments, which showed that the BLEU scores shows similar increases both when the contextual tokens are randomly shuffled and when the exact same token is used for each sentence (last two rows of Tab. 1). This confirms that BLEU is not at all suited to evaluate how well a model is adapting to context and is subject to length-based biases.

	BLEU	
	src	*src+trg*
gender	30.35	**30.67****
age	30.27	30.44*
genre	30.25	30.45*
year	30.42	**30.65****
pre-trained	30.17	
baseline	30.31	
same-gender-all	30.27	**30.56****
shuffled-gender	30.20	30.46*

Table 1: BLEU scores on our test set for each of the features used as pseudo-tokens. Pre-trained and baseline models are included for comparison, as are two control models where the same token is used for all sentences or the genders are randomly shuffled.

Impact of multiple tokens Tab. 2 shows that contrary to our expectations, adding multiple pseudo-tokens (rows 1-5) does not seem to have a negative impact on how well context is exploited, as shown by very little fluctuation in the models' scores. In fact the highest scoring model in terms of percentage matching includes 3 features (row 4) and the highest according to BLEU score includes all 4 features (row 5). All contextual models show a very high capacity to control for speaker gender, especially compared to the pre-trained and baseline models, which are not gender-aware.

Impact of order Similarly, altering the relative order of tokens (rows 2 and 3) does not appear to have a major impact on the way in which context is exploited, both in terms of our targeted evaluation and BLEU score. But this would need to be confirmed by testing with the addition of a larger number of features and combinations.

Multi-encoder approach Our multi-encoder models appear to be able to translate using more compatible gender marking, with higher percentages for both male and female speaker than the pseudo-token approaches, particularly for the non-sequential model. However, the overall quality of the models is significantly lower (by more than 2 BLEU points). This is most probably due to the pre-training of these models: since null tokens were used in the pre-training step for the additional encoder, it is likely that this provides an inadequate initialisation of the additional encoder's parameters. This would have to be remedied in the future if these models are to be used in real settings. However, it does show the importance of looking at both a targeted evaluation and overall translation quality.

Features	%match male	%match female	BLEU gold
Pre-trained and baseline models (no features)			
pre-trained	77.5	19.5	30.17
baseline	65.8	41.1	30.31
Pseudo-tokens (src+trg)			
1: gender	95.9	91.4	**30.67****
2: **gender**-genre	95.5	91.4	**30.71****
3: genre-**gender**	96.1	90.8	30.61
4: **gender**-year-genre	**96.5**	**91.6**	30.52**
5: **gender**-age-year-genre	96.0	91.1	**30.81****
Multi-encoder (src+trg) (gender-age-year-genre)			
6: Multi-enc.	**96.3**	**92.5**	28.26
7: Multi-enc. (non-seq)	**97.3**	**93.2**	28.63

Table 2: Model performance for gender adaptation (as % of gender-marked sentences that match each of the specified genders) (cf. Sec. 3) and BLEU scores (using gold feature labels).

7 Conclusion and future work

We have compared two approaches for the integration of multiple extra-linguistic factors into NMT. The first approach, encoding context as pseudo-tokens, appears to extend well to the use of multiple features, producing high scores both in terms of BLEU and agreement with speaker gender. It would be interesting to test if our conclusions hold with a larger number of features: at what point does an increased sentence length have a detrimental impact, and how would this affect the encoding of many multiple context values? The second approach, encoding context using an additional encoder, works well for speaker adaptation too, but suffers in terms of translation quality. Future work will look into better methods of pre-training the additional encoder.

References

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating Discourse Phenomena in Neural Machine Translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics; Human Language Technologies*, NAACL-HLT'18, pages 1304–1313, New Orleans, Louisiana, USA.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM Model 2. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT'13, pages 644–648, Denver, Colorado, USA.

Mostafa Elaraby, Ahmed Y. Tawfik, Mahmoud Khaled, Hany Hassan, and Aly Osama. 2018. Gender Aware Spoken Language Translation Applied to English-Arabic. In *Proceedings of the 2nd International Conference on Natural Language and Speech Processing*, ICNLSP'18, Algiers, Algeria.

George Foster and Roland Kuhn. 2007. Mixture-Model Adaptation for SMT. In *Proceedings of the 2nd Workshop on Statistical Machine Translation*, WMT'07, pages 128–135, Prague, Czech Republic.

Laura Jehl and Stefan Riezler. 2017. Document Information as Side Constraints for Improved Neural Patent Translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas*, AMTA'18, pages 1–12, Boston, Massachusetts, USA.

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

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

Pierre Lison and Raveesh Meena. 2016. Automatic turn segmentation for movie and TV subtitles. In *Proceedings of the 2016 Spoken Language Technology Workshop*, pages 245–252, San Diego, California, USA.

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

Xiaoyi Ma. 2006. Champollion: A robust parallel text sentence aligner. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, LREC'06, pages 489–492, Genoa, Italy.

Paul Michel and Graham Neubig. 2018. Extreme Adaptation for Personalized Neural Machine Translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, ACL'18, Melbourne, Australia.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ACL'02, pages 311–318, Philadelphia, Pennsylvania, USA.

Benoît Sagot. 2010. The Le*fff*, a Freely Available and Large-coverage Morphological and Syntactic Lexicon for French. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, LREC'10, pages 2744–2751, Valletta, Malta.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Controlling Politeness in Neural Machine Translation via Side Constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT'16, pages 35–40, San Diego, California, USA.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL'16, pages 1715–1725, Berlin, Germany.

Shunsuke Takeno, Masaaki Nagata, and Kazuhide Yamamoto. 2017. Controlling Target Features in Neural Machine Translation via Prefix Constraints. In *Proceedings of the 4th Workshop on Asian Translation*, AFNLP'17, pages 55–63, Taipei, Taiwan.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting Gender Right in Neural Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP'18, pages 3003–3008, Brussels, Belgium.

Longyue Wang, Xiaojun Zhang, Zhaopeng Tuy, Andy Way, and Qun Liu. 2016. Automatic construction

of discourse corpora for dialogue translation. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, LREC'16, pages 2748–2754, Portorož, Slovenia.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2016. Measuring the Effect of Conversational Aspects on Machine Translation Quality. In *Proceedings of the 26th International Conference on Computational Linguistics*, COLING'16, pages 2571–2581, Osaka, Japan.

Appendices

A Training setup

Data pre-processing:

- Tokenisation, cleaning (length of 1-80) and true-casing with the Moses toolkit (Koehn et al., 2007)
- Subword segmentation with BPE (Sennrich et al., 2016b): 90,000 joint operations, and threshold=50

Hyper-parameters for all models:

- Embedding layer dimension=512, hidden layer dimension=1024, batch size=80, tied decoder embeddings and layer normalisation, maximum sentence length=50

For pre-trained models:

- Filtering out of parallel sentences in which fewer than 80% of tokens are aligned (after running FastAlign (Dyer et al., 2013))
- Training continued until convergence
- Model with the best BLEU score on the dev set is used to continue fine-trained on the annotated data

For fine-trained models:

- Final model is an ensemble of 3 models produced after 30k, 60k and 90k updates.

B Data statistics

	English			French	
	#sents	#tokens	#tokens/#sents	#tokens	#tokens/#sents
Unannotated Opensubtitles2016 data (for pre-training)					
pre-train	24,140,225	174,593,562	7.2	175,432,942	7.3
Annotated Opensubtitles2016 data (for adapted training)					
train	1,696,040	13,462,830	7.9	13,268,188	7.8
dev	3,000	24,131	8.0	23,524	7.8
test	50,000	399864	8.0	394212	7.9

Table 3: Corpus statistics for each dataset, including the unannotated data used for pre-training (pre-train) and the annotated data described in this section (train, dev and test). Token numbers are calculated on pre-processed sentences (to which subword segmentation has been applied)

C Distribution of gender values

	#subtitles				
	MASC SG	FEM SG	MASC PL	FEM PL	MASC/FEM PL
train	815,951 (48.1%)	423,249 (25.0%)	1,416 (0.0%)	2,485 (0.1%)	19,657 (1.2%)
dev	1,143 (38.1%)	587 (19.6%)	0 (0.0%)	0 (0.0%)	29 (1.0%)
test	24,440 (48.9%)	13,064 (25.1%)	51 (0.1%)	96 (0.2%)	958 (1.9%)

Table 4: The distribution of the gender labels in the train, dev and test sets. Percentages of the total number of subtitles per dataset are given in brackets. Note that speaker gender annotations are not available for all subtitles.

		#sents annotated for...			
Set	#sents	film genre	film year	speaker gender	speaker age
train	1,696,040	508,928	1,696,040	1,262,758	241,439
dev	3,000	0	3,000	1,759	0
test	50,000	15,004	50,000	38,609	7,283

Table 5: Corpus statistics per dataset: numbers of sentences and numbers of annotated sentences for each type of extra-linguistic context.

D Multi-encoder models

Figure 1 gives a detailed illustration of the non-sequential multi-encoder model. Whereas the first model (sequential) uses an additional encoder of an identical form to the main sentence encoder, this non-sequential version has a lighter encoder for the context and the attention mechanism applies over projected embeddings.

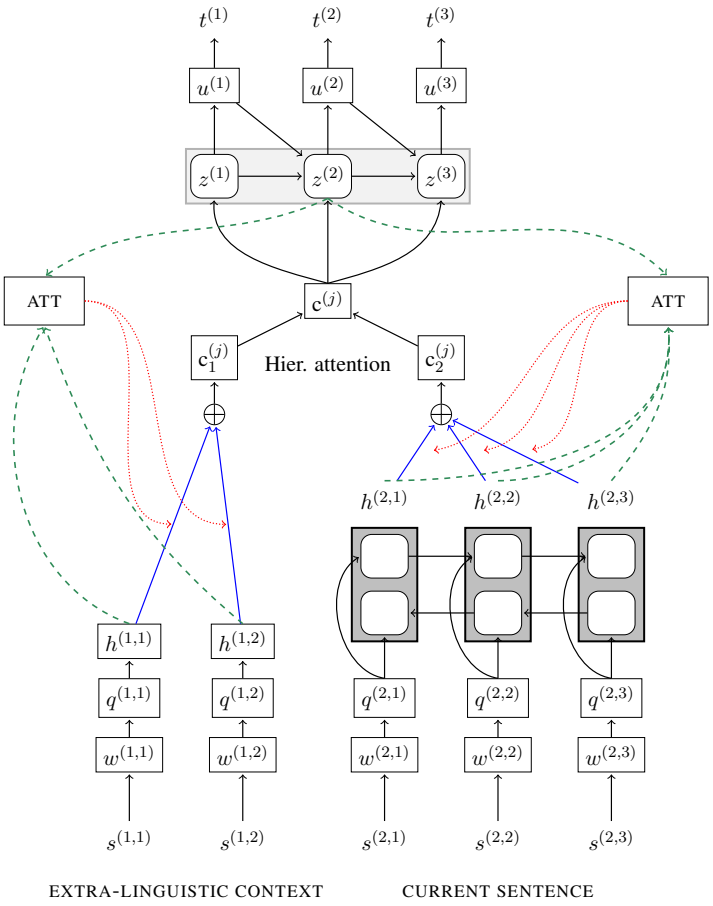

Figure 1: Our proposition for a multi-encoder model in which the sequential element of the first encoder (for the encoding of contextual tokens) is removed. The attention mechanism applies to projected versions of the token embeddings in order to calculate $c_1^{(j)}$.