Supplementary Material: Document-level Machine Translation For Scientific Texts

Ziqian Peng [1]
François Yvon [2]

This material consists of some results showing that COMET cannot penalize incomplete short translations, and supplementary tables to the main document, especially some document-level scores and complementary results of the evaluation with length constraint.

Table 3 and 4 reported the document-level score evaluated with BLEU and COMET respectively, for sentence-level models, full-context models, and architectures with context masks.

Table 5 compares various values of the attention factor applied to all, future or past source contexts during cross-attention when translating `THE_doc` test set, measured at document-level. Table 6 contains the same evaluation on `TAL_doc` test set.

We also reported the BLEU score of documents grouped by their range of length in sentences or in tokens, on `THE_doc` test set (in Table 7) and `TAL_doc` (Table 8) respectively.

In table 9, we compare the translation quality of the first 5 sentences from documents in `THE` and `TAL` test sets using `FTsent` and `FTdoc`. Results show that translating at document-level is the best for `TAL`. While for `THE`, `FTsent` still has better BLEU score than `FTdoc`, the gap between them is much less significant than that between the translation quality of the whole documents, with 2 points of BLEU score instead of around 7 points. Table 10 reports the relevant document-level scores.

At the end, Table 11 and 12 provides more details about the quality of translated sentences grouped by their location in documents.

1. Evaluation of incomplete translations with COMET

Noticing that COMET can highly score an incomplete but accurate translation, we tested this phenomenon with our document-level test sets `THE_doc` and `TAL_doc`.

To begin, the first one to three sentences from each reference document are extracted as translation hypothesis, denoted as $crop1$, $crop2$, and $crop3$. We evaluated them with BLEU and COMET. As shown in Table 1, COMET cannot penalize the incomplete translations, so as to give scores of around 0.6 to $crop1$ and around 0.7 to $crop2$ for `THE_doc`.

Table 1. Evaluation of the first one to three reference sentences (crop1, 2, 3 resp.) with BLEU/COMET

	crop1	crop2	crop3
THE_doc	0.0 / 0.576	1.8 / 0.683	9.7 / 0.749
TAL_doc	4.8 / 0.643	34.9 / 0.771	64.2 / 0.837

Table 2. Evaluation of the first one to three reference sentences of `THE_doc`, which are completed with random words from the whole reference test set (N_{corpus}, first and last row), or from the current cropped document ($N_{document}$, second row), with BLEU/COMET, until the reference length (the first 2 rows) or at most 20 tokens (the last row)

.

	noise1	noise2	noise3
N_{corpus}	14.2 / 0.337	24.4 / 0.380	34.4 / 0.455
$N_{document}$	13.2 / 0.306	23.8 / 0.352	34.2 / 0.423
$N_{corpus20}$	0.6 / 0.405	5.4 / 0.512	15.6 / 0.604

We continued the test with `THE_doc`, which contains longer documents on average. We introduced random words to complete the first one to three sentences from the reference until the reference length for each document. These types of hypotheses are denoted as $noise1$, $noise2$, and $noise3$. We randomly sampled words from the whole reference (N_{corpus}), or the cropped documents ($N_{document}$). For comparison, we created $N_{corpus20}$ by sampling at most 20 random tokens from the whole reference.

Table2 shows that scores of COMET are reduced to around half than before, which demonstrates that COMET is more sensible to the precision than the completion of translations.

2. Tables

Please check the following pages.

Table 3. BLEU score at document-level. BP denotes the brevity penalty. The <sep> tags are always excluded for evaluation.

Score	Model	TED	IWSLT2023	THE_sent2doc	THE_doc	TAL_sent2doc	TAL_doc
Score	baseline	27.8	48.6	43.9	-	34.4	-
	FTsent	28.5	47.1	45.4	-	35.9	-
	FTdoc	24.0	41.2	43.1	38.5	34.2	34.6
	FTdoc_MR	21.2	40.2	41.8	37.2	33.8	33.4
	FTdoc_maskAll	24.4	42.8	43.7	37.3	35.2	33.7
	FTdoc_maskFuture	26.5	44.7	44.4	30.9	35.5	25.3
	FTdoc_maskPast	25.0	43.5	43.5	37.9	35.0	34.5
BP	baseline	0.953	0.967	0.999	-	0.98	-
	FTsent	0.977	0.973	1.0	-	0.985	-
	FTdoc_sep	0.984	0.984	1.0	0.959	0.983	0.98
	FTdoc_MR	0.977	0.97	1.0	0.953	0.986	0.982
	FTdoc_maskAll	0.991	0.976	1.0	0.998	0.981	0.948
	FTdoc_maskFuture	0.975	0.967	1.0	0.871	0.984	0.721
	FTdoc_maskPast	1.0	0.979	1.0	0.976	0.988	0.979

Table 4. Evaluation with COMET at document-level. The <sep> tags are all excluded.

Model	THE_doc	THE_sent2doc	TAL_doc	TAL_sent2doc
baseline	0.643	0.855	0.647	0.813
FTsent	0.713	0.860	0.670	0.816
FTdoc	0.833	0.852	0.810	0.809
FTdoc_MR	0.831	0.844	0.805	0.806
FTdoc_maskAll	0.822	0.855	0.804	0.814
FTdoc_maskFuture	0.790	0.856	0.744	0.814
FTdoc_maskPast	0.832	0.855	0.812	0.813

Table 5. Document-level BLEU score of Transformer with attention factor on THE_doc test set

Score	**FTdoc_factor**	**All**	**Future**	**Past**
BLEU	0.1	37.4	33.5	38.4
	0.2	37.4	33.1	38.0
	0.3	36.8	33.9	38.6
	0.4	36.1	33.8	38.1
	0.5	37.1	35.3	38.3
	0.6	37.8	36.8	38.3
	0.7	37.7	37.7	38.6
	0.8	38.1	37.9	38.3
	0.9	37.9	38.7	38.7
BP	0.1	0.979	0.930	0.980
	0.2	0.981	0.921	0.978
	0.3	0.949	0.889	0.996
	0.4	0.931	0.907	0.979
	0.5	1.000	0.956	0.973
	0.6	0.971	0.962	0.977
	0.7	1.000	0.990	0.990
	0.8	0.977	0.991	0.964
	0.9	0.975	0.991	0.990

Table 6. Document-level BLEU score of Transformer with attention factor on TAL_doc test set

Score	**FTdoc_factor**	**ALL**	**Future**	**Past**
BLEU	0.1	32.1	25.4	34.3
	0.2	32.5	26.1	34.4
	0.3	32.6	26.2	34.0
	0.4	33.0	29.0	33.9
	0.5	34.3	33.9	34.2
	0.6	34.1	34.4	34.3
	0.7	34.2	34.8	33.9
	0.8	35.0	34.8	34.4
	0.9	34.7	34.3	34.1
BP	0.1	0.913	0.776	0.988
	0.2	0.939	0.774	0.982
	0.3	0.931	0.737	0.985
	0.4	0.925	0.792	0.979
	0.5	0.976	0.939	0.986
	0.6	0.980	0.968	0.983
	0.7	0.984	0.984	0.982
	0.8	0.986	0.985	0.982
	0.9	0.985	0.989	0.986

Table 7. Evaluation of THE_doc documents translated using FTdoc, grouped by their ranges of length in tokens (top) and in sentences (bottom).

Length (in tokens)	Bleu	BP	Average length	Count (document)
0-100	41.7	1.000	69.0	5
100-200	43.3	0.988	161.5	24
200-300	51.9	1.000	250.8	32
300-400	41.9	0.974	353.2	16
400-max	38.7	0.914	499.5	23

Length (in sentences)	BLEU	BP	Average length	Count (document)
0-5	35.7	1.000	104.7	10
5-10	40.5	0.997	224.44	43
10-15	39.9	0.959	341.0	32
15-20	40.4	0.933	470.5	8
20-25	28.4	0.866	573.14	7

Table 8. Evaluation of TAL_doc documents translated using FTdoc, grouped by their ranges of length in tokens (top) and in sentences (bottom).

Length (in tokens)	BLEU	BP	Average length	Count (document)
0-100	34.1	0.964	69.64	91
100-200	41.2	0.986	140.49	152
200-300	40.5	0.960	237.0	2
300-400	34.9	1.000	325.0	1

Length (in sentences)	BLEU	BP	Average length	Count (document)
0-5	32.5	0.976	95.55	151
5-10	36.7	0.983	148.03	95

Table 9. Evaluation of the **first 5 sentences** in THE_doc, THE_sent,TAL_doc,TAL_sent with BLEU at sent-level

	THE_sent	THE_doc2sent	TAL_sent	TAL_doc2sent
FTdoc	39.9	**39.8**	32.5	**35**
FTsent	41.9	-	34.2	-

Table 10. Evaluation of the **first 5 sentences** in THE_doc, THE_sent,TAL_doc,TAL_sent with BLEU at document-level

	THE_sent2doc	THE_doc	TAL_sent2doc	TAL_doc
FTdoc	41.4	**41.8**	33.9	**36.7**
FTsent	43.4	-	35.6	-

Table 11. Evaluation by sentence position with THE_doc using sacreBLEU

Sentence position	FTdoc	FTdoc_MR	FTdoc_maskAll	FTdoc_maskFuture	FTdoc_maskPast
0	42.9	39.5	42.7	27.4	41.7
1	35.2	34.1	38.6	37.1	36.8
2	39.2	36.1	39.2	39.6	38.8
3	40.0	36.4	40.7	38.3	38.9
4	38.1	36.5	40.9	40.3	39.9
5	38.1	36.7	40.9	37.4	40.3
6	40.8	39.1	33.2	37.6	38.1
7	31.4	33.7	28.2	28.3	33.7
8	31.0	32.3	34.4	28.5	32.3
9	31.3	33.2	29.7	23.2	32.3
10	25.0	28.0	24.1	21.2	24.1
11	22.4	20.7	21.1	17.6	22.9
12	12.8	14.8	11.5	16.3	17.0
13	17.4	17.8	21.5	13.7	24.3
14	10.5	11.5	13.6	10.5	15.2
15	8.1	11.7	13.2	8.5	15.4
16	7.1	8.0	8.5	2.7	9.2
17	5.8	4.5	19.3	5.5	14.2
18	6.5	2.3	11.6	9.2	8.9
19	2.8	5.5	13.2	6.6	10.8

Table 12. Evaluation by sentence position with TAL_doc using sacreBLEU

Sentence position	FTdoc	FTdoc_MR	FTdoc_maskAll	FTdoc_maskFuture	FTdoc_maskPast
0	34.2	33.1	34.2	30.5	34.0
1	30.5	29.5	31.2	31.8	30.8
2	34.6	33.6	35.6	36.3	34.1
3	30.9	28.6	29.3	31.9	30.1
4	36.9	36.2	37.1	34.6	36.8
5	33.8	33.8	36.5	32.0	35.3
6	40.6	39.9	42.3	37.7	41.7
7	19.3	18.2	22.6	20.3	22.3