A User Study of the Incremental Learning in NMT

Miguel Domingo[1], Mercedes García-Martínez[2], Álvaro Peris[3],
Alexandre Helle[2], Amando Estela[2], Laurent Bié[2],
Francisco Casacuberta[1], Manuel Herranz[2]

{midobal, fcn}@prhlt.upv.es,
{m.garcia, a.helle, a.estela, l.bie, m.herranz}@pangeanic.com,
lvapeab@gmail.com

[1]PRHLT Research Center - Universitat Politècnica de València
[2]Pangeanic / B.I Europa - PangeaMT Technologies Division
[3]Independent Researcher

EAMT 2020

Online conference, November 3, 2020

Contributions

- User study on adaptive NMT under an online learning paradigm.

- Conducted with the help of professional post-editors.

- Human evaluation to verify the quality of the post-editions generated in the user study.

- Study of the sporadic appearance of made-up words.

Corpora

Small medico-technical (description of medical equipment) corpus from our production scenario. Training data comes from WMT, UFAL and a technological corpus.

Corpus	#Sentences	# Tokens		# Types		Average length	
		En	Es	En	Es	En	Es
Training	23.4M	702M	786M	1.8M	1.9M	30.0	33.6
Document 1	150	1.7K	-	618	-	11.3	-
Document 2	150	2.6K	-	752	-	17.3	-

Table: Corpora statistics in terms of number of sentences, number of tokens, number of types (vocabulary size) and average sentence length. K denotes thousands and M, millions.

Post-editors

User	Sex	Age	Professional experience
User 1	Male	24	1.5 years
User 2	Female	25	5 years
User 3	Female	30	5 years
User 4	Female	24	1 month
User 5	Female	22	1 year
User 6	Male	48	22 years

Table: Information about the participants.

User	Document 1	Document 2
User 1	Static	Adaptive
User 2	Adaptive	Static
User 3	Static	Adaptive
User 4	Adaptive	Static
User 5	Static	Adaptive
User 6	Adaptive	Static

Table: Distribution of users, document sets and scenarios. All users conducted first the experiment which involved post-editing document 1 and then document 2.

Translation Quality

Test	System	hTER [↓]	hBLEU [↑]
Document 1	Static	39.5	47.9
	Adaptive	32.8†	55.9†
Document 2	Static	36.2	42.9
	Adaptive	34.3†	50.5†

Table: Results of the user experiments, in terms of translation quality. *Adaptive system* refers to post-editing in an environment with online learning. † indicates statistically significant differences between the static and the adaptive systems.

User Productivity

| User	System	Time [↓]	Words per hour [↑]
User 1	Static	37.9	1685
	Adaptive	33.0[†]	1935[†]
User 2	Static	30.5	2091
	Adaptive	30.4	2097[†]
User 3	Static	38.0	1678
	Adaptive	27.0[†]	2364[†]
User 4	Static	37.5	1701
	Adaptive	47.4[†]	1346[†]
User 5	Static	80.2	795
	Adaptive	46.7[†]	1367[†]
User 6	Static	53.7	1188
	Adaptive	49.7[†]	1284[†]

Table: *Static system* stands for conventional post-editing, without adaptation. *Adaptive system* refers to post-editing in an environment with online learning. Users 4 to 6 has less experience, in this particular domain, than users 1 to 3.

Quality of the Post-edits

A human evaluation was conducted with the help of two professional translators.
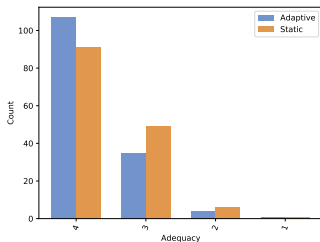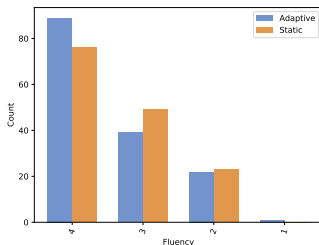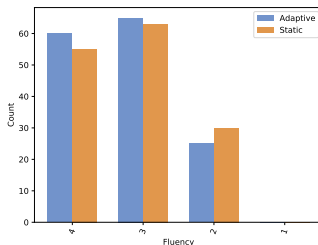

(a) Document 1.
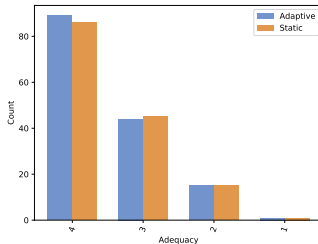
(b) Document 2.

Figure: Sentence-level adequacy scores. Count values are the average between both evaluators.
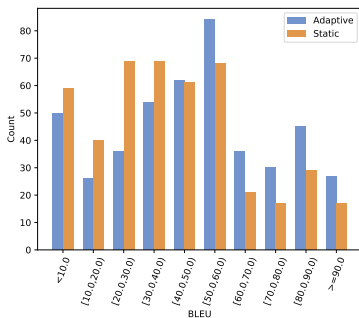
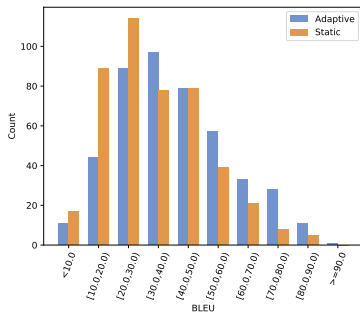# Quality of the Post-edits

(a) Document 1.

(b) Document 2.

Figure: Sentence-level fluency scores. Count values are the average between both evaluators.

Static vs Adaptive

(a) Document 1.

(b) Document 2.

Figure: Histogram of sentence-level BLEU scores. The counts are distributed in buckets of range 10.

Static vs Adaptive

Phenomenon	System	Example
Acronyms	Source	QSE Number
	Post-edit	Número **de ESC**
	Adaptive	Número **de ESC**
	Static	Número QSE
Entities	Source	Show the R Series ALS
	Post-edit	Mostrar **la serie R** ALS
	Adaptive	Mostrar **la serie R** ALS
	Static	Mostrar el R Series ALS
Terminology	Source	There are several steps involved with sidestream end tidal CO_2 setup.
	Post-edit	La configuración del CO_2 **espiratorio final** de flujo lateral se realiza en varios pasos.
	Adaptive	Hay varias etapas de la configuración del CO_2 **espiratorio final** del ajuste.
	Static	Hay varias etapas que involucran la configuración del CO_2 maremoto del CO_2 maremoto

Figure: Examples of the *n*-gram differences between adaptive and static systems. In **boldface** we highlight the differences introduced by adaptive systems.

Made-up Words

User	System	Words
User 1	Static	3
	Adaptive	6
User 2	Static	8
	Adaptive	5
User 3	Static	3
	Adaptive	17
User 4	Static	8
	Adaptive	5
User 5	Static	3
	Adaptive	14
User 6	Static	8
	Adaptive	4

Table: Total made-up words generated per user.

Document	System	Words
Document 1	Static	5
	Adaptive	4
Document 2	Static	8
	Adaptive	12

Table: Average of made-up words generated per document for all users.

Made-up Words

1. La zona verde es para **pacio**.

2. Roll al paciente a su lado, y luego rodar el electrodo hacia la espalda del paciente a la izquierda de su columna y debajo de la **escaga**.

3. Presione la tecla del **softón**.

4. Sin embargo, el metrónomo **absolvido** si las compresiones son inferiores a las directrices.

5. Que el dispositivo puede hacer un choque de prueba de 30 **jojuelas**.

Figure: Example of made-up words (in **bold**) from the static system. The first word should have been *estimulación*, the second one *omóplato*, the third one *RCP*, the fourth one *sonará* and the fifth one *julios*.

Made-up Words

1. Al mover el Selector de modo a Pacer se activará la puerta del **pidante** para abrir.

2. Coloque el sensor con el adaptador instalado fuera de todas las fuentes de CO_2 (incluidos los **válvulos** de aire de respiración y respiratorio) exhalado.

3. Las **marcapasas** de estimulación deben producirse aproximadamente cada centímetro en la tira.

4. El conector de autoprueba funciona solo cuando el envase del electrodo es **inabierto** y conectado a la serie R Series.

5. Para aplicar los electrodos OneStep, introduzca primero el electrodo trasero para evitar la **herración** del electrodo delantero.

Figure: Example of made-up words (in **bold**) from the adaptive systems. The first word should have been *marcapasos*, the second one *válvulas*, the third one *marcadores*, the fourth one *cerrado* and the fifth one *deformación*.

Conclusions

- Significant increase of the user's productivity (in terms of post-editing time and number of words generated).

- The users were pleased with the system.

- All post-edits generated during the study were of high-quality.

- The sporadic appearance of made-up words seems to be related with the increase of out-of-vocabulary words.

- Sometimes these made-up words are very similar, in morphological terms, to the correct words.