How Much Does Tokenization Affect Neural Machine Translation?

Authors: Miguel Domingo[1], Mercedes García-Martínez[2], Alexandre Helle[2], Francisco Casacuberta[1], Manuel Herranz[2]

[1] Pattern Recognition and Human Language Technology Research Center
Universitat Politècnica de València
{midobal, fcn}@prhlt.upv.es

[2] Pangeanic / B.I Europa
PangeaMT Technologies Division
{m.garcia, a.helle, m.herranz}@pangeanic.com

Introduction

- In this work, we use tokenization referring to separating punctuation and splitting tokens into words or subwords.
- Tokenizing words has proven to be helpful to reduce vocabulary and increase the number of examples of each word.
- It is extremely important for languages in which there is no separation between words.
- In this study, we aim to find the impact of tokenization on the quality of the final translation produced using neural machine translation.

Neural Machine Translation

Given a source sentence $x_1^J = x_1, \ldots, x_J$ of length J, NMT aims to find the best translated sentence $\hat{y}_1^{\hat{I}} = \hat{y}_1, \ldots, \hat{y}_{\hat{I}}$ of length $\hat{I}$:

$$\hat{y}_1^{\hat{I}} = \arg\max_{I, y_1^I} Pr(y_1^I \mid x_1^J)$$

Corpora

Languages: Japanese (Ja), English (En), Russian (Ru), Chinese (Zh), German (De) and Arabic (Ar).

Specific Domain	Language				
	Ja–En	Ru–En	Zh–En	De–En	Ar–En
Computer Software - Instructions for use		X			X
Medical Equipment and Supplies	X	X	X	X	X
Consumer Electronics	X		X	X	X
Industrial Electronics		X		X	
Stores and Retail Distribution	X	X	X		
Healthcare		X			

Metrics

- **BiLingual Evaluation Understudy (BLEU)**: geometric average of the modified n-grams precision, with a penalty factor for short sentences.
- **Translation Error Rate (TER)**: word edit operations normalized by the number of words in the final translation.

Tokenizers

SentencePiece (Kudo and Richardson, 2018): an unsupervised text tokenizer and detokenizer mainly for Neural Network-based text generation systems.

OpenNMT tokenizer (Klein et al., 2017): the tokenizer included with the *OpenNMT* toolkit.

Moses tokenizer (Koehn et al., 2007): the tokenizer included with the *Moses* toolkit.

Mecab (Sim, 2013): an open source morphological analysis engine for Japanese, based on conditional random fields.

Stanford Word Segmenter (Tseng et al., 2005): a Chinese word segmenter based on conditional random fields.

Tokenization

SentencePiece
Original: *In a browser window (Internet Explorer or Firefox) browse to www.dellconnect.com.*
Segmented: *_In _a _browser _window _(Internet _Explorer _or _Firefox) _browse _to _www . dell connect . com .*

OpenNMT tokenizer
Original: *In a browser window (Internet Explorer or Firefox) browse to www.dellconnect.com.*
Segmented: *In a browser window (Internet Explorer or Firefox) browse to www . dellconnect . com .*

Moses tokenizer
Original: *In a browser window (Internet Explorer or Firefox) browse to www.dellconnect.com.*
Segmented: *In a browser window (Internet Explorer or Firefox) browse to www.dellconnect.com.*

Mecab
Original: ブラウザウィンドウ(*Internet Explorer*または*Firefox*)で、*www.dellconnect.com*にアクセスします。
Segmented: ブラウザウィンドウ (*Internet Explorer* または *Firefox*) で 、 *www . dellconnect . com* に アクセス し ます 。

Stanford Word Segmenter
Original: 到 *http://www.kace.com/trial*，然后“下 *K1000* 用版”，将的 *OVF*（放虚化格式）文件下到 *vSphere* 系。
Segmented: 到 *http : //www.kace.com/trial* ， 然后 “ 下 *K1000* 用版 ” ， 将 的 *OVF* （ 放 虚化 格 式 ） 文件 下 到 *vSphere* 系 。

Results

Language	SentencePiece		OpenNMT tokenizer		Moses tokenizer		Mecab		Stanford	
	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
Ja–En	32.0 ± 1.3	51.1 ± 1.5	29.1 ± 1.4	54.7 ± 1.4	$\mathbf{36.3 \pm 1.4}$	$\mathbf{47.5 \pm 1.3}$	$\mathbf{36.0 \pm 1.5}$	$\mathbf{48.6 \pm 1.4}$	-	-
En–Ja	26.5 ± 1.4	62.5 ± 1.9	25.0 ± 4.4	89.9 ± 4.1	33.6 ± 2.3	61.0 ± 2.5	$\mathbf{45.8 \pm 1.3}$	$\mathbf{43.7 \pm 1.3}$	-	-
Ru–En	12.9 ± 0.9	72.7 ± 1.1	11.9 ± 0.9	74.9 ± 1.3	$\mathbf{15.3 \pm 1.0}$	$\mathbf{68.6 \pm 1.2}$	-	-	-	-
En–Ru	12.2 ± 0.8	75.0 ± 1.0	11.3 ± 0.9	77.3 ± 1.1	$\mathbf{16.3 \pm 1.2}$	$\mathbf{70.4 \pm 1.6}$	-	-	-	-
Zh–En	20.5 ± 1.1	64.8 ± 1.2	23.1 ± 1.3	64.8 ± 1.3	$\mathbf{27.5 \pm 1.3}$	$\mathbf{59.8 \pm 1.2}$	-	-	$\mathbf{26.0 \pm 1.3}$	$\mathbf{59.3 \pm 1.2}$
En–Zh	17.1 ± 1.2	71.2 ± 1.2	10.4 ± 3.9	101.1 ± 3.1	21.4 ± 2.0	65.8 ± 1.7	-	-	$\mathbf{29.9 \pm 1.2}$	$\mathbf{55.6 \pm 1.2}$
De–En	21.4 ± 0.8	67.8 ± 2.1	$\mathbf{29.6 \pm 0.9}$	$\mathbf{54.2 \pm 0.9}$	$\mathbf{30.3 \pm 0.9}$	$\mathbf{52.8 \pm 0.9}$	-	-	-	-
En–De	16.1 ± 0.7	76.4 ± 2.3	$\mathbf{22.5 \pm 0.9}$	$\mathbf{65.0 \pm 1.5}$	$\mathbf{23.6 \pm 0.9}$	$\mathbf{62.9 \pm 1.0}$	-	-	-	-
Ar–En	$\mathbf{17.9 \pm 0.8}$	$\mathbf{66.9 \pm 1.3}$	14.8 ± 0.8	71.3 ± 1.1	$\mathbf{19.1 \pm 0.9}$	$\mathbf{65.4 \pm 1.9}$	-	-	-	-
En–Ar	10.1 ± 0.6	75.3 ± 1.3	9.2 ± 0.6	77.2 ± 0.9	$\mathbf{12.4 \pm 0.7}$	$\mathbf{69.8 \pm 0.9}$	-	-	-	-

Example

Example 1	
Source	Revalidation of single-pilot single-engine class ratings
Reference	verlängerung von klassenberechtigungen für einmotorige flugzeuge mit einem piloten
SentencePiece	**verlängerung** *der einzelantriebsklasse einmotorischer motorklasse*
OpenNMT tokenizer	*zur validierung der einmotorik-einzelmaschine* **mit** *einzelantrieb*
Moses tokenizer	**verlängerung von klassenberechtigungen für einmotorige flugzeuge mit einem piloten**
Example 2	
Source	Cold drawing of wire
Reference	herstellung von kaltgezogenem draht
SentencePiece	*kalt zeichnung des drahtes*
OpenNMT tokenizer	*kaltbildzeichnung*
Moses tokenizer	**herstellung von kaltgezogenem draht**

Conclusions

- Gains of up to 12 BLEU points and 15 TER points due to tokenization impact.
- Each tokenizer is more suitable for specific languages.
- *Moses tokenizer* seems to be the most suitable for European languages.

Future Work

- Evaluation of the corpus since impact using *SentencePiece*.
- Comparison of more segmentation strategies.
- Experimentation with generic data or bigger corpus size.

Acknowledgements

The research leading to these results has received funding from the Centro para el Desarrollo Tecnológico Industrial (CDTI) and the European Union through Programa Operativo de Crecimiento Inteligente (EXPEDIENT: IDI-20170964). We gratefully acknowledge the support of NVIDIA Corporation with the donation of a GPU used for part of this research.