Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3752–3760
Marseille, 11–16 May 2020

To Case or not to case: Evaluating Casing Methods for Neural Machine Translation

Thierry Etchegoyhen and Harritxu Gete Ugarte
Department of Speech and Natural Language Technologies, Vicomtech
Mikeletegi Pasalekua, 57, Donostia, Gipuzkoa, Spain
{tetchegoyhen, hgete}@vicomtech.org

Abstract

We present a comparative evaluation of casing methods for Neural Machine Translation, to help establish an optimal pre- and post-processing methodology. We trained and compared system variants on data prepared with the main casing methods available, namely translation of raw data without case normalisation, lowercasing with recasing, truecasing, case factors and inline casing. Machine translation models were prepared on WMT 2017 English-German and English-Turkish datasets, for all translation directions, and the evaluation includes reference metric results as well as a targeted analysis of case preservation accuracy. Inline casing, where case information is marked along lowercased words in the training data, proved to be the optimal approach overall in these experiments.

Keywords: Neural Machine Translation, Casing, Evaluation

1. Introduction

Neural Machine Translation (NMT) (Cho et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017) has become the dominant paradigm of machine translation (MT) research and development in recent years. As a data-driven method, where translation knowledge is induced from multilingual corpora, NMT modelling typically involves preprocessing steps that notably include tokenisation, case normalisation and word segmentation into subword units, as demonstrated for instance by the system descriptions in the annual WMT translation shared tasks (see, e.g., (Bojar et al., 2017) and references therein).

Whereas different approaches to tokenisation and subword generation have been proposed and evaluated in different studies, the optimal handling of casing has not been systematically evaluated. In this work, we evaluate the main methods to handle capitalisation, namely training without case normalisation, lowercasing with recasing, truecasing, case factors and inline casing. The latter method, which is based on inserting case tags along lowercased words and does not require any extension of NMT infrastructures is shown to obtain the best results overall.

The different approaches to casing are evaluated on two standard datasets, namely WMT 2017 for German-English and Turkish-English, in both translation directions, thus covering translation cases which differ in terms of capitalisation rules. Translation results are evaluated on the WMT test sets and on additional datasets that include newspaper titles, which feature extended capitalisation usage in English, and subtitles, which involve specific uses of capitalisation as well.

Our main contribution is thus the first systematic comparison between the main casing methods available for Neural Machine Translation, measuring the impact of the different variants and helping determine optimal data preparation pipelines for NMT system development.

The remainder of the paper is organised as follows: Section 2. describes related work in data preprocessing for machine translation; Section 3. presents the different methods to be evaluated in this work; Section 4. describes the experimental setup and results; finally, Section 5. draws conclusions from this work.

2. Related work

Pre- and post-processing of textual data have been standard steps in data-driven approaches to machine translation, such as Example-Based Machine Translation EBMT (Nagao, 1984) or Statistical Machine Translation SMT (Brown et al., 1990; Koehn, 2010). Work on the latter in particular has brought standard pipelines consisting in language-specific rule-based tokenisation, followed by full or partial case normalisation, performed with the tools made available in the Moses toolkit (Koehn et al., 2007).

Different approaches to tokenisation have been proposed over the years, to optimise data preparation for machine translation, notably via unsupervised approaches (Kudo and Richardson, 2018). This component of preprocessing has received specific attention in particular to solve segmentation issues in Asian languages (Xiao et al., 2010; Chung and Gildea, 2009).

To handle casing for machine translation, a standard approach in SMT has involved training a first system on lowercased data, followed by recasing the data with a monolingual translation system without reordering and a cased language model on the target side. Over the year, this approach has been replaced in practice by truecasing, i.e. preserving case information except for sentence-initial words, which are converted to their most frequent form according to a frequency model trained on the relevant cased monolingual data; detruecasing simply consists then in capitalising sentence initial words as a post-processing step. This approach follows (Mikheev, 1999) in converting only those instances of capitalised words in contexts where capitalisation is expected, such as the beginning of a sentence or after quotes. As shown by the system description papers of the WMT translation shared tasks, this version of truecasing has been the default approach to casing in machine translation, for both SMT and NMT.

Other methods have been proposed to improve over simple 1-gram tagging, where all words are converted to their most frequent case. Lita et al. (2003) use a trigram language model to determine optimal capitalisation tag sequences along with contextual information such as sentence-initial position as well. In (Chelba and Acero, 2006), capitalisation is performed with maximum entropy Markov models over tag sequences conditioned on the word sequences, an approach which is also shown to improve the handling of casing over 1-gram capitalisation. Wang et al. (2006) proposed a probabilistic bilingual capitalisation model based on conditional random fields, defining a series of feature functions to model capitalisation knowledge; their approach improved over a strong baseline consisting of a monolingual capitaliser based on a trigram language model. Recently, Berard et al. (2019) proposed inline casing for neural machine translation, a simple approach where case tags indicating either title case or uppercase are added next to lowercased words. The tags are handled by the encoder and decoder as additional tokens in the input, and a simple post-processing step reconstructs casing based on sequences of lowercased forms and tags in the output translation. Recent approaches have also explored using raw data directly, without case-related preprocessing, and reported similar results to those obtained with truecasing (Bawden et al., 2019).

Although several approaches are available to handle case in neural machine translation, so far no systematic comparative evaluation has been performed on the main methods currently employed in the field. In the following sections, we describe the results of such an evaluation on different translation pairs and datasets representative of case handling challenges.

3. Methods

To perform our evaluation of case handling methods for NMT, we included the approaches described below. Tokenisation was performed first with the Moses tokeniser in its default form in all cases, to avoid introducing a separate variable in the evaluation.

Baseline. As a baseline, we left the corpora in its natural casing. Although this approach could be thought of as adding spurious ambiguity, considering that sentence initial words are capitalised by default in most corpora, its use has been reported in NMT system configurations (Bawden et al., 2019) and the impact of maintaining natural casing needs to be evaluated as well in a systematic comparison between casing methods. We refer to this method as RAW in what follows.

Truecasing. For this approach, we used the truecasing script available in the Moses toolkit (*op. cit.*), by first generating a truecasing model from the monolingual training corpora then applying case conversion to words in contexts where capitalisation is expected, mainly sentence-initial position, using the Moses truecaser script.[1] Although the identification of sentence initial positions may be further adapted to cover additional cases of what may be considered as delayed sentence initial positions, we used the default version of the script to facilitate the reproduction of our results. We refer to this method as MTC, standing for Moses truecasing.

Recasing. To implement this approach, we trained monolingual SMT models on the lowercased and natural case corpora in the target language, without reordering and with a trigram language model trained with KenLM (Heafield, 2011). We refer to this method as LRC in the remainder of this paper. Although this method is seldom used in current practice, it has not been systematically evaluated against competing approaches for NMT and we included it for completeness.

Case factors. Under this term, we refer to the method of lowercasing words and concatenating an embedding vector denoting case information to the lowercased word embedding. We used the implementation provided in the Sockeye toolkit (Hieber et al., 2017), which was used to train all NMT models in our experiments. Since only source factors are supported in this toolkit,[2] we used the data with their original casing in the target language. Case factors were implemented as embeddings of dimension 8, concatenated to each input word to indicate casing. This method will be referred to as CFT in the remainder of this paper, and its inclusion aimed to evaluate the accuracy of modelling case information in the parameter space of the translation models directly.

Inline casing. As a final method we implemented inline casing (Berard et al., 2019), a simple approach where case tags indicating either title case or uppercase are added next to lowercased words. For each title cased word, we add the tag ^{o} to the left of the lowercased word, with a single white space separating the two; for uppercase word, we use the tag ^{oo}; for mixed case, we used a third tag, namely ˘. All three tags are taken to be right-associative by definition and tagging was applied prior to BPE segmentation (Sennrich et al., 2016).[3] Mixed case words were segmented at cased character boundaries during the lowercasing and tagging step. We will refer to this approach as ILC in what follows.

These additional symbols are processed along other elements of the vocabulary, without any further indication of their role, and thus are mapped to their corresponding em-

[1] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/recaser/truecase.perl

[2] To our knowledge, only OpenNMT (Klein et al., 2017) supported target factors as well, but only in the now deprecated Lua implementation, which does not include the state-of-the-art Transformer models we aimed to evaluate. Implementing target factors is a non-trivial task, as it involves important changes in the decoder, and we left their inclusion for future experiments.

[3] Note that our implementation differs from that of Berard et al. (2019), who use the tags <U> and <T> for uppercase and title case, respectively, added to the right of the lowercased word, and perform tagging after BPE segmentation. The only reason for these differences comes from our implementation of this tagging method preceding our knowledge of theirs. Time constraints did not permit an evaluation of the impact of these implementation differences, i.e. right-association and pre-BPE tagging, as in our implementation, versus left-association and post-BPE tagging, as in theirs.

]

CORPUS	WMT		OPENSUBS		GLOBALVOICES	
	EN-DE	EN-TR	EN-DE	EN-TR	EN-DE	EN-TR
TRAIN	5,852,458	207,373	-	-	-	-
DEV	2,999	3,000	-	-	-	-
TEST	3,004	3,007	10,000	10,000	1,465	160

Table 1: Corpora statistics, in number of sentence pairs

beddings. At post-processing time, words with these symbols to their left are converted into title casing or uppercasing, depending on the symbol generated by the decoder; sequences with a mixed cased symbol are further joined during the post-processing step.

As a simple example, the sentence in 1a would be tokenised as in 1b and augmented with inline casing as in 1c.

(1) a. This is an EXAMPLE with WiFi.
b. This is an EXAMPLE with WiFi .
c. ^{o} this is an ^{oo} example with ^{o} wi $^{\smile}$ ^{o} fi .

4. Experiments

In this section, we first describe the experimental setup, including datasets and system parameters, then present and discuss the results obtained on all test sets.

4.1. Experimental setup

We trained Transformer models (Vaswani et al., 2017) on two language pairs, namely English-German and English-Turkish, in both translation directions. Selecting these two language pairs was based on several factors. First, since all nouns are capitalised in German, but only proper names are in English, translation in both directions presents interesting challenges: from German to English, methods such as inline casing may for instance be tested on their ability to properly handle source case tags when translating into lowercased forms; from English to German, the evaluation would reflect case tag generation at decoding time for this type of approach. For English-Turkish, the agglutinative nature of Turkish morphology presents an interesting test bench to measure the accuracy of the different methods in higher data sparseness conditions; additionally, for this language pair, parallel data are scarce and the robustness of the different methods can be tested against comparatively lower amounts of training data.

As training and development sets, we selected the WMT 2017 datasets (Bojar et al., 2017), in the preprocessed form provided by the organisers to facilitate reproduction of results.[4] Since this preprocessing involved both tokenisation and truecasing, performed with the Moses toolkit, the original casing was reconstructed by detruecasing the provided datasets; this reconstruction is error-free as it only involves reversing the truecasing transformation by capitalising the first word in the same contexts defined for the truecasing operation. For all truecasing operations on test sets, we used the truecasing models provided for the WMT 2017 shared task. Development sets were also those provided for the shared task, in preprocessed form as well.

As test sets, we included the official WMT 2017 datasets for the two selected language pairs. Additionally, we prepared a test set from the Open Subtitles 2018 corpus (Lison et al., 2018), by randomly sampling 10,000 parallel subtitles in each language pair; this test set aimed to provide a large evaluation basis in a domain where casing differs from texts in the news domain, with multiple or delayed sentence starts within single subtitles, for instance. Finally, we prepared a third test set based on titles sampled from the Global Voices corpus in the version published in the OPUS repository (Tiedemann, 2012), to evaluate the accuracy of the different casing approaches on instances where open vocabulary words are usually capitalised in English, but not necessarily in other languages. For this dataset, we selected all sentence pairs where the proportion of words with title casing in the English sentence was above 80% of the total number of words. Corpora statistics are shown in Table 1.

All translation models were of type Transformer-small, composed of 8 attention heads and 6 layer of encoder and decoder, each with 2048 units. We used the Adam optimiser with an initial learning rate $\alpha = 0.0002$, which is reduced by a factor of 0.7 after 8 checkpoints with no improvement. Dropout was set to 0.1 for attention layers, preprocessing and postprocessing blocks, and before activation in feed-forward layers. Each batch contained 4096 tokens and the maximum sequence length was set to 99. The validation data was evaluated every 5000 steps for EN-DE models and every 1500 steps for EN-TR models. The training process ended if there was no improvement in the perplexity of 10 consecutive checkpoints. Source and target vocabularies are shared by the network and all datasets were segmented with BPE, using 30,000 operations.

4.2. Results

We present evaluation results along two main lines. First, the quality of the systems trained with the selected casing methods was evaluated in terms of BLEU scores (Papineni et al., 2002), both case-sensitive and case-insensitive, to measure the overall impact of case handling variants on the generated translations. All BLEU scores were computed with sacreBLEU (Post, 2018).

We then performed a targeted evaluation of the selected methods in terms of their ability to generate the casing forms provided in the reference sets. For each word in the reference, we thus counted cased and uncased matches in the machine-translated output generated by each method,

[4] These datasets are available at the following address: `http://data.statmt.org/wmt17/translation-task/preprocessed/`

METHOD	WMT		OPENSUBS		GLOBALVOICES	
	CASED	DECASED	CASED	DECASED	CASED	DECASED
RAW	33.9	35.2	21.9	23.2	18.9	29.9
LRC	32.6	35.3	21.1	23.2	11.9	30.7
MTC	33.8	35.2	21.9	23.3	18.9	30.7
ILC	**34.2**	**35.6**	**22.4**	**23.8**	**20.6**	**31.0**
CFT	33.8	35.1	22.0	23.3	19.1	30.3

Table 2: Case sensitive and insensitive BLEU results for German to English

METHOD	WMT		OPENSUBS		GLOBALVOICES	
	CASED	DECASED	CASED	DECASED	CASED	DECASED
RAW	27.6	28.2	18.2	19.0	20.1	20.7
LRC	26.1	28.0	17.3	19.0	18.1	**23.2**
MTC	28.0	28.6	18.2	19.0	18.8	19.3
ILC	28.1	28.2	**18.6**	**19.4**	**22.7**	23.0
CFT	**28.2**	**28.7**	18.4	19.1	20.8	21.3

Table 3: Case sensitive and insensitive BLEU results for English to German

and computed the percentages of matches on each dataset for each method.[5]

4.2.1. Reference metrics

The results in terms of cased and decased BLEU for German to English and English to German are presented in tables 2 and 3, respectively.

For German to English, inline casing proved optimal across the board, with the best scores in both cased and decased evaluations on all test sets. These first results indicate that this simple method, which requires neither the preparation of language-specific truecasing models, nor the extension of NMT modelling toolkits, offers a solid basis for case preservation along with the benefits of reduced data sparseness via lowercasing of the original data. Among the other methods, training on raw data proved similarly effective to truecasing in this translation direction, in line with previous results (Bawden et al., 2019). The approach based on case factors was only slightly better than truecasing, with worse results on the decased GLOBALVOICES test set, which might be due to the use of source factors only, as the use of raw data on the target side was detrimental to both the RAW and CFT methods for the generation of cased titles in English. Finally, in this translation direction, the LRC approach proved significantly worse as a casing method, when compared to the other approaches, as indicated by the relatively good results obtained on decased output but significantly lower metric results when measuring against cased references.

On the English to German test sets, relative results were comparable, although a few interesting differences could be observed. Inline casing performed well on these datasets as well, obtaining first or second place results overall, with minor differences in the latter case. Case factors performed better overall than in German to English, outperforming all methods but ILC overall, although it performed worse than the latter in two of the three scenarios. The benefits of using raw data against truecasing did not translate in the WMT scenario, with significantly worse results when compared to all methods but LRC, although it was comparable to, or better than, truecasing on the other datasets. Interestingly, the recasing method obtained the best results on decased title translation in the GLOBALVOICES scenario, indicating that all methods that include some form of casing within the translation process, may still underperform in terms of accurate translation.

As shown by the results obtained on decased output, the ILC method improves in most cases over the benefits of lowercase translation, considering the comparatively lower scores obtained by the LRC method, which reduces to strictly lowercase translation when measured on decased BLEU. This indicates that inline casing provides an effective means to combine lowercase-based translation benefits with case information exploitation. Overall, in both translation directions, the still popular truecasing approach was matched or outperformed by most variants, with inline casing proving optimal across the board.

The results for Turkish to English and English to Turkish are presented in tables 4 and 5, respectively. In English to Turkish, results were more uniform than with the previous language pair, which may be correlated with the overall lower quality of the translation models trained on relatively lower volumes of data. Overall, in this translation direction, truecasing and inline casing obtained slightly better results.

[5] Since different casing methods have an impact on the actual translations, where reference words may be generated or not by one method or another, there were several options to perform the targeted evaluation. For instance, only reference words produced by all methods, ignoring case variation, could have been evaluated, although this approach would take the lowest common denominator amongst methods and penalise those which generated larger amounts of translations that match reference words. Alternatively, the methods could be evaluated in terms of individual performance, measuring for instance precision and recall on the reference casing for common words in each system's translations and the reference; however, such an approach would have made it difficult to compare the different methods.

METHOD	WMT		OPENSUBS		GLOBALVOICES	
	CASED	DECASED	CASED	DECASED	CASED	DECASED
RAW	15.1	15.6	7.6	8.0	9.5	10.1
LRC	13.7	15.8	7.4	8.2	3.4	**16.3**
MTC	15.1	15.8	7.8	8.3	10.7	12.3
ILC	**16.1**	**16.7**	**8.4**	**8.9**	**14.2**	14.9
CFT	14.8	15.4	7.3	7.7	9.6	11.7

Table 4: Case sensitive and insensitive BLEU results for Turkish to English

METHOD	WMT		OPENSUBS		GLOBALVOICES	
	CASED	DECASED	CASED	DECASED	CASED	DECASED
RAW	10.4	10.7	4.7	5.2	9.4	9.5
LRC	9.0	**10.8**	4.2	5.0	3.3	10.6
MTC	**10.5**	**10.8**	4.7	**5.3**	**10.9**	**11.0**
ILC	10.4	**10.8**	**4.8**	**5.3**	9.8	9.9
CFT	10.3	10.6	4.4	4.8	7.2	7.6

Table 5: Case sensitive and insensitive BLEU results for English to Turkish

For Turkish to English, inline casing was the markedly better approach on all test sets but decased GLOBALVOICES, with truecasing performing only slightly better than training on raw data and source case factors underperforming on both cased and decased output.

The approach based on source case factors was less effective for this language pair overall, as shown in particular by the systematically lower results obtained on decased output when compared to the other approaches. This may be due to the relatively lower amounts of training data coupled with the added task of modelling with the additional dimensions of the factors themselves. Recasing was confirmed to be sub-optimal for cased translation, although it obtained competitive results on decased output, outperforming most other approaches but ILC on the selected titles of the GLOBALVOICES test set. Inline casing also proved to be robust with the low amounts of training data available in this language pair, as it obtained the best results overall on cased and decased output.

In terms of reference metrics, inline casing was thus the most robust approach across datasets and language pairs overall. Truecasing and training on raw data obtained similar results, while recasing was consistently outperformed by all other methods on cased output and case factors gave inconsistent results depending on the language pair, being usually outperformed by the ILC approach.

4.2.2. Targeted evaluation

The results of the targeted evaluation for English to German and German to English are presented in tables 6 and 7, respectively.[6]

In English to German, on the WMT test set all methods performed similarly on all categories, to the exception of LRC which obtained significantly lower results, with up to 10 percentage points lower accuracy on matching reference casing. A similar tendency can be observed on the OPENSUBS test set, with an even larger drop for the LRC method, which is not unexpected given that it relies on a language model trained on a different domain. For this test set, inline casing provided better case translation overall for title casing and uppercasing. On the GLOBALVOICES test set, it is worth noting that using truecasing or raw data, entirely or on the target side as with the CFT method, was significantly detrimental; this is expected as title capitalisation increases data sparseness issues for these methods. The ILC approach does not face the same issues and obtained the best results overall.

For German to English, the results were similar for the different methods on all three test set, except for LRC on uppercasing and title casing, with ILC obtaining only marginally better results in most cases. The results on the GLOBALVOICES test set were significantly higher for lowercasing and uppercasing than in the opposite direction, for all methods, with reversed results for title casing. This may be simply due to the difference in the number of references in each case, with larger numbers of references correlating with lower accuracy across the board. On WMT, the results were balanced across methods, with significantly higher results on title casing overall than for English to German, although the larger number of references in the latter case may also account for these differences. Finally, for OPENSUBS, all methods obtained similar results as well, with slightly higher marks for inline casing.

For all methods, title casing proved the most difficult in terms of casing, with the largest drops in accuracy for this category from the case insensitive results to the case sensitive ones, in both translation directions. This effect was less notable for English to German, with nouns capitalised by default, but significant in the opposite directions, across test sets and on the GLOBALVOICE test set in particular, as expected since it is composed of capitalised titles on the English side.

Overall, inline casing was systematically better at title cas-

[6] In all tables, %CS denotes the percentage of case-sensitive matches and CS-CI the difference in percentage points between case-sensitive and case-insensitive matches.

CASING	#REFS	RAW		LRC		MTC		ILC		CFT	
		%CS	CS-CI	%CS	CS-CI	CS	CS-CI	%CS	CS-CI	%CS	CS-CI
						WMT					
LOWER	33,411	53.45	-1.07	53.2	-1.32	53.65	-4.09	53.64	-1.02	**53.89**	-1.06
UPPER	437	**81.69**	-0.23	73.68	-8.47	**81.69**	-0.46	81.01	-0.23	80.09	0.0
TITLE	18,402	58.32	-2.06	55.12	-5.25	58.62	-2.19	**59.38**	-2.12	58.69	-2.13
MIXED	76	**77.63**	0.0	42.11	-31.57	75.0	-1.32	76.32	0.0	**77.63**	0.0
						OPENSUBS					
LOWER	43,658	41.26	-1.46	41.64	-1.71	41.75	-1.5	41.57	-1.57	**41.82**	-1.48
UPPER	253	38.34	-3.56	23.72	-20.94	-30.43	-9.89	**41.9**	-3.95	37.55	-2.77
TITLE	22,214	46.28	-4.93	43.45	-7.56	46.55	-5.25	**47.44**	-4.86	46.46	-4.87
MIXED	45	**15.56**	0.0	13.33	-2.23	**15.56**	0.0	**15.56**	0.0	13.33	0.0
						GLOBALVOICES					
LOWER	3,445	27.11	-1.74	**36.43**	-1.38	24.62	-1.74	34.17	-1.82	30.33	-2.04
UPPER	38	39.47	-7.9	28.95	-15.79	39.47	-7.9	**42.11**	-7.89	**42.11**	-7,89
TITLE	5,636	54.01	-0.64	52.38	-7.38	53.39	-0.76	**58.73**	-0.64	54.68	-0.64
MIXED	21	**23.81**	-4.76	9.52	-9.53	**23.81**	0.0	14.29	-4.76	**23.81**	-4.76

Table 6: Percentage of case sensitive word reference matches and case-insensitive differences for English to German

CASING	#REFS	RAW		LRC		MTC		ILC		CFT	
		%CS	CS-CI	%CS	CS-CI	CS	CS-CI	%CS	CS-CI	%CS	CS-CI
						WMT					
LOWER	48,063	59.77	-1.46	**60.02**	-1.35	59.7	-1.45	59.89	-1.63	59.64	-1.52
UPPER	657	81.43	-1.98	75.34	-7.31	**81.74**	-2.89	81.13	-1.37	80.82	-1.83
TITLE	7,928	74.95	-6.19	67.12	-14.26	74.86	-6.61	**76.08**	-5.62	75.01	-5.83
MIXED	70	**94.29**	-1.42	65.71	-28.58	90.0	-4.29	92.86	-1.43	**94.29**	-1.42
						OPENSUBS					
LOWER	59,349	43.32	-1.76	43.38	-1.63	43.31	-1.82	**43.56**	-1.88	43.44	-1.8
UPPER	3,370	66.29	-1.96	65.13	-3.27	66.17	-2.52	**66.94**	-2.11	66.26	-1.87
TITLE	12,027	47.38	-7.22	42.29	-13.08	46.86	-8.24	**47.98**	-7.53	47.19	-7.84
MIXED	52	**15.38**	0.0	13.46	-1.92	**15.38**	0.0	**15.38**	0.0	11.54	0.0
						GLOBALVOICES					
LOWER	855	**58.36**	-1.17	57.54	-0.94	57.54	-1.06	56.37	-1.29	58.13	-1.05
UPPER	36	**88.89**	0.0	72.22	-16.67	86.11	-2.78	86.11	0.0	**88.89**	0.0
TITLE	9,573	36.53	-19.54	28.56	-28.24	36.17	-20.17	**39.35**	-17.76	37.24	-19.34
MIXED	780	0.51	-0.26	0.13	-0.64	0.51	-0.13	0.51	-0.13	**0.64**	-0.13

Table 7: Percentage of case sensitive word reference matches and case-insensitive differences for German to English

ing than all other approaches, and comparable or slightly better the the alternatives for uppercasing and lowercasing. It is worth noting that this approach in particular is penalised by cases where a word was uppercased in the source but not in the target reference, as the model learns to apply this particular casing to the output in the general case and cannot predict deviations from the uppercasing norm found in the training datasets.

Excepting the generally underperforming recasing method, at least on cased references, the other methods obtained similar results overall. This is not entirely surprising considering that raw and truecased data mainly differ on the form of sentence initial words, although these results confirm that the impact of truecasing is rather minor overall, and sometimes detrimental. The even closer similarity of results between the approaches based on raw data and case factors is also notable, and may be derived from the use of raw data on the target side as the dominant factor in case handling for our implementation of the approach.

The results of the targeted evaluation for English to Turkish and Turkish to English are presented in tables 8 and 9, respectively. Overall, the results for this language pair are similar to those obtained in English-German, with inline casing as the most robust method overall, performing either similarly to, or notably better than, the alternative approaches. The ILC method was notably the optimal approach for title casing on all test cases, as was the case for English-German. Using raw or truecased data gave comparable results, in particular on the categories with more reference points, i.e. lowercase and title case.

Also similar in this language pair were the comparable, and sometimes identical, results obtained with raw data and case factors, in all categories except uppercasing, which may also be attributed to the use of raw data on the target side as a dominant trait with the source factor approach. The recasing approach also proved similarly deficient for

CASING	#REFS	RAW		LRC		MTC		ILC		CFT	
		%CS	CS-CI	%CS	CS-CI	CS	CS-CI	%CS	CS-CI	%CS	CS-CI
					WMT						
LOWER	36,511	31.01	-0.89	31.2	-1.13	31.61	-0.9	31.81	-1.05	**32.13**	-0.94
UPPER	675	47.85	-0.45	31.56	-16.29	45.78	-3.55	**52.44**	-2.52	41.48	-2.67
TITLE	8,576	51.15	-2.93	38.61	-14.14	51.35	-3.28	**54.16**	-4.03	49.27	-2.96
MIXED	40	30.0	0.0	2.5	-35.0	27.5	0.0	**50.0**	0.0	35.0	0.0
					OPENSUBS						
LOWER	34,417	17.22	-1.23	17.67	-1.65	17.1	-1.63	17.3	-1.53	**17.96**	-1.37
UPPER	410	21.46	-9.03	13.66	-15.36	21.46	-11.71	16.1	-11.7	**22.44**	-6.83
TITLE	12,533	17.18	-6.45	15.13	-8.96	18.22	-6.24	**19.37**	-6.38	16.06	-6.53
MIXED	4	**25.0**	0.0	0.0	0.0	**25.0**	0.0	0.0	0.0	0.0	0.0
					GLOBALVOICES						
LOWER	77	14.29	0.0	**24.68**	0.0	14.29	0.0	18.18	0.0	19.48	0.0
UPPER	2	**50.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TITLE	711	34.32	-0.84	20.39	-20.1	32.91	-1.55	**40.51**	-1.12	33.19	-2.11
MIXED	0	-	-	-	-	-	-	-	-	-	-

Table 8: Percentage of case sensitive word reference matches and case-insensitive differences for English to Turkish

CASING	#REFS	RAW		LRC		MTC		ILC		CFT	
		%CS	CS-CI	%CS	CS-CI	CS	CS-CI	%CS	CS-CI	%CS	CS-CI
					WMT						
LOWER	49,539	43.8	-1.33	**44.08**	-1.5	43.67	-1.44	43.97	-1.51	**44.08**	-1.21
UPPER	1,033	52.08	-3.0	40.66	-15.49	49.27	-4.75	**52.37**	-3.58	47.05	-4.06
TITLE	9,069	51.23	-5.52	37.8	-18.95	50.89	-6.17	**54.57**	-5.06	48.27	-6.29
MIXED	41	53.66	0.0	7.32	-48.78	31.71	0.0	**70.73**	0.0	53.66	-4.88
					OPENSUBS						
LOWER	55,070	25.22	-1.35	26.42	-1.63	25.54	-1.6	**26.55**	-1.89	26.37	-1.35
UPPER	2,796	66.27	-0.83	65.02	-1.97	67.73	-0.76	**68.56**	-0.86	68.49	-0.82
TITLE	11,102	25.38	-5.93	21.57	-10.67	25.87	-6.62	**28.14**	-6.48	23.82	-7.0
MIXED	244	**0.41**	0.0	0.0	0.0	**0.41**	0.0	**0.41**	0.0	0.0	0.0
					GLOBALVOICES						
LOWER	74	47.3	0.0	50.0	0.0	**51.35**	-1.35	44.59	-4.06	44.59	-1.36
UPPER	4	25.0	0.0	0.0	-25.0	**50.0**	0.0	25.0	0.0	0.0	0.0
TITLE	844	30.21	-2.73	19.08	-24.4	29.86	-6.28	**42.65**	-1.66	29.62	-4.03
MIXED	54	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 9: Percentage of case sensitive word reference matches and case-insensitive differences for Turkish to English

this language pair, in both translation directions.

In general, models based on raw data obtained better results than those relying on truecasing on the WMT test sets, although the results were reversed overall in the other scenarios. The CFT results on title casing were markedly worse for this language pair, which may be due to the lower amounts of training data for a method the relies on encoded representations of case.

In terms of differences between translation directions, translation into Turkish was worse with inline casing in the uppercase category with OPENSUBS, although this was mainly due to the larger number of casing mismatches between the source and target references. Case factors also performed on a par or better than the other methods in English to Turkish, except on title casing, but significantly worse in WMT in the other translation direction. Aside from these differences, the results in both translation directions for this language pair also point to inline casing as the more robust approach.

4.3. Summary of results

Although further specific analyses could be made on the results of the previous experiments, the following general points can be derived from these results.

First, the impact of casing can be significant, as demonstrated notably by the results in terms of BLEU for English-German on the OPENSUBS and GLOBALVOICES test sets, or with Turkish to English translation on the WMT test set. The results of the targeted evaluations also demonstrate the importance of case handling in terms of generating the correct word forms in the target language. These results are consistent over four translation directions and three different test sets covering different aspects of casing.

Secondly, the variety of test scenarios and metrics supports the comparative results obtained between the different casing methods currently or previously employed in machine translation. Among these, the use of raw data was confirmed to be somewhat equivalent to truecasing, on all metrics and most test cases. The use of source case factors was inconclusive, with similar results or minor improvements over other methods in some cases, e.g. WMT English to German, but with significant negative impact in others, e.g. WMT Turkish to English; further analysis could be warranted, notably by including target factors, although the latter requires non-trivial extensions to NMT toolkits, which might not be necessary to handle casing considering the overall results of our experiments.

Thirdly, considering the generalised drop in translation accuracy for all methods when comparing case-sensitive and case-insensitive reference word matches, it appears that none of the evaluated approaches fully manages to handle casing in an appropriate manner. Further research on this topic might thus be warranted to improve NMT translation accuracy.

Finally, inline casing proved the most robust approach overall across test sets and translation pairs, in terms of BLEU scores as well as accuracy in cased form generation. This can be seen as a welcome result, given the simplicity of the approach, which only involves simple offline tagging of the data and does not require any particular change to the NMT architecture, nor the use of external truecasing models. It may be worth exploring in more details its impact on the translation process, in particular on attention parameters and decoding efficiency, considering the additional tokens introduced in the source sentences, but current results strongly favour the adoption of this method at present to handle casing in NMT.

5. Conclusions

In this paper, we presented an evaluation of casing methods for Neural Machine Translation, which included the use of data in their original form, truecasing, recasing, case factors and inline casing. The different approaches were evaluated against different datasets, covering news data, on which all systems were trained, subtitles and newspaper titles, thus providing the first comprehensive evaluation of casing methods for current machine translation in varied scenarios.

The still-popular truecasing approach was shown to obtain similar results to simply using the data in their original form, in most cases. Source case factors were shown to be only slightly more effective than either approach in some cases, although a more complete evaluation including target case factors would be needed to fully assess the potential of case handling via dedicated embeddings. Finally, recasing was consistently suboptimal across test sets and inline casing proved significantly more robust across the board, resulting in higher accuracy in terms of reference case matching and BLEU scores.

The use of different case handling methods was shown to impact translation quality, in terms of BLEU scores as well as proportions of correctly translated words depending on casing. The results of our evaluations also indicate that, although inline casing was demonstrably the optimal approach in these experiments, none of the examined methods handled casing correctly in all cases, which leaves the matter open for future research. The results discussed in this article may nonetheless help consolidate current data processing pipelines and help optimise the development of more accurate machine translation systems.

6. Acknowledgements

This work was supported by the Department of Economic Development and Competitiveness of the Basque Government, via the MODENA and COMETA projects. We would like to thank the anonymous LREC reviewers for their insightful comments and suggestions.

7. Bibliographical References

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*.

Bawden, R., Bogoychev, N., Germann, U., Grundkiewicz, R., Kirefu, F., Miceli Barone, A. V., and Birch, A. (2019). The University of Edinburgh's Submissions to the WMT19 News Translation Task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 103–115, Florence, Italy, August. Association for Computational Linguistics.

Berard, A., Calapodescu, I., and Roux, C. (2019). Naver Labs Europe's Systems for the WMT19 Machine Translation Robustness Task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 526–532, Florence, Italy, August. Association for Computational Linguistics.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., and Turchi, M. (2017). Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214. Association for Computational Linguistics, September.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A Statistical Approach to Machine Translation. *Computational linguistics*, 16(2):79–85.

Chelba, C. and Acero, A. (2006). Adaptation of maximum entropy capitalizer: Little data can help a lot. *Computer Speech & Language*, 20(4):382–399.

Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111. Association for Computational Linguistics.

Chung, T. and Gildea, D. (2009). Unsupervised tokenization for machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 718–726. Association for Computational Linguistics.

Heafield, K. (2011). Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on*

Statistical Machine Translation, pages 187–197. Association for Computational Linguistics.

Hieber, F., Domhan, T., Denkowski, M., Vilar, D., Sokolov, A., Clifton, A., and Post, M. (2017). Sockeye: A toolkit for neural machine translation. *CoRR*, abs/1712.05690.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of the 20th Annual Conference of the European Association for Machine Translation*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 177–180. Association for Computational Linguistics.

Koehn, P. (2010). *Statistical machine translation*. Cambridge University Press.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71. Association for Computational Linguistics, November.

Lison, P., Tiedemann, J., and Kouylekov, M. (2018). Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Lita, L. V., Ittycheriah, A., Roukos, S., and Kambhatla, N. (2003). Truecasing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 152–159. Association for Computational Linguistics.

Mikheev, A. (1999). A knowledge-free method for capitalized word disambiguation. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 159–166. Association for Computational Linguistics.

Nagao, M. (1984). A Framework for a Mechanical Translation Between Japanese and English by Analogy Principle. In *Proceedings of the International NATO Symposium on Artificial and Human Intelligence*, pages 173–180, New York, NY, USA. Elsevier North-Holland, Inc.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*, pages 1715–1725.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th Language Resources and Evaluation Conference*, pages 2214–2218.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

Wang, W., Knight, K., and Marcu, D. (2006). Capitalizing machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 1–8.

Xiao, X., Liu, Y., Hwang, Y.-S., Liu, Q., and Lin, S. (2010). Joint tokenization and translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1200–1208. Association for Computational Linguistics.