mathematics

MDPI

Article

On the Use of Morpho-Syntactic Description Tags in Neural Machine Translation with Small and Large Training Corpora

Gregor Donaj * and **Mirjam Sepesy Maučec**

Faculty of Electrical Engineering and Computer Science, University of Maribor,
SI-2000 Maribor, Slovenia; mirjam.sepesy@um.si
* Correspondence: gregor.donaj@um.si; Tel.: +386-2-220-7205

Abstract: With the transition to neural architectures, machine translation achieves very good quality for several resource-rich languages. However, the results are still much worse for languages with complex morphology, especially if they are low-resource languages. This paper reports the results of a systematic analysis of adding morphological information into neural machine translation system training. Translation systems presented and compared in this research exploit morphological information from corpora in different formats. Some formats join semantic and grammatical information and others separate these two types of information. Semantic information is modeled using lemmas and grammatical information using Morpho-Syntactic Description (MSD) tags. Experiments were performed on corpora of different sizes for the English–Slovene language pair. The conclusions were drawn for a domain-specific translation system and for a translation system for the general domain. With MSD tags, we improved the performance by up to 1.40 and 1.68 BLEU points in the two translation directions. We found that systems with training corpora in different formats improve the performance differently depending on the translation direction and corpora size.

Keywords: neural machine translation; POS tags; MSD tags; inflected language; data sparsity; corpora size

MSC: 68T50

check for updates

Citation: Donaj, G.; Sepesy Maučec, M. On the Use of Morpho-Syntactic Description Tags in Neural Machine Translation with Small and Large Training Corpora. *Mathematics* **2022**, *10*, 1608. https://doi.org/10.3390/math10091608

Academic Editors: Florentina Hristea and Cornelia Caragea

Received: 31 January 2022
Accepted: 6 May 2022
Published: 9 May 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

In the last decade, research in machine translation has seen the transition from statistical models to neural net-based models for most mainstream languages and also some other languages. At the same time, researchers and developers got access to more and more parallel corpora, which are essential for training machine translation systems. However, many languages can still be considered low-resource languages.

Before the widespread use of neural machine translation (NMT), statistical machine translation (SMT) was the predominant approach. Additional linguistic information was added to deal with data sparsity or the morphological complexity of some languages. Often part-of-speech (POS) or morpho-syntactic description (MSD) tags were included in SMT systems in some way or the other. The tags can be included on the source side, the target side, or on both sides of the translation direction. This can be done either in the alignment or the training and translation phase.

Since the emergence of NMT, relatively few studies have explored the use of additional linguistic information for machine translation.

In this paper, we want to give an overview of the available studies and present experiments on using MSD tags for the translation between English and Slovene, a morphologically complex language. In order to provide a more comprehensive look at the usefulness of MSD tags, we perform several sets of experiments on a general domain corpus and a domain-specific corpus by using different training corpora sizes and methods to include MSD tags. In addition, we explore possibilities to reduce MSD tags, which can become rather complex in morphologically rich languages.

The rest of the paper is organized as follows. In Section 2 we present related work in machine translation and morphology. In Section 3 we present our experimental system, the used corpora, all corpora pre-processing steps, and the designed experiments. The results of the experiments are presented in Section 4 and discussed in Section 5. Our conclusions are presented in Section 6.

2. Related Work

2.1. Machine Translation

At first, machine translation was guided by rules. Corpus-based approaches followed. They were notably better and easier to use than the preceding rule-based technologies. Statistical machine translation (SMT) [1] is a corpus-based approach, where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual and monolingual text corpora. For a long time, SMT was the dominant approach with the best results. Migration from SMT to neural machine translation (NMT) started in 2013. In [2], a first attempt was made, where a class of probabilistic continuous translation models was defined that was purely based on continuous representations for words, phrases, and sentences and did not rely on alignments or phrasal translation units like in SMT. Their models obtain a perplexity with respect to gold translations that was more than 43% lower than that of the state-of-the-art SMT models. Using NMT was, in general, a significant step forward in machine translation quality. NMT is an approach to machine translation that uses an artificial neural network. Unlike SMT systems, which have separate knowledge models (i.e., language model, translation model, and reordering model), all parts of the NMT model are trained jointly in a large neural model. Since 2015, NMT systems have been shown to perform better than SMT systems for many language pairs [3,4]. For example, NMT generates outputs that lower the overall post-edit effort with respect to the best PBMT system by 26% for the German–English language pair [5]. NMT outperformed SMT in terms of automatic scores and human evaluation for German, Greek, and Russian [6]. In recent years NMT has also been applied to inflectional languages such as Slovene, Serbian, and Croatian. Experiments in [7] showed that on a reduced training dataset with around two million sentences, SMT outperformed the NMT neural models for those languages. In [8], NMT regularly outperformed SMT for the Slovene–English language pair. In the present paper, we will more systematically analyze NMT performance using training corpora of different sizes.

2.2. Neural Machine Translation

Neural machine translation is based on the recurrent neural network (RNN), which is a generalization of feedforward neural networks to sequences. Given a sentence in the source language as an input sequence of words $(x_1, \ldots, x_T)$, the RNN computes an output sequence of words $(y_i, \ldots, y_{T'})$ by using the equations:

$$\begin{aligned} h_t &= \mathrm{sigm}(W^{hx} x_t + W^{hh} h_{t-1}) \\ y_t &= W^{yh} h_t, \end{aligned} \tag{1}$$

where W^{hx}, W^{hh}, and W^{yh} are parameter matrices from the input layer to the hidden layer, from the hidden layer to itself, and from the hidden layer to the output layer; while x_t, h_t, and y_t denote the input, hidden, and output layer vectors at word t respectively.

The output sequence is a translation in the target language. The RNN presumes an alignment between the input and the output sequence, which is unknown. Furthermore, input and output sequences have different lengths. The solution is to map the input sequence to a fixed-sized vector using one RNN and then another RNN to map the vector to the target sequence. Another important phenomenon in natural languages is long-term dependencies which have been successfully modeled by using long short-term memory (LSTM) [9].

The goal of the LSTM is to estimate the conditional probability $p(y_1, \ldots, y_{T'}|x_1, \ldots, x_T)$, where $(x_1, \ldots, x_T)$ is an input sequence, and $(y_i, \ldots, y_{T'})$ is an output sequence, which can be of a different length than the input sequence. First, the LSTM obtains the fixed-dimensional representation v of the input sequence $(x_1, \ldots, x_T)$ from the last hidden state. Then, the probability of $(y_i, \ldots, y'_T)$ is computed with another LSTM whose initial hidden state is v:

$$p(y_1, \ldots, y_{T'}|x_1, \ldots, x_T) = \prod_{t=1}^{T'} p(y_t|v, y_1, \ldots, y_{t-1}). \qquad (2)$$

Here, $p(y_t|v, y_1, \ldots, y_{t-1})$ is represented with a softmax function over all words in the vocabulary.

Not all parts of the input sequence are relevant to producing an output word. The concept of attention was introduced to identify parts of the input sequence that are relevant [10]. The context vector c_t captures relevant input information to help predict the current target word y_t. Given the target hidden state h_t and the context vector c_t, a simple concatenation layer is used to combine the information from both vectors to produce a hidden attentional state:

$$\tilde{h_t} = \tanh(W_c[c_t, h_t]). \qquad (3)$$

The attentional vector $\tilde{h_t}$ is then fed through the softmax layer:

$$p(y_t|v, y_1, \ldots, y_{t-1}) = \mathrm{softmax}(W_s\tilde{h_t}). \qquad (4)$$

LSTM helps to deal with long sequences, but still, it fails to maintain the global information of the source sentence. Afterward, the transformer architecture was proposed, which handles the entire input sequence at once and does not iterate only word by word. The superior performance of transformer architecture was reported in [11].

2.3. Morphology of Inflected Language

Morphology is very important in machine translation as it directly affects the complexity and performance of a translation system. Among morphologically rich languages are inflected languages. In an inflected language, the lexical information for each word form may be augmented with information concerning the grammatical function, pronominal clitics, and inflectional affixes. As a result, the number of different word forms grows. Due to the high level of morphological variation, we may not find all morphological variants of a word even in very large corpora. When developing machine translation systems, we have to deal with data sparsity and high out-of-vocabulary rates.

The rich morphology in an inflected language permits a flexible word order because grammatical information is encoded within a word form. In English, grammatical information is expressed implicitly by word order and adjacency.

The complexity of the linguistic patterns found in inflected languages was shown to challenge machine translation in many ways. Translation results for inflected languages were much worse than for morphologically poor languages. Machine translation systems have been developed that take morphological information into account in one way or another. There is a big difference between translation from an inflected language and translation to an inflected language. If we are translating from an inflected language to English, the sparsity caused by the rich morphology of the source language is reduced in the translation process. However, if we are translating to an inflected language, the morphology of the target language needs to be generated in the translation process. Machine translation systems perform better in the former case. In this paper, we analyze these differences.

2.4. Morphological Information in SMT

Morphologically rich languages were extensively studied in the scope of phrase-based SMT. A factored translation model was proposed [1] as an extension of phrase-based SMT for incorporating any additional annotation to words, including morphological information, at decoding time. In factored models, a word is not a token but a vector of factors. Usually, factors in each word are the surface form of a word, lemma, POS tag, additional morphological information, etc. Factored models use more than one translation step and one generation step. Each translation step corresponds to one factor, e.g., lemma, POS factor, MSD feature vector, etc. Translation from the sequence of factors on the source side to the sequence of factors on the target side is trained. In the generation step, translated factors are combined into the final surface form of a word. Factored model training uses factors the same way as words in word-based models. In [1], gains of up to 2 BLEU points were reported for factored models over standard phrase-based models for the English–German language pair. Grammatical coherence was also improved.

In [12] a noisier channel is proposed, which extends the usual noisy channel metaphor by suggesting that an "English" source signal is first distorted into a morphologically neutral language. Then morphological processes represent a further distortion of the signal, which can be modeled independently. Noisier channel implementation uses the information extracted from surface word forms, lemmatized forms, and truncated forms. Truncated forms are generated using a length limit of six characters. This means that the maximum of the first six characters of each word is taken into account, and the rest is discarded. Hierarchical grammar rules are then induced for each type of information. Finally, all three grammars are combined, and a hierarchical phrase-based decoder is built. The authors reported a 10% relative improvement of the BLEU metric when translating into English from Czech, a language with considerable inflectional complexity.

In [13] the focus is on the three common morphological problems: gender, number, and the determiner clitic. The translation is performed in multiple steps, which differ from the steps used in factored translation. The decoding procedure is not changed. Only the training corpus is enriched with new features: lemmas, POS tags, and morphological features. Classic phrase-based SMT models are trained on this enriched corpus. They afterward use conditional random fields for morphological class prediction. They reported improvement of BLEU by 1 point on a medium-size training set (approximately 5 million words) for translation from English to Arabic.

In [14,15] the quality of SMT was improved by applying models that predict word forms from their stems using extensive morphological and syntactic information from both the source and target languages. Their inflection generation model was the maximum entropy Markov model, and it was trained independently of the SMT system. The authors reported a small BLEU improvement (<0.5) when translating from English to Russian and a larger BLEU improvement of 2 when translating English to Arabic. A similar approach was used in [16]. They trained a discriminative model to predict inflections of target words from rich source-side annotations. They also used their model to generate artificial word-level and phrase-level translations, that were added to the translation model. The authors reported BLEU improvement by 2 points when translating from English to Russian.

2.5. Morphological Information in NMT

In 2013, NMT emerged [2]. The NMT models did not achieve a good performance at the beginning because they were unable to handle the long-distance dependencies. With the introduction of long short-term memory [9] the results improved because, thanks to the gate mechanism, long-distance dependencies in a sentence were captured much better. Although translation results improved considerably, morphologically rich languages remain a challenge. The better translation capability of NMT models does not make linguistic features redundant. In some papers, they were incorporated to provide further performance improvements.

Factored representations of words were considered as input features [17]. The linguistic features improved results by 1.5 BLEU for Germain to English, 0.6 BLEU for English to German, and 1.0 BLEU for English to Romanian translation. The authors used 4.2 million sentence pairs for training the systems between English and German, and 0.6 million sentence pairs between English and Romanian. Using additional synthetic parallel data, they again achieved comparable improvements.

Lemma and morphological factors were also used as output features [18,19]. Results for English to French translation were improved by more than 1.0 BLEU using a training corpus with 2 million sentence pairs. They also used an approach that generates new words controlled by linguistic knowledge. They halved the number of unknown words in the output. Using factors in the output reduces the target side vocabulary and allows the model to generate word forms that were never seen in the training data [18]. A factored system can support a larger vocabulary because it can generate words from the lemma and factors vocabularies, which is advantageous when data is sparse.

The decoder learns considerably less morphology than the encoder in the NMT architecture. In [20], the authors found that the decoder needs assistance from the encoder and the attention mechanism to generate correct target morphology. Three ways to explicitly inject morphology in the decoder were explored: joint generation, joint-data learning, and multi-task learning. Multi-task learning outperformed the other two methods. A 0.2 BLEU improvement was reported for English to Czech translation. Authors argued that by having larger corpora, the improvement would be larger.

Unseen words can also be generated by using character-level NMT or sub-word units, determined by the byte-pair encoding (BPE) algorithm [21]. Character-level NMT outperformed NMT with BPE sub-words when processing unknown words, but it performed worse than BPE-based systems when translating long-range phenomena of morpho-syntactic agreement. Another possibility is to use morpheme-based segmentations [22]. A new architecture was proposed to enrich character-based NMT with morphological information in a morphology table as an external knowledge source [23]. Morphology tables increase the capacity of the model by increasing the number of network parameters. The proposed extension improved BLEU by 0.55 for English to Russian translation using 2.1 million sentence pairs in the training corpus.

The authors in [24] also investigate different factors on the target side. Their experiments include three representations varying in the quantity of grammatical information they contain. In the first representation, all the lexical and grammatical information was encoded in a single factor. In the second representation, only a well-chosen subset of morphological features was kept in the first factor, and the second factor corresponded to the POS tag. In the third representation, the first factor was a lemma, and the second was the POS tag. In some experiments, BPE splitting was also used. They reported that carefully selected morphological information improved the translation results by 0.56 BLEU for English to Czech translation and 0.89 BLEU for English to Latvian translation.

In [25], the authors focused on information about the subject's gender. For regular source language, words were annotated with grammatical gender information of the corresponding target language words. Information about gender improved the BLEU results by 0.4.

Most often, NMT systems rely only on large amounts of raw text data and do not use any additional knowledge source. Linguistically motivated systems can help to overcome data sparsity, generalize, and disambiguate, especially when the dataset is small.

2.6. Aim and Research Contribution

A literature review shows the potential of linguistic information for many highly inflected languages. However, we are not aware of any such research for the Slovene–English language pair in the scope of neural machine translation.

Some researchers noticed the correlation between the training corpus size and the improvement brought by linguistic features. Others may compare only a few different formats of adding linguistic features. However, they did not go into detail or perform a systematic comparison.

This paper aims to provide a more extensive analysis of the use of MSD tags in NMT between English and highly inflected Slovene language. The idea of our approach is to focus on data preparation rather than on NMT architecture. We used the same NMT architecture through all experiments. We only modify the data used for training the NMT systems, similarly to some systems presented in [20]. The advantage of our approach is that it can be easily transferred to other language pairs, as appropriate tools for lemmatization and MSD tagging are already available for several morphologically complex languages.

The literature review also shows that often changes in the architecture of the models are required, or that special translation tools are used, which enable a factored representation of the input tokens. However, the most widely used NMT tools do not enable such a representation. Thus, our approach can be more easily included in practical applications, such as the use of the OPUS-MT plugin for translation tools, which uses Marian NMT [26].

The main contributions of this paper are: (1) a review of literature on using morphological information in SMT and NMT; (2) an empirical comparison of the influence of different types of morphologically tagged corpora on translation results in both translation directions separately; (3) a comparison of the effect of morphologically pre-processed training corpora with regard to the training corpus size; (4) a comparison of results obtained on domain-specific corpus and general corpus.

3. Methods

3.1. Corpora

In our experiments, we used the freely available Europarl and ParaCrawls corpora for the English–Slovene language pair. Both can be downloaded at the OPUS collection website (https://opus.nlpl.eu/, accessed on 28 January 2022).

Europarl is extracted from the proceedings of the European Parliament and is considered a domain-specific corpus. It consists of approximately 600,000 aligned parallel segments of text. ParaCrawl was built by Web Crawling and automatic alignment. We consider it to be a more general domain corpus. It consists of approximately 3.7 million aligned segments of text. Both corpora are available for several language pairs. The above sizes refer to the English–Slovene language pair. Experiments were performed in the same way separately on both corpora. Therefore, all processing steps apply the same way for both corpora.

We divided the corpora into three parts by using 2000 randomly selected segment pairs as the development set, another 2000 randomly selected segment pairs as the evaluation set and the remaining segments as the training set. For some later experiments, we further divided the training set into ten equally sized parts.

3.2. Standard Pre-Processing

We performed the standard corpus pre-processing steps for machine translation training: cleanup, punctuation normalization, tokenization, and truecasing. During these steps, some segments were excluded due to their excessive length. The final sizes of all datasets are shown in Table 1.

Table 1. Number of segments (sentences) in each dataset.

Dataset Part	Europarl	ParaCrawl
Training	618,516	3,714,473
Development	1993	1987
Evaluation	1991	1990
Total	622,500	3,718,450

Punctuation normalization mainly ensures the correct placing of spaces around punctuation symbols and replaces some UTF-8 punctuation symbols with their ASCII counterparts. The most significant effect from the pre-processing steps can be seen between the normalized text (which after post-processing is also called de-truecased and de-tokenized text) and the tokenized and truecased text. A couple of examples from the ParaCrawl corpus are presented in Table 2. From them, we see that tokenization separates punctuation symbols into separate tokens by inserting spaces around them. Note, that in the second example, the point (".") is not tokenized, as it is used as a decimal symbol. Truecasing then converts all words to their lowercase forms, unless they are proper names.

De-truecasing and de-tokenization is the reverse process, i.e., rejoining punctuation symbols and capitalizing the first word in a segment (sentence). These steps are done in the post-processing stage in order to obtain a translation in the normalized form according to the grammar of the target language.

Table 2. Comparison between normalized segments (NM), and tokenized and truecased segments (TC).

Form	Segment
NM	After the fourth hour, Moda plays a major role.
TC	After the fourth hour, Moda plays a major role .
NM	User manual PDF file, 2.6 MB, published 19 October 2018
TC	User manual PDF file, 2.6 MB, published 19 October 2018

3.3. MSD Tagging and Lemmatization

Part-of-speech (POS) tagging is the process of assigning POS tags to each word in a text. In morphologically rich languages, these tags are often referred to as morpho-syntactic description (MSD) tags. POS tags for English usually convey the part of speech of a word, and for some words type, degree (comparative, superlative), and number (singular or plural). MSD tags in Slovene, however, convey much more information: POS, type, number, case, gender, person, degree, etc. Consequently, MSD tags in Slovene are more complex and numerous.

For simplicity, we will refer to the tags in both languages as MSD tags.

Lemmatization is the process of assigning a lemma (the word's canonical form) to each word in a text. For example, the lemma "be" is the lemma for: am, are, is, was, were, etc. In English, words have a relatively small number of different surface word forms, mainly varying by number and tense. In Slovene, on the other hand, the main parts of speech (nouns, adjectives, and verbs) and pronouns vary by case, gender, number, person, and tense. This results in a higher number of different surface word forms.

We tagged and lemmatized all datasets using a statistical tagger and lemmatizer. For English, the tagger was trained on the Penn treebank using a modified Penn treebank tagset (https://www.sketchengine.eu/modified-penn-treebank-tagset, accessed on 28 January 2022), and for Slovene, the tagger was trained on the ssj500k training corpus [27] using the tagset defined in the JOS project [28]. Since the tagger was developed to work on standard text input, it has its own tokenizer built in. Therefore, we used the normalized text as the input for the tagger.

Table 3 shows the same sentence in English and Slovene in three forms: surface word forms, lemmatized text, and MSD tags. The complexity of the Slovene MSD tags can already be seen by their length, as each letter in the tag corresponds to one attribute of the given word.

Table 3. Example of a MSD tagged and lemmatized sentence in English (EN) and Slovene (SL).

Language	Form	Text		
EN	Words	You	went	away
	Lemmas	you	go	away
	MSD Tags	PP	VVD	RB
SL	Words	Ti	si	odšel
	Lemmas	ti	biti	oditi
	MSD Tags	Zod-ei	Gp-sde-n	Ggdd-em

The English tagset defines 58 different MSD tags for words and punctuations, all of which appear in the used corpora. The Slovene tagset defines 1902 different MSD tags.

However, only 1263 of them appear in the Europarl training corpus, and 1348 of them appear in the ParaCrawl training corpus.

3.4. Corpora Formats and Translation Combinations

The surface word-form corpora serve us as the basic format. We then used the tagged and lemmatized corpora to build corpora in five additional formats for each dataset by using different methods for constructing the final segments. The formats are as follows.

1. W: Surface word forms after tokenization and truecasing.
2. W+M: Words and MSD tags were written alternately as separate tokens.
3. WM: Words and MSD tags were written concatenated as one token.
4. L+M: Lemmas and MSD tags were written alternately as separate tokens
5. LM: Lemmas and MSD tags were written concatenated as one token.
6. WW-MM: Words were written consecutively, followed by a special tag, and then MSD tags were written consecutively.

We used the second format (W+M) as the most straightforward format to add morphological information. We then also wanted to explore the difference if the MSD tag is a separate token in the training and translation processes or not. However, from a combination of the lemma and the MSD tag of a word, we can replicate the surface word form. Additionally, in inflected languages, different lemmas can give the same surface form, e.g., there are some such pairs of adjectives and adverbs. Thus, the combination of a lemma and MSD tag can give the same amount of information as the combination of a word and MSD tag. Lastly, we explored the possibility of writing the MSD tags as an almost separate string and the end of the sentence, similar to one of the systems in [20].

Table 4 shows an example segment from the English Europarl corpus in all six formats. Surface word forms were unchanged, while lemmas and MSD tags were tagged with "LEM:" and "MSD:", respectively. Concatenation in the formats WM and LM were done with a hyphen. Surface word forms and MSD tags in the format WW-MM were divided with the tag "<MSD>".

Table 4. An example segment in the different formats of the corpora.

Format	Segment
W	let us be honest .
W+M	let MSD:VV us MSD:PP be MSD:VB honest MSD:JJ . MSD:.
WM	let-MSD:VV us-MSD:PP be-MSD:VB honest-MSD:JJ .-MSD:.
L+M	LEM:let MSD:VV LEM:us MSD:PP LEM:be MSD:VB LEM:honest MSD:JJ LEM:. MSD:SENT
LM	LEM:let-MSD:VV LEM:us-MSD:PP LEM:be-MSD:VB LEM:honest-MSD:JJ LEM:.-MSD:SENT
WW-MM	let us be honest . <MSD> MSD:VV MSD:PP MSD:VB MSD:JJ MSD:.

We build word-based vocabularies from both training corpora, containing 60,000 words on each translation side for the experiments with the Europarl corpus and 200,000 words on each side in the experiments with the ParaCrawl corpus.

The reason for the different sizes lies in the different text coverages given a particular vocabulary size. Table 5 shows the out-of-vocabulary (OOV) rates on the evaluation set in both corpora. We see that those rates are far lower in the domain-specific corpus Europarl. The remaining OOV rate at the larger vocabularies is due to words that do not appear in the training set at all. On the other hand, the OOV rates are higher in the general-domain corpus ParaCrawl, where a more diverse vocabulary is to be expected. Therefore, a more extensive vocabulary is needed to obtain reasonable results. From the data, we also see that OOV rates are significantly greater in Slovene, which is due to the high number of word forms.

Table 5. Out-of-vocabulary rates on both corpora for differently sized vocabularies.

Vocabulary Size	ParaCrawl OOV Rate (%)		Europarl OOV Rate (%)	
	Slovene	English	Slovene	English
60k	6.66	2.57	1.05	0.16
100k	4.44	1.77	0.55	0.14
200k	2.53	1.09	0.34	0.14
300k	1.82	0.85	0.34	0.14

For the systems that use MSD tags as separate tokens (W+M, WW-MM), the vocabulary was simply extended with the MSD tags. In the system which uses the words and MSD tags concatenated to one token (WM), the vocabulary was extended so far that it includes all words in the original vocabulary in combination with all possible corresponding MSD tags that appear in the training set. Some word forms can have different grammatical categories. For example, feminine gender words ending in "-e" typically can have four different tags given their case and number: (1) singular and dative, (2) singular and locative, (3) dual and nominative, and (4) dual and accusative. This is similarly true for most words (primarily other nouns, adjectives, and verbs). Hence, the size of the vocabulary increases substantially for concatenated combinations.

In the systems using lemmas instead of words, the vocabularies are similarly built. The final sizes of all used vocabularies are presented in Table 6.

Table 6. Final vocabulary sizes on all tagged formats of the corpora.

Corpora Format	Europarl Vocabulary Size		ParaCrawl Vocabulary Size	
	Slovene	English	Slovene	English
Baseline	60,000	60,000	200,000	200,000
W+M	61,263	60,058	201,348	200,058
WM	140,738	73,332	500,000	242,589
L+M	57,518	50,200	201,373	200,058
LM	211,499	71,943	500,000	313,135
WW-MM	61,264	60,059	201,349	200,059

Often BPE or other data-driven methods for splitting words into subword units are used to improve the results of NMT. However, since those methods are not linguistically informed, they can produce erroneous surface forms by concatenating incompatible subword units. Also, the generation of MSD tags is based on full word forms and cannot be performed on subword units. For this reason, we decided to avoid using data-driven segmentation in our paper.

3.5. Training and Evaluation

Having six formats of the corpora with the accompanying vocabularies on each translation side, we were able to build 36 translation systems for all possible combinations of any format on the source side and any format on the target side. The system, which contains only the surface word forms on both the source and target side, is the baseline system. The other 35 systems will be evaluated and compared to it.

The system was a neural net-based translation system, using a recurrent neural network-based encode–decoder architecture with 512 dimensions for the encoding vector and 1024 as the dimension for the hidden RNN states. The models were trained on all the possible training data format combinations. The training duration was limited to 20 epochs, and the development set was used for verification during the training process. Finally, the model update which gave the best results on the development set was kept. The model hyperparameters and other options for the training tools were kept at their default values for all experiments. The only exception was the maximal allowed sentence length which had to be doubled in all experiments that used MSD tags as separate tokens.

We used BLEU for the validation during training and the final evaluation of the models. BLEU is a widely used metric to evaluate the performance of machine translation systems, and it is defined by

$$BLEU = \min\left\{1, \frac{len_o}{len_r}\right\} \cdot \left(\prod_{i=1}^{4} prec_i\right)^{\frac{1}{4}}, \tag{5}$$

where len_o is the length of the evaluated translation, len_r is the length of the reference translation, and $prec_i$ is the ratio between the number matched n-grams and the number of total n-grams of order i between the reference translation and the evaluated translation.

The machine translation output was post-processed: de-truecased and de-tokenized during validation and final evaluation. Post-processing also included removing MSD tags, the special tag in format WW-MM. In the two systems that generated the lemma and MSD tag on the target side, we also had to employ a model that would determine the correct worm form.

Additionally, we performed statistical significance tests on all results compared to the baseline results. We used the paired bootstrap resampling test [29], a widely used test in machine translation with a p-value threshold value of 0.05 [12,13,17,23,25].

3.6. MSD Tag Reduction

Due to the high number of different MSD tags in Slovene, we also designed experiments using reduced MSD tags, that retain only the most important information. This was done only on the Slovene part of the corpora.

To construct the rules for MSD reduction, we used grammatical knowledge. For the main parts of speech, information was retained based on which the correct word form can be determined. These are the case, number, gender, and person. MSD tags for the minor parts of speech were reduced to the mere POS tag and sometimes the type. Table 7 shows a segment with full and reduced MSD tags.

Table 7. An example segment tagged with full and reduced MSD tags.

Form	Segment						
Words	Gospod	predsednik	,	hvala	za	besedo	.
Full MSD tags	Ncmsn	Ncmsn	,	Ncfsn	Sa	Ncfsa	.
Reduced MSD tags	Nmsn	Nmsn	,	Nfsn	Sa	Nfsa	.

The reduction of the MSD tag complexity resulted in a decreased number of different tags. In the Europarl training corpus, the number was reduced from 1263 to 418, and in the ParaCrawl training corpus from 1348 to 428.

3.7. Tools

In the pre-processing and post-processing steps, we used several scripts, which are part of the MOSES statistical machine translation toolkit [30]. Those steps are cleanup, tokenization, truecasing, de-truecasing, and de-tokenization. We used the SacreBLEU [31] evaluation tool for scoring and statistical tests.

MSD tagging and lemmatization were performed using TreeTagger [32] and the accompanying pre-trained models for Slovene and English.

Translation model training and translation were performed using Marian NMT [33] on NVIDIA A100 graphical processing units.

4. Results

4.1. Translation Systems

Table 8 shows translation results from English to Slovene for all combinations of corpora formats, and Table 9 shows the results for the same experiments from Slovene to

English on the Europarl corpus. In both cases, the result in the first line and first column is the baseline result (surface word form).

Below all results, we show p-values for the paired bootstrap resampling test between the given result and the baseline result.

Table 8. BLEU scores and p-values for the translations from English to Slovene for different systems on the Europarl corpus.

Source Form	**Target Form**					
	W	**W+M**	**WM**	**L+M**	**LM**	**WW-MM**
W	35.97	37.17 $p < 0.001$	36.77 $p = 0.003$	36.96 $p = 0.002$	34.78 $p = 0.001$	36.78 $p = 0.004$
W+M	37.12 $p < 0.001$	37.37 $p < 0.001$	37.18 $p < 0.001$	36.88 $p = 0.003$	35.90 $p = 0.319$	37.28 $p < 0.001$
WM	36.48 $p = 0.037$	37.21 $p < 0.001$	36.14 $p = 0.195$	36.16 $p = 0.188$	35.15 $p = 0.008$	36.67 $p = 0.012$
L+M	37.24 $p < 0.001$	36.57 $p = 0.033$	36.73 $p = 0.011$	36.31 $p = 0.117$	34.92 $p = 0.001$	36.66 $p = 0.011$
LM	35.90 $p = 0.305$	36.20 $p = 0.158$	35.77 $p = 0.202$	35.52 $p = 0.071$	33.64 $p < 0.001$	36.01 $p = 0.342$
WW-MM	23.83 $p < 0.001$	27.84 $p < 0.001$	25.03 $p < 0.001$	36.71 $p = 0.014$	23.91 $p < 0.001$	28.51 $p < 0.001$

Table 9. BLEU scores and p-values for the translations from Slovene to English for different systems on the Europarl corpus.

Source Form	**Target Form**					
	W	**W+M**	**WM**	**L+M**	**LM**	**WW-MM**
W	40.35	40.48 $p = 0.222$	40.33 $p = 0.366$	39.14 $p < 0.001$	39.10 $p < 0.001$	40.35 $p = 0.421$
W+M	41.16 $p = 0.002$	40.55 $p = 0.177$	40.97 $p = 0.012$	40.12 $p = 0.146$	39.46 $p = 0.001$	40.54 $p = 0.167$
WM	40.19 $p = 0.193$	39.95 $p = 0.069$	40.15 $p = 0.153$	38.86 $p < 0.001$	38.76 $p < 0.001$	39.60 $p = 0.003$
L+M	42.03 $p < 0.001$	41.80 $p < 0.001$	41.53 $p < 0.001$	40.40 $p = 0.347$	40.27 $p = 0.288$	41.01 $p = 0.006$
LM	39.66 $p = 0.004$	39.87 $p = 0.040$	39.70 $p = 0.007$	38.79 $p < 0.001$	38.50 $p < 0.001$	39.83 $p = 0.029$
WW-MM	35.03 $p < 0.001$	39.73 $p = 0.012$	39.80 $p = 0.022$	38.93 $p < 0.001$	38.39 $p < 0.001$	39.59 $p = 0.002$

When comparing baseline results with other models, we see that some combinations give better results and others give worse results than the baseline model.

In the direction from English to Slovene, the best results are achieved when using words and MSD tags as separate tokens on both translation sides. On the other hand, in the direction from Slovene to English, the best results are achieved with lemmas and MSD tags as separate tokens on the source side, and only the surface words form on the target side. We will refer to these two systems as the improved systems for their respective translation direction.

Tables 10 and 11 show the same results as the previous two tables, albeit on the ParaCrawl corpus. The results show that the best results are obtained with the same combination of formats as on the Europarl corpus.

Table 10. BLEU scores and p-values for the translations from English to Slovene for different systems on the ParaCrawl corpus.

Source Form	Target Form					
	W	**W+M**	**WM**	**L+M**	**LM**	**WW-MM**
W	44.09	44.51 $p = 0.103$	44.19 $p = 0.306$	41.78 $p < 0.001$	41.79 $p < 0.001$	44.65 $p = 0.059$
W+M	44.41 $p = 0.149$	44.80 $p = 0.025$	39.35 $p < 0.001$	42.00 $p < 0.001$	41.50 $p < 0.001$	44.65 $p = 0.053$
WM	44.01 $p = 0.315$	43.96 $p = 0.255$	44.64 $p = 0.071$	41.82 $p < 0.001$	41.88 $p < 0.001$	44.28 $p = 0.213$
L+M	42.95 $p = 0.001$	41.89 $p < 0.001$	42.50 $p < 0.001$	39.52 $p < 0.001$	39.75 $p < 0.001$	42.58 $p < 0.001$
LM	42.41 $p < 0.001$	41.91 $p < 0.001$	42.63 $p < 0.001$	40.41 $p < 0.001$	40.04 $p < 0.001$	42.34 $p < 0.001$
WW-MM	44.35 $p = 0.167$	43.50 $p = 0.052$	44.29 $p = 0.227$	41.83 $p < 0.001$	41.06 $p < 0.001$	44.15 $p = 0.354$

Table 11. BLEU scores and p-values for the translations from Slovene to English for different systems on the ParaCrawl corpus.

Source Form	Target Form					
	W	**W+M**	**WM**	**L+M**	**LM**	**WW-MM**
W	47.89	47.45 $p = 0.088$	47.34 $p = 0.045$	43.63 $p < 0.001$	42.88 $p < 0.001$	47.67 $p = 0.179$
W+M	47.84 $p = 0.340$	48.02 $p = 0.259$	47.44 $p = 0.087$	43.73 $p < 0.001$	42.70 $p < 0.001$	47.53 $p = 0.124$
WM	47.81 $p = 0.314$	47.67 $p = 0.187$	46.88 $p = 0.005$	43.12 $p < 0.001$	43.23 $p < 0.001$	47.57 $p = 0.136$
L+M	48.19 $p = 0.153$	48.06 $p = 0.221$	48.00 $p = 0.285$	43.98 $p < 0.001$	43.11 $p < 0.001$	47.60 $p = 0.181$
LM	46.79 $p = 0.002$	47.06 $p = 0.016$	46.84 $p = 0.003$	42.82 $p < 0.001$	43.23 $p < 0.001$	46.87 $p = 0.002$
WW-MM	47.53 $p = 0.127$	47.53 $p = 0.126$	41.99 $p < 0.001$	42.74 $p < 0.001$	42.76 $p < 0.001$	47.46 $p = 0.085$

4.2. Training Corpora Size

For the second set of experiments, we selected three combinations: the baseline system, the system that performed best in the given translation direction with full MSD tags, and the same two systems with reduced MSD tags. We explored the effect of the training corpora size by repeating the experiments with these three systems on smaller corpora from 10% of the full training corpora size to 100% of the size in increments of 10% for the experiments with the Europarl corpus. We then repeated the same set of experiments on the ParaCrawl corpus, albeit with training corpora sizes from 5% of the full training corpora size to 100%. The results are presented in Figures 1–4. For better readability, the x-axis is on a logarithmic scale.

We again used the paired bootstrap resampling test between the baseline results and the results with the full MSD tags and all corpora sizes. For the results on the Europarl corpus in Figures 1 and 2, we found that all differences are statistically significant at a threshold value of 0.05. The differences in Figure 3 are statistically significant, up to 30% of the full training corpora size, and in Figure 4, up to 70% of the full training corpora size.

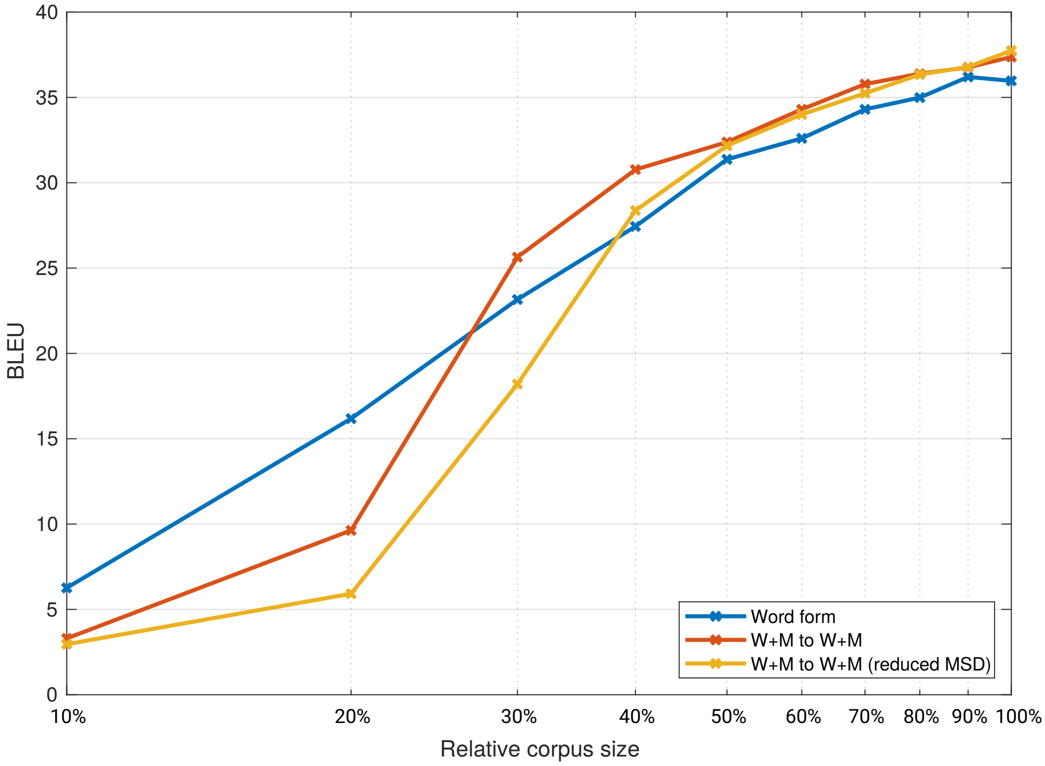


Figure 1. Translation results for the selected translation systems from English to Slovene using Europarl with respect to relative training corpora size.

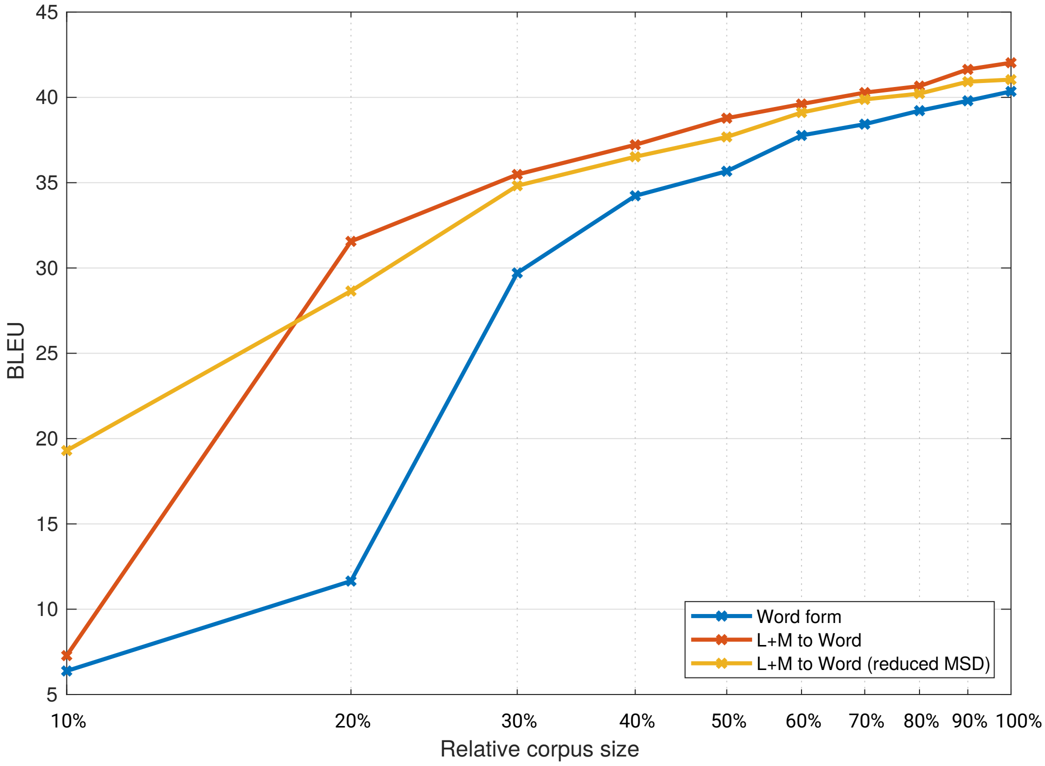


Figure 2. Translation results for the selected translation systems from Slovene to English using Europarl with respect to relative training corpora size.

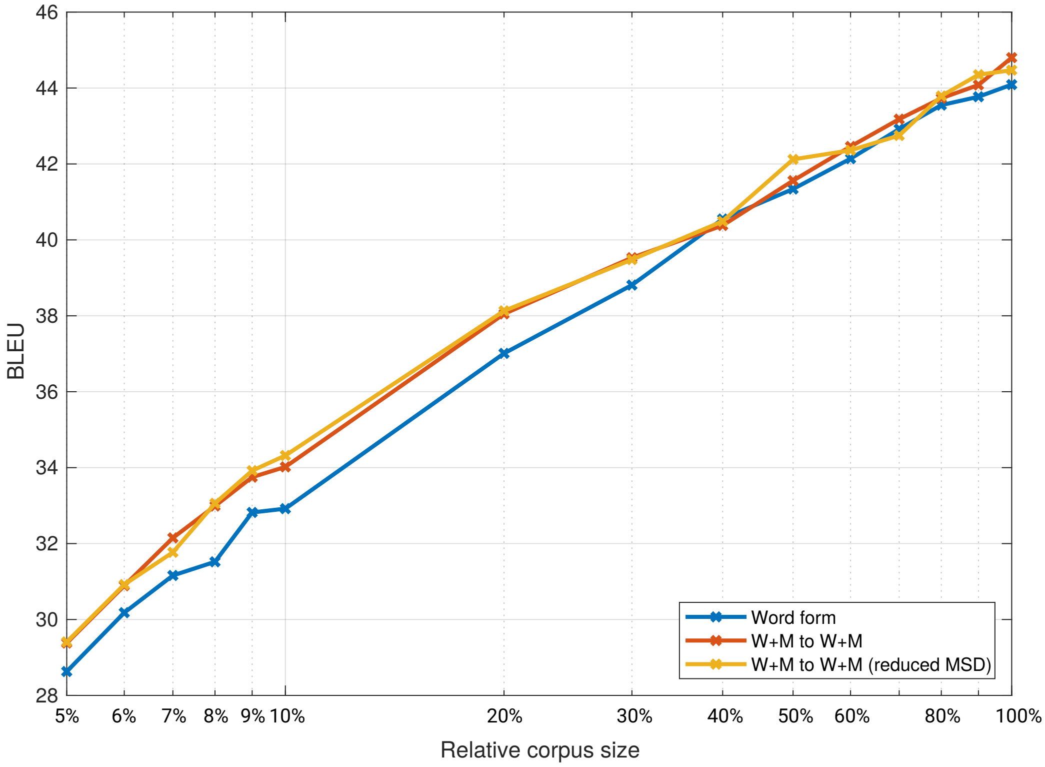


Figure 3. Translation results for the selected translation systems from English to Slovene using ParaCrawl with respect to relative training corpora size.

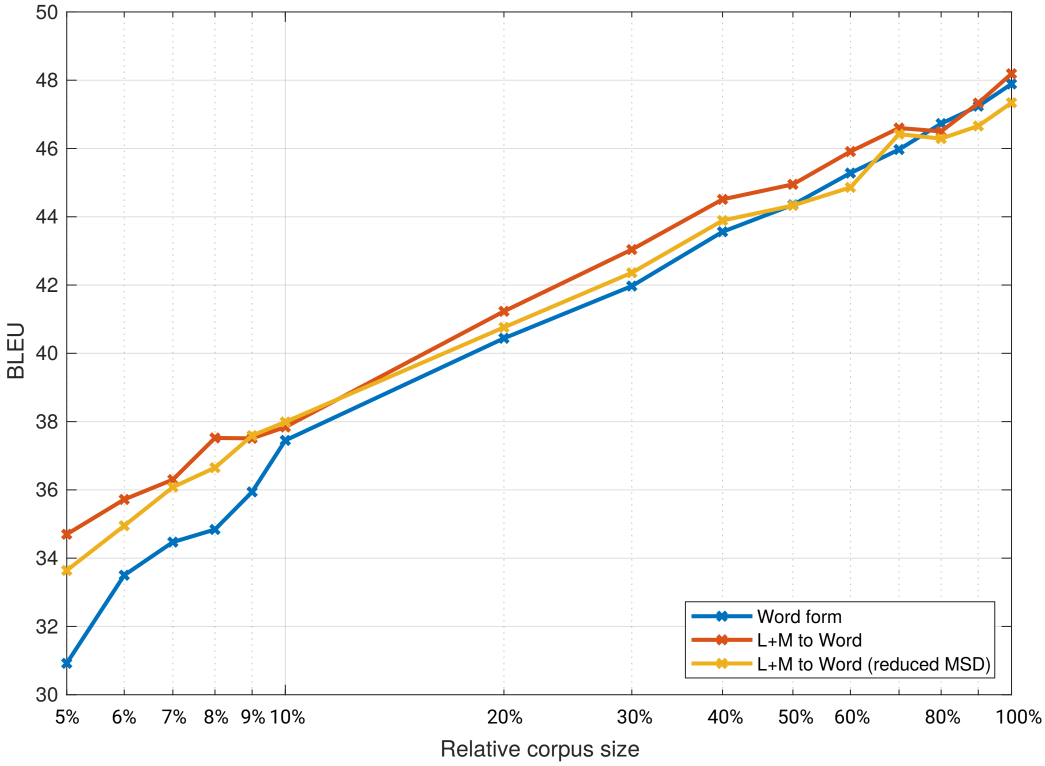


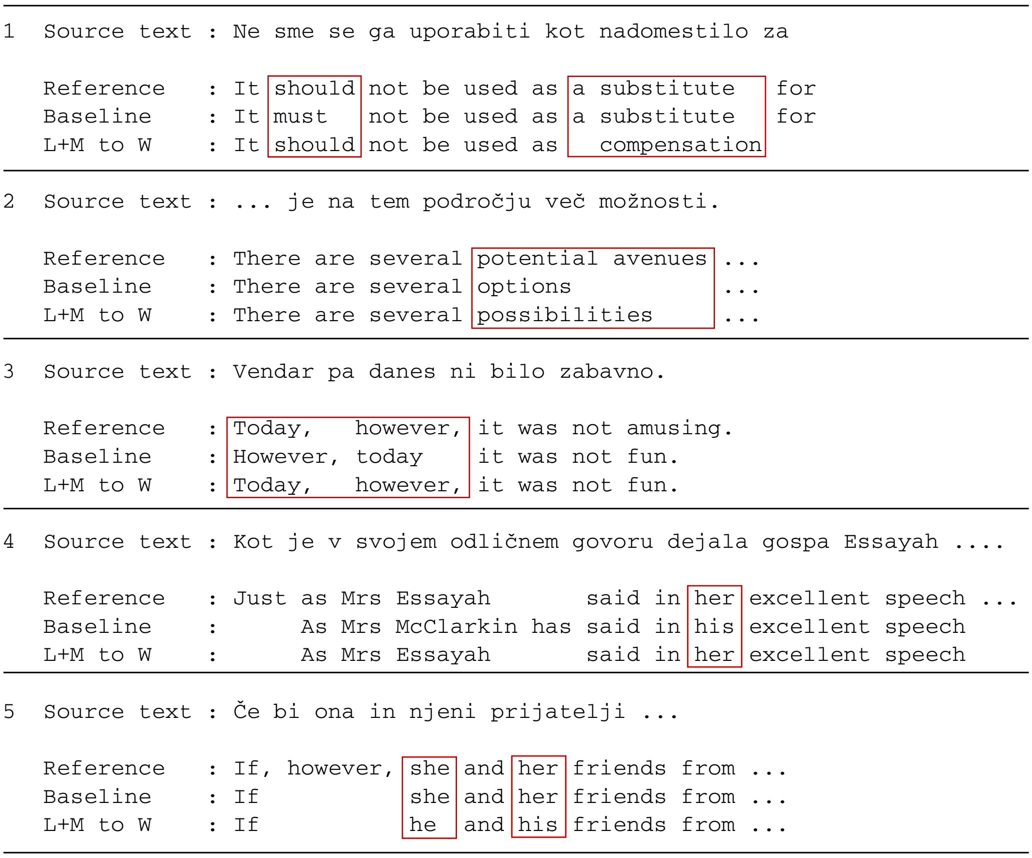
Figure 4. Translation results for the selected translation systems from Slovene to English using ParaCrawl with respect to relative training corpora size.

4.3. Transformer Models

The above-presented results were obtained using LSTM models. The transformer architecture is often found to give better results in NMT. However, our findings were that the results using transformers were slightly worse. When using transformers to translate from English to Slovene, the baseline systems gave a BLEU score of 30.90, and the improved system (W+M to W+M) gave a score of 32.64. When translating from Slovene to English, the baseline systems gave a BLEU score of 34.97, and the improved system (L+M to Words) gave a score of 36.21. Overall, the results were approximately 5 points below the results of the LSTM models. However, we can observe similar improvements and trends as with the LSTM models. Thus, the choice of model architecture does not impact our conclusions. Finally, we decided to use the better-performing models.

4.4. Qualitative Analysis

A comparison between the baseline translations and the translations from the improved systems can give us more insight into the different performances of the two systems. Examples from the translation from Slovene to English can be seen in Figure 5.



```
1  Source text : Ne sme se ga uporabiti kot nadomestilo za

   Reference   : It should not be used as a substitute   for
   Baseline    : It must   not be used as a substitute   for
   L+M to W    : It should not be used as   compensation

2  Source text : ... je na tem področju več možnosti.

   Reference   : There are several potential avenues ...
   Baseline    : There are several options            ...
   L+M to W    : There are several possibilities      ...

3  Source text : Vendar pa danes ni bilo zabavno.

   Reference   : Today,   however, it was not amusing.
   Baseline    : However, today    it was not fun.
   L+M to W    : Today,   however, it was not fun.

4  Source text : Kot je v svojem odličnem govoru dejala gospa Essayah ....

   Reference   : Just as Mrs Essayah       said in her excellent speech ...
   Baseline    :      As Mrs McClarkin has said in his excellent speech
   L+M to W    :      As Mrs Essayah       said in her excellent speech

5  Source text : Če bi ona in njeni prijatelji ...

   Reference   : If, however, she and her friends from ...
   Baseline    : If           she and her friends from ...
   L+M to W    : If           he  and his friends from ...
```

Figure 5. Selected translation examples from Slovene to English with the source text, reference translation, baseline system translation and improved system translation using lemmas and MSD tags (L+M to W).

In many cases, the difference consists of some words in one translation being replaced with synonyms or words with only slightly different meanings. Such are the first and second examples in Figure 5. In the first example, the first marked difference gives a better score to the improved system, while the second difference gives a better score to the baseline system. Other examples we found are this–that, final–last, manage–deal with, firstly–first of all. Similar differences can be seen in the translation from English to Slovene.

In the second example, we see that both systems generated different translations than what is in the reference translation.

In the third example, we see a restructuring of the sentence that results in a better score with the improved system, although both systems give a correct translation.

Examples that refer specifically to morphology in English translation are mostly found in prepositions. This can be seen in the last two examples. In the fourth example, the improved system produces a better translation, and in the fifth example, the baseline system produces a better translation.

With regard to morphology, a more interesting comparison can be done when translating from English to Slovene. Figure 6 shows four examples to illustrate some differences between the baseline and the improved system.

```
1  Source text : ... important incidents are a lot for one man.

   Reference   : ... pomembni pripetljaji so veliko za enega človeka.
   Baseline    : ... pomembni incidenti   so        za en    človek  veliko.
   W+M to W+M  : ... pomembni incidenti   so veliko za enega človeka.

2  Source text : I was very surprised by this judgment ...

   Reference   : To       me je zelo presenetilo ...
   Baseline    : To sodbo me je zelo presenetilo ...
   W+M to W+M  : Ta sodba me je zelo presenetila ...

3  Source text : I could not agree more ...

   Reference   : S tem se povsem            strinjam ...
   Baseline    :       Se ne bi mogla bolj strinjati ...
   W+M to W+M  : S tem se ne bi mogel bolj strinjati ...

4  Source text : The US is one of our international partners ...

   Reference   : ZDA so ena od    naših mednarodnih ...
   Baseline    : ZDA je ena izmed naših mednarodnih ...
   W+M to W+M  : ZDA so ena izmed naših mednarodnih ...
```

Figure 6. Selected translation examples from English to Slovene with the source text, reference translation, baseline system translation and improved system translation using words and MSD tags (W+M to W+M).

In the first example, we have a sentence that ends with "a lot for one man". We can see that the improved system produces a translation that is identical (in the last part of the sentence) to the reference translation. The baseline system, on the other hand, places the word "veliko" (engl: much) at the end of the sentence, which is still grammatically correct. However, the baseline system translates "one man" to "en človek", which wrongly is the nominative case instead of the correct accusative case "enega človeka".

In the second example, we have a sentence with a reference translation that does not include the word for judgment but instead refers only to "this". Both translation systems added the phrase for "this judgment". However, it was added in different cases. The baseline system produced "To sodbo", which can be the accusative or instrumental case (the last of the six cases in Slovene grammar), both of which are grammatically incorrect. The improved system produced the correct nominative form "Ta sodba". There is another difference in the translations—the word "presenetilo" (engl: to surprise) is in the incorrect neutral gender form in the baseline system and in the correct feminine gender form in the improved system.

The third example is interesting from a different angle. The reference translation literally translates to "I agree completely", while both systems translate the phrase "I could not agree more", and the improved system also adds "S tem" (engl: with this). Interestingly, the word "could" is in the baseline system translated to the feminine form "mogla", and in the improved system to the masculine form "mogel". Both translations are grammatically correct, and since the source text is gender neutral, both translations can be considered correct. The reference translation is also gender neutral.

In the fourth example, we have the abbreviation "ZDA" (engl: USA), which is in Slovene used in the plural form. The baseline system produced the singular verb "je" (engl: is), while the improved system gave the correct plural form "so" (engl: are). This example clearly shows the usefulness of MSD tags, as they more often lead to grammatically correct translations. We speculate that this particular translation is improved since the MSD tag for "ZDA" included the information that it is a noun in female form. The baseline translation, on the other hand, may be wrong because "US" is in the source text (and in English in general) treated as a singular noun. Hence the English singular form "is" was translated to the Slovene singular form.

5. Discussion

The results obtained in experiments indicate the role of MSD tags in translation between English and Slovene. The first set of experiments, when translating from English to Slovene, showed that the best performance was achieved with models using words and MSD tags on both sides (W+M to W+M). To generate correct Slovene word forms, morphological information is needed. In Table 8, we see that the improvement on the Europarl corpus is 1.40 BLEU points. Examining Figures 1 and 3, we can see these models start outperforming the baseline models when using 30% of the Europarl training corpora (187,000 segments). They also outperformed the baseline models on the ParaCrawl corpus from the smallest tested size, 5% (186,000 segments), and up to 30% of the full size (1.1 million segments). The results on those corpora sizes are also statistically significant. At larger corpora sizes, these models at some data points also outperform the baseline models. However, the results are no longer statistically significant.

In the first set of experiments, for the translation from Slovene to English, we found the best performance with the models that use lemmas and MSD tags on the source side and words on the target side (L+M to W). To separate the meaning (i.e., lemma) and the morphology (i.e., MSD) on the Slovene side is beneficial, as almost no morphological information is needed to generate the correct English translations. Many different Slovene words are having the same lemma translate to the same English word. Using lemmas and MSD tags instead of words also reduce the data sparsity to a great extent. In Table 9, we see that the improvement on the Europarl corpus is 1.68 BLEU points. Examining Figures 2 and 4, we can see these models outperformed the baseline models at all data points on the Europarl corpus and on the ParaCrawl corpus up to 70% of its full size (2.6 million segments). These results are statistically significant, while the results from 80% of the ParaCrawl corpus upwards are not.

We can also examine the results in Tables 8–11 for format combinations other than the best. We see that models using lemmas on the target side generally perform the worst, regardless of the source side. Additionally, such models require the conversion from lemmas and MSD tags to surface word forms in the post-processing steps, making them more difficult to use. However, when lemmas are used as separate tokens with MSD tags on the source side, the models perform better than the baseline model, except on the ParaCrawl corpus when translating from English to Slovene.

We can also compare results where the MSD tags are separate tokens with the results where they are concatenated either to words or to lemmas. With a few exceptions, combinations with separate tokens perform better. We assume that this is due to the vocabulary sizes. When MSD tags are concatenated with words or lemmas, the vocabulary size increases significantly. Consequently, the model might need to learn more parameters to attain a comparable performance. However, this would mean that the models are no longer comparable and thus not suited for our research.

Next, we can compare results from models that use full MSD tags and their counterparts with reduced MSD tags. The differences are mostly small. However, we can see that in most cases, models with full MSD tags outperform models with reduced MSD tags. The most noticeable exception is in Figure 2 at 10% of the training corpora size. We speculate that this result is due to data sparsity with the very small training corpora size

(60,000 segments). Such corpora sizes might still be found for very specific domains. These results indicate that MSD reduction does not result in significant further improvements in translation performance, or might even lower the performance. Still, a reduction in the complexity of MSD tags to the most important features could have benefits, as such a tagging might be more precise and faster. This would be of importance in practical translation systems, where speed is important.

Our approach to this research consisted of data preparation rather than a modification of the translation architecture and available tools. This has the advantage that the approach can easily be transferred to other language pairs, as appropriate tools for lemmatization and MSD tagging are already available for several morphologically complex languages.

Several examples show that differences between translation systems often consist of producing synonyms or sentences with a different structure. Systems can generate grammatically correct and meaningful translations but receive worse scores with an automatic evaluation metric such as BLEU, as it compares them against gold standard translation. One method to alleviate this problem is the use of manual evaluation. The drawback here is cost and time. This further emphasizes the need for evaluation sets with several possible reference translations for each sentence on the source side.

It is more difficult to draw a conclusion when comparing the results in Figure 1 with the results in Figure 3, i.e., the same translation direction in different domains. All results show a dependency on corpora size, but the size of the Europarl training corpus is about 17% of the size of the ParaCrawl training corpus. We selected the first data point for the ParaCrawl corpus to be 5% of the full size, which is fairly similar to 30% of the Europarl corpus. In both figures, we can see a similar trend in the result up to the full size of the Europarl corpus, which is approximately 17% of the ParaCrawl corpus. Comparing those results in Figures 1 and 3 we can see that the improved system (W+M) outperforms the baseline system by approximate the same amount. The same can be seen in the other translation direction from Figures 2 and 4. Since we compare different domains with different test sets, it is not reasonable to compare exact numbers. From our results, we can only conclude that we found no strong indications for a difference in the inclusion of linguistic knowledge between the domain-specific system and the general domain system.

6. Conclusions

In this research, we presented a systematic comparison of translation systems trained on corpora in a variety of formats and of different sizes. Such a systematic approach for empirical evaluation of a large number of NMT systems has only recently become possible due to the availability of high-performance computing systems. Such systems employ a large number of graphical processing units, which are needed for training NMT systems.

In this research, we were able to show that NMT systems can benefit from additional morphological information if one of the languages in the translation pair is morphologically complex. We were also able to show that those benefits depend on the form in which morphological information is added to the corpora and the translation direction.

We were able to show which combination of corpora formats gives the best results when translating to or from a morphologically complex language. Those conclusions may apply to other language pairs of English and inflected languages. However, for translation pairs consisting of two complex languages, the experiments would have to be repeated to determine the best format on the source and the target side.

Mainly, we were able to confirm that the benefits heavily depend on the size of the training corpora. In the paper, we present empirical evidence for corpora of different sizes. Thus, we were able to give specific corpora sizes, at which the improvements cease to be statistically significant. On the other hand, we found that the same systems that outperform the baseline system on the small domain-specific corpus also outperform the baseline system on the larger general-domain corpus.

Hence, we would argue that the inclusion of morphological information into NMT is mostly beneficial for specific domains or, in general, for language pairs with a small amount of parallel data. However, even when a larger amount of training data is available, translation performance can still be improved.

Our qualitative analysis also showed that not all differences between the systems are recognized as improvements. Further work may include testing such systems with evaluation sets that have several reference translations or include a manual evaluation.

The presented approach for reducing the complexity of MSD tags was based on grammatical knowledge, and it brought only slightly improved translation accuracy. We would argue, however, that the tagging speed in practical applications would benefit from simpler tags. Here, further work may consist of testing data-driven approaches to reduce their complexity, retain translation performance, and increase the tagging speed.

One of the challenges of machine translation is the vocabulary sizes, especially in highly inflected languages. There are methods to alleviate this problem by using word splitting. Future work in this area might include combining the presented addition of linguistic knowledge with data-driven and knowledge-driven approaches for word splitting, e.g., BPE and stem-ending splitting.

Author Contributions: Conceptualization, G.D.; methodology, G.D. and M.S.M.; software, G.D.; formal analysis, G.D. and M.S.M.; writing, G.D. and M.S.M.; visualization, G.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Slovenian Research Agency (research core funding No.P2-0069-Advanced Methods of Interaction in Telecommunications).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. The data can be found here: https://opus.nlpl.eu/Europarl.php (accessed on 22 January 2021) and https://opus.nlpl.eu/ParaCrawl.php (accessed on 22 January 2021).

Acknowledgments: The authors thank the HPC RIVR (www.hpc-rivr.si, accessed on 31 January 2022) consortium for the use of the HPC system VEGA on the Institute of Information Science (IZUM). They also want to thank the authors of the Europarl and ParaCrawl parallel corpora.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Koehn, P.; Hoang, H. Factored translation models. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prague, Czech Republic, 28–30 June 2007; pp. 868–876.
2. Kalchbrenner, N.; Blunsom, P. Recurrent continuous translation models. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 1700–1709.
3. Cettolo, M.; Jan, N.; Sebastian, S.; Bentivogli, L.; Cattoni, R.; Federico, M. The IWSLT 2015 evaluation campaign. In Proceedings of the 12th International Workshop on Spoken Language Translation, Da Nang, Vietnam, 3–4 December 2015; pp. 2–14.
4. Junczys-Dowmunt, M.; Dwojak, T.; Hoang, H. Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions. In Proceedings of the 9th International Workshop on Spoken Language Translation (IWSLT), Hong Kong, China, 6–7 December 2016.
5. Bentivogli, L.; Bisazza, A.; Cettolo, M.; Federico, M. Neural versus Phrase-Based Machine Translation Quality: A Case Study. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 257–267. [CrossRef]
6. Castilho, S.; Moorkens, J.; Gaspari, F.; Calixto, I.; Tinsley, J.; Way, A. Is neural machine translation the new state of the art? *Prague Bull. Math. Linguist.* **2017**, *108*, 109–120. [CrossRef]
7. Arčan, M. A Comparison of Statistical and Neural Machine Translation for Slovene, Serbian and Croatian. In Proceedings of the Conference on Language Technologies & Digital Humanities 2018, Ljubljana, Slovenia, 20–21 September 2018; pp. 3–10.
8. Vintar, Š. Terminology Translation Accuracy in Phrase-Based versus Neural MT: An Evaluation for the English-Slovene Language Pair. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018; Du, J., Arcan, M., Liu, Q., Isahara, H., Eds.; European Language Resources Association (ELRA): Paris, France, 2018.

9. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to Sequence Learning with Neural Networks. In Proceedings of the 27th International Conference on Neural Information Processing Systems, NIPS'14, Montreal, QC, Canada, 8–13 December 2014; MIT Press: Cambridge, MA, USA, 2014; Volume 2, pp. 3104–3112.
10. Luong, T.; Pham, H.; Manning, C.D. Effective Approaches to Attention-based Neural Machine Translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics, Lisbon, Portugal, 17–21 September 2015; pp. 1412–1421. [CrossRef]
11. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
12. Dyer, C. The "noisier channel": Translation from morphologically complex languages. In Proceedings of the Second Workshop on Statistical Machine Translation, Prague, Czech Republic, 23 June 2007; pp. 207–211.
13. El Kholy, A.; Habash, N. Translate, predict or generate: Modeling rich morphology in statistical machine translation. In Proceedings of the 16th Annual Conference of the European Association for Machine Translation, Trento, Italy, 28–30 May 2012; pp. 27–34.
14. Minkov, E.; Toutanova, K.; Suzuki, H. Generating complex morphology for machine translation. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, 23–30 June 2007; pp. 128–135.
15. Toutanova, K.; Suzuki, H.; Ruopp, A. Applying morphology generation models to machine translation. In Proceedings of the ACL-08: HLT, Columbus, OH, USA, 12–14 June 2008; pp. 514–522.
16. Chahuneau, V.; Schlinger, E.; Smith, N.A.; Dyer, C. Translating into morphologically rich languages with synthetic phrases. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 1677–1687.
17. Sennrich, R.; Haddow, B. Linguistic Input Features Improve Neural Machine Translation. In Proceedings of the First Conference on Machine Translation, Berlin, Germany, 11–12 August 2016; Research Papers; Association for Computational Linguistics: Berlin, Germany, 2016; Volume 1, pp. 83–91. [CrossRef]
18. García-Martínez, M.; Barrault, L.; Bougares, F. Factored Neural Machine Translation Architectures. In Proceedings of the 13th International Conference on Spoken Language Translation; International Workshop on Spoken Language Translation, Seattle, WA, USA, 8–9 December 2016.
19. Garcia-Martinez, M.; Barrault, L.; Bougares, F. Neural machine translation by generating multiple linguistic factors. In Proceedings of the International Conference on Statistical Language and Speech Processing, Le Mans, France, 23–25 October 2017; pp. 21–31.
20. Dalvi, F.; Durrani, N.; Sajjad, H.; Belinkov, Y.; Vogel, S. Understanding and improving morphological learning in the neural machine translation decoder. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Taipei, Taiwan, 27 November–1 December 2017; pp. 142–151.
21. Sennrich, R.; Haddow, B.; Birch, A. 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), Berlin, Germany, 7–12 August 2016; Association for Computational Linguistics: Berlin, Germany, 2016; pp. 1715–1725. [CrossRef]
22. Passban, P. Machine Translation of Morphologically Rich Languages Using Deep Neural Networks. Ph.D. Thesis, Dublin City University, Dublin, Ireland, 2018.
23. Passban, P.; Liu, Q.; Way, A. Improving Character-Based Decoding Using Target-Side Morphological Information for Neural Machine Translation. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, LA, USA, 1–6 June 2018; Association for Computational Linguistics: New Orleans, LA, USA, 2018; pp. 58–68. [CrossRef]
24. Burlot, F.; Garcia-Martinez, M.; Barrault, L.; Bougares, F.; Yvon, F. Word representations in factored neural machine translation. In Proceedings of the Second Conference on Machine Translation, Copenhagen, Denmark, 7–8 September 2017; pp. 20–31.
25. Stafanovičs, A.; Bergmanis, T.; Pinnis, M. Mitigating gender bias in machine translation with target gender annotations. *arXiv* **2020**, arXiv:2010.06203.
26. Tiedemann, J.; Thottingal, S. OPUS-MT—Building open translation services for the World. In Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT), Lisbon, Portugal, 3–5 November 2020.
27. Krek, S.; Dobrovoljc, K.; Erjavec, T.; Može, S.; Ledinek, N.; Holz, N.; Zupan, K.; Gantar, P.; Kuzman, T.; Čibej, J.; et al. Training Corpus ssj500k 2.3, 2021. Slovenian Language Resource Repository CLARIN.SI. Available online: http://hdl.handle.net/11356/1434 (accessed on 22 January 2021).
28. Erjavec, T.; Fišer, D.; Krek, S.; Ledinek, N. The JOS Linguistically Tagged Corpus of Slovene. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, 17–23 May 2010; Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D., Eds.; European Language Resources Association (ELRA): Valletta, Malta, 2010.
29. Koehn, P. Statistical Significance Tests for Machine Translation Evaluation. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 25–26 July 2004; Association for Computational Linguistics: Barcelona, Spain, 2004; pp. 388–395.

30. Koehn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R.; et al. Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, Prague, Czech Republic, 25–27 June 2007; Association for Computational Linguistics: Prague, Czech Republic, 2007; pp. 177–180.
31. Post, M. A Call for Clarity in Reporting BLEU Scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, Brussels, Belgium, 31 October–1 November 2018; Association for Computational Linguistics: Brussels, Belgium, 2018; pp. 186–191. [CrossRef]
32. Schmid, H. Probabilistic Part-of-Speech Tagging Using Decision Trees. In Proceedings of the International Conference on New Methods in Language Processing, Yokohama, Japan, 18–22 September 1994.
33. Junczys-Dowmunt, M.; Grundkiewicz, R.; Dwojak, T.; Hoang, H.; Heafield, K.; Neckermann, T.; Seide, F.; Germann, U.; Fikri Aji, A.; Bogoychev, N.; et al. Marian: Fast Neural Machine Translation in C++. In Proceedings of the ACL 2018, System Demonstrations, Melbourne, Australia, 15–20 July 2018; Association for Computational Linguistics: Melbourne, Australia, 2018; pp. 116–121.