Global Under-Resourced MEdia Translation (GoURMET)

H2020 Research and Innovation Action
Number: 825299
D2.1 – Initial progress report on modelling morphological structure

Nature	Report	**Work Package**	WP2
Due Date	30/6/2020	**Submission Date**	30/6/2020
Main authors	Rachel Bawden (UEDIN)		
Co-authors	Wilker Aziz (UVA), Víctor Sánchez-Cartagen (UA), Alexandra Birch (UEDIN)		
Reviewers	Ivan Titov		
Keywords	morphology, alignment, latent modelling		
Version Control			
v0.1	**Status**	Draft	25/6/2020
v1.0	**Status**	Final	29/6/2020

Contents

List of Figures

Abstract

This deliverable reports the work carried out in the first half of the GoURMET project under Work Package 3: Modelling morphological structure. The work package is divided into three main tasks, looking at different aspects of morphology or different techniques that can be used to exploit morphological structure to improve the quality of machine translation (MT) systems, particularly for low resource and morphologically rich languages. Task 1 is dedicated to linguistically-informed NMT models using morphological information and intuitions. Task 2 is dedicated to future work on jointly learning alignments between parallel sentences and morphology, as a way of inducing subword segmentation. Finally, Task 3 comprises work on encoding features of morphology as latent features within the model as a way of discovering and exploiting such information.

1 Introduction

In low resource translation settings, learning to better exploit available data is crucial to improving the performance of machine translation (MT) systems. Two major consequences of limited training data are (i) a lack of vocabulary coverage (i.e. many words may not have previously been seen during training) and (ii) limited learning for vocabulary items that have been seen few times during training. This situation is exacerbated when dealing with languages that are highly inflectional (i.e. they have rich morphology with many inflected forms for a single lemma); they have a higher type-token ratio, meaning that there are more distinct vocabulary items to cover and therefore generally more training data is required than for languages that are morphologically poor.

A way of reducing this problem is to work with the compositionality and patterns of the inflected forms of a language to provide a better capacity to generalise across wordforms and ultimately provide a better coverage of the languages' vocabulary and of the role of words within a sentence. There are two main ways in which morphology can be used to improve the generalisation capacity of MT models: (i) through linguistically informed segmentation of words into their component parts (e.g. corresponding to linguistic morphs) such that subwords are efficiently shared amongst inflected versions of the same lemma,[1] and (ii) using or inducing representations of morphological information associated with words as a way of abstracting away from the wordforms and learning patterns. Within the GoURMET project, we look at how both types of strategy can be used to enhance the translation of low resource language pairs when applied to neural MT (NMT).

The work completed so far can be structured into the following three tasks (corresponding to the tasks drawn up in the project proposal):

Task 1 **Developing linguistically informed NMT models using morphology**
This task is dedicated to the development of strategies to induce linguistically plausible strategies of segmentation and to the integration of morphological information or morphologically guided intuitions in either the source or target language. The scope of this task has expanded slightly since the initial proposal and has more emphasis on using external morphological information and intuitions to improve NMT.

Task 2 **Jointly learning alignments and morphology**
The goal of this task is to approach the learning of morphological segmentation by exploiting parallel data in multiple languages. By jointly learning to align words or subwords in the sentence-aligned parallel data, morphological alignments can be induced to develop patterns of usage across languages.

Task 3 **Exploit factors encoding latent features of morphology**
Within this task, we aim to induce morphological patterns from the training data during the training of the MT models by encoding morphological features as latent factors. Instead of being provided as additional features as in Task 1, here, the features are induced during the learning of the NMT model.

There has been progress particularly in Tasks 1 and 3, for which there are three pieces of work each, some of which have led to peer-reviewed publications (See Section 5 for publications associated

[1] This is in opposition to current statistically motivated subword segmentation strategies such as byte pair encoding (BPE; Sennrich et al. 2016), which are not morphologically grounded.

with this work package). Research related to Task 2 has started recently and will therefore be expanded on in the final review of this work package. Individual directions for future work are presented in the individual sections where relevant.

2 Task 1: Linguistically informed NMT models using morphology

The research carried out in the context of this task looks at how to exploit linguistically motivated information and structure in NMT models, through the use of external tools and in the design of the model architecture. The first two pieces of research (Sections 2.1 and 2.2) are related to segmentation strategies in NMT; the challenge of representing the theoretically infinite vocabulary of a language using a finite NMT vocabulary. The third piece of research (Section 2.3) looks at the injection of morphological information into NMT: which type of information is most useful and how to best integrate it into the model to improve translation.

2.1 Morphological segmentation using Apertium (UA)

It is common practice in NMT to segment words into smaller units in order to represent any word in a language using a fixed-size vocabulary. One of the most commonly used strategies is byte pair encoding (BPE) (Sennrich et al., 2016), consisting in merging sequences of characters that more commonly appear in succession. However, it is a statistically driven strategy rather than a linguistically motivated one, potentially limiting its capacity to effectively create valid patterns across subword units. Morphological segmentation on the other hand is a strategy for segmenting words into subword units that consists in splitting them into a *stem*, which carries the meaning of the word, and a *suffix* (or sequence of suffixes) containing morphological and syntactic information. This strategy has been found to outperform the more commonly used BPE segmentation strategy when used to train NMT models for highly inflected languages such as Finnish (Sánchez-Cartagena and Toral, 2016), German (Huck et al., 2017) and Basque (Sánchez-Cartagena, 2018). In these cases, morphological segmentation can allow the NMT system to better generalise the observed evidence, since the core meaning of words and their grammatical properties are encoded in different tokens.

Unfortunately, morphological segmenters may not be available for many under-resourced languages. Although morphological analysers, which are more generally available, could be modified to perform morphological segmentation, they may have poor coverage for under-resourced languages, i.e. they may be able to provide an analysis/segmentation for only a small proportion of words in a given text. We address these issues in the Universitat d'Alacant's submission to the WMT 2019 news translation shared task (Barrault et al., 2019) for the English–Kazakh language pair (Sánchez-Cartagena et al., 2019), where we use morphological segmentation for Kazakh. We follow a hybrid morphological segmentation algorithm that combines (i) an existing morphological analyser with (ii) knowledge from a corpus. This hybrid strategy enables us to overcome the coverage issues of the morphological analyser and provide morphological segmentation for all words in the corpus.

Hybrid morphological approach to word segmentation We use the Apertium Kazakh morphological analyser[2] to generate candidates for subword segmentation based on the morphological analyses generated. These analyses are then disambiguated using a semi-supervised learning approach and the Morfessor toolkit (Virpioja et al., 2013).

For each word, the Apertium morphological analyser provides a set of candidate analyses, each made up of a lemma and morphological information. Those analyses in which the lemma is a prefix of the word are considered valid analyses for segmentation. When this is the case, it means that the word can be morphologically segmented into two parts: the lemma and the remainder of the word, a strategy that can safely be applied in Kazakh since the stem usually corresponds to the lemma.[3] For instance, in the first example in Table 1, the Kazakh word университетінің has a single analysis whose lemma is университет 'university'. As the lemma is a prefix of the word, it is morphologically segmented as университет plus інің. When a word has no valid analysis for segmentation, we generate as many segmentation candidates as there are known suffixes that match the word (plus the empty suffix, since a possible option could be no segmentation at all). Known suffixes are extracted in advance from the words for which a single valid analysis is found.

When there are multiple segmentation candidates (either because they come from multiple valid analyses or from suffix matching), they need to be disambiguated. We decide on the best segmentation for these words by applying semi-supervised morphology learning methods to a monolingual corpus. We choose to use Morfessor, a family of methods for automatically learning morphological segmentation based on the principle of minimum description length (**?**); the words in a corpus are split into morphs such that the size of the morph vocabulary and the length in tokens of the corpus are minimised. We use a semi-supervised variant of Morfessor in which the segmentation model can be estimated from a raw corpus and a set of already segmented words (Kohonen et al., 2010). A similar approach was followed by Sánchez-Cartagena (2018) to exploit a spellchecker for morphological segmentation. A Morfessor model was trained on all available Kazakh corpora for the WMT shared task with the supervision of those words from the corpus with a single valid analysis. Note that the words used as a supervised input to Morfessor do not need to be segmented by the model, since their segmentation obtained from the Apertium morphological analysis is always preferred.

Finally, as suggested by Huck et al. (2017), BPE splitting is applied on top of morphological segmentation with a model learned on the concatenation of all training corpora. Applying BPE to further split the subword units obtained after morphological segmentation helps the system to transliterate proper nouns, translate compounds and to control vocabulary size. As the suffixes are rather regular and frequent, usually only the lemmas are affected by BPE splitting. As an estimation of the amount of segmentation carried out in each step, we provide token counts for the Kazakh corpus used to train the system submitted to the WMT shared task: it contained 219 million tokens before any segmentation, 298 million tokens after morphological segmentation, and 319 million tokens after the BPE splitting applied on top of morphological segmentation. The BPE model was trained with 50,000 operations.

[2] Available at https://wiki.apertium.org/wiki/Apertium-kaz. The size of the dictionary of this morphological analyser is smaller than that of other languages available in the Apertium project: it contains around 35 000 entries, while the English one contains 55 000.

[3] This cannot be used in all languages, for example Romance languages. Consider, for instance, the verb *cantar* 'sing' in Spanish. The first person singular form of the present tense of *cantar* is *canto*, whose stem is *cant-* and not *cantar*.

Results and analysis Preliminary experiments carried out using the English–Kazakh parallel corpus made available for the shared task showed that the proposed morphological segmentation strategy outperforms plain BPE segmentation by over 1 chrF++ point (Popović, 2017). These results are supported by the fact that the final submitted system was ranked 1st according to the shared task human evaluation.

Word	Analyses	Morph. seg.	Plain BPE
университетінің	университет-n.px3sp.gen	университет@@ інің	университетінің
жасалмайды	жаса-v.tp.n.p3 жасал-v.i.n.p3*	жасал@@ майды	жас@@ алмайды

Table 1: Examples of Kazakh words, their morphological analyses according to Apertium and their segmentation.

Some examples of Kazakh words with their morphological analysis according to Apertium and morphological segmentation (prior to BPE splitting) are given in Table 1 (subword boundaries are marked using @@).The first word is the genitive form of университет 'university'. Morphological segmentation allows the NMT system to generalise to other inflected forms of the same word (with the same stem), whereas BPE does not split it because it is a frequent term in the corpus. The second word is an inflected form of the verb жаса 'to do'. It is also analysed as a inflected form of жасал, a word that does not exist, due to an error in the Apertium analyser. The Morfessor model preferred the wrong analysis. However the plain BPE segmentation makes translation even more difficult for the MT system by choosing the prefix жас 'young', introducing more ambiguity, as the token жас can encode both the verb *to do* and the adjective *young*.

2.2 Hierarchical modelling of word boundaries for NMT (UEDIN, UVA)

In neural machine translation it is standard practise to break words into sub-word units to avoid problems with limits on vocabulary size. Recent studies have shown that the idea of segmenting words into consistent components can be done directly at the character level (i.e. using a much smaller vocabulary consisting of individual characters) (Cherry et al., 2018; Luong and Manning, 2016). It has been shown to deliver translation accuracy on par with subword-based segmentation, with the caveat that it comes at a cost of having to use deeper networks and endure longer training times. In order to provide a more computationally efficient solution, we investigate the importance of maintaining word boundaries in character-level NMT by means of a hierarchical decoding architecture, where translations are subsequently generated at the level of words and characters. We evaluate our method against conventional open-vocabulary NMT methods from English into five morphologically rich languages, and show that our model can reach higher translation accuracy using significantly fewer parameters, while demonstrating a better capacity to learn to represent longer distance context and grammatical dependencies. The full description of this work can be found in the publication that we presented at the Workshop on Neural Generation and Translation 2019 (Ataman et al., 2019).

Hierarchical decoding We propose character-level decoding in NMT by modelling translation through a hierarchical architecture (Luong and Manning, 2016). In this architecture, the input

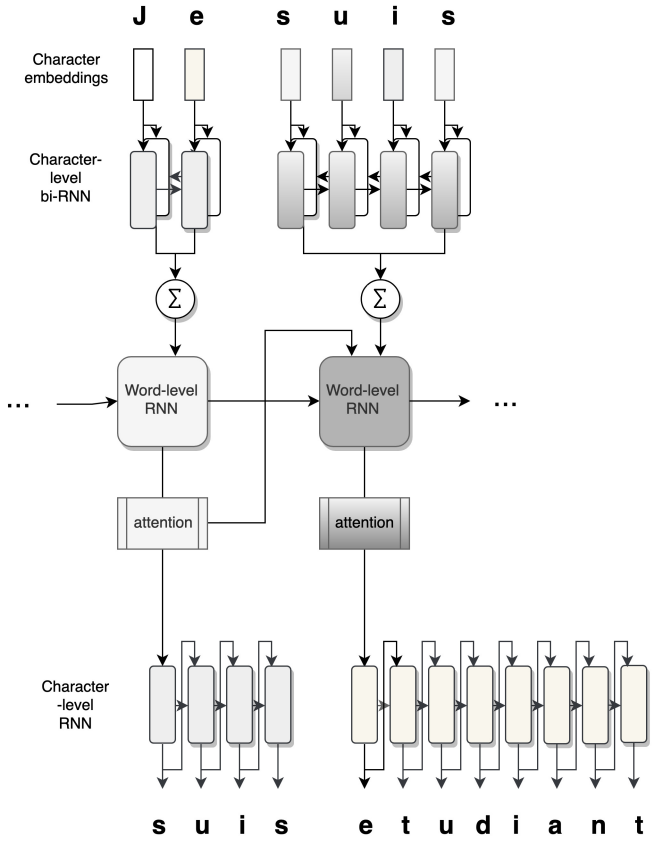

Figure 1: Hierarchical NMT decoder: input words are encoded as character sequences and the translation is predicted at the level of words. The output words are generated as character sequences. Character-level NMT decoder: the next token in the sentence is predicted by computing the attention weights and the target context successively for each character in the sentence.

embedding layer of the decoder is augmented with a character-level bi-RNN, which estimates a composition function over the embeddings of the characters in each word to compute distributed representations of target words in the sentence, as illustrated in Figure 1.

We test the architecture on the translation of English into five languages from different language families and exhibiting distinct morphological typologies: Arabic (*templatic*), Czech (*mostly fusional, partially agglutinative*), German (*fusional*), Italian (*fusional*) and Turkish (*agglutinative*). We use the TED Talks corpora (Cettolo et al., 2012), which range from 110K to 240K sentences. The low-resource setting for the training data allows us to examine the quality of the internal representations learned by each decoder under high data sparseness.

Results and analysis The models are evaluated using the standard automatic evaluation metric BLEU (Papineni et al., 2002). The results of the experiments given in Table 2 show that the hierarchical decoder can reach a translation performance comparable to or better than the NMT model based on subword units in all languages, while using almost three times fewer parameters. The improvements are especially evident in Arabic (AR) and Turkish (TR), where the hierarchical decoder improves by +0.82 and +0.89 BLEU respectively. In Czech (CS), Italian (IT) and German (DE), which are fusional languages, the performance of the two decoders is generally comparable.

Model	Output Units	BLEU					Avg.	Avg.
		AR	CS	DE	IT	TR	# Params	Conv. Time
Linear	Subwords	14.67	16.60	**24.29**	26.23	8.85	22M	10.63
Linear	Characters	12.72	**16.94**	22.23	24.33	**10.65**	7.3M	23.40
Hierarchical	Characters	**15.55**	16.79	23.91	**26.64**	9.74	7.3M	19.60

Table 2: Results of the experiment under low-resource settings The average convergence time is the average number of epochs until convergence. The average numbers of parameters are calculated only for the decoders of the NMT models (in millions). The best scores for each translation direction are in bold font. All improvements over the baselines are statistically significant (p-value < 0.01).

In Czech (CS), the hierarchical model outperforms the subword-based model by 0.19 BLEU and in Italian by 0.41 BLEU. The subword-based NMT model achieves the best performance in German, a language that is rich in compounding, where explicit word segmentation is likely to be be useful in increasing the translation accuracy.

The fully character-level NMT model, on the other hand, outperforms the hierarchical model in Turkish by 0.91 BLEU and in Czech by 0.15 BLEU. These two directions constitute the most sparse settings. The improvements are proportional to the amount of sparsity in two languages, as shown by the type-token ratios in the training corpora; Turkish has the highest amount of sparsity, followed by Czech. In the case of high lexical sparsity, learning to translate based on character-level representations might aid to reduce contextual sparsity, allowing to translate rare or unseen words more accurately.

2.3 Integration of linguistic information into NMT (UA)

In under-resourced settings, additional sources of information in the form of relevant linguistic *factors* (word-level annotations, such as part-of-speech (PoS), morphological or syntactic tags) have proved their usefulness in boosting translation performance. These factors may be integrated in the source language (Sennrich and Haddow, 2016) or in the target language (García-Martínez et al., 2016), the latter opening the door to the introduction of multi-task learning techniques (McCann et al., 2018; Luong et al., 2016; Niehues and Cho, 2017).

Currently, when building an NMT system for a new language pair, it is difficult to know which type of linguistic factor is the most appropriate and which mechanism is the most effective for integrating it into the system. The literature only provides partial results and the conclusions are often contradictory. For instance, while Nadejde et al. (2017) successfuly combine factors with surface forms in the target language, Tamchyna et al. (2017) claim that the introduction of PoS and morphological information in the target language is only useful when it is combined with lemmatisation. Yang et al. (2019) concludes that target language PoS information boosts translation quality with a carefully designed architecture, but (Wagner, 2019) find that target language PoS and morphological information does not bring any advantage.

Systematic comparison of the use of morphological information With the aim of clarifying the role of the linguistic annotation of words in NMT, we carried out a systematic study (Sánchez-Cartagena et al., 2020). The study covers the following dimensions: eight language

pairs, three training corpus sizes, two NMT architectures: Transformer (Vaswani et al., 2017) and recurrent (Bahdanau et al., 2015), three types of tag: dummy (with no linguistic information at all), PoS tags and morphological tags. Moreover, the impact of each type of tag is studied when it is integrated either in the source or in the target language. The linguistic tags are integrated by means of the *interleaving* approach (Nadejde et al., 2017) (tags are treated as additional tokens following the word they qualify), which provides a straightforward framework for comparing the different types of linguistic information and NMT architectures.

Main results and conclusions We evaluate the resulting translations using automatic evaluation metrics and analyse the translation performance with respect to sentence length and automatic error classification. In line with existing results in the literature, the study shows that source language tags, which help to obtain more accurate representations of the source language sentence, are helpful regardless of training corpus size, language pair, and NMT architecture. Moreover, no consistent differences were found between PoS and morphological information. On the contrary, results for target language tags were more meaningful: PoS tags systematically outperform morphological tags in terms of automatic evaluation metrics, despite the fact that the addition of morphology leads to a more grammatical output. Finally, the combination of tags in both the source language and the target language further improves the results.

The cause of the performance degradation observed when introducing morphological information in the target language seems to be related to data sparseness; additional experiments show that predicting target language PoS and target language morphological information together in a single tag degrades the accuracy of the PoS tag prediction. This suggests that predicting the target language PoS tag of each target language word should be a different task from predicting its morphological information in order to optimise the use of target language morphological information in NMT.

Results also show that interleaved tags are not the optimal way of introducing target language linguistic information into Transformer systems, as the introduction of tags without content (i.e. an identical tag is introduced before every word) degrades translation quality. That degradation is not observed for recurrent systems, which suggest that they are more tolerant to the introduction of additional information in the target-language stream. This is compatible with the results of the automatic error analyses, which show that transformer systems make consistently more reordering errors. Another empirical observation that supports this hypothesis is the fact that the gain brought by the added linguistic information only scales to large data scenarios (e.g. the English–German pair of the WMT news translation shared task) when using the recurrent architecture.

Future work We are currently following a line of research aiming to extend these analyses. It is based on the following problem formalisation: training an NMT system with interleaved target language tags can be seen as a multi-task learning problem. The two tasks are: prediction of target language surface forms and prediction of target language tags. The interleaving approach involves: i) sharing all the parameters of the network between both tasks; and ii) giving exactly the same weight to both tasks. It is worth studying whether there is an optimal separation between the tasks by specialising some parts of the network on a specific task and dynamically adapting the weight of the two subtasks (Chen et al., 2018; Jean et al., 2018).

3 Task 2: Jointly learning alignments and morphology

The research directions proposed in this second task concern using information from alignments between parallel sentences to guide morphological segmentation, by training a joint generative model of segmentation and parallel data. This research has just been started and will therefore be presented in the final report. However, we present a summary of the proposal in Section 3.1.

3.1 Segmentation models informed by alignment (UVA)

We aim to use translation data (i.e. sentence-aligned parallel corpora) to induce subword units without direct supervision for segmentation. The idea is to combine models of unsupervised segmentation, such as the model of Kawakami et al. (2019), and models of unsupervised alignment (such as the model presented in D3.1, Task 2).

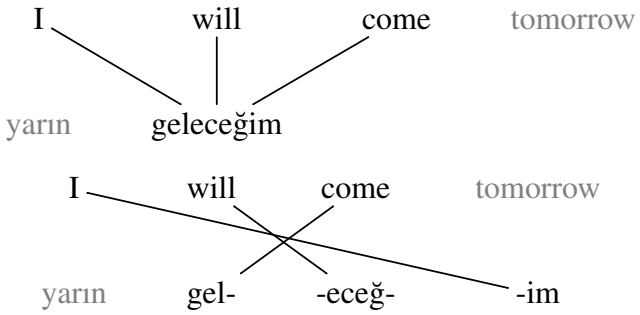

Figure 2: Two ways of generating a complex Turkish wordform from its English translation. Top: we condition on the English tokens and generate the Turkish word *geleceğim* by transforming the 3 English tokens it aligns to (suppose we are given alignment information). This generation can be realised, for example, with a character-based model. Alternatively (bottom), we can generate the Turkish word one segment at a time, in which case, each segment is independently supported by a smaller set of English token (one each, in this case). Clearly, neither segmentation nor alignments are directly observable. The example is meant to illustrate that word segmentation and word alignment are unsupervised problems that interact when modelling translation data.

Take the example of parallel data involving an agglutinative language, such as Turkish, and an analytic language, such as English (see Figure 2 for an example). Due to differences in morphology, we expect a Turkish token to align to multiple English tokens: this fine-grained alignment information is a valuable inductive bias for segmentation of Turkish words. Intuitively, the translation bias injected by the alignment component helps ground segmentation decisions leading to subword units that are more likely to capture linguistic features. Using an alignment model rather than a full MT system will allow us to explore richer inductive bias with smaller models.

4 Task 3: Factors encoding latent features of morphology

This final task is dedicated to models in which morphological attributes are induced as latent variables while training towards a downstream learning signal such as word generation or translation. This lies in contrast to the models in Task 1, for which the morphological information was provided more explicitly to the models.

Three pieces of research correspond to this task, concerning the latent modelling of morphological features. The first (Section 4.1) involves designing a generative model of inflected wordforms in a semi-supervised way, applied to a single language. This work is then revisited and extended in Section 4.2 to apply it to MT (rather than in a monolingual setting), where the decoder must inflect every single word in the target sequence. In this second work, latent morphology is induced in an entirely unsupervised way, rather than being partly supervised as in the first work. Finally, we present a variational NMT model with morphological priors (Section 4.3), which introduces latent modelling of morphology as in the second work, but is also designed to benefit from some degree of supervision for the morphological features, rather than them being entirely unsupervised.

4.1 Generative models of inflected wordforms (UVA)

The goal of this research is to design models that can generated complex inflected wordforms. In this section, we study models that generate words in isolation and in a monolingual setting, and in Section 4.2 we extend this work to NMT, where words must be generated in context. To study word generation in a monolingual setting, in this section, we approach the task of morphological reinflection.

Morphological reinflection We use data from the *morphological reinflection task* organised by SIGMORPHON 2016 (Cotterell et al., 2016). In this task, for a given language, we are given a *source* word $x^{(s)}$ and *target* word $x^{(t)}$ (corresponding to two inflected forms of the same lemma), and a collection of discrete features y that describe the morphological attributes of $x^{(t)}$. Two such examples are shown in Table 3. The task corresponds to learning to map $\langle x^{(s)}, y \rangle$ to $x^{(t)}$, i.e. learning to reinflect the source into $x^{(t)}$. Note that $x^{(s)}$ is a form of indirect supervision for lemmas, that is, we can think of this problem as learning to represent the *lemma* these wordforms share, which we can do from $x^{(s)}$ alone, and transform this lemma representation into $x^{(t)}$ by realisation of the morphological features in y.

Lang	Source form ($x^{(s)}$)	Features (y)	Target form ($x^{(t)}$)
English	ran	PresPart	running
Spanish	digo	Future2S	dirás

Table 3: Two examples (one English and one Spanish) from the SIGMORPHON 2016 reinflection task. $x^{(s)}$ corresponds to an inflected form of the same lemma as $x^{(t)}$ to be predicted and the features y refer to the morphological attributes that define the inflected target word.

Motivation In NMT, which is the end goal of this work, target-language wordforms are generated one BPE subword at a time. Instead, we envision generating wordforms with a morphologically inspired model that infers a representation of a word's lemma from the available information (source-language sentence and target-language prefix) as well as the morphological attributes that should govern the generation of the inflected wordform one character at a time. In NMT we will not have a related wordform $x^{(s)}$, but rather the source-language sentence and the target prefix which implicitly constrain the possible lemmas, similarly, we will not have a vector of observed morphological attributes y, but rather we will have to infer them from the available (unlabelled)

data. While we do not apply the morphological model to NMT in this section (See Section 4.2 for its application to NMT), we approach the task with this final goal in mind, and therefore we adapt an existing model (MSVED; Zhou and Neubig, 2017), making it more convenient for NMT.

MSVED Zhou and Neubig (2017) introduce the multi-space variational auto-encoder (MSVED), a generative model of inflected wordforms with continuous latent variables to represent lemmas (or the lexical semantics of a word) and (approximately) discrete variables to capture morphological attributes. MSVED uses $x^{(s)}$ to predict a continuous representation z of the lemma of $x^{(t)}$, and then uses the lemma along with morphological attributes y to generate $x^{(t)}$. Where y is not available, for example because it is not annotated in the dataset, MSVED predicts it from $x^{(t)}$ using a variational auto-encoder objective (VAE; Kingma and Welling, 2014). This combines supervised and unsupervised learning into a semi-supervised VAE (Kingma et al., 2014). The model contains a number of components, namely a generator of inflected wordforms $p(x^{(t)}|z, y, \theta)$, where y correspond to a vector of morphological attributes and z a continuous embedding of the lemma of $x^{(t)}$, a classifier $q(y|x^{(t)}, \phi)$ that recognises morphological attributes y in the inflected form $x^{(t)}$, and a mechanism $q(z|x^{(s)}, \phi)$ to infer representation of lemmas. Note that for the lemma, MSVED uses the related wordform $x^{(s)}$ rather than $x^{(t)}$ directly, thus MSVED can also be seen as a combination of an encoder-decoder and a variational auto-encoder. The model is trained via variational inference (VI; Jordan et al., 1999) by maximising the evidence lowerbound (ELBO) given a dataset $\mathcal{D}$ of labelled observations, each of the form $\langle x^{(s)}, x^{(t)}, y\rangle$, and a dataset $\mathcal{U}$ of unlabelled observations, each of the form $\langle x^{(s)}, x^{(t)}\rangle$. In the supervised case, it uses the reparameterisation trick of Kingma and Welling (2014) to backpropagate through lemma samples. In the unsupervised case, additionally, we need to backpropagate through samples from a discrete distribution, namely, the distribution over morphological attributes of $x^{(t)}$. As this is not generally possible, Zhou and Neubig (2017) use a continuous relaxation known as Concrete distribution (Maddison et al., 2017; Jang et al., 2017), or Gumbel-Softmax, together with a biased proxy gradient known as straigh-through estimator (STE; Bengio et al., 2013). Figure 6 illustrates the architecture (though with some components adapted as we discuss next).

Sparse relaxation to a latent factor model The design of MSVED is rather elegant, yet it makes a few assumptions that we revisit. First, the classifier exploits knowledge about the morphological attribute space that we wish to remove. That is because as we shall see, this knowledge will not be available when we adapt this component to be an integral part of an NMT model. Essentially, the classifier network in MSVED treats each type of morphological attribute separately (e.g. verb features are separate from nominal features, etc.). Instead, we would like to model all features together using a collection of binary attributes as in classic latent feature models (Ghahramani and Griffiths, 2006). This sidesteps the need for careful specification of morphological attributes, which is convenient when we have to discover them unsupervisedly from data.[4] Secondly, the inference model in MSVED is trained with a biased proxy gradient known as straight-through estimator (STE; Bengio et al., 2013). In order to make this component an integral part of a large and complex NMT architecture, we prefer to work with unbiased gradient estimates, as biased gradients violate a formal requirement of stochastic optimisation (Robbins and Monro, 1951; Bottou and Cun, 2004). Here we use the sparse relaxation to binary variables developed in WP3 (see D3.1, Task 2) and presented in (Bastings et al., 2019), which admits unbiased and differentiable

[4] Of course, should enough information be available, we can split features into different collections as they do.

sampling. Thirdly, we would also like to investigate how well we can do without the related source form $x^{(s)}$, since this would be closer to the realistic scenario where we inflect target-language wordforms in NMT.

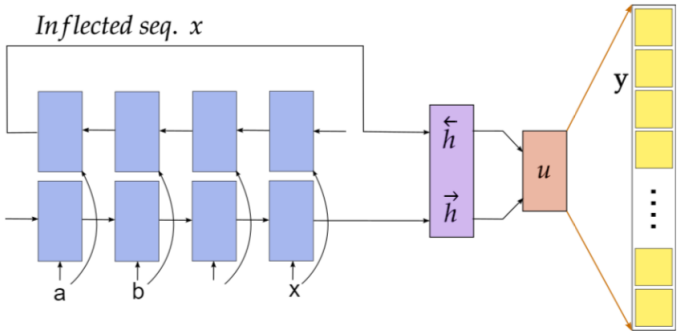

Figure 3: Architecture of the classifier for morphological attributes: we encode the inflected word form x using a BiLSTM and parameterise a distribution $q(y|x, \phi)$ over D morphological attributes. Each $q(y_d = 1|x, \phi)$ indicates the likelihood of recognising the dth morphological feature.

As there are many aspects to this model, here we report on a subset of the complete investigation.[5]

MSVED has several components: a model that infers a lemma representation, a model that predicts morphological attributes and a model that generates inflected words one character at a time. We therefore investigate the impact of our modifications on all of them. First, we concentrate on the subtask of predicting morphological attributes from a given wordform. This is initially framed as a fully supervised learning problem, i.e. learning to predict y from $x^{(t)}$. Next, we investigate whether unlabelled data ($x^{(t)}$ missing y) can help improve classification performance via semi-supervised learning. Finally, we evaluate our models in terms of the generation of inflected forms $x^{(t)}$ where we are given $x^{(s)}$ and y. Here we compare our approach to a fully supervised baseline (i.e. an encoder-decoder model that learns from labelled data only) and the original MSVED (which learns from labelled and unlabelled data).

Learning to predict morphological attributes We start by assessing whether we can learn to predict morphological attributes using a relaxation to binary variables developed in WP3. The setup is as follows: we are given a dataset $\mathcal{D}$ of pairs $\langle x, y \rangle$ and learn to map x to a distribution over its likely morphological features. In this case, we train a fully supervised model by maximising

$$\sum_{\langle x,y \rangle \in \mathcal{D}} \log q(y|x, \phi) \tag{1}$$

the likelihood of observations. Table 4 shows results for Turkish (SIGMORPHON Task 3) with the model illustrated in Figure 3. In particular, we vary the choice of likelihood, namely Bernoulli (properly discrete) versus HardKuma (approximately discrete).

Note that the relaxation does hurt classification performance and that predicting features independently is bad for precision. This makes sense: since we put all feature types together in one long list of binary attributes, modelling correlations becomes essential. We model correlations by using a MADE (Germain et al., 2015) in the output layer shown in Figure 3.[6] See Figure 4 for some

[5] This work appears in Gupta's 2019 MSc thesis, which contains many more details, experiments, and analyses.

[6] A MADE is a feed-forward neural network designed such that the dth output depends only on inputs $k < d$. MADEs are very powerful and retain the scalability of feed-forward networks.

Architecture	Distribution	Precision	Recall	F1
Independent	Bernoulli	71.85	95.88	82.14
	HardKuma	56.73	97.44	71.71
MADE	Bernoulli	78.32	92.12	84.66
	HardKuma	67.81	94.10	78.82

Table 4: Test classification performance for Turkish morphology (SIGMORPHON Task 3).

examples. The HardKuma seems to sacrifice precision for recall. This is not necessarily bad as our goal is generation rather than classification. Having a good classifier is a plus because it helps disentangle morphological inflections in the generator, but having low-variance gradients is also important for maximisation of the ELBO. Plugging the truly binary (Bernoulli) model into the VAE would lead to difficulties with regards to gradient estimation and the need for the rather noisy score function estimator (SFE; Rubinstein, 1986; Williams, 1992). To avoid that kind of complication, we decide continue with HardKuma in its stronger (MADE) parameterisation.

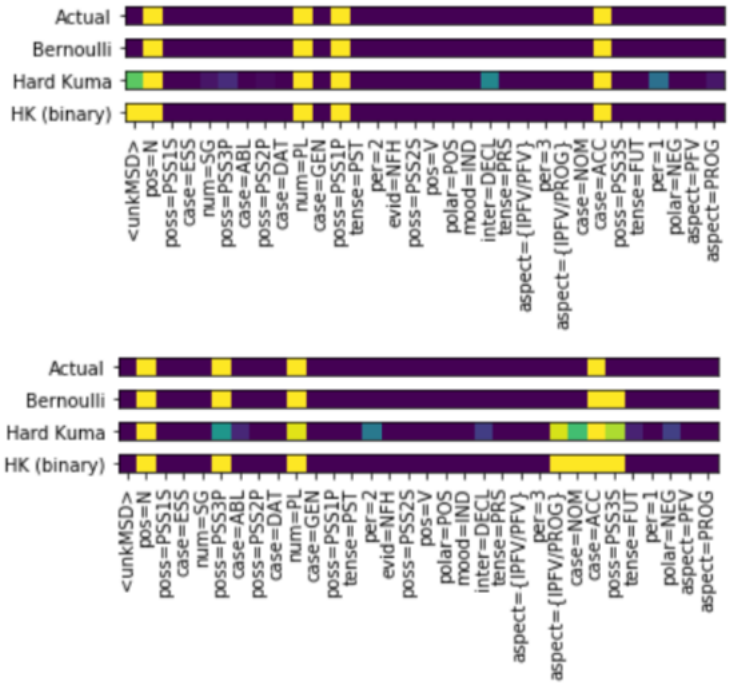

Figure 4: Two example words from the Turkish dataset showing the morphological feature: (top) `haritalarimizi` and (bottom) `lezzetlerini`. The first row shows the target features. The second row shows predictions with truly binary attributes. The third and fourth show predictions with HardKuma attributes. In the last row we discretise HardKuma draws (not for training, only for predictions) with threshold 0.5.

Semi-Supervised Learning Next we attempt to also learn from data that lacks supervision for morphological attributes. To do so, we use a semi-supervised VAE objective, i.e. given a wordform x, we predict a continuous representation z of its lemma and its morphological attributes y, for which we use the model $q(y|x, \phi)$ just discussed. The setup is as follows: either we are given a

labelled word $\langle x, y \rangle \in \mathcal{D}$ or an unlabelled word $x \in \mathcal{U}$. We therefore optimise the unsupervised objective:

$$\sum_{x \in \mathcal{U}} \mathbb{E}_{q(z|x,\phi)q(y|x,\phi)} \left[\log p(x|z, y, \theta) \right] - \mathrm{KL}(q(z|x, \phi) \| p(z)) - \mathrm{KL}(q(y|x, \phi) \| p(y)) \,, \tag{2}$$

along with the supervised objective

$$\sum_{\langle x,y \rangle \in \mathcal{D}} \mathbb{E}_{q(z|x,\phi)} \left[\log p(x|z, y, \theta) \right] - \mathrm{KL}(q(z|x, \phi) \| p(z)) + \alpha \log q(y|x, \phi) \,. \tag{3}$$

These objectives combine a standard sequence VAE (Bowman et al., 2016) (though using character-level generation) with the classifier for morphological attributes that we previously discussed, which plays the role of an inference model for missing y. First, we faced the problem of posterior collapse (Bowman et al., 2016; Alemi et al., 2018), whereby the recurrent character-level decoder $p(x|z, y, \theta) = \prod_{i=1}^{|x|} p(x_i|z, y, x_{<i}, \theta)$ learns to model the data without conditioning on z. This makes the latent space useless in the sense that neighbourhood in latent space does not correspond to any aspect of structural similarity in data space (let alone morphological or lexical semantics similarity). To counter that effect we employ the minimum desired rate constraint (MDR; Pelsmaeker and Aziz, 2020) developed in WP3 (see D3.2, Task 2). Table 5 compares our model, which employs HardKuma distributions and thus admits unbiased gradient estimates, to biased estimates from the Concrete distribution (Maddison et al., 2017; Jang et al., 2017), which is used in MSVED. We can see that dealing with the cases where we lack supervision via unbiased gradients leads to better classification performance. Figure 5 shows that words with similar lemmas are roughly grouped together, though pronounced clusters are not yet formed.

Distribution	Precision	Recall	F1
Concrete	65.13	91.96	76.25
HardKuma	68.57	95.22	79.72

Table 5: Classification performance (Turkish) with a VAE that does not receive indirect supervision for lemma.

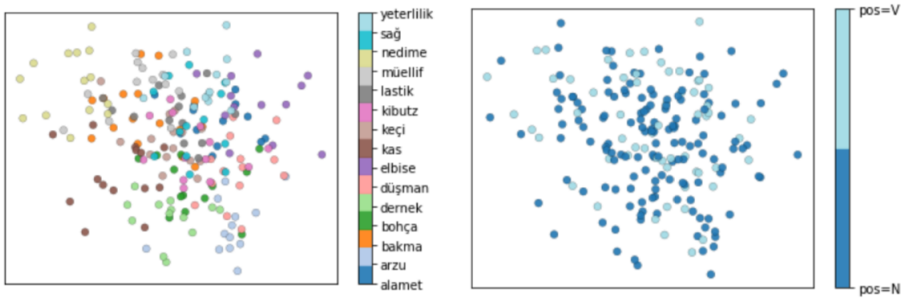

Figure 5: tSNE plot of the latent space of an unsupervised model of Turkish wordforms: each point is the Gaussian posterior mean for a given wordform. We show verb and noun forms: on the left data points are coloured according to lemma, on the right according to POS.

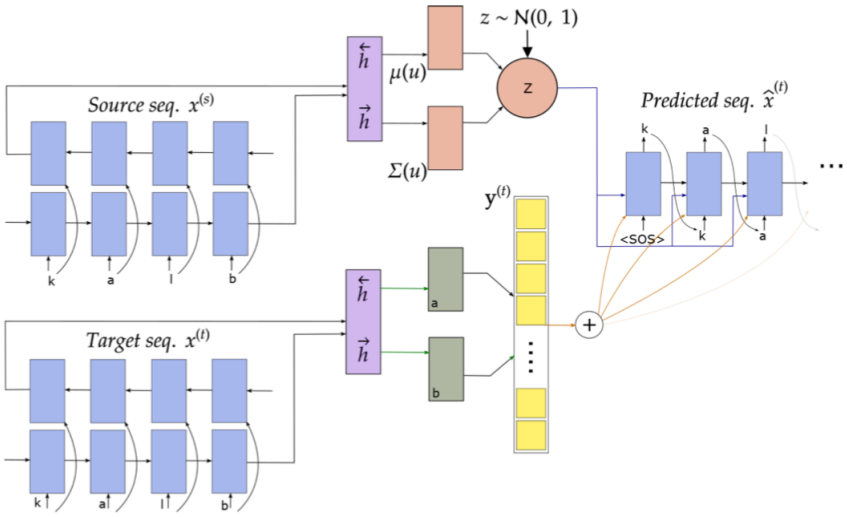

Figure 6: Architecture of the complete VAE used to generate inflected wordforms. We encode the related source wordform $x^{(s)}$ using a BiLSTM, and parameterise a Gaussian distribution over latent continuous representations z, and a collection of HardKuma distributions over sparse attributes y, after sampling z adn y with differentiable reparameterisations, we use them to initialise a recurrent decoder that generates the inflected form $x^{(t)}$ one character at a time while attending to the attributes y.

Reinflection Finally, we put all components together into a VAE that learns from unlabelled paired wordforms, i.e. $\langle x^{(s)}, x^{(t)} \rangle$, as well as from paired wordforms where the target wordform is annotated with morphological attributes, i.e. $\langle x^{(s)}, x^{(t)}, y \rangle$. Figure 6 illustrates the complete architecture. Again, posterior collapse was a problem and again we resorted to MDR. Figure 7 shows how MDR leads to a latent space where neighbouring words share lemma. This time clusters are very pronounced, showing the importance of the indirect supervision provided by $x^{(s)}$.

We now assess the model on the complete *reinflection* task, that is, where we are supposed to read $x^{(s)}$ and reinflect it according to a choice of y. In this case we compare our model to a fully supervised neural baseline (MED) and the MSVED model of Zhou and Neubig (2017). Table 6 shows the results. Note that semi-supervision leads to appreciable improvements. Our approach is superior to MED (as intended) and performs very close to MSVED, though the difference is not too large. Note that MSVED has more linguistic prior knowledge, since its morphological attributes are modelled separately depending on the type of feature. The success of this model suggests that the technology is ripe for use in an NMT architecture, though the situation in NMT is quite a bit more complex, as previously discussed (See the following section).

4.2 Latent modelling of morphology for character-based NMT (UEDIN, UVA)

From independent word generation to sequence generation The model we developed in Section 4.1 generates inflected words in isolation (i.e. not in a particular context) and conditions rather crucially as we saw on a wordform that shares the lemma of the wordform being generated. Extending that model to NMT is not trivial. First of all, we have to model an entire sequence of inflected words, which we do by extending the stochastic decoder of Schulz et al. (2018), a VAE with a sequence of continuous latent variables $z = \langle z_1, \ldots, z_{|y|} \rangle$, one per generation step. We see the sequence z as a sequence of lemma embeddings and augment that model with an additional

	Supervised			Semi-Supervised	
Language	**MED**	**MSVED**	**Ours**	**MSVED**	**Ours**
Arabic	71.47	78.13	76.99	**92.25**	88.47
Finnish	**91.15**	75.59	77.75	89.55	90.20
Georgian	92.06	88.10	87.42	**93.83**	92.91
German	**88.11**	74.48	73.90	87.59	87.54
Hungarian	95.46	95.94	95.99	**98.21**	96.61
Maltese	79.49	82.18	80.12	85.67	**86.00**
Navajo	60.67	84.24	93.72	95.28	**96.74**
Russian	80.23	74.55	73.65	**82.98**	80.12
Spanish	**93.28**	85.41	88.98	92.99	92.99
Turkish	89.00	21.21	93.87	96.62	**97.10**
Average	84.09	83.08	84.44	**91.40**	90.87

Table 6: Aggregate accuracy for morphological reinflection. We report average across 3 independent runs. For semi-supervision we use 1000 unlabelled data points from Task 1 and Task 2 datasets (note that we do not use the soruce morphology information available in Task 2).

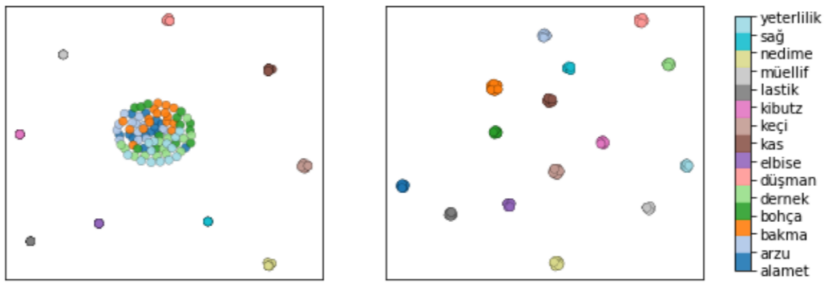

Figure 7: Addressing posterior collapse in models of Turkish word generation: tSNE plot showing latent space of two VAEs, one trained with KL annealing (left), one trained with MDR (right). Each point represents the Gaussian mean for a given wordform, same colour implies sharing the lemma.

sequence of latent variables, namely, $f = \langle f_1, \ldots, f_{|y|} \rangle$, each of which annotates the corresponding token y_j with discrete attributes. This takes the form:

$$p(y|x,\theta) = \int_z \sum_f \prod_{j=1}^{|y|} \underbrace{p(z_j|x, y_{<j}, z_{<j}, \theta)}_{\text{represent lemma}} \underbrace{p(f_j|x, y_{<j}, z_{<j}, z_j, \theta)}_{\text{predict morphology}} \underbrace{p(y_j|x, y_{<j}, z_{<j}, z_j, f_j, \theta)}_{\text{generate wordform}} \mathrm{d}\, z\,. \quad (4)$$

At each step we generate a complete target wordform y_j with a model similar to our latent factor model of Section 4.1. Instead of conditioning on a related wordform to infer the distribution over lemma embeddings z_j, we condition on the source sentence x and the prefix of already generated target words $y_{<j}$, which we represent with an attention-based NMT architecture. Morphological features receive no supervision this time, since translation data are not typically annotated with morphological attributes, and therefore the binary switches are entirely latent.[7] However, exploit-

[7] We need to decide on a fixed number of switches, as learning this number would require complex non-parametric priors, and we do so by trying a handful of values. Note that for any number D, we have as many as 2^D combinations

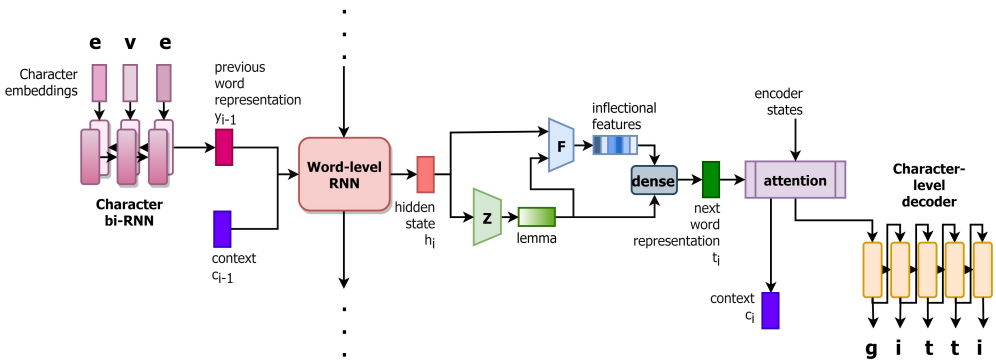

Figure 8: The latent morphology model for computing word representations while translating the sentence *'... went home'* into Turkish (*'eve gitti'*). The character-level decoder is initialised with the attentional vector $\mathbf{h}_i$ computed by the attention mechanism using current context $\mathbf{c}_i$ and the word representation $\mathbf{t}_i$ as in Luong and Manning (2016).

	(only in-domain)						*(multi-domain)*	
Model	**AR**		**CS**		**TR**		**TR**	
	BLEU	**chrF3**	**BLEU**	**chrF3**	**BLEU**	**chrF3**	**BLEU**	**chrF3**
Subwords	14.67	0.5625	16.60	0.5583	8.85	0.5225	10.65	0.5489
Char.s	12.72	0.5265	16.94	0.5608	10.63	0.5349	8.94	0.5265
Hierarch.	15.55	0.5609	16.79	0.5479	9.74	0.5127	10.35	0.5218
Hierarch. with LMM	16.06	0.5687	16.97	0.5575	10.93	0.5383	11.48	0.5575

Table 7: MT accuracy in Arabic (AR), Czech (CS) and Turkish (TR) under low-resource settings using in-domain training data *(middle column)* and multi-domain training data *(rightmost column)*. *LMM* represents the Latent Morphology Model. All improvements over the baselines are statistically significant (p-value < 0.05).

ing noisy supervision from a morphological analyser is an interesting direction for future work. Each word y_j is therefore represented by a continuous vector z_j and a collection of discrete attributes f_j. Here we again make use of sparse relaxations to binary random variables via the HardKuma distribution (Bastings et al., 2019). Architecture details, the complete specification of the parameterisation of the model, and its training algorithm can be found in our ICLR paper (Ataman et al., 2020) along with software to reproduce the model.[8]

Results The experimental results given in Table 7 show the performance of each model in translating English into Arabic, Czech and Turkish. In Turkish, the most sparse target language in our benchmark, using character-based decoding shows to be more advantageous compared to the subword-level and hierarchical models, due to the fact that reduced granularity in the vocabulary units might aid in better predicting words under conditions of high data sparsity. In Arabic, on the other hand, using a hierarchical decoding model shows to be advantageous compared to the character-level decoder, as it might be useful in better learning syntactic dependencies, whereas it also outperforms the subword-level decoder. Using the latent morphology model provides im-

of features, thus even a relatively a small value (e.g., 32 or 64) increases the model capacity considerably.

[8] https://github.com/d-ataman/lmm

provements of 0.51 and 0.30 BLEU in Arabic and Turkish over the best performing baselines, respectively. The fact that our model can efficiently work in both Arabic and Turkish suggests that it can handle the generation of both concatenative and non-concatenative morphological transformations. The results for English-to-Czech suggest that there might not be a specific advantage of using either method for generating fusional morphology, where morphemes are already optimised at the surface level, although our model is still able to achieve translation accuracy comparable to the character-level model.

4.3 Variational NMT with morphological priors (UEDIN)

The model of Section 4.1 exploits some amount of linguistic annotation to learn how to inflect, or reinflect, words in isolation. While that is a skill we want to endow a generator with, in a realistic setting, some inflectional features will capture morphosyntactic agreement and constraints and will depend on linguistic context. The model of Section 4.2 addresses the later limitation, but gives up on morphological supervision altogether. The benefit of semi-supervision is that we can bootstrap the model with the little morphological data we *do* have, and also train on the vast amount of unlabelled data. In this section we aim to learn a generator of complex wordforms during MT, thus inflecting words in context, but also exploiting some amount of supervision, and thus benefit from semi-supervised learning. In Section 4.1 we learnt from incomplete supervision by stochastically completing the data with an inference model, for efficient gradient-based learning we used a relaxation to binary variables that admits unbiased and differentiable sampling. In this section, semi-supervised learning is more challenging because we need to account for a sequence of missing morphological attributes. To tackle this more complex problem, we turn to the approach of Wolf-Sonkin et al. (2018), which we extend to translation data. This section reports on work that is not yet published.

Morphological priors Our goal is to train a morphological inflector on unlabelled data. We do so by first generating the morphological tags $m = \langle m_1, \ldots, m_n \rangle$ and the lemmata $l = \langle l_1, \ldots, l_n \rangle$ and, from a complex transformation of the two, the inflected words $x = \langle x_1, \ldots, x_n \rangle$ that compromise a sentence. Hence, l and m may be regarded as latent variables that can be marginalised out when no morphological data is available:

$$p(x|\theta) = \sum_{l} \sum_{m} \prod_{i=1}^{|x|} p_\theta(x_i|x_{<i}, l_{\leq i}, m_{\leq i}, \theta) p(l_i|x_{<i}, l_{<i}, m_{\leq i}, \theta) p(m_i|x_{<i}, l_{<i}, m_{<i}, \theta) \; . \tag{5}$$

Sadly, this marginalisation is intractable: one would need to consider an exponential number of morphological tags and a potentially infinite number of lemmata sequences. Hence, we have to fall back on approximate inference to actually learn such a model. Note that unlike the models of Section 4.1 and 4.2, here the lematta are *discrete categories*. Gradient estimation for categorical variables is far more difficult than for binary variables, and unbiased relaxations much harder to design (Mohamed et al., 2019). Wolf-Sonkin et al. (2018) choose to approach the problem via the wake-sleep algorithm (WS; Hinton et al., 1995), which circumvents the need for gradient estimation by alternating two related objectives. The parameters θ of the generative model are estimated to maximise

$$\sum_{x \in \mathcal{D}} \log p(x, m, l|\theta) \; . \tag{6}$$

For labelled data, m and l are observed along with x. For unlabelled data, m and l are sampled from an independent approximation $q(l, m|x, \phi)$ to the model's true posterior distribution $p(l, m|x, \theta)$. This corresponds to gradient-based maximum likelihood learning with imputed data. The parameters ϕ of the approximate posterior are estimated to maximise

$$\sum_{\langle m,l,x\rangle \sim p_\star} \log q(m, l|x, \phi) \; . \tag{7}$$

For labelled data, $\langle m, l, x\rangle$ is a sample from the data, and the objective corresponds to maximum-likelihood learning for a morphological analyser. For unlabelled data, wake-sleep (WS) approximates the objective using $\langle m, l, x\rangle$ generated from the model distribution $p(m, l, x|\theta)$ via ancestral sampling. This corresponds to gradient-based maximum likelihood learning for *dream data*, in the terminology of the original paper (Hinton et al., 1995).

Wake-Sleep Since the wake-sleep algorithm does not require a differentiable reparametrisation, there is some flexibility regarding the selection of a variational posterior. The model must adhere to the following qualities:

- It must allow for sampling of lemmata and morphological tags given an inflected sequence.
- It must be able to assess the likelihood of *external* samples, i.e. it must be able to return the log-probability of provided lemmata and morphological tags given an inflected sequence.
- Its parameters must allow for optimisation through backpropagation.

We selected LEMMING (Thomas et al.), a joint morphological tagger and lemmatiser, as a variational posterior. This model is a higher-order, linear-chain conditional random field, which adheres to the criteria outlined above. The authors claim that this model outperforms neural alternatives in a low-resource setting and is therefore a natural choice as variational posterior.

Variational Inference While the Wake-Sleep algorithm circumvents the need for back-propagation through Monte Carlo (MC) samples, and thus cannot suffer from problems such as noisy gradients, the two networks are trained on two different objectives. The second objective, in particular, which updates the posterior approximation, is based on updates towards dream data, which may be particularly bad early on in training when the model is not yet very good, or throughout training if we do not have enough labelled data. A unified objective can be found in variational inference (VI; Jordan et al., 1999), in particular, in the form of a variational auto-encoder (VAE; Kingma and Welling, 2014). VAEs, however, require gradient estimation through samples that can be done via the score function method (Rubinstein, 1986; Williams, 1992) or via reparameterised gradients. The former is very noisy and calls for complex variance reduction techniques, the latter is unavailable for categorical variables (such as l and m). For now we settled for a biased proxy gradient known as straight-through estimator (STE; Bengio et al., 2013) in combination with a relaxation to categorical variables known as Concrete distribution (a.k.a. Gumbel-Softmax; Maddison et al., 2017; Jang et al., 2017).

These models are trained on a set of approximately 57,000 English-Dutch sentences, and approximately 50,000 monolingual morphologically annotated Dutch data. We test four models: standard BPE-to-Char translation (BPE2Char), BPE-to-Char translation into interleaved lemmas and

morphological tags (BPE2LT), a Wake-Sleep model (WS) with the BPE2LT model as prior and a jointly trained morphological inflector that outputs the target words character-by-character and a variational autoencoder with a BPE2LT prior, morphological inflector and a jointly trained posterior annotator that predicts lemmas and morphological tags from inflected forms (VAE). Note that this VAE is trained in a semi-supervised manner, where the posterior receives supervision from the separate monolingual morphologically annotated Dutch data. The WS and BPE2LT model were trained on the En-Du dataset only, which was morphologically annotated by an external Hidden Markov Model trained on the Dutch dataset (LEMMING).

Model	W.Bound.	Schedule	KL	REC	Prior				Posterior	
					BLEU	CHRF	L-BLEU	L-CHRF	BLEU	CHRF
BPE2Char	Implicit	-	-	-	**13.28**	**40.98**	-	-	-	-
BPE2LT	Implicit	-	-	-	-	-	**16.27**	**42.33**	-	-
WS	Implicit	-	58.65	8.70	10.68	38.67	15.84	41.32	67.50	91.54
VAE	Explicit	-	-	-	-	-	-	-	-	-
	Implicit	Fixed	146.50	41.88	3.84	33.45	5.53	35.71	41.03	72.62
		No-Joint	122.81	18.75	5.85	38.68	7.11	40.96	55.43	86.02
	Explicit	Fixed	-	-	-	-	-	-	-	-
		No-Joint	-	-	-	-	-	-	-	-

Table 8: English to Dutch translation on a small dataset.

Results and discussion Results can be found in Table 8. For the WS and VAE model we report the KL-divergence between the prior and posterior, and the reconstruction error of the morphological inflector. Note that the KL of the WS model is not exact but a single sample estimate based on the data annotated by the external posterior. For all models we report the SacreBLEU and SacreCHRF scores, based on three different output levels. The scores under *Prior* refer to scores obtained when translating from English to Dutch using the prior/BPE2LT/BPE2Char model. We differentiate between scores at the level of inflected forms and scores at the level of lemmas (with L- prefix). The scores under *Posterior* are obtained by encoding Dutch with the posterior and decoding with the morphological inflector and can be interpreted as estimates of the quality of the auto-encoder.

It can be seen that neither the WS nor the VAE model can beat the BPE2LT and BPE2Char baselines. The Wake-Sleep model achieves similar quality to the BPE2LT model, but performs sub-par after inflection when compared to the BPE2Char model. This implies that there is a loss of quality during the inflection step. This can also be seen from the good, but not perfect, scores that the model attains on posterior BLEU/CHRF and reconstruction error. It therefore seems that the complication of the model structure by translating using this two-step procedure based on noisy annotation only hurts the final translation quality, at least in the low-resource setting under consideration.

The VAE model performs much worse. Neither BLEU nor CHRF come close to the baselines, and the auto-encoder reconstruction is also sub-par when compared to the WS model. The most likely cause is model optimisation. It could be that the fixed schedule is inadequate and more care is needed to balance the various components of the model. One observation that could be made from the tensorboard logs is that during the joint training phase BLEU and CHRF always deteriorate,

even though the KL-divergence improves. This indicates that the posterior is collapsing towards the prior, whereas ideally the prior would 'collapse' towards the posterior. This might not be surprising, as the task of the prior (translation) is more challenging than the task of the posterior (annotating). There are no current plans to continue with this particular piece of research.

5 Publications

The following papers are the result of research related to morphological structure carried out in the GoURMET project:

- Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. The Universitat d'Alacant Submissions to the English-to-Kazakh News Translation Task at WMT 2019. In *Proceedings of the 4th Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 356–363, Florence, Italy, 2019
- Víctor Manuel Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. Understanding the effect of morphological tags in under-resourced neural machine translation. In *Submitted to the International Conference on Computational Linguistics*, 2020
- Duygu Ataman, Orhan Firat, Mattia A. Di Gangi, Marcello Federico, and Alexandra Birch. On the Importance of Word Boundaries in Character-level Neural Machine Translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 187–193, Hong Kong, 2019
- Duygu Ataman, Wilker Aziz, and Alexandra Birch. A Latent Morphology Model for Open-Vocabulary Neural Machine Translation. In *Proceedings of the 8th International Conference on Learning Representations*, Virtual Conference, Formerly Addis Ababa, Ethiopia, 2020
- Akash Raj Komarlu Narendra Gupta. Semi-supervised morphological reinflection using rectified random variables. Master's thesis, University of Amsterdam, 2019

6 Software and code

- Morphological segmentation using Apertium:
  https://github.com/transducens/smart-segmentation
- Code accompanying (Ataman et al., 2019):
  https://github.com/d-ataman/Char-NMT
- Code accomanying (Ataman et al., 2020):
  https://github.com/d-ataman/lmm

7 Conclusion

Having described the various research associatd with this work package, we conclude with a summary of the work carried out and the current status of the work package as a whole.

With respect to the initial proposal, all three tasks have progressed well, with a number of lines of research being pursued. Tasks 1 and 3 have received most attention (with three research outputs), with a little less to Task 2, which is more specific in scope. A summary of the activities of each task and plans for future work can be found below:

Task 1 **Developing linguistically informed NMT models using morphology**
The scope of Task 1 has expanded slightly with respect to its original definition to now include the impact of using linguistically informed models on the quality of MT. Whereas the initial focus was on morphological models of segmentation, it now also includes work on integrating morphological information from analysers, such as PoS tags and morphological tags (Section 2.3). This work, applied to a range of different languages, showed that both PoS and morphological tags are useful on both the source- and target-side of the data, but particularly on the target-side when using the coarser grained PoS tags. The two other lines of research looked into segmentation strategies. The first (Section 2.1), applied to English-Kazakh translation, found that using morphologically guided subword segmentation, using an Apertium morphological analyser, improves translation qualtiy over using the more commonly used and less linguistically inpsired BPE strategy. The second work on segmentation (Section 2.2) takes the idea of subword segmentation further to find a compromise between whole word and character-level decoding by proposing a hierarchical decoder, whereby individual words are generated character by character. The model can often reach comparable scores (although this is language-dependent) to the standard subword based models and uses approximately three times fewer parameters. There is some future work planned for this work package, notably for the inclusion of morphological information in a multi-task setting. However, progress has been made more quickly than initially set out in the proposal and therefore fewer resources are likely to be dedicated to this task in the second half of the project.

Task 2 **Jointly learning alignments and morphology**
The second task will be a focus of the second half of the project, as there is currently no completed work associated with it. A proposal has been drawn up for research in this direction and work on it has already begun.

Task 3 **Exploit factors encoding latent features of morphology**
Task 3 has progressed faster than initially proposed in the project (it was expected to start later than the other two tasks). There are therefore more contributions than expected at the half-way stage. The three contributions look at designing models, either of morphology or applied to NMT, that induce morphological features through the use of latent variables. The first work cited (Section 4.1) investigates morphological form inflection as a task in itself, applied to the SIGMORPHON task, and this work is then extended to the case of NMT in Section 4.2, where morphological features are learnt in an entirely unsupervised fashion within the model. The third work is an MT model using latent morphology that looks at reintroducing some supervision back into training.

References

Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken ELBO. In *Proceedings of the 35th International Conference on Machine Learning*, pages 159–168, Stockholm, Sweden, 2018.

Duygu Ataman, Orhan Firat, Mattia A. Di Gangi, Marcello Federico, and Alexandra Birch. On the Importance of Word Boundaries in Character-level Neural Machine Translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 187–193, Hong Kong, 2019.

Duygu Ataman, Wilker Aziz, and Alexandra Birch. A Latent Morphology Model for Open-Vocabulary Neural Machine Translation. In *Proceedings of the 8th International Conference on Learning Representations*, Virtual Conference, Formerly Addis Ababa, Ethiopia, 2020.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, USA, 2015.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the 4th Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy, 2019.

Joost Bastings, Wilker Aziz, and Ivan Titov. Interpretable Neural Predictions with Differentiable Binary Variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2963–2977, Florence, Italy, 2019.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Léon Bottou and Yann L. Cun. Large Scale Online Learning. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 217–224. MIT Press, 2004.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating Sentences from a Continuous Space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany, 2016.

Mauro Cettolo, Girardi Christian, and Federico Marcello. Wit3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy, 2012.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 794–803, Stockholmsmässan, Stockholm Sweden, 2018.

Colin Cherry, George Foster, Ankur Bapna, Orhan Firat, and Wolfgang Macherey. Revisiting character-based neural machine translation with capacity and compression. In *Proceedings of*

the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4295–4305, 2018.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. The SIGMORPHON 2016 Shared Task—Morphological Reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin, Germany, 2016.

Mercedes García-Martínez, Loïc Barrault, and Fethi Bougares. Factored neural machine translation architectures. In *Proceedings of the 13th International Workshop on Spoken Language Translation*, 2016.

Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. MADE: Masked autoencoder for Distribution Estimation. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 881–889, Lille, France, 2015.

Zoubin Ghahramani and Thomas L. Griffiths. Infinite latent feature models and the Indian buffet process. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 475–482. MIT Press, 2006.

Akash Raj Komarlu Narendra Gupta. Semi-supervised morphological reinflection using rectified random variables. Master's thesis, University of Amsterdam, 2019.

G. E. Hinton, P. Dayan, B. J. Frey, and R. M. Neal. The Wake-Sleep Algorithm for Unsupervised Neural Networks. *Science*, 268:1158–1161, 1995.

Matthias Huck, Simon Riess, and Alexander Fraser. Target-side word segmentation strategies for neural machine translation. In *Proceedings of the 2nd Conference on Machine Translation*, pages 56–67, Copenhagen, Denmark, 2017.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical Reparameterization with Gumbel-Softmax. In *Proceedings of the 5th International Conference on Learning Representations*, Toulon, France, 2017.

Sébastien Jean, Orhan Firat, and Melvin Johnson. Adaptive Scheduling for Multi-Task Learning. In *Continual Learning Workshop at the 32nd Conference on Neural Information Processing Systems*, Montréal, Canada, 2018.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37(2):183–233, 1999.

Kazuya Kawakami, Chris Dyer, and Phil Blunsom. Learning to Discover, Ground and Use Words with Segmental Neural Language Models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6429–6441, Florence, Italy, 2019.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised Learning with Deep Generative Models. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3581–3589. Curran Associates, Inc., 2014.

Oskar Kohonen, Sami Virpioja, and Krista Lagus. Semi-supervised Learning of Concatenative Morphology. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 78–86, Uppsala, Sweden, 2010.

Minh-Thang Luong and Christopher D. Manning. Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1054–1063, Berlin, Germany, 2016.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. Multi-task Sequence to Sequence Learning. In *Proceedings of the 4th International Conference on Learning Representations*, San Juan, Puerto Rico, 2016.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The Concrete Distribution: A Continous Relaxation of Discrete Random Variables. In *Proceedings of the 5th International Conference on Learning Representations*, Toulon, France, 2017.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. The Natural Language Decathlon: Multitask Learning as Question Answering. *arXiv preprint arXiv:1806.08730*, 2018.

Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. Monte Carlo Gradient Estimation in Machine Learning. *CoRR*, abs/1906.10652, 2019. URL http://arxiv.org/abs/1906.10652.

Maria Nadejde, Siva Reddy, Rico Sennrich, Tomasz Dwojak, Marcin Junczys-Dowmunt, Philipp Koehn, and Alexandra Birch. Predicting Target Language CCG Supertags Improves Neural Machine Translation. In *Proceedings of the 2nd Conference on Machine Translation, Volume 1: Research Papers*, pages 68–79, Copenhagen, Denmark, 2017.

Jan Niehues and Eunah Cho. Exploiting Linguistic Resources for Neural Machine Translation Using Multi-task Learning. In *Proceedings of the 2nd Conference on Machine Translation, Volume 1: Research Papers*, pages 80–89, Copenhagen, Denmark, 2017.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, 2002.

Tom Pelsmaeker and Wilker Aziz. Effective estimation of deep generative language models. In *Proceedings of the 58th Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Seattle, Washington, USA, 2020.

Maja Popović. chrF++: words helping character n-grams. In *Proceedings of the 2nd Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark, 2017.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

Reuven Y Rubinstein. The score function approach for sensitivity analysis of computer simulation models. *Mathematics and Computers in Simulation*, 28(5):351–379, 1986.

Víctor M Sánchez-Cartagena. Prompsit's Submission to the IWSLT 2018 Low Resource Machine Translation Task. In *Proceedings of the 15th International Workshop on Spoken Language Translation*, Bruges, Belgium, 2018.

Víctor M Sánchez-Cartagena and Antonio Toral. Abu-matran at WMT 2016 translation task: Deep learning, morphological segmentation and tuning on character sequences. In *Proceedings of the 1st Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 362–370, Berlin, Germany, 2016.

Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. The Universitat d'Alacant Submissions to the English-to-Kazakh News Translation Task at WMT 2019. In *Proceedings of the 4th Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 356–363, Florence, Italy, 2019.

Philip Schulz, Wilker Aziz, and Trevor Cohn. A Stochastic Decoder for Neural Machine Translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1243–1252, Melbourne, Australia, 2018.

Rico Sennrich and Barry Haddow. Linguistic Input Features Improve Neural Machine Translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany, 2016.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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, Berlin, Germany, 2016.

Víctor Manuel Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. Understanding the effect of morphological tags in under-resourced neural machine translation. In *Submitted to the International Conference on Computational Linguistics*, 2020.

Aleš Tamchyna, Marion Weller-Di Marco, and Alexander Fraser. Modeling Target-Side Inflection in Neural Machine Translation. In *Proceedings of the 2nd Conference on Machine Translation*, pages 32–42, Copenhagen, Denmark, 2017.

Muller Thomas, Cotterell Ryan, Fraser Alexander, and Schutze Hinrich. Joint lemmatization and morphological tagging with lemming. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2268–2274.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, and Mikko Kurimo. Morfessor 2.0: Python Implementation and Extensions for Morfessor Baseline. D4 julkaistu kehittämis- tai tutkimusraportti tai -selvitys, Aalto University, 2013.

Martin Wagner. *Target Factors for Neural Machine Translation*. PhD thesis, KIT Department of Informatics Institute for Anthropomatics and Robotics (IAR), 2019.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Reinforcement Learning*, pages 5–32. Springer, 1992.

Lawrence Wolf-Sonkin, Jason Naradowsky, Sebastian J Mielke, and Ryan Cotterell. A Structured Variational Autoencoder for Contextual Morphological Inflection. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2631–2641, Melbourne, Australia, 2018.

Xuewen Yang, Yingru Liu, Dongliang Xie, Xin Wang, and Niranjan Balasubramanian. Latent Part-of-Speech Sequences for Neural Machine Translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 780–790, Hong Kong, China, 2019.

Chunting Zhou and Graham Neubig. Multi-space Variational Encoder-Decoders for Semi-supervised Labeled Sequence Transduction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 310–320, Vancouver, Canada, 2017.

ENDPAGE

GoURMET

H2020-ICT-2018-2 825299

D2.1 Initial progress report on modelling morphological structure