Alignment verification to improve NMT translation towards highly inflectional languages with limited resources

George Tambouratzis
ILSP, Athena R.C.
6 Artemidos Str.,
Maroussi, 15125, Greece
giorg_t@athenarc.gr

Marina Vassiliou
ILSP, Athena R.C.
6 Artemidos Str.,
Maroussi, 15125, Greece
mvas@athenarc.gr

Abstract

The present article studies translation quality when limited training data is available to translate towards morphologically rich languages. The starting point is a neural MT system, used to train translation models with only publicly available parallel data. An initial analysis of the translation output has shown that quality is sub-optimal, mainly due to the insufficient amount of training data. To improve translation, a hybridized solution is proposed, using an ensemble of relatively simple NMT systems trained with different metrics, combined with an open source module designed for low-resource MT that measures the alignment level. A quantitative analysis based on established metrics is complemented by a qualitative analysis of translation results. These show that over multiple test sets, the proposed hybridized method confers improvements over (i) both the best individual NMT and (ii) the ensemble system provided in the Marian-NMT package. Improvements over Marian-NMT are in many cases statistically significant.

1 Introduction

The state of the art in MT involves corpus-based systems developed with machine-learning methods. These methods learn from corpora the models needed for translation. A key strength of this approach is that the system is adapted specifically towards the data it is trained with.

For many years, the most successful data-driven approaches were phrase-based and syntax-based Statistical MT (SMT; Koehn, 2009). However, lately Neural MT (NMT) based on the encoder-decoder architecture and the concept of attention (Sutskever et al., 2014; Bahdanau et al., 2016) has become very popular. Indeed, since 2015, in MT shared tasks (Cettolo et al., 2015; Bojar et al., 2015; Bojar et al., 2016) most top-performing systems have been NMT systems. This trend is confirmed in the most recent MT shared task (Barrault et al., 2019), where 80% of participating systems are of NMT type. Though NMT represents the state of the art for MT, specific weaknesses have been reported:

- NMT performance suffers from the lack of data resources (Koehn and Knowles, 2017), giving lower translation performance, especially when training with out-of-domain rather than in-domain data.
- Recent advances in NMT models have been shown (Sennrich and Zhang, 2019) to allow good translations to be achieved with smaller parallel corpora of typically 10^5 sentences, though substantial improvements are achieved when the corpus size reaches 10^6 sentences. However, training sets of such sizes are not available for all languages.
- Translation performance is affected by non-parallel texts and non-literal translations (Carpuat et al., 2017).
- The integration of multiple algorithms into an NMT system does not necessarily improve translation (Denkowski and Neubig, 2017).
- The time complexity of training a new NMT system can be very high, with training sessions of the order of weeks.

NMT requires very large amounts of parallel data, measured in millions of parallel sentences. This is reflected by the separate studies carried out for MT with limited resources, which includes initiatives such as Lorelei [1]. In the case of morphologically-rich languages, the requirements for parallel corpora are further exacerbated.

[1] https://www.darpa.mil/program/low-resource-languages-for-emergent-incidents

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 1841–1851
April 19 - 23, 2021.

Proposed approaches for translating towards low-resource and morphologically-rich languages have included transfer learning (Zoph et al., 2016) as well as multilingual and multi-way NMT (Rikters et al., 2018).

In this paper, an effort to improve the translation quality is presented, when translating towards a morphologically-rich language, while reducing the training time. This approach combines the output of multiple NMT systems with an NLP module developed for an example-based MT paradigm, resulting in a hybridized solution. The latter module is fast and runs independently of its original MT system and thus the computational complexity of the proposed hybrid solution is not substantially increased over the base NMT system.

The idea of combining multiple MT models to produce a higher performing MT system has been studied extensively in the area of MT. For instance in the recent shared task (Barrault et al., 2019) more than 20 entries consist of ensembles of multiple NMT systems. Ensembles of weaker NMT systems of the same general architecture have been proposed by Freitag et al. (2017) to train a higher performing NMT system. In addition ensembles of factored NMT models have been proposed for automatic post-editing and quality estimation (for example Hokamp, 2017).

This base NMT system is described in section 2. The training data used is reported in section 3. The proposed hybridization is presented in section 4, whilst the improvements attained are presented in section 5. Future developments are discussed in section 6.

2 Overview of the Base NMT System

Since NMT systems have achieved the highest translation quality in recent evaluation contests, the Marian-NMT package (Junczys-Dowmunt et al., 2018) is adopted for experimentation here. Marian-NMT development was funded by the European Commission to consolidate NMT research and incorporates the most recent advances in NMT. Its code is optimized to reduce the CPU/GPU time required to complete the simulations of NMT systems.

For creating NMT systems, three of the models provided by Marian-NMT were chosen, termed as the "transformer", "amun" and "s2s" models. The "transformer" model has been based on the work of Vaswani et al. (2017) and uses a simple structure incorporating attention mechanisms and dispensing with recurrence to implement a fast NMT system. The other two models are more conventional, using a recurrent neural network to implement the translation. The "amun" model follows the approach of Bahdanau et al. (2016), employing a recurrent neural network but allowing the model to automatically search for wider ranges of the source language (SL) to connect with the target language side (TL) words. Finally, "s2s" implements a recurrent neural network-based encoder-decoder model with attention mechanism, using the architecture proposed in (Sennrich et al., 2017). Hereafter, the three models are identified via the names used within Marian-NMT, which are also used in evaluations (cf. Bojar et al., 2018).

The main configuration parameters used for each model are depicted in Table 1, to enable replication of experiments. For each model, different optimization options from Marian-NMT during the validation phase are used to create three NMT variants of each model, namely optimizing with (i) BLEU, (ii) entropy and (iii) word-wise normalized cross-entropy (denoted as "ce-mean" and representing the default optimization for Marian-NMT).

Regarding the main NMT parameters, all recurrent networks comprise 1,024 units in the hidden layer, an encoder depth of 6 layers and an embedding size of 512. All cells used both in the encoder and decoder side are gated recurrent units (GRU). The transformer dimension is set to 2,048. To reduce the lexicon size, a total of 85,000 merge operations are allowed using the BPE (Byte Pair Encoding) method proposed in (Sennrich et al., 2016), this being the default setting for marian-nmt applications.

Initially, the three Marian-NMT models are trained to provide the base NMT systems. Typically, for a single-GPU system (equipped with an NVIDIA Titan XP GTX1080 GPU card driven by an Intel i-9700K CPU), 24 hours are required for training the transformer, 130 hours for amun and 308 hours for s2s. This is equivalent to a ratio of 1:5:12 to train the respective systems.

3 Experimental Set-up

The experiments aim to improve the translation accuracy of an NMT system, taking into account limited training data and constrained computing resources. In order to investigate translation into a lesser-used and highly inflectional language, we

Common to all 3 models	
layer-normalization	yes
exponential smoothing	yes
beam-size	6
normalize	0.6
early-stopping	5
Transformer-specific	
transformer-dropout	0.1
transformer-dropout-attention	0.1
transformer-dropout-ffn	0.1
Amun and s2s-specific	
dropout-rnn	0.2
dropout-src	0.1
dropout-trg	0.1

Table 1: Key NMT hyper-parameters used

corpus	senten.	wordEn	wordGr
raw(Europarl)	1.23M	31.8M	31.9M
raw(DGT)	4.90M	97.8M	87.2M
train(DGT + Europarl)	6.13M	129.6M	119.1M
devel(Eparl)	3,000	77,681	78,610
testset2(Eparl)	1,000	27,712	27,630
testset1(Pres.)	200	2,873	2,757

Table 2: Corpora for training and evaluation

have chosen the English-to-Greek language pair.

When selecting the training corpora, it has been decided to refrain from using expensive language resources such as specialized or hand-built parallel corpora. Instead, only standard publicly available parallel corpora have been adopted, namely the Europarl and DGT-Acquis corpora[2], as listed in Table 2.

The largest part of the Europarl corpus and the entire DGT-Acquis corpus are used to train the NMT system. Three small portions of the Europarl corpus have been reserved for test and validation purposes. More specifically, two independent sets of approx. 3,000 Europarl sentences each are excluded, to ensure that the NMT evaluation is unbiased. In the present experiment, one of these sets is used for in-training validation. The other set is reserved to allow additional cross-evaluation of experiments in the future, without invalidating the previously trained models. Finally, a sample of 1,000 sentences from Europarl (Testset2) is retained to provide an unseen in-domain test set.

Another independent test set was drawn from the PRESEMT project resources, comprising 200 sentences which have not been used to either train an MT model or create any resources used herewith (denoted as Testset1).

A preliminary analysis of the NMT outputs has shown that translations are commendably fluent, though errors are evident. A sample of amun translations is shown in Figure 1. In sentence #1, the term "Αμερικανοί" (Transl. "Americans") is erroneously used as a translation of the terms "American", "European", and "Japanese monopolies". Similarly, in sentence #2, the phrase "η κατ-απολέμηση της φτώχειας" (meaning "the reduction of poverty") is used to translate semantically diverse phrases, including "genetically modified organisms", and "the negative social effects of unbridled, unregulated globalization". Repetition is a widely reported weakness of NMT systems, most frequently attributed to insufficient training data.

An additional problem concerns the translation of rare words (i.e. words with low frequency in the corpus), due to the limited vocabulary that NMT systems can directly handle. This is especially severe when translating towards languages with complex morphology, which increases the effective vocabulary size. For example the word "ostensibly" is translated into Greek as "ostenfigher" (ungrammatical). Similarly the word "room" is translated as "δωματείο" instead of the correct "δωμάτιο" (meaning room), whilst the word "indistinct" is translated as "άχωρος" which is not a valid Greek word.

Another issue is that entire phrases present in the source text may be omitted in the translation. For instance the sentence "Businesses have undertaken the education" is translated by a transformer NMT as "Η εκπαίδευση έχει αναλάβει», [meaning "education has undertaken"]. Hence, the subject "business" has been deleted.

4 Improving NMT via the Alignment Verification Method (AVM)

4.1 Aim of AVM

To improve translation accuracy, the main errors need to be identified in an automated manner. The idea is that a poor alignment between source text and translation indicates substantial loss of meaning during translation. On the contrary a high alignment score is indicative of a high likelihood that

[2]The Europarl corpus (ver.7) was retrieved from https://www.statmt.org/europarl. The DGT-Acquis corpus was retrieved from https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory

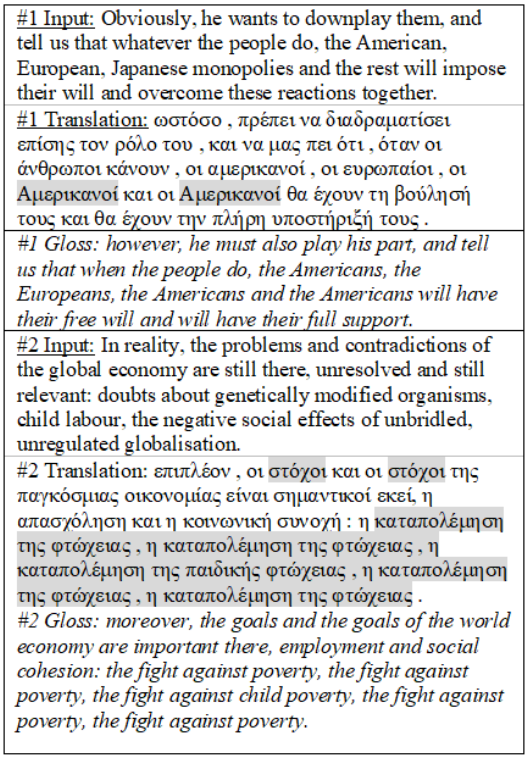

Figure 1: Example translations generated by amun, with repetitions of texts highlighted in grey

the NMT output is an accurate translation.

To this end a module will be added to implement alignment verification (AVM), by determining the match between the input sentence and its translation. The establishment of representative alignment scores allows in turn the combination of multiple NMT models, using AVM to evaluate the accuracy of each candidate translation and thus select the best translation on a sentence-by-sentence basis. For this research, MT software and tools released via open-source code have been surveyed and the Phrase Aligner Module (PAM, cf. Troullinos, 2013) has been selected. The architecture of the proposal hybrid NMT is depicted in Figure 2.

4.2 PRESEMT essentials

PAM was developed as part of the PRESEMT hybrid MT methodology (Tambouratzis et al., 2017 (Tambouratzis et al., 2017)). PRESEMT was designed to create MT systems requiring only very limited amounts of specialized, expensive linguistic resources. Frequently, the most expensive resource is the parallel corpus of SL – TL sentences. PRESEMT uses parallel corpora of only a few hundred sentences, augmented by very extensive but comparatively inexpensive monolingual corpora.

Within the PRESEMT methodology, the small parallel corpus serves to establish the transformation from the SL structure to the TL one, using the Phrase Aligner module. This module, handling sentence pairs from this parallel corpus, identifies the correspondence of words and phrases from SL to TL, to determine the translation accuracy.

4.3 Description of the PAM module

PAM utilizes a limited-size bilingual lexicon (of typically 30 to 40 thousand token pairs) together with a publicly available parser. Details on these resources are reported in section 4.4, as their choices are language-specific. Based on these resources, PAM establishes for the set of parallel sentences the alignment of both words and phrases from SL to TL, in three hierarchically ordered stages:

1. Within the first stage, the alignment of words is based on equivalences provided by the bilingual lexicon. Dedicated PAM processes resolve cases where (i) words have multiple appearances within a sentence and (ii) multiple potential translations of an SL word exist in the TL side.

2. Within the second stage, words are aligned by establishing statistical correspondences between grammatical features across the SL and TL pair. These correspondences are automatically extracted from the lexicon.

3. Within the third stage, any remaining words are aligned and grouped into phrases on the basis of the alignments of their neighboring words that are successfully aligned. To implement this, the principle of locality across languages is adopted (words at a small distance to each other in SL also tend to be located close to each other in TL).

The key PAM principle is that decisions made at a later stage have a lower degree of confidence than those made at an earlier stage (Troullinos, 2013).

4.4 Using PAM for Alignment Verification

In the current application, PAM determines the suitability of each candidate translation, based on its match with the source sentence. Thus, the assumption made is that the input sentence and the candidate translation represent the corresponding SL and TL entries of a parallel corpus and PAM determines their level of parallelism.

As the requirement is to grade various translations, the PAM operation is reversed, to identify the quality of match between the input sentence and the generated translations. When PAM was used in PRESEMT, sentence pairs from the parallel corpus with a very low percentage of successful alignments were discarded without measuring their degree of parallelism, as poor exemplars of the structural transformations from SL to TL. Here, PAM is modified so that for all pairs of input sentence and NMT-translation the word alignments and assignments of words to phrases are calculated. This allows the re-roled PAM to grade any source/translation pair, no matter how poor the match of the two sentences is.

Two metrics have been established to calculate divergence between the SL sentence and its NMT-derived translations. The first metric (*Uscore*) calculates the number of unaligned words of the source sentence, after PAM is applied. The aim is to have as few unaligned words as possible, so the lower *Uscore* is, the better the translation is.

$$Uscore = \#unaligned\_words \quad (1)$$

The second metric (*Wscore*) is a weighted combination of several indicators of alignment between source sentence and candidate translation. This summarizes in one measurement the type of alignments and the stage at which they were achieved. Hence, for a sentence with *K* words, *Wscore* is defined as:

$$Wscore = \sum_{i=1}^{K} (w_i * align\_stage_i) \quad (2)$$

In equation (2), *align-stage*$_i$ denotes the stage (cf. section 4.3 for the different stages) at which the i-th SL word is aligned successfully to a TL word, and w_i denotes the relevant weight for this stage. In the case of the weighted metric Wscore, the higher the score, the more accurate the corresponding translation is. The actual weight values must reward the establishment of alignments at an earlier rather than a later stage. Thus, w_i should be larger than w_j , for *i* smaller than *j*. For the purposes of the present article, w_i is set to integer values of 5, 2 and 1 for the first, second and third stage respectively (other sets of weight values that follow this reasoning produce similar results to those reported here). The code of PAM has been modified to integrate *Wscore* and *Uscore* calculation, though the actual alignment algorithms within PAM have remained intact. The associated PAM resources (the TL-side parser and bilingual lexicon) remain unchanged. For processing the SL language, Treetagger (Schmid, 1994) is used, with a reported tagging accuracy exceeding 96%.

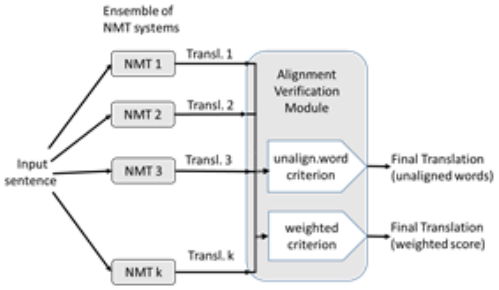

Figure 2: Proposed hybrid NMT approach

5 Experimental Results

To determine the quality of the NMT-based translations (amun-, s2s- and transformer-based models), two widely used MT evaluation metrics are utilised, namely BLEU (Papineni et al., 2002) and NIST (Doddington, 2002). To calculate both these metrics, the mt-eval package (version 13a) is used.

For PAM, the PRESEMT bilingual lexicon from Greek to English is used, which contains approx. 8,000 lemmas and 40,000 Greek-English token pairs. This lexicon is from the same domain as testset1 and is thus out-of-domain for testset2, providing a more limited coverage for this testset.

Two different types of experiments are possible, depending on whether the ensemble comprises multiple NMT architectures, or only one type of architecture. The first experiment reported here involves NMT ensembles that all share the same architecture, but are optimized with different criteria. The second type of experiment studies ensembles which consist of systems with different architectures, to investigate if their combination results in a better translation quality.

5.1 NMT Ensembles of a single architecture

The results obtained for testset1 of the English-to-Greek translation pair are depicted in Table 3, when running the single transformer, amun and s2s models respectively, as well as their ensembles. The corresponding results for testset2 are depicted in Table 4.

In Tables 3 and 4, the first 3 rows correspond to

single NMT models generated when Marian-NMT is trained to optimise (i) BLEU, (ii) entropy and (iii) the word-wise normalised cross-entropy (this is denoted as "ce-mean").

The final three rows of Tables 3 and 4 report the accuracy of translations obtained by NMT ensembles. Marian-NMT implements a standard ensemble method, which allows the user to combine different models provided they use the same lexicon. The user may specify weighting factors to boost selection of the models deemed to be better. For this article, this ensemble combines the three aforementioned NMT models (i), (ii) and (iii), with equal weights for all NMTs. The last two rows report the accuracy of ensembles using PAM with (i) Uscore and (ii) Wscore, respectively.

A key difference of the Marian-NMT ensemble is that it is able to recombine partial results of the translation process from each NMT model and thus may generate a new translation that is different from all the translations of single-NMT systems. On the contrary, Uscore and Wscore grade the translations generated by single-NMT models in the ensemble, and then select the highest-scoring translation to be the translation produced by the PAM-based ensemble.

To evaluate the quality of translations produced by the PAM-based ensembles, two baselines are selected. The first baseline is the "ce-mean" option of the Marian-NMT translation system.The second, and stronger, baseline is the Marian-NMT ensemble (referred to as "Marian-ensemble" hereafter). Entries that exceed the first baseline are depicted in bold. Entries with scores that exceed the stronger Marian-ensemble baseline are annotated with an asterisk.

Based on Table 3, for testset1 the best BLEU scores are achieved by the Marian-NMT ensemble in comparison to single-NMT models. The PAM-Wscore ensemble gives a higher accuracy than the Marian-NMT ensemble, whilst the accuracy of PAM-Uscore is lower than PAM-Wscore. On the whole, it is Marian-ensemble and PAM-Wscore that generate the best NIST and BLEU scores.

A broadly similar situation is found when using testset2 (Table 4). Here, the improvement conferred by the ensemble methods over the three base models is much more marked. For instance, for BLEU, the score is only 19.0 to 20.0 for single NMT models, but rises to more than 28.0 for the ensembles, which equates to more than eight BLEU percentage points of improvement.

5.2 Statistical analysis of ensemble results

One question is whether the improvements conferred by the ensembles are statistically significant. To that end, the BLEU and NIST scores of all the independent sentences are assembled, forming two populations of scores (one for BLEU and one for NIST) for each experimental run. Then the Wilcoxon and sign tests are used to determine if these populations have significant differences.

For testset1, the scores of the single NMT systems and the NMT-ensembles are relatively close, differing by less than 2 BLEU points. Applying the sign and Wilcoxon tests, Marian-ensemble produces statistically better NIST scores (at a 0.05 level) than the default Marian-NMT output for amun and s2s models, but not for the transformer model.

For the transformer and s2s models, the scores generated by PAM-Wscore are significantly better that those of single-model Marian-NMT, according to both the Wilcoxon and sign tests (at a 0.05 level). Similarly, PAM-Uscore gives statistically superior results to Marian-NMT (ce-mean optimization) for the s2s model (at a significance level of 0.05).

Comparing the ensembles to each other, Wscore consistently produces higher scores than Uscore. This superiority is statistically significant at a 0.05 level according to both Wilcoxon and sign tests, for the transformer and the amun models.

PAM-Wscore achieves consistently higher translation scores than Marian-ensemble for both BLEU and NIST. According to the Wilcoxon test, these differences are statistically significant, at a 0.05 level, only for the s2s (BLEU score) and the transformer model (both BLEU and NIST scores).

Turning to testset2, the results are more clearly separated. All three ensembles (i.e. PAM-Wscore, PAM-Uscore and Marian-ensemble) have statistically superior scores to Marian (optimised with ce-mean) for both BLEU and NIST, at a significance level of 0.01. This extends to all three NMT models (amun, transformer and s2s), and indicates that both Marian-ensemble and the two PAM-based ensembles give substantially higher scores than single Marian-NMT models.

On the other hand, when comparing PAM-Wscore to PAM-Uscore for testset2, no statistically significant difference (at a 0.05 level of signifi-

cance) between the two systems is discerned by either the Wilcoxon or sign test. Similarly, no statistically significant differences at a 0.05 level are found between the PAM-based ensembles and the Marian-ensemble and only small differences at a 0.10 level. Thus, even though PAM-based ensembles achieve scores higher than Marian-ensemble, differences are not significant.

5.3 Measuring improvement over baselines

To quantize the improvements achieved by the proposed PAM-Wscore approach, in this section the computational requirements posed by each NMT system are also considered. To this end, the most accurate NMT system is defined for each dataset and metric combination. Two baselines are chosen, namely the most accurate NMT model and the most accurate Marian-ensemble.

We focus on the transformer model, which is the least expensive model to train. For each ensemble using transformers, the aim is to determine how close to the Marian-ensemble baseline this is. Results are shown in Table 5, where the accuracy of each transformer NMT is expressed as a fraction of the Marian-ensemble score.

The best single transformer model achieves for testset1 88.7% of the baseline BLEU score and 93.1% of the NIST score. Using the Wscore ensembling method, this rises to 90.7% for BLEU and 95.2% for NIST, showing a gain of 2%.

Turning to dataset2, the single transformer scores just 70.5% in comparison to the baseline BLEU score and 73.4% of the NIST score (therefore it is 27% to 30% lower). The Wscore ensemble improves relative scores, reaching 92.7% and 94.5% of the baseline scores for BLEU and NIST respectively. This equates to an increase of ca. 22% in both scores, making the final result directly comparable to s2s, though GPU training requirements are reduced by a factor of 12.

5.4 Subjective studies

A second type of evaluation moves away from metrics to focus on analysing the translation errors by different models, with subjective methods. For instance, when transformer NMT models are tasked to translate testset1, the BLEU-optimised NMT generates 26 ungrammatical words, the entropy-optimised NMT generates 24 ungrammatical words and the cross-entropy optimised model produces 23 ungrammatical words. The Wscore-ensemble reduces the ungrammatical words to 21, improving

Sent :648	**Source text:** That amendment establishes three or four very important changes: firstly, the Canary Islands were incorporated into the Common agricultural policy and the Common fisheries policy; secondly, they were incorporated into the Community Customs Union, and are therefore subject to ordinary customs duties; and thirdly, to protect local industry .
	Marian-ensemble translation: Η τροποποίηση του κανονισμού (ΕΟΚ) αριθ.
	Wscore-ensemble translation: Η τροπολογία για τη σύσταση τριών ή τεσσάρων πολύ σημαντικών αλλαγών: Πρώτον, οι Καναρίους Νήσους ενσωματώθηκαν στην κοινή γεωργική πολιτική και στην κοινή αλιευτική πολιτική · δεύτερον, ενσωματώθηκαν στην τελωνειακή ένωση της Κοινότητας και, συνεπώς, υπόκεινται σε συνήθεις τελωνειακούς δασμούς · και τρίτον, την προστασία της τοπικής βιομηχανίας.
Sent 774	**Source text:** Thirdly, who gets the money ? **Marian-ensemble translation:**<NULL>
	Wscore-ensemble translation: Ποιος λαμβάνει τα χρήματα
Sent 962	**Source text:** Furthermore, we have been abolishing subsidies in the shipbuilding sector, yet it has continued to fall apart.
	Marian-ensemble translation: Ωστόσο, καταργήσαμε τις επιδοτήσεις στον τομέα της ναυπηγικής βιομηχανίας
	Wscore-ensemble translation: Για παράδειγμα, καταργήσαμε τις επιδοτήσεις στον ναυπηγικό τομέα, αλλά εξακολουθεί να υφίσταται.

Figure 3: Examples of poor translations produced by Marian-ensemble (omitted parts are underlined in source).

translation. The ungrammatical words were determined in all cases by visual inspection of the body of translations complemented by spell-checking tools to aid detection.

Further inspection of translation quality has involved comparing the Marian-ensemble and Wscore-ensemble outputs. The length (in words) of translations per test sentence is found to differ substantially between the two ensembles, with the difference being more than 1/10 for 9% of sentences, more than 1/4 for 2.5% of sentences and more than 1/2 for 1% of sentences (close to identical results are obtained for testset1 and testset2). As such deviations are unexpectedly large, an analysis was performed, with typical examples being shown in Figure 3. As can be seen, PAM assists the Wscore-ensemble in retaining all phrases of the sentence. On the contrary, Marian-ensemble fails to ensure this, and frequently discards portions of the input sentence. In one case (sentence #774) Marian-ensemble results in a null-length translation, and in another (sentence #648) the final translation covers less than 10% of the input text, radically distorting meaning. Both PAM-ensembles are unaffected by such phenomena.

6 Conclusions and Future Work

This article has studied the creation of translation systems towards highly inflectional languages, when the amount of in-domain training data is limited. Emphasis has been placed on improving the translation accuracy of NMT models that can be trained more rapidly and cost-effectively (in terms of CPU processing power) and rendering this performance comparable to that of more complex models. The Marian-NMT package has been chosen as the starting point to create NMT models for the English to Greek language pair. Using only publicly available text corpora, the NMT models produce commendably fluent translations. Identified errors in the NMT translations are typical of a lack of training data.

A hybrid methodology has been proposed that samples an ensemble of NMT models to select the final translation, chosen by a module calculating the alignment level between the input sentence and each translation. This module was developed for resource-poor MT systems.

The proposed hybrid approach has resulted in higher BLEU and NIST scores, compared to those of single NMT models. Improvements are in many cases statistically significant even over the ensemble system provided within the Marian-NMT package, indicating the promising nature of the hybrid approach. Also, the translation process is found to be more robust, giving more consistent translations in comparison to the Marian-NMT ensemble system, which occasionally omits large portions of the input text from the translation.

One of the advantages of the proposed method is that it is general-purpose and does not rely on the use of ensembles of Neural MT systems with a specific architecture. Instead, it can be used to combine the results of different types of Neural MT systems, or MT systems that belong to different paradigms, or even to combine human translations.

In addition the proposed method can be used to clean up a corpus of parallel sentences or several such corpora, by removing sentence pairs for which the source and target-language texts do not have a high degree of parallelism. Similarly, the proposed method may be used to filter a corpus consisting of original text and its MT-derived translation, to produce a parallel corpus for training of other MT systems, fulfilling a role similar to that proposed by (Rikters and Fishel, 2017). One point for future research is how effective a filtering system based on PAM would be, in comparison to already proposed systems.

Future work involves some relatively simple activities that can be imminently implemented, such as releasing the modified version of PAM for experimentation by interested parties. Another short term activity involves using the proposed method with sacreBLEU instead of the BLEU and NIST metrics provided by mt-eval. Future experiments will investigate the effectiveness of this hybrid approach for other language pairs. One area of interest would be to determine the effectiveness of the PAM-based method when very limited dictionaries are available as well as the limitations when the accuracy of the parser used is relatively low. All these represent issues for the future.

It is also planned to study the approach using systematic optimisation of the PAM parameters, to identify in more detail configurations that produce more accurate translations. Another possibility is to use PAM to detect sub-sentential parts of the translated sentences with particularly poor alignments between input and translation and seek better translations of only these specific parts.

Another direction is to investigate more extensively cases where the translation is not sufficiently close to the input sentence. Then, comparisons to other low-scored translations are more difficult and result in a reduced level of confidence of the chosen translation. Such a line of study will evaluate more thoroughly the robustness of the proposed method.

Acknowledgements

The authors acknowledge support of this work by the project "DRASSI" (MIS5002437) which is implemented under the Action "Reinforcement of the Research and Innovation Infrastructure", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF2014-2020) and co-financed by Greece and the European Commission (European Regional Development Fund). The authors wish to acknowledge the contribution of NVIDIA who donated for research purposes in the area of Machine Translation a Titan XP GPU card under the NVIDIA Academic Support Programme to the MT group of ILSP/Athena R.C.

	BLEU			NIST		
criterion	transformer	amun	s2s	transformer	amun	s2s
BLEU-optimised (1)	**35.22**	36.92	35.28	**6.238**	6.440	6.459
Entropy-optimised (2)	34.80	**37.91**	35.25	6.213	**6.532**	6.428
Ce-mean-optimised (3)	34.96	37.77	35.58	6.222	6.524	6.476
Marian-ensemble (1,2,3)	**35.63**	**38.38**	**39.70**	**6.305**	**6.590**	**6.707**
PAM–ensemble + Uscore	33.94	37.25	**39.57**	6.128	6.463	**6.721***
PAM–ensemble + Wscore	**35.99***	**38.56***	**39.79***	**6.384***	**6.584**	**6.758***

Table 3: Translation accuracy of NMT models and ensembles, where each ensemble consists of identically structured NMTs that have been optimized with different criteria (using testset1)

	BLEU			NIST		
criterion	transformer	amun	s2s	transformer	amun	s2s
BLEU-optimised (1)	18.27	**19.11**	18.92	4.100	**4.271**	3.894
Entropy-optimised (2)	**20.41**	**19.07**	20.04	**4.944**	**4.260**	4.396
Ce-mean-optimised (3)	18.84	18.83	20.48	4.254	4.129	4.526
Marian-ensemble (1,2,3)	**26.48**	**26.95**	**28.79**	**6.454**	**6.507**	**6.735**
PAM–ensemble + Uscore	**26.35**	**27.61***	**28.96***	**6.407**	**6.467**	**6.684**
PAM–ensemble + Wscore	**26.68***	**27.45***	**28.85***	**6.363**	**6.513***	**6.716**

Table 4: Translation accuracy of NMT models and ensembles, where each ensemble consists of identically structured NMTs that have been optimized with different criteria (using testset2)

	Testset1	Testset1	Testset2	Testset2
Model	BLEU	NIST	BLEU	NIST
Best NMT (single model)	95.5% (amun)	97.4% (amun)	71.1% (s2s)	73.4% (transf)
Best NMT (Marian-ensemble)	100% (s2s)	100% (s2s)	100% (s2s)	100% (s2s)
Transformer (single model)	88.7%	93.1%	70.9%	73.4%
Transformer (Marian-ensem)	89.7%	94.0%	92.0%	95.8%
Transf PAM + Uscore	85.5%	91.4%	91.5%	95.2%
Transf PAM + Wscore	90.7%	95.2%	92.7%	94.5%

Table 5: Scores achieved for testset1 by different transformer models in comparison to the two baseline models, reported in the first two rows. Scores are normalized over the Marian-ensemble score (cf. row 2).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate. *Computing Research Repository*, arXiv:1409.0473. Version 7.

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. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Marine Carpuat, Yogarshi Vyas, and Xing Niu. 2017. Detecting cross-lingual semantic divergence for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 69–79, Vancouver. Association for Computational Linguistics.

Mauro Cettolo, Jan Niehues, Sebastian Stuker, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2015. The iwslt 2015 evaluation campaign. In *IWSLT Proceedings*, Da Nang.

Michael Denkowski and Graham Neubig. 2017. Stronger baselines for trustable results in neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 18–27, Vancouver. Association for Computational Linguistics.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *HLT '02: Proceedings of the Second International Conference on Human Language Technology Research*, pages 138–145.

Chris Hokamp. 2017. Ensembling factored neural machine translation models for automatic post-editing and quality estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 647–654, Copenhagen, Denmark. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Philipp Koehn. 2009. *Statistical Machine Translation*. Cambridge University Press, Cambridge, UK.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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. Association for Computational Linguistics.

Matiss Rikters and Mark Fishel. 2017. Confidence through attention. *CoRR*, abs/1710.03743.

Matīss Rikters, Mārcis Pinnis, and Rihards Krišlauks. 2018. Training and adapting multilingual NMT for less-resourced and morphologically rich languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nădejde. 2017. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*,

pages 65–68, Valencia, Spain. Association for Computational Linguistics.

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

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Ilia Sutskever, Oriol Vinyals, and Quok V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the NIPS Conference*, page 3104–3112. NIPS.

George Tambouratzis, Marina Vassiliou, and Sokratis Sofianopoulos. 2017. *Machine Translation with Minimal Reliance on Parallel Resources*. Springer-Verlag, Berlin.

Michalis Troullinos. 2013. Phrase aligner. technical report. *Faculty of Informatics, Masaryk University Brno, FI MU Report Series FIMU-RS-2013-2*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.