# Neural Machine Translation into Language Varieties

**Surafel M. Lakew**[†⋆], **Aliia Erofeeva**[†], **Marcello Federico**[⋆+]
[†]University of Trento, [⋆]Fondazione Bruno Kessler, [+]MMT Srl, Trento, Italy
[†]name.surname@unitn.it, [⋆]surname@fbk.eu

## Abstract

Both research and commercial machine translation have so far neglected the importance of properly handling the spelling, lexical and grammar divergences occurring among language varieties. Notable cases are standard national varieties such as Brazilian and European Portuguese, and Canadian and European French, which popular online machine translation services are not keeping distinct. We show that an evident side effect of modeling such varieties as unique classes is the generation of inconsistent translations. In this work, we investigate the problem of training neural machine translation from English to specific pairs of language varieties, assuming both labeled and unlabeled parallel texts, and low-resource conditions. We report experiments from English to two pairs of dialects, European-Brazilian Portuguese and European-Canadian French, and two pairs of standardized varieties, Croatian-Serbian and Indonesian-Malay. We show significant BLEU score improvements over baseline systems when translation into similar languages is learned as a multilingual task with shared representations.

## 1 Introduction

The field of machine translation (MT) is making amazing progress, thanks to the advent of neural models and deep learning. While just few years ago research in MT was struggling to achieve *useful* translations for the most requested and high-resourced languages, the level of translation quality reached today has raised the demand and interest for less-resourced languages and the solution of more subtle and interesting translation tasks (Bentivogli et al., 2018). If the goal of machine translation is to help worldwide communication, then the time has come to also cope with dialects or more generally language varieties[1]. Remarkably, up to now, even standard national language varieties, such as Brazilian and European Portuguese, or Canadian and European French, which are used by relatively large populations have been quite neglected both by research and industry. Prominent online commercial MT services, such as Google Translate and Bing, are currently not offering any variety of Portuguese and French. Even worse, systems offering such languages tend to produce inconsistent outputs, like mixing lexical items from different Portuguese (see for instance the translations shown in Table 1). Clearly, in the perspective of delivering high-quality MT to professional post-editors and final users, this problem urges to be fixed.

While machine translation from many to one varieties is intuitively simpler to approach[2], it is the opposite direction that presents the most relevant problems. First, languages varieties such as dialects might significantly overlap thus making differences among their texts quite subtle (e.g., particular grammatical constructs or lexical divergences like the ones reported in the example). Second, parallel data are not always labeled at the level of language variety, making it hard to develop specific NMT engines. Finally, training data might be very unbalanced among different varieties, due to the population sizes of their respective speakers or for other reasons. This clearly makes it harder to model the lower-resourced varieties (Koehn and Knowles, 2017).

In this work we present our initial effort to systematically investigate ways to approach NMT from English into four pairs of language varieties:

[1] In sociolinguistics, a variety is a specific form of language, that may include dialects, registers, styles, and other forms of language, as well as a standard language. See Wardhaugh (2006) for a more comprehensive introduction.

[2] We will focus on this problem in future work and disregard possible varieties in the source side, such as American and British English, in this work.



*Proceedings of the Third Conference on Machine Translation (WMT)*, Volume 1: Research Papers, pages 156–164

https://doi.org/10.18653/v1/W18-6316

| | |
|---|---|
| English (source) | I'm going to the gym before breakfast. No, I'm not going to the gym. |
| pt (Google Translate) | Eu estou indo para a academia antes do café da manhã. Não, eu não vou ao ginásio. |
| pt-BR (M-C2) | Eu vou á academia antes do café da manhã. Não, eu não vou à academia. |
| pt-EU (M-C2) | Vou para o ginásio antes do pequeno-almoço. Não, não vou para o ginàsio. |
| pt-BR (M-C2_L) | Vou à academia antes do café da manhã. Não, não vou à academia. |
| pt-PT (M-C2_L) | Vou ao ginásio antes do pequeno-almoço. Não, não vou ao ginásio. |

Table 1: MT from English into Portuguese varieties. Example of mixed translations generated by Google Translate (as of 20th July, 2018) and translations generated by our variety-specific models. For the underlined English terms both their Brazilian and European translation variants are shown.

Portuguese European - Portuguese Brazilian, European French - Canadian French, Serbian - Croatian, and Indonesian - Malay[3]. For each couple of varieties, we assume to have both parallel text labeled with the corresponding couple member, and parallel text without such information. Moreover, the considered target pairs, while all being mutually intelligible, present different levels of linguistic similarity and also different proportions of available training data. For our tasks we rely on the WIT[3] TED Talks collection[4], used for the International Workshop of Spoken Language Translation, and OpenSubtitles2018, a corpus of subtitles available from the OPUS collection[5].

After presenting related work (Section 2) on NLP and MT of dialects and related languages, we introduce (in Section 3) baseline NMT systems, either language/dialect specific or generic, and multilingual NMT systems, either trained with fully supervised (or labeled) data or with partially supervised data. In Section 4, we introduce our datasets, NMT set-ups based on the Transformer architecture, and then present the results for each evaluated system. We conclude the paper with a discussion and conclusion in Sections 5 and 6.

## 2 Related work

### 2.1 Machine Translation of Varieties

Most of the works on translation between and from/to written language varieties involve rule-based transformations, e.g., for European and Brazilian Portuguese (Marujo et al., 2011), Indonesian and Malay (Tan et al., 2012), Turkish and Crimean Tatar (Altintas and Çiçekli, 2002); or phrase-based statistical MT (SMT) systems, e.g., for Croatian, Serbian, and Slovenian (Popović et al., 2016), Hindi and Urdu (Durrani et al., 2010), or Arabic dialects (Harrat et al., 2017). Notably, Pourdamghani and Knight (2017) build an unsupervised deciphering model to translate between closely related languages without parallel data. Salloum et al. (2014) handle mixed Arabic dialect input in MT by using a sentence-level classifier to select the most suitable model from an ensemble of multiple SMT systems. In NMT, however, there have been fewer studies addressing language varieties. It is reported that an RNN model outperforms SMT when translating from Catalan to Spanish (Costa-jussà, 2017) and from European to Brazilian Portuguese (Costa-Jussà et al., 2018). Hassan et al. (2017) propose a technique to augment training data for under-resourced dialects via projecting word embeddings from a resource-rich related language, thus enabling training of dialect-specific NMT systems. The authors generate spoken Levantine-English data from larger Arabic-English corpora and report improvement in BLEU scores compared to a low-resourced NMT model.

### 2.2 Dialect Identification

A large body of research in dialect identification stems from the DSL shared tasks (Zampieri et al., 2014, 2015; Malmasi et al., 2016; Zampieri et al., 2017). Currently, the best-performing methods include linear machine learning algorithms such as SVM, naïve Bayes, or logistic regression, which use character and word $n$-grams as features and are usually combined into ensembles (Jauhiainen et al., 2018). Tiedemann and Ljubeši (2012) present the idea of leveraging parallel corpora for language identification: content comparability allows capturing subtle linguistic differences between dialects while avoiding content-related biases. The problem of ambiguous sentences, i.e., those for which it is impossible to decide upon the dialect tag, has been demonstrated for Portuguese by Goutte et al. (2016) through inspection of disagreement between human annotators.

[3]According to Wikipedia, Brazilian Portuguese is a dialect of European Portuguese, Canadian French is a dialect of European French, Serbian and Croatian are standardized registers of Serbo-Croatian, and Indonesian is a standardized register of Malay.

[4]http://wit3.fbk.eu/

[5]http://opus.nlpl.eu/

### 2.3 Multilingual NMT

In a *one-to-many* multilingual translation scenario, Dong et al. (2015) proposed a multi-task learning approach that utilizes a single encoder for source languages and separate attention mechanisms and decoders for every target language. Luong et al. (2015) used distinct encoder and decoder networks for modeling language pairs in a *many-to-many* setting. Firat et al. (2016) introduced a way to share the attention mechanism across multiple languages. A simplified and efficient multilingual NMT approach is proposed by Johnson et al. (2016) and Ha et al. (2016) by prepending language tokens to the input string. This approach has greatly simplified multi-lingual NMT, by eliminating the need of having separate encoder/decoder networks and attention mechanism for every new language pair. In this work we follow a similar strategy, by incorporating an artificial token as a unique *variety flag*.

## 3 NMT into Language Varieties

Our assumption is to translate from language $E$ (English) into each of two varieties $A$ and $B$. We assume to have parallel training data $D_{E\to A}$ and $D_{E\to B}$ for each variety as well as unlabeled data $D_{E\to A\cup B}$. For the sake of experimentation we consider three application scenarios in which a fixed amount of parallel training data $E$-$A$ and $E$-$B$ is partitioned in different ways:

- *Supervised*: all sentence pairs are respectively put in $D_{E\to A}$ and $D_{E\to B}$, leaving $D_{E\to A\cup B}$ empty;
- *Unsupervised*: all sentence pairs are jointly put in $D_{E\to A\cup B}$, leaving $D_{E\to A}$ and $D_{E\to B}$ empty;
- *Semi-supervised*: two-third of $E$-$A$ and $E$-$B$ are, respectively, put in $D_{E\to A}$ and $D_{E\to B}$, and the remaining sentence pairs are put in $D_{E\to A\cup B}$.

**Supervised and Unsupervised Baselines.** For each translation direction we compare three baseline NMT systems. The first system is an unsupervised generic (`Gen`) system trained on the union of the language varieties training data. Notice that `Gen` makes no distinction between $A$ and $B$ and uses all data in an unsupervised way. The second is a supervised variety-specific system (`Spec`) trained on the corresponding language variety training set. The third system (`Ada`) is obtained by adapting the `Gen` system to a specific variety.[6] Adaptation is carried out by simply restarting the training process from the generic model using all the available variety specific training data.

**Supervised Multilingual NMT.** We build on the idea of multilingual NMT (`Mul`), where one single NMT system is trained on the union of $D_{E\to A}$ and $D_{E\to B}$. Each source sentence both at training and inference time is prepended with the corresponding target language variety label ($A$ or $B$). Notice that the multilingual architecture leverages the target forcing symbol both as input to the encoder to build its states, and as initial input to the decoder to trigger the first target word.

**Semi-Supervised Multilingual NMT.** We consider here multilingual NMT models that make also use of unlabeled data $D_{E\to A\cup B}$. The first model we propose, named `M-U`, uses the available data $D_{E\to A}$, $D_{E\to B}$ and $D_{E\to A\cup B}$ as they are, by not specifying any label at training time for entries from $D_{E\to A\cup B}$. The second model, named `M-C2`, works similarly to `Mul`, but relying on a language variety identification module (trained on the target data of $D_{E\to A}$ and $D_{E\to B}$) that maps each unlabeled data point either to $A$ or $B$. The third model, named `M-C3`, can be seen as an enhancement of `M-U`, as the unlabeled data is automatically classified into one of three classes: $A$, $B$, or $A\cup B$. For the third class, like with `M-U`, no label is applied on the source sentence.

## 4 Experimental Set-up

### 4.1 Dataset and Preprocessing

The experimental setting consists of eight target varieties and English as source. We use publicly available datasets from the WIT[3] TED corpus (Cettolo et al., 2012). The summary of the partitioned training, dev, and test sets are given in Table 2, where Tr. 2/3 is the labeled portion of the training set used to train the semi-supervised models, while the other 1/3 are either held out as unlabeled (`M-U`) or classified automatically (`M-C2`, `M-C3`). In the preprocessing stages, we tokenize the corpora and remove lines longer than 70 tokens. The Serbian corpus written in Cyrillic is transliterated into Latin script with CyrTranslit[7]. In addition, to also run a large-data experiment,

[6] We test this system only on the Portuguese varieties.
[7] https://pypi.org/project/cyrtranslit

| | Train | Ratio (%) | Tr. 2/3 | Dev | Test |
|---|---|---|---|---|---|
| pt-BR | 234K | 58.23 | 156K | 1567 | 1454 |
| pt-EU | 168K | 47.77 | 56K | 1565 | 1124 |
| fr-CA | 18K | 10.26 | 12K | 1608 | 1012 |
| fr-EU | 160K | 89.74 | 106K | 1567 | 1362 |
| hr | 110K | 54.20 | 73K | 1745 | 1222 |
| sr | 93K | 45.80 | 62K | 1725 | 1214 |
| id | 105k | 96.71 | 70K | 932 | 1448 |
| ms | 3.6K | 3.29 | 2.4k | 1024 | 738 |
| pt-BR_L | 47.2M | 64.91 | 31.4M | 1567 | 1454 |
| pt-EU_L | 25.5M | 35.10 | 17M | 1565 | 1124 |

Table 2: Number of parallel sentences of the TED Talks used for training, development and testing. At the bottom, the large-data set-up which uses the OpenSubtitles (pt-BR_L and pt-PT_L) as additional training set.

we expand the English−European/Brazilian Portuguese data with the corresponding OpenSubtitles2018 datasets from the OPUS corpus. Table 2 summarizes the augmented training data, while keeping the same dev and test sets.

### 4.2 Experimental Settings

We trained all systems using the Transformer model[8] (Vaswani et al., 2018). We use the Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 0.2 and a dropout also set to 0.2. A shared source and target vocabulary of size 16k is generated via sub-word segmentation (Wu et al., 2016). The choice for the vocabulary size follows the recommendations in Denkowski and Neubig (2017) regarding training of NMT systems on TED Talks data. Overall we use a uniform setting for all our models, with a 512 embedding dimension and hidden units, and 6 layers of self-attention encoder-decoder network. The training batch size is of 6144 sub-word tokens and the max length after segmentation is set to 70. Following Vaswani et al. (2017) and for a fair comparison, experiments are run for 100k training steps, i.e., in the low-resource settings all models are observed to converge within these steps. Adaptation experiments are run to convergence, which requires roughly half of the steps (i.e., 50k) required to train the generic low-resource model. On the other hand, large-data systems are trained for up to 800k steps, which also showed to be a convergence point. For the final evaluation we take the best performing checkpoint on the dev set. All models are trained using Tesla V100-pcie-16gb on a single GPU.

[8]https://github.com/tensorflow/tensor2tensor

| | pt | sr-hr | fr | id-ms | pt_L |
|---|---|---|---|---|---|
| ROC AUC | 82.29 | 88.12 | 80.99 | 81.99 | 52.75 |

Table 3: Performance of language identification on the low-resource and high-resource (pt_L) settings

### 4.3 Language Variety Identification

To automatically identify the language variety of unlabeled target sentences, we train a fastText model (Joulin et al., 2017), a simple yet efficient linear bag of words classifier. We use both word- and character-level $n$-grams as features. In the low-resource condition, we train the classifier on the 2/3 portion of the labeled training data. For the large-data experiment, instead, we used a relatively smaller and independent corpus consisting of 3.3 million pt-BR−pt-EU parallel sentences extracted from OpenSubtitles2018 after filtering out identical sentences pairs and sentences occurring (in any of the two varieties) in the NMT training data. Additionally, low-resource training sentences (fr-CA and ms) are randomly oversampled to mitigate class imbalance.

For each pair of varieties, we train five base classifiers differing in random initialization. In the M−C2 experiments, prediction is determined based on soft fusion voting, i.e., the final label is the argmax of the sum of class probabilities. Due to class skewness in the evaluation set, we report binary classification performance in terms of ROC AUC (Fawcett, 2006) instead of accuracy in Table 3. For M−C3 models, we handle ambiguous examples using the majority voting scheme: in order for a label to be assigned, its softmax probability should be strictly higher than fifty percents according to the majority of the base classifiers, otherwise no tag is applied. On average, this resulted in $<$1% of unlabeled sentences for the small data condition, and about 2% of unlabeled sentences for the large data condition.

## 5 Results and Discussion

We run experiments with all the systems introduced in Section 3, on four pairs of languages varieties. Results are reported in Table 4 for the low-resource setting and in Table 5 for the large data setting.

### 5.1 Low-resource setting

Among the supervised models, which are using all the available training data, the multilingual NMT model Mul outperforms the variety-specific

| | | pt-BR | pt-EU | average |
|---|---|---|---|---|
| Unsuper. | Gen | ↓36.52 | ↓33.75 | 35.14 |
| Supervis. | Spec | ↓35.85 | ↓35.84 | 35.85 |
| ” | Ada | ↓36.54 | ↓36.59 | 36.57 |
| ” | Mul | **37.86** | **38.42** | **38.14** |
| Semi-sup. | M-U | ↓37.09 | 37.59 | 37.34 |
| ” | M-C2 | **37.70** | **38.35** | **38.03** |
| ” | M-C3 | 37.59 | 38.31 | 37.95 |

| | | fr-EU | fr-CA | average |
|---|---|---|---|---|
| Unsuper. | Gen | **33.91** | ↓30.91 | 32.41 |
| Supervis. | Spec | 33.52 | ↓17.13 | 25.33 |
| ” | Mul | 33.40 | **37.37** | **35.39** |
| Semi-sup. | M-U | 33.28 | 37.96 | 35.62 |
| ” | M-C2 | 33.79 | ↑38.60 | 36.20 |
| ” | M-C3 | ↑**34.16** | ↑**39.30** | **36.73** |

| | | hr | sr | average |
|---|---|---|---|---|
| Unsuper. | Gen | ↓21.71 | ↓19.20 | 20.46 |
| Supervis. | Spec | ↓22.50 | ↓19.92 | 21.21 |
| ” | Mul | **23.99** | **21.37** | **22.68** |
| Semi-sup. | M-U | **24.30** | 21.53 | 22.91 |
| ” | M-C2 | 24.14 | 21.26 | 22.70 |
| ” | M-C3 | 24.22 | **21.97** | **23.10** |

| | | id | ms | average |
|---|---|---|---|---|
| Unsuper. | Gen | 26.56 | ↓13.86 | 20.21 |
| Supervis. | Spec | 26.20 | ↓2.73 | 14.47 |
| ” | Mul | **26.66** | **15.77** | **21.22** |
| Semi-sup. | M-U | **26.52** | 15.58 | 21.05 |
| ” | M-C2 | 26.36 | **16.31** | **21.34** |
| ” | M-C3 | 26.40 | 15.23 | 20.82 |

Table 4: BLEU scores of the presented models, trained with unsupervised, supervised and semi-supervised data, from English to Brazilian Portuguese (pt-BR) and European Portuguese (pt-EU), Canadian French (fr-CA) and European French (fr-EU), Croatian (hr) and Serbian (sr), and Indonesian (id) and Malay (ms). Arrows ↓↑ indicate statistically significant differences calculated against `Mul` using bootstrap resampling with $\alpha = 0.05$ (Koehn, 2004).

models on all considered directions. Remarkably, the `Mul` model also outperforms the adapted `Ada` model on the available translation directions. The unsupervised generic model `Gen`, that mixes together all the available data, as expected tends to perform better than the supervised specific models of the less resourced varieties. Particularly, this improvement is observed for Malay (ms) and Canadian French (fr-CA), which respectively represent the 3.3% and 10% of the overall training data used by their corresponding (`Gen`) systems. On the contrary, a degradation is observed for European Portuguese (pt-Eu) and Serbian (sr), which represent 42% and 45% of their respective training sets. Even though very low-resourced varieties can benefit from the mix, it is also evident that the `Gen` model can easily get biased because of the imbalance between the datasets.

In the semi-supervised scenario, we report results with three multilingual systems that integrate the 1/3 of unlabeled data to the training corpus in three different ways: *(i)* without labels (`M-U`), *(ii)* with automatic labels forcing one of two possible classes (`M-C2`), *(iii)* with automatic labels of one of the two options or no label in case of low confidence of the classifier (`M-C3`).

Results show that on average automatic tagging of the unlabeled data is better than leaving them unlabeled, although `M-U` still remains a better choice than using specialized and generic systems. The best between `M-C2` and `M-C3` performs on average from very close to better than the best supervised method.

If we look at the single language variety, the obtained figures are not showing a coherent picture. In particular, in the Croatian-Serbian and Indonesian-Malay pairs the best resourced language seems to benefit more from keeping the data unlabeled (`M-U`). Interestingly, even the worst semi-supervised model performs very close or even better than the best supervised model, which suggests the importance of taking advantage of all available data even if they are not labeled.

Focusing on the statistically significant improvements, the best supervised (`Mul`) is better than the unsupervised (`Gen`), whereas the best semi-supervised (`M-C2` or `M-C3`) is either comparable or better than the best supervised.

### 5.2 High-resource setting

Unlike what observed in the low-resource setting, where `Mul` outperforms `Spec` in the supervised scenario, in the large data condition, variety specific models apparently seem the best choice. Notice, however, that the supervised multilingual system `Mul` provides just a slightly lower level of performance with a simpler architecture (one network in place of two). The unsupervised generic model `Gen`, trained with the mix of the two varieties datasets, performs significantly worse than the other two supervised approaches, this is particularly visible for the pt-EU direction. Very

| | | pt-BR | pt-EU | average |
|---|---|---|---|---|
| Unsuper. | Gen | ↓ 39.78 | ↓ 36.13 | 37.96 |
| Supervis. | Spec | **41.54** | **40.42** | **40.98** |
| ” | Mul | 41.28 | 40.28 | 40.78 |
| Semi-sup. | M-U | 41.21 | 39.88 | 40.55 |
| ” | M-C2 | 41.20 | 40.02 | 40.61 |
| ” | M-C3 | **41.56** | **40.22** | **40.89** |

Table 5: BLEU score on the test set of models trained with large-scale data, from English to Brazilian Portuguese (pt-BR) and European Portuguese (pt-EU). Arrows ↓↑ indicate statistically significant differences calculated against the `Mul` model.

likely, in addition to the ambiguities that arise from naively mixing the data of the two different dialects, there is also a bias effect towards pt-BR which is due to the very unbalanced proportions of data between the two dialects (almost 1:2).

Hence, in the considered high-resource setting, the `Spec` and `Mul` models result as best possible solutions against which comparing our semi-supervised approaches.

In the semi-supervised scenario, the obtained results confirm that our approach of automatically classifying the unlabeled data $D_{E \to A \cup B}$ improves over using the data as they are (`M-U`). Nevertheless, `M-U` still confirms to perform better than the fully unlabeled `Gen` model. In both translation directions, `M-C2` and `M-C3` get quite close to the performance of the supervised `Spec` model. In particular, `M-C3` shows to outperform the `M-C2` model, and even outperforms on average the supervised `Mul` model. In other words, the semi-supervised model leveraging three-class automatic labels (of $D_{E \to A \cup B}$) seems to perform better than the supervised model with two dialect labels. Besides the comparable BLEU scores, the supervised (`Spec` and `Mul`) perform in statistically insignificant way against the best semi-supervised (`M-C3`), although outperforming the unsupervised (`Gen`) model.

This result raises the question if relabeling all the training data can be a better option than using a combination of manual and automatic labels. This issue is investigated in the next subsection.

**Unsupervised Multilingual Models**

As discussed in Section 4.3, the language classifier for the large-data condition is trained on dialect-to-dialect parallel data that does not overlap with the NMT training data. This condition permits hence to investigate a fully unsupervised training condition. In particular, we assume that all the available training data is unlabeled and create automatic language labels for all 47.2M sentences of pt-BR and 25.5M sentences of pt-EU (see Table 2). In a similar way as in Table 5, we keep the experimental setting of `M-C2` and `M-C3` models.

| | | pt-BR | pt-EU | average |
|---|---|---|---|---|
| Unsuper. | M-C2 | 41.50 | **40.21** | 40.86 |
| ” | M-C3 | **41.66** | 40.13 | **40.90** |

Table 6: BLEU scores on the test set by large scale multi-lingual models trained under an unsupervised condition, where all the training data are labeled automatically.

Table 6 reports the results of the multilingual models trained under the above described unsupervised condition. In comparison with the semi-supervised condition, both `M-C2` and `M-C3` show a slight performance improvement. In particular, the three-label `M-C3` performs on average slightly better than the two-label `M-C2` model. Actually, the little difference is justified by the fact that the classifier used the “third” label only for 6% of the data. Remarkably, despite the relatively low performance of the classifier, average score of the best unsupervised model `M-C2` is almost on par with the supervised model `Mul`.

### 5.3 Translation Examples

Finally, in Table 7, we show an additional translation example produced by our semi-supervised multilingual models (both under low and high resource conditions) translating into the Portuguese varieties. For comparison we also include output from Google Translate which offers only a generic English-Portuguese direction. In particular, the examples contain the word *refrigerator* that has specific dialect variants. All our variety-specific systems show to generate consistent translations of this term, while Google Translate prefers to use the Brazilian translation variants for these sentences.

## 6 Conclusions

We presented initial work on neural machine translation from English into dialects and related languages. We discussed both situations where parallel data is supplied or not supplied with target language/dialect labels. We introduced and compared different neural MT models that can be

| | |
|---|---|
| English (source) | We offer a considerable number of different refrigerator models. We have also developed a new type of refrigerator. These include American-style side-by-side refrigerators. |
| pt (Google Translate) | ferecemos um número considerável de modelos diferentes de refrigeradores. Nós também desenvolvemos um novo tipo de geladeira. Estes incluem refrigeradores lado a lado estilo americano. |
| **Low-resource models** | |
| pt-BR (M-C2) | Nós oferecemos um número considerável de diferentes modelos de refrigerador. Também desenvolvemos um novo tipo de refrigerador. Eles incluem o estilo americano nas geladeiras lado a lado. |
| pt-EU (M-C2) | Oferecemos um número considerável de modelos de refrigeração diferentes. Também desenvolvemos um novo tipo de frigorífico. Também desenvolvemos um novo tipo de frigorífico. |
| **High-resource models** | |
| Spec-pt-BR | Oferecemos um nmero considerável de modelos de geladeira diferentes. Também desenvolvemos um novo tipo de geladeira. Isso inclui o estilo americano lado a lado refrigeradores. |
| Spec-pt-PT | Oferecemos um número considerável de modelos de frigorífico diferentes. Também desenvolvemos um novo tipo de frigorfico. Estes incluem frigoríficos americanos lado a lado. |
| pt-BR (M-C3_L) | Oferecemos um número considerável de diferentes modelos de geladeira. Também desenvolvemos um novo tipo de geladeira. Estes incluem estilo americano lado a lado, geladeiras. |
| pt-PT (M-C3_L) | Oferecemos um número considerável de diferentes modelos frigoríficos. Também desenvolvemos um novo tipo de frigorfico. Estes incluem estilo americano lado a lado frigoríficos. |

Table 7: English to Portuguese translation generated by Google Translate (as of 20th July, 2018) and translations into Brazilian and European Portuguese generated by our semi-supervised multilingual (M-C2 and M-C3_L) and supervised Spec models. For the underlined English terms both their Brazilian and European translation variants are shown.

trained under unsupervised, supervised, and semi-supervised training data regimes. We reported experimental results on the translation from English to four pairs of language varieties with systems trained under low-resource conditions. We show that in the supervised regime, best performance is achieved by training a multilingual NMT system. For the semi-supervised regime, we compared different automatic labeling strategies that permit to train multilingual neural MT systems with performance comparable to the best supervised NMT system. Our findings were also confirmed by large scale experiments performed on English to Brazilian and European Portuguese. In this scenario, we have also shown that multilingual NMT fully trained on automatic labels can perform very similarly to its supervised version.

In future work, we plan to extend our approach to language varieties in the source side, as well as investigate the possibility of applying transfer-learning (Zoph et al., 2016; Nguyen and Chiang, 2017) for language varieties by expanding our Ada adaptation approach.

## Acknowledgments

This work has been partially supported by the EC-funded project ModernMT (H2020 grant agreement no. 645487). We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research. Moreover, we thank the Erasmus Mundus European Program in Language and Communication Technology.

## References

Kemal Altintas and lyas Çiçekli. 2002. A Machine Translation System Between a Pair of Closely Related Languages. In *Proceedings of the 17th International Symposium on Computer and Information Sciences (ISCIS 2002)*, pages 192–196.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2018. Neural versus phrase-based mt quality: An in-depth analysis on english-german and english-french. *Computer Speech & Language*, 49:52–70.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit$^3$: Web inventory of transcribed and

translated talks. In *Proceedings of the 16$^{th}$ Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy.

Marta R Costa-jussà. 2017. Why Catalan-Spanish Neural Machine Translation? Analysis, comparison and combination with standard Rule and Phrase-based technologies. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 55–62.

Marta R Costa-Jussà, Marcos Zampieri, and Santanu Pal. 2018. A Neural Approach to Language Variety Translation. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 275–282.

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.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *ACL (1)*, pages 1723–1732.

Nadir Durrani, Hassan Sajjad, Alexander Fraser, and Helmut Schmid. 2010. Hindi-to-urdu machine translation through transliteration. In *Proceedings of the 48th Annual meeting of the Association for Computational Linguistics*, pages 465–474. Association for Computational Linguistics.

Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv:1601.01073*.

Cyril Goutte, Serge Léger, Shervin Malmasi, and Marcos Zampieri. 2016. Discriminating Similar Languages: Evaluations and Explorations. In *Proceedings of Language Resources and Evaluation (LREC)*, pages 1800–1807.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *arXiv preprint arXiv:1611.04798*.

Salima Harrat, Karima Meftouh, and Kamel Smaili. 2017. Machine translation for Arabic dialects (survey). *Information Processing & Management*, pages 1–12.

Hany Hassan, Mostafa Elaraby, and Ahmed Y Tawfik. 2017. Synthetic Data for Neural Machine Translation of Spoken-Dialects. In *Proceedings of the 14th International Workshop on Spoken Language Translation*.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2018. Automatic Language Identification in Texts: A Survey.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 427–431.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, volume 4, pages 388–395.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubeši, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016. Discriminating Between Similar Languages and Arabic Dialect Identification: A Report on the Third DSL Shared Task. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–14.

Luis Marujo, Nuno Grazina, Tiago Luis, Wang Ling, Luisa Coheur, and Isabel Trancoso. 2011. BP2EP - Adaptation of Brazilian Portuguese texts to European Portuguese. In *Proceedings of the 15th International Conference of the European Association for Machine Translation (EAMT)*, May, pages 129–136.

Toan Q Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 296–301.

Maja Popović, Mihael Arcan, and Filip Klubička. 2016. Language Related Issues for Machine Translation between Closely Related South Slavic Languages. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 43–52.

Nima Pourdamghani and Kevin Knight. 2017. Deciphering Related Languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2503–2508.

Wael Salloum, Heba Elfardy, Linda Alamir-Salloum, Nizar Habash, and Mona Diab. 2014. Sentence Level Dialect Identification for Machine Translation System Selection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 772–778.

Tien-Ping Tan, Sang-Seong Goh, and Yen-Min Khaw. 2012. A Malay Dialect Translation and Synthesis System: Proposal and Preliminary System. In *2012 International Conference on Asian Language Processing*, pages 109–112. IEEE.

Jörg Tiedemann and Nikola Ljubeši. 2012. Efficient Discrimination Between Closely Related Languages. In *Proceedings of COLING 2012: Technical Papers*, pages 2619–2634.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2tensor for neural machine translation. *CoRR*, abs/1803.07416.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

Ronald Wardhaugh. 2006. *An Introduction to Sociolinguistcs*. Blackwell Publishing.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubeši, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial Evaluation Campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–15.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Nikola Ljube. 2014. A Report on the DSL Shared Task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, 2013, pages 58–67.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. 2015. Overview of the DSL Shared Task 2015. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, 2014, pages 1–9.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.