Improved Neural Machine Translation using Side Information

Cong Duy Vu Hoang† and **Gholamreza Haffari**‡ and **Trevor Cohn**†
† University of Melbourne, Melbourne, VIC, Australia
‡ Monash University, Clayton, VIC, Australia
vhoang2@student.unimelb.edu.au, gholamreza.haffari@monash.edu,
t.cohn@unimelb.edu.au

Abstract

In this work, we investigate whether side information is helpful in the context of neural machine translation (NMT). We study various kinds of side information, including topical information and personal traits, and then propose different ways of incorporating these information sources into existing NMT models. Our experimental results show the benefits of side information in improving the NMT models.

1 Introduction

Neural machine translation is the task of generating a target language sequence given a source language sequence, framed as a neural network (Sutskever et al., 2014; Bahdanau et al., 2015, *inter alia*). Most research efforts focus on inducing more prior knowledge (Cohn et al., 2016; Zhang et al., 2017; Mi et al., 2016, *inter alia*), incorporating linguistics factors (Hoang et al., 2016b; Sennrich and Haddow, 2016; García-Martínez et al., 2017) or changing the network architecture (Gehring et al., 2017b,a; Vaswani et al., 2017; Elbayad et al., 2018) in order to better exploit the source representation. Consider a different direction, situations in which there exists other modality other than the text of the source sentence. For instance, the WMT 2017 campaign[1] proposed to use additional information obtained from *images* to enrich the neural MT models, as in (Calixto et al., 2017; Matusov et al., 2017; Calixto and Liu, 2017). This task, also known as multi-modal translation, seeks to leverage images which can contain cues representing the perception of the image in source text, and potentially can contribute to resolve ambiguity (e.g., lexical, gender), vagueness, out-of-vocabulary terms, and topic relevancy.

Inspired from the idea of multi-modal translation, in our work, we propose the use of another modality, namely metadata or side information. Previously, Hoang et al. (2016a) have shown the usefulness of side information for neural language models. This work will investigate the potential usefulness of side information for NMT models. In our work, we target towards *unstructured and heterogeneous* side information which potentially can be found in practical applications. Specifically, we investigate different kinds of side information, including topic keywords, personality information and topic classification. Then we study different methods with minimal efforts for incorporating such side information into existing NMT models.

2 Machine Translation Data with Side Information

First, let's explore some realistic scenarios in which the side information is potentially useful for NMT.

TED Talks The TED Talks website[2] hosts technical videos from influential speakers around the world on various topics or domains, such as: education, business, science, technology, creativity, etc. Thanks to users' contributions, most of such videos are subtitled in multiple languages. Based on this website, Cettolo et al. (2012) created a parallel corpus for the MT research community. Inspired by this, Chen et al. (2016) further customised this dataset and included an additional sentence-level topic information.[3] We consider such topic information as side information. Fig-

[1] http://www.statmt.org/wmt17/multimodal-task.html

[2] https://www.ted.com/talks

[3] https://github.com/wenhuchen/iwslt-2015-de-en-topics

Cong Duy Vu Hoang, Gholamreza Haffari and Trevor Cohn. 2018. Improved Neural Machine Translation using Side Information. In *Proceedings of Australasian Language Technology Association Workshop*, pages 6–16.

ure 1 illustrates some examples of this dataset. As can be seen, the keywords (second column, treated as side information) contain additional contextual information that can provide complementary cues so as to better guide the translation process. Let's take an example in Figure 1 (TED Video Id 172), the keyword "art" provides cues for words and phrases in target sequence such as: "place, design"; whereas the keyword "tech" refers to "Media Lab, computer science".

Personalised Europarl For the second dataset, we evaluate our proposed idea in the context of personality-aware MT. Mirkin et al. (2015) explored whether translation preserves personality information (e.g., demographic and psychometric traits) in statistical MT (SMT); and further Rabinovich et al. (2017) found that personality information like author's gender is an obvious signal in source text, but it is less clear in human and machine translated texts. As a result, they created a new dataset for personalised MT[4] partially based on the original Europarl. The personality such as author's gender will be regarded as side information in our setup. An excerpt of this dataset is shown in Figure 2. As can be seen from the figure, there exist many kinds of side information pertaining to authors' traits, including identification (ID, name), native language, gender, date of birth/age, and plenary session date. Here, we will focus on the "gender" trait and evaluate whether it can have any benefits in the context of NMT complementing the work of Rabinovich et al. (2017) attempted a similar idea as part of a SMT, rather than NMT, system.

Patent MT Collection Another interesting data is patent translation which includes rich side information. PatTR[5] is a sentence-parallel corpus which is a subset of the MAREC Patent Corpus (Wäschle and Riezler, 2012a). In general, PatTR contains millions of parallel sentences collected from all patent text sections (e.g., title, abstract, claims, description) in multiple languages (English, French, German) (Wäschle and Riezler, 2012b; Simianer and Riezler, 2013). An appealing feature of this corpus is that it provides a labelling at a sentence level, in the form of IPC (International Patent Classification) codes. The IPC codes explicitly provide a hierarchical classification of patents according to various different areas of technology to which they pertain. This kind of side information can provide a useful signal for MT task – which has not yet been fully exploited. Figure 3 gives us an illustrating excerpt of this corpus. We can see that each of sentence pair in this corpus is associated with any number of IPC label(s) as well as other metadata, e.g., patent ID, patent family ID, publication date. In this work, we consider only the IPC labels. The full meaning of all IPC labels can be found on the official IPC website,[6] however we provide in Figure 3 the glossess for each referenced label. Note that those IPC labels form a WordNet style hierarchy (Fellbaum, 1998), and accordingly may be useful in many other deep models of NLP.

[4] http://cl.haifa.ac.il/projects/pmt/index.shtml

[5] http://www.cl.uni-heidelberg.de/statnlpgroup/pattr/

3 NMT with Side Information

We investigate different ways of incorporating side information into the NMT model(s).

3.1 Encoding of Side Information

In this work, we propose the use of unstructured *heterogeneous* side information, which is often available in practical datasets. Due the heterogeneity of side information, our techniques are based on a bag-of-words (BOW) representation of the side information, an approach which was shown to be effective in our prior work (Hoang et al., 2016a). Each element of the side information (a label, or word) is embedded using a matrix $\boldsymbol{W}_e^s \in \mathcal{R}^{H_s \times |V_s|}$, where $|V_s|$ is the vocabulary of side information and H_s the dimensionality of the hidden space. These embedding vectors are used for the input to several different neural architectures, which we now outline.

3.2 NMT Model Formulation

Recall the general formulation of NMT (Sutskever et al., 2014; Bahdanau et al., 2015, *inter alia*) as a conditional language model in which the generation of target sequence is conditioned on the source sequence (Sutskever et al., 2014; Bahdanau et al., 2015, *inter alia*), formulated as:

$$\begin{aligned} \boldsymbol{y}_{t+1} &\sim \mathrm{p}_{\boldsymbol{\Theta}}\left(y_{t+1}|\boldsymbol{y}_{<t}, \boldsymbol{x}\right) \\ &= \mathrm{softmax}\left(\mathrm{f}_{\boldsymbol{\Theta}}\left(\boldsymbol{y}_{<t}, \boldsymbol{x}\right)\right); \end{aligned} \quad (1)$$

[6] http://www.wipo.int/classifications/ipc/en/

TED Video Id	Keywords	German Sentence	English Sentence
172	arts,tech	Aber das Media Lab ist ein interessanter Ort, und es ist wichtig für mich, denn ich studierte ursprünglich Computerwissenschaften und erst später in meinem Leben habe ich Design entdeckt.	But the Media Lab is an interesting place, and it's important to me because as a student, I was a computer science undergrad, and I discovered design later on in my life.
645	politics,issues,business	In anderen Worten, ich glaube, dass der französische Präsident Sarkozy recht hat, wenn er über eine Mittelmeer Union spricht.	So in other words, I believe that President Sarkozy of France is right when he talks about a Mediterranean union.
1193	recreation,arts,issues	Eine andere Welt tat sich ungefähr zu dieser Zeit auf: Auftreten und Tanzen.	Another world was opening up around this time: performance and dancing.
692	politics,arts,issues,env	Dieses Gebäude beinhaltet die Weltgrößte Kollektion von Sammlungen und Artefakten die der Rolle der USA im Kampf auf der Chinesischen Seite gedenken. In diesem langen Krieg -- die "fliegenden Tiger".	This building contains the world's largest collection of documents and artifacts commemorating the U.S. role in fighting on the Chinese side in that long war -- the Flying Tigers.
1087	politics,education	Es erlaubt uns, Kunst, Biotechnologie, Software und all solch wunderbaren Dinge zu schaffen.	It allows us to do the art, the biotechnology, the software and all those magic things.
208	recreation,education,arts,issues	Ich liebe Bartóks Musik, so wie Herr Teszler, und er hatte wirklich jede Aufnahme Bartóks die es gab.	I love Bartok's music, as did Mr. Teszler, and he had virtually every recording of Bartok's music ever issued.

Fig. 1 An example with side information (e.g., keywords) for MT with TED Talks dataset.

English Sentence: Accordingly , I consider it essential that both the identification of cattle and the labelling of beef be introduced as quickly as possible on a compulsory basis .
German Sentence: Entsprechend halte ich es auch für notwendig , daß die Kennzeichnung möglichst schnell und verpflichtend eingeführt wird , und zwar für Rinder und für Rindfleisch .
Meta Info: EUROID="2209" NAME="Schierhuber" LANGUAGE="DE" GENDER="FEMALE" DATE_OF_BIRTH="31 May 1946" SESSION_DATE="97-02-19" AGE="50"

English Sentence: Can the Commission say that it will seek to have sugar declared a sensitive product ?
German Sentence: Kann die Kommission sagen , dass sie danach streben wird , Zucker zu einem sensiblen Produkt erklären zu lassen ?
Meta Info: EUROID="22861" NAME="Ó Neachtain (UEN)." LANGUAGE="EN" GENDER="MALE" DATE_OF_BIRTH="22 May 1947" SESSION_DATE="03-09-02" AGE="56"

English Sentence: For example , Brazil has huge concerns about the proposals because the poor and landless there will suffer if sugar production expands massively , as is predicted .
German Sentence: So hegt beispielsweise Brasilien bezüglich der Vorschläge enorme Bedenken , denn wenn die Zuckerproduktion , wie vorhergesagt , massiv expandiert , wird das die Not der Armen und Landlosen dort noch verstärken .
Meta Info: EUROID="28115" NAME="McGuinness (PPE-DE)." LANGUAGE="EN" GENDER="FEMALE" DATE_OF_BIRTH="13 June 1959" SESSION_DATE="05-02-22" AGE="45"

English Sentence: The European citizens ' initiative should be seen as an opportunity to involve people more closely in the EU 's decision-making process .
German Sentence: Die Europäische Bürgerinitiative ist als Chance zu werten , um die Menschen stärker in den Entscheidungsprozess der EU miteinzubeziehen .
Meta Info: EUROID="96766" NAME="Ernst Strasser" LANGUAGE="DE" GENDER="MALE" DATE_OF_BIRTH="29 April 1956" SESSION_DATE="10-12-15-010" AGE="54"

Fig. 2 An example with side information (e.g., author's gender highlighted in red) for MT with personalised Europarl dataset.

English Sentence: In the case of the actual value coinciding with and/or deviating from the desired value input, the device emits audible signals.
German Sentence: Bei Übereinstimmung und/oder Abweichung des Istwertes von der Sollwerteingabe gibt die Vorrichtung akustische Signale ab.
Meta Info: EP-0017737-A1 6068117 19801029 G01D,G01P

English Sentence: The invention relates to a feed device for teeth (21) which forms part of a device for connecting a saw-blade base body to teeth (22, 23) that are subdivided into a metallic and a non-metallic material area (25, 26).
German Sentence: Eine Zahnzuführeinrichtung (21) ist Teil einer Vorrichtung zum Verbinden eines Sägeblattgrundkörpers mit Zähnen (22, 23), die in einen metallischen und einen nicht-metallischen Materialbereich (25, 26) unterteilt sind.
Meta Info: WO-2001002130-A1 7913309 20010111 B23K,B23D

G –> PHYSICS
G01 –> MEASURING; TESTING
G01D –> MEASURING NOT SPECIALLY ADAPTED FOR A SPECIFIC VARIABLE; ARRANGEMENTS FOR MEASURING TWO OR MORE VARIABLES NOT COVERED BY A SINGLE OTHER SUBCLASS; TARIFF METERING APPARATUS; TRANSFERRING OR TRANSDUCING ARRANGEMENTS NOT SPECIALLY ADAPTED FOR A SPECIFIC VARIABLE; MEASURING OR TESTING NOT OTHERWISE PROVIDED FOR
G01P –> MEASURING LINEAR OR ANGULAR SPEED, ACCELERATION, DECELERATION OR SHOCK; INDICATING PRESENCE OR ABSENCE OF MOVEMENT; INDICATING DIRECTION OF MOVEMENT

B –> PERFORMING OPERATIONS; TRANSPORTING
B23 –> MACHINE TOOLS; METAL-WORKING NOT OTHERWISE PROVIDED FOR
B23K –> SOLDERING OR UNSOLDERING; WELDING; CLADDING OR PLATING BY SOLDERING OR WELDING; CUTTING BY APPLYING HEAT LOCALLY, e.g. FLAME CUTTING; WORKING BY LASER BEAM
B23D –> PLANING; SLOTTING; SHEARING; BROACHING; SAWING; FILING; SCRAPING; LIKE OPERATIONS FOR WORKING METAL BY REMOVING MATERIAL, NOT OTHERWISE PROVIDED FOR

Fig. 3 An example with side information (e.g., IPC highlighted in red) for MT with PatTR dataset.

where the probability $\mathrm{p}_{\boldsymbol{\Theta}}(.)$ of generating the next target word y_{t+1} is conditioned on the previously generated target words $\boldsymbol{y}_{<t}$ and the source sequence $\boldsymbol{x}$; f is a neural network which can be framed as an encoder-decoder model (Sutskever et al., 2014) and can use an attention mechanism (Bahdanau et al., 2015; Luong et al., 2015). In this model, the encoder encodes the information of the source sequence; whereas, the decoder decodes the target sequence sequentially from left-to-right. The attention mechanism controls which parts of the source sequence where the decoder should attend to in generating each symbol of target sequence. Later, advanced models have been proposed with modifications of the encoder and decoder architectures, e.g., using the 1D (Gehring et al., 2017b,a) and 2D (Elbayad et al., 2018) convolutions; or a transformer network (Vaswani et al., 2017). These advanced models have led to significantly better results in terms of both performance and efficiency via different benchmarks (Gehring et al., 2017b,a; Vaswani et al., 2017; Elbayad et al., 2018).

Regardless of the NMT architecture, we aim to explore in which case side information can be useful, as well as the effective and efficient way of incorporating them with minimal modification of the NMT architecture. Mathematically, we formulate the NMT problem given the availability of side information $\boldsymbol{e}$ as follows:

$$\begin{aligned}\boldsymbol{y}_{t+1} &\sim \mathrm{p}_{\boldsymbol{\Theta}}\left(y_{t+1}|\boldsymbol{y}_{<t}, \boldsymbol{x}, \boldsymbol{e}\right) \\ &= \mathrm{softmax}\left(\mathrm{f}_{\boldsymbol{\Theta}}\left(\boldsymbol{y}_{<t}, \boldsymbol{x}, \boldsymbol{e}\right)\right); \end{aligned} \quad (2)$$

where $\boldsymbol{e}$ is the representation of additional side information we would like to incorporate into NMT model.

3.3 Conditioning on Side Information

Keeping in mind that we would like a generic incorporation method so that only minimal modification of NMT model is required, we propose and evaluate different approaches.

Side Information as Source Prefix/Suffix The most simple way to include side information is to add the side information as a string prefix or suffix to the source sequence, and letting the NMT model learn from this modified data. This method requires no modification of the NMT model. This method was firstly proposed by Sennrich et al. (2016a) who added the side constraints (e.g., honorifics) as suffix of the source sequence for controlling the politeness in translated outputs.

Side Information as Target Prefix Alternatively, we can add the bag of words as a target prefix, inspired from Johnson et al. (2017) who introduces an artificial token as a prefix for specifying the required target language in a multilingual NMT system. Note that this method leads to additional benefits in the following situations: a) when the side information exists, the model takes them as inputs and then does its translation task as normal; b) when the side information is missing, so the model first generates the side information itself and subsequently uses it to proceed with translation.

Output Layer Similar to Hoang et al. (2016a) – who considers side information in the model focusing on the output side which worked well in LM, this method involves in two phases. First, it transforms the representation of the side information into a *summed* vector representation, $\boldsymbol{e} = \sum_{m\in[1,M]} \boldsymbol{e}_{w_m^s}$. We also tried the *average* operator in our preliminary experiments but observed no difference in end performance.

Next, the side representation vector, $\boldsymbol{e}$, is added to the *output layer* before the $\mathrm{softmax}$ transformation of the NMT model, e.g.,

$$\begin{aligned} y_{t+1} &\sim \mathrm{softmax}\left(\boldsymbol{W}_o \cdot \mathrm{f}_t^{dec}\left(\ldots\right) + \boldsymbol{b}_e + \boldsymbol{b}_o\right) \\ \boldsymbol{b}_e &= \boldsymbol{W}_e \cdot \boldsymbol{e}; \end{aligned} \quad (3)$$

where $\boldsymbol{W}_e \in \mathcal{R}^{|V_T|\times H_s}$ is an additional weight matrix (learnable model parameters) for linear projection of side information representation onto the target output space (H_s is a predefined dimension for embedding side information). The rationale behind this method is to let the model learn to control the importance of the existing side information contributed to the generation. The function $\mathrm{f}_t^{dec}(\ldots)$ is specific to our chosen network reparameterisation, based on RNN (Sutskever et al., 2014; Bahdanau et al., 2015; Luong et al., 2015), or convolution (Gehring et al., 2017b,a; Elbayad et al., 2018), or transformer (Vaswani et al., 2017). Although we make an effort for modification of the NMT model, we believe that it is minimally simple, and generic to suit many different styles of NMT model.

Multi-task Learning Consider the case where we would like to use existing side information to

improve the main NMT task. We can define a generative model $\mathrm{p}(\boldsymbol{y}, \boldsymbol{e}|\boldsymbol{x})$, formulated as:

$$\mathrm{p}(\boldsymbol{y}, \boldsymbol{e}|\boldsymbol{x}) := \underbrace{\mathrm{p}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{e})}_{\text{translation model}} \cdot \underbrace{\mathrm{p}(\boldsymbol{e}|\boldsymbol{x})}_{\text{classification model}} ; \quad (4)$$

where $\mathrm{p}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{e})$ is a translation model conditioned on the side information as explained earlier; $\mathrm{p}(\boldsymbol{e}|\boldsymbol{x})$ can be regarded as a classification model – which predicts the side information given the source sentence. Note that side information can often be represented as individual words – which can be treated as labels, making the classification model feasible.

Importantly, the above formulation of a generative model would require summing over "$\boldsymbol{e}$" at test/decode time, which might be done by decoding for all possible label combinations, then reporting the sentence with the highest model score. This may be computationally infeasible in practice. We resort this by approximating the NMT model as $\mathrm{p}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{e}) \approx \mathrm{p}(\boldsymbol{y}|\boldsymbol{x})$, resulting in

$$\mathrm{p}(\boldsymbol{y}, \boldsymbol{e}|\boldsymbol{x}) \approx \mathrm{p}(\boldsymbol{y}|\boldsymbol{x}) \cdot \mathrm{p}(\boldsymbol{e}|\boldsymbol{x}) ; \quad (5)$$

and thus force the model to encode the shared information in the encoder states.

Our formulation in Equation 5 gives rise to multi-task learning (MTL). Here, we propose the joint learning of two different but related tasks: NMT and multi-label classification (MLC). Here, the MLC task refers to predicting the labels that possibly represent words of the given side information. This is interesting in the sense that the model is capable of not only generating the translated outputs, but also explicitly predicting what the side information is. Here, we adopt a simple instance of MTL for our case, called soft parameter sharing similar to (Duong et al., 2015; Yang and Hospedales, 2016). In our MTL version, the NMT and MLC tasks share the parameters of the encoders. The difference between the two is at the decoder part. In the NMT task, the decoder is kept unchanged. For the MLC task, we define its objective function (or loss), formulated as:

$$\mathcal{L}_{MLC} := -\sum_{m=1}^{M} \mathbb{1}_{w_m^s}^T \log \mathrm{p}_s ; \quad (6)$$

where p_s is the probability of predicting the presence or absence of each element in the side information, formulated as:

$$p_s = \mathrm{sigmoid}\left(\boldsymbol{W}_s \left[\frac{1}{|\boldsymbol{x}|}\sum_i \mathrm{g}'(x_i)\right] + \boldsymbol{b}_s\right) ; \quad (7)$$

where $\boldsymbol{x}$ is the source sequence, comprising of $x_1, \ldots, x_i, \ldots, x_{|\boldsymbol{x}|}$ words. Here, we denote a generic function term $\mathrm{g}'(.)$ which refers to a vectorised representation of a specific word depending on designing the network architecture, e.g., stacked bidirectional (forward and backward) networks over the source sequence (Bahdanau et al., 2015; Luong et al., 2015); or a convolutional encoder (Gehring et al., 2017b,a) or a transformer encoder (Vaswani et al., 2017).[7] Further, $\boldsymbol{W}_s \in \mathcal{R}^{|V_s| \times H_x}$ and $\boldsymbol{b}_s \in \mathcal{R}^{|V_s|}$ are two additional model parameters for linear transformation of the source sequence representation (where H_x is a dimension of the output of the $\mathrm{g}'(.)$ function, it will differ from network architectures as discussed earlier).

Now, we have two objective functions at the *training* stage, including the NMT loss $\mathcal{L}_{NMT}$ and the MLC loss $\mathcal{L}_{MLC}$. The total objective function of our joint learning will be:

$$\mathcal{L} := \mathcal{L}_{NMT} + \lambda \mathcal{L}_{MLC}; \quad (8)$$

where: λ is the coefficient balancing the two task objectives, whose value is fine-tuned based on the development data to optimise for NMT accuracy measured using BLEU (Papineni et al., 2002).

The idea of MTL applied for NLP was firstly explored by (Collobert and Weston, 2008), later attracts increasing attentions from the NLP community (Ruder, 2017). Specifically, the idea behind MTL is to leverage related tasks which can be learned jointly — potentially introducing an inductive bias (Feinman and Lake, 2018). An alternative explanation of the benefits of MTL is that joint training with multiple tasks acts as an additional regulariser to the model, reducing the risk of overfitting (Collobert and Weston, 2008; Ruder, 2017, *inter alia*).

4 Experiments

4.1 Datasets

As discussed earlier, we conducted our experiments using three different datasets including TED Talks (Chen et al., 2016), Personalised Europarl

[7] Here, to avoid repeating the materials, we will not elaborate their formulations.

	No. of labels	Examples
TED Talks	11	tech business arts issues education health env recreation politics others
Personalised Europarl	2	male female
PatTR-1 (deep)	651	G01G G01L G01N A47F F25D C01B ...
PatTR-2 (shallow)	8	G A F C H B E D

Table 1 Side information statistics for the three datasets, showing the number of types of the side information label, and the set of tokens (display truncated for PatTR-1 (deep)).

(Rabinovich et al., 2017), and PatTR (Wäschle and Riezler, 2012b; Simianer and Riezler, 2013), translating from German (de) to English (en). The statistics of the training and evaluation sets can be shown in Table 2. For the TED Talks and Personalised Europarl datasets, we followed the same sizes of data splits since they are made available on the authors' github repository and website. For the PatTR dataset, we use the *Abstract* sections for patents from 2008 or later, and the development and test sets are constructed to have 2000 sentences each, similar to (Wäschle and Riezler, 2012b; Simianer and Riezler, 2013).

It is important to note the labeling information for side information. We extracted all kinds of side information from three aforementioned datasets in the form of individual words or labels. This makes the label embeddings much easier. Their relevant statistics and examples can be found in Table 1.

We preprocessed all the data using Moses's training scripts[8] with standard steps: punctuation normalisation, tokenisation, truecasing. For training sets, we set word-based length thresholds for filtering long sentences since they will not be useful when training the seq2seq models as suggested in the NMT literature (Sutskever et al., 2014; Bahdanau et al., 2015; Luong et al., 2015, *inter alia*). We chose 80, 80, 150 length thresholds for TED Talks, Personalized Europarl, and PatTR datasets, respectively. Note that the 150 threshold indicates that the sentences in the PatTR dataset is in average much longer than in the others. For better handling the OOV problem, we segmented all the preprocessed data with subword units using byte-pair-encoding (BPE) method proposed by Sennrich et al. (2016b). We already know that languages such English and German share an alphabet (Sennrich et al., 2016b), hence learning BPE on the concatenation of source and target languages (hence called shared BPE) increases the consistency of the segmentation. We applied 32000 operations for learning the shared BPE by using the open-source toolkit.[9] Also, we used dev sets for tuning model parameters and early stopping of the NMT models based on the perplexity. Table 2 shows the resulting vocabulary sizes after subword segmentation for all datasets.

4.2 Baselines and Setups

Recall that our method for incorporating the additional side information into the NMT models is generic; hence, it is applicable to any NMT architecture. We chose the transformer architecture (Vaswani et al., 2017) for all our experiments since it arguably is currently the most robust NMT models compared to RNN and convolution based architectures. We re-implemented the transformer - based NMT system using the C++ Neural Network Library - DyNet[10] as our deep learning backend toolkit. Our re-implementation results in the open source toolkit.[11]

In our experiments, we use the same configurations for all transformer models and datasets, including: 2 encoder and decoder layers; 512 input embedding and hidden layer dimensions; sinusoid positional encoding; dropout with 0.1 probability for source and target embeddings, sub-layers (attention + feedforward), attentive dropout; and label smoothing with weight 0.1. For training our neural models, we used early stopping based on development perplexity, which usually occurs after 20-30 epochs.[12]

We conducted our experiments with various incorporation methods as discussed in Section 3. We

[8] https://github.com/moses-smt/mosesdecoder/tree/master/scripts

[9] https://github.com/rsennrich/subword-nmt

[10] https://github.com/clab/dynet/

[11] https://github.com/duyvuleo/Transformer-DyNet/

[12] The training process of transformer models is much faster than the RNN and convolution - based ones, but requires more epochs for convergence.

dataset	# tokens (M)		# types (K)		# sents	# length limit
TED Talks de→en						
train	3.73	3.75	19.78	14.23	163653	80
dev	0.02	0.02	4.03	3.15	567	n.a.
test	0.03	0.03	6.07	4.68	1100	n.a.
Personalised Europarl de→en						
train	8.46	8.39	21.15	14.04	278629	80
dev	0.16	0.16	14.67	9.83	5000	n.a.
test	0.16	0.16	14.76	9.88	5000	n.a.
PatTR de→en						
train	33.07	32.52	24.97	13.28	656352	150
dev	0.13	0.13	13.50	6.88	2000	n.a.
test	0.13	0.12	13.35	6.89	2000	n.a.

Table 2 Statistics of the training & evaluation sets from datasets including TED Talks, Personalised Europarl, and PatTR; showing in each cell the count for the source language (left) and target language (right); "#types" refers to subword-segmented vocabulary sizes; "n.a." is not applicable, for development and test sets. Note that all the "#tokens"' and "#types"' are approximated.

Method	TED Talks	Personalised Europarl	PatTR-1	PatTR-2
base	29.48	31.12	45.86	
si−src−prefix	29.28	30.87	**45.99**	**45.97**
si−src−suffix	29.36	31.03	**46.01**	45.83
si−trg−prefix−p	29.06	30.89	**45.97**	45.85
si−trg−prefix−h	29.28	30.93	**46.03**	**45.92**
output−layer	**29.99**†	**31.22**	**46.32**†	**46.09**
w/o *side info*	**29.62**	31.10	**46.14**	**45.99**
mtl	**29.86**†	31.12	**46.14**	**46.01**

Table 3 Evaluation results with BLEU scores of various incorporation variants against the baseline; **bold**: better than the baseline, †: statistically significantly better than the baseline.

denote the system variants as follows:

base refers to the baseline NMT system using the transformer without using any side information.

si-src-prefix and *si-src-suffix* refer to the NMT system using the side information as respective prefix or suffix of the source sequence (Jehl and Riezler, 2018), applied to both training and decoder/inference.

si-trg-prefix refers to the NMT system using the side information as prefix of the target sequence. There are two variants, including "*si-trg-prefix-p*" means the side information is generated by the model itself and is then used for decoding/inference; "*si-trg-prefix-h*" means the side information is given at decoding/inference runtime.

output-layer refers to the method of incorporating side information in the final output layer.

mtl refers to the multi-task learning method.

It's worth noting that the dimensional value for the *output-layer* method was fine-tuned over the development set, using the value range of $\{64, 128, 256, 512\}$. Similarly, the balancing weight in the *mtl* method is fine-tuned using the value range of $\{0.001, 0.01, 0.1, 1.0\}$. For evaluation, we measured the end translation quality with case-sensitive BLEU (Papineni et al., 2002). We averaged 2 runs for each of the method variants.

4.3 Results and Analysis

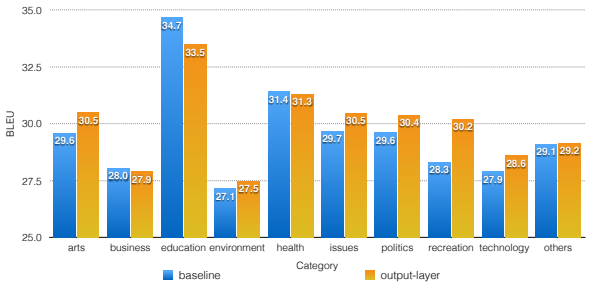
The experimental results can be seen in Table 3. Overall, we obtained limited success for the method of adding side information as prefix or suffix for TED Talks and Personalised Europarl datasets. On the PatTR dataset, small improvements (0.1-0.2 BLEU) are observed. We experimented two sets of side information in the PatTR dataset, including PatTR-1 (651 deep labels) and PatTR (8 shallow labels).[13] The possible reason for this phenomenon is that the multi-head attention mechanism in the transformer may have some confusion given the existing side information, ei-

[13]The shallow setting takes the first character of each label code, which denotes the highest level concept in the type hierarchy, e.g., G01P (measuring speed) → G (physics), with definitions as shown in Fig 3.

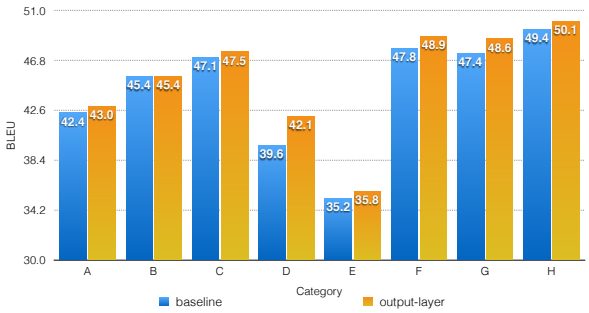
ther in source or target sequences. In some ambiguous cases, the multi-head attention may not know where it should pay more attention. Another possible reason is that the implicit ambiguity of side information that may exist in the data.

Contrary to these variants, the *output-layer* variant was more consistently successful, obtaining the best results across datasets. In the TED Talks and PatTR datasets, this method also provides the statistically significant results compared to the baselines. Additionally, we conducted another experiment by splitting the TED Talks and coarse PatTR-2 datasets by the meta categories, then observed the individual effects when incorporating the side information with *output-layer* variant, as shown in Figure 4a and 4b. In the TED Talks dataset, we observed improvements for most categories, except for "business, education". In the coarse PatTR-2 dataset, the improvements are obtained across all categories. The key behind this success of the *output-layer* variant is that the representation of existing side information is added in the final output layer and controlled by additional learnable model parameters. In that sense, it results in a more direct effect on lexical choice of the NMT model. This resembles the success in the context of language modelling as presented in Hoang et al. (2016a). Further, we also obtained the promising results for the *mtl* variant although we did implement a very simple instance of MTL with a sharing mechanism and no side information given at a test time. For a fair comparison with the *output-layer* method, we added an additional experiment in which the *output-layer* method does not have the access of side information. As expected, its performance has been dropped, as can be seen in the second last row in Table 3. In this case, the *mtl* method without the side information at a test time performs better. We believe that more careful design of the *mtl* variant can lead to even better results. We also think that the hybrid method combining the *output-layer* and *mtl* variants is also an interesting direction for future research, e.g., relaxing the approximation as shown in Equation 5.

Given the above results, we can find that the characteristics of side information plays an important role in improving the NMT models. Our empirical experiments show that topical information (as in the TED Talks and PatTR datasets) is more useful than the personal traits (as in the Person-



(a) The TEDTalks dataset.



(b) The coarse PatTR-2 dataset.

Fig. 4 Effects on individual BLEU scores for each of categories in the TEDTalks and coarse PatTR-2 datasets, with the NMT model enhanced with the *output-layer* variant.

alised Europarl dataset). However, sometimes it is still good to reserve the personal traits in the target translations (Rabinovich et al., 2017) although their BLEU scores are not necessarily better.

5 Related Work

Our work is mainly inspired from Hoang et al. (2016a) who proposed the use of side information for boosting the performance of recurrent neural network language models. We further apply this idea for a downstream task in neural machine translation.

We've adapted different methods in the literature for this specific problem and evaluated using different datasets with different kinds of side information.

Our methods for incorporating side information as *suffix*, *prefix* for either source or target sequences have been adapted from (Sennrich et al., 2016a; Johnson et al., 2017). Also working on the same patent dataset, Jehl and Riezler (2018) proposed to incorporate document meta information as special tokens, similar to our source prefix/suffix method, or by concatenating the tag with each source word. They report an improvements, consistent with our findings, although the changes

they observe are larger, of about 1 BLEU point, albeit from a lower baseline.

Also, Michel and Neubig (2018) proposed to personalise neural MT systems by taking the variance that each speaker speaks/writes on his own into consideration. They proposed the adaptation process which takes place in the "output" layer, similar to our *output layer* incorporation method.

The benefit of the proposed MTL approach is not surprising, resembling from existing works, e.g., jointly training translation models from/to multiple languages (Dong et al., 2015); jointly training the encoders (Zoph and Knight, 2016) or both encoders and decoders (Johnson et al., 2017).

6 Conclusion

In this work, we have presented various situations to which extent the side information can boost the performance of the NMT models. We have studied different kinds of side information (e.g. topic information, personal trait) as well as present different ways of incorporating them into the existing NMT models. Though being simple, the idea of utilising the side information for NMT is indeed feasible and we have proved it via our empirical experiments. Our findings will encourage practitioners to pay more attention to the side information if exists. Such side information can provide valuable external knowledge that compensates for the learning models. Further, we believe that this idea is not limited to the context of neural LM or NMT, but it may be applicable to other NLP tasks such as summarisation, parsing, reading comprehension, and so on.

Acknowledgments

We thank the reviewers for valuable feedbacks and discussions. Cong Duy Vu Hoang is supported by Australian Government Research Training Program Scholarships at the University of Melbourne, Australia.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proc. of 3rd International Conference on Learning Representations (ICLR2015)*.

Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. pages 992–1003. https://aclanthology.info/papers/D17-1105/d17-1105.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*. pages 1913–1924. https://doi.org/10.18653/v1/P17-1175.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT^3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*. Trento, Italy, pages 261–268.

Wenhu Chen, Evgeny Matusov, Shahram Khadivi, and Jan-Thorsten Peter. 2016. Guided alignment training for topic-aware neural machine translation. *CoRR* abs/1607.01628. http://arxiv.org/abs/1607.01628.

Trevor Cohn, Cong Duy Vu Hoang, Ekaterina Vymolova, Kaisheng Yao, Chris Dyer, and Gholamreza Haffari. 2016. Incorporating Structural Alignment Biases into an Attentional Neural Translation Model. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 876–885. http://www.aclweb.org/anthology/N16-1102.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*. ACM, New York, NY, USA, ICML '08, pages 160–167. https://doi.org/10.1145/1390156.1390177.

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

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, pages 845–850. https://doi.org/10.3115/v1/P15-2139.

Maha Elbayad, Laurent Besacier, and Jakob Verbeek. 2018. Pervasive attention: 2d convolutional neural networks for sequence-to-sequence prediction. *CoRR* abs/1808.03867. http://arxiv.org/abs/1808.03867.

Reuben Feinman and Brenden M. Lake. 2018. Learning inductive biases with simple neural networks. *CoRR* abs/1802.02745. http://arxiv.org/abs/1802.02745.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

Mercedes García-Martínez, Loïc Barrault, and Fethi Bougares. 2017. Neural machine translation by generating multiple linguistic factors. *CoRR* abs/1712.01821. http://arxiv.org/abs/1712.01821.

Jonas Gehring, Michael Auli, David Grangier, and Yann Dauphin. 2017a. A convolutional encoder model for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*. pages 123–135. https://doi.org/10.18653/v1/P17-1012.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017b. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*. pages 1243–1252. http://proceedings.mlr.press/v70/gehring17a.html.

Cong Duy Vu Hoang, Trevor Cohn, and Gholamreza Haffari. 2016a. Incorporating Side Information into Recurrent Neural Network Language Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 1250–1255. http://www.aclweb.org/anthology/N16-1149.

Cong Duy Vu Hoang, Reza Haffari, and Trevor Cohn. 2016b. Improving Neural Translation Models with Linguistic Factors. In *Proceedings of the Australasian Language Technology Association Workshop 2016*. Melbourne, Australia, pages 7–14. http://www.aclweb.org/anthology/U16-1001.

Laura Jehl and Stefan Riezler. 2018. Document-level information as side constraints for improved neural patent translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas, AMTA 2018, Boston, MA, USA, March 17-21, 2018 - Volume 1: Research Papers*. pages 1–12. https://aclanthology.info/papers/W18-1802/w18-1802.

Melvin Johnson, Mike Schuster, Quoc Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernand a Vigas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics* 5:339–351. https://transacl.org/ojs/index.php/tacl/article/view/1081.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 1412–1421. http://aclweb.org/anthology/D15-1166.

Evgeny Matusov, Andy Way, Iacer Calixto, Daniel Stein, Pintu Lohar, and Sheila Castilho. 2017. Using images to improve machine-translating e-commerce product listings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*. pages 637–643. https://aclanthology.info/papers/E17-2101/e17-2101.

Haitao Mi, Baskaran Sankaran, Zhiguo Wang, and Abe Ittycheriah. 2016. Coverage Embedding Models for Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 955–960.

Paul Michel and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, pages 312–318. http://aclweb.org/anthology/P18-2050.

Shachar Mirkin, Scott Nowson, Caroline Brun, and Julien Perez. 2015. Motivating personality-aware machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1102–1108. https://doi.org/10.18653/v1/D15-1130.

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 on Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '02, pages 311–318. https://doi.org/10.3115/1073083.1073135.

Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, pages 1074–1084. http://aclweb.org/anthology/E17-1101.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *CoRR* abs/1706.05098. http://arxiv.org/abs/1706.05098.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation.

In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*. Association for Computational Linguistics, pages 83–91. https://doi.org/10.18653/v1/W16-2209.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 35–40. https://doi.org/10.18653/v1/N16-1005.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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)*. Association for Computational Linguistics, Berlin, Germany, pages 1715–1725. http://www.aclweb.org/anthology/P16-1162.

Patrick Simianer and Stefan Riezler. 2013. Multi-task learning for improved discriminative training in SMT. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Sofia, Bulgaria, pages 292–300. http://www.aclweb.org/anthology/W13-2236.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*. MIT Press, Cambridge, MA, USA, NIPS'14, pages 3104–3112.

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, Curran Associates, Inc., pages 5998–6008. http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Katharina Wäschle and Stefan Riezler. 2012a. Analyzing Parallelism and Domain Similarities in the MAREC Patent Corpus. *Multidisciplinary Information Retrieval* pages 12–27. http://www.cl.uni-heidelberg.de/ riezler/publications/papers/IRF2012.pdf.

Katharina Wäschle and Stefan Riezler. 2012b. Structural and Topical Dimensions in Multi-Task Patent Translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012), Avignon, France*. http://www.aclweb.org/anthology-new/E/E12/E12-1083.pdf.

Yongxin Yang and Timothy M. Hospedales. 2016. Trace norm regularised deep multi-task learning. *CoRR* abs/1606.04038. http://arxiv.org/abs/1606.04038.

Jiacheng Zhang, Yang Liu, Huanbo Luan, Jingfang Xu, and Maosong Sun. 2017. Prior knowledge integration for neural machine translation using posterior regularization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1514–1523. https://doi.org/10.18653/v1/P17-1139.

B. Zoph and K. Knight. 2016. Multi-source neural translation. In *Proc. NAACL*. http://www.isi.edu/natural-language/mt/multi-source-neural.pdf.