Flexible Customization of a Single Neural Machine Translation System with Multi-dimensional Metadata Inputs

Evgeny Matusov, Patrick Wilken, Christian Herold
{ematusov,pwilken,cherold}@apptek.com
AppTek, Aachen, Germany

Abstract

Advances in neural machine translation (NMT) technology not only significantly raised machine translation quality for general-purpose, out-of-the-box systems, but also provided a way for additional input signals to an NMT model to effectively influence its output for a given translation unit, so that a single NMT system can serve different customer needs. At the same time, language service providers, media companies, and other businesses started to systematically store metadata associated with their translatable content. In this work, we show how these metadata can be used both for training and at inference time to flexibly customize a given NMT system to produce somewhat different translations of the same translation unit. The metadata as an extra input to NMT can enable such customization across multiple dimensions and at different levels of input granularity: for individual documents or their collections, for a given post-editing session of a professional translator, or even for individual sentences.

1 Introduction

The following meta-information can be "mixed in" to influence the translation output of a single neural machine translation system:

- Domain, genre, and topic can be provided either in terms of fixed labels (e.g. "patents", "contracts", "news"), or can be inferred as topic embeddings from the content of the given or similar document(s).
- In a multilingual NMT model, the language (variety) or dialect metadata not only augments the representation of corresponding input documents or sentences, but can also specify the desired target language or dialect. Post-editing tools can implement a flexible switch between supported languages/dialects for mixed-language text or speech input.
- Document-level context of different size (e.g. previous/next N sentences) can be "turned on" for better word disambiguation and pronoun resolution.
- Machine translation (MT) output length can be influenced without significant information loss. This is important in applications like subtitling and software localization where translations sometimes have to fit into a given fixed-size template.
- Translation style can be adjusted with a simple "switch" between (binary) classes (e.g. with/without profanities; informal vs. formal "you" forms in languages like German).

- The gender of the speaker/author may be important for a correct, unbiased translation in target languages like Czech where past tense verb forms have different endings depending whether a male or a female is talking about his or her actions. In speech translation, speaker gender labels can be inferred automatically from the upstream speech recognition. Post-editing applications can pre-fetch translations of all styles and genders so that a post-editor can instantaneously switch from e.g. a formal to informal translation with a single click.

- Finally, document- or user-specific terminology glossary entries can accompany each translation request to the NMT system so that for any matched source-language glossary entry it has to produce the translation from the glossary. The challenge here is how to generate this translation in a grammatically correct form, which is especially difficult for morphologically rich target languages.

All of these customizations of a single NMT model are very much suitable for commercial settings. Instead of deploying multiple different NMT models for each domain, style, length, dialect, etc., ideally we deploy a single system. Thus, not only we save on computational resources, reducing the environmental footprint of the MT technology. We also save time and machine and human power necessary for fine-tuning or otherwise adapting each of these customized systems, save on measures to counteract over-fitting, organization of parallel deployment and elaborate load balancing, etc.

In the following section, we will give an overview in which ways additional meta-information can flow into the training and inference of a single NMT system. In Section 3 we will revisit the above meta-information types and show, in many cases supported by experimental findings and/or examples, as well as citations of related work, the positive influence of meta-information on translation quality. We will also provide tips for a practical implementation of metadata-based "switches" in MT applications such as post-editing tools.

2 Using Meta-information for NMT Customization

The meta-information accompanying a source sentence or document can be incorporated into the NMT training in different ways.

The most straightforward way that does not require any changes to the NMT architecture is the use of source-side pseudo-tokens, usually in the beginning of a sentence, that correspond to a (discrete) meta-information. Pseudo-tokens are most widely used in multilingual systems (Johnson et al., 2017; Ha et al., 2016) with multiple target languages: the pseudo-token with the language code, such as `@es@`, signals that a translation into a particular language, in this case Spanish, is desired. Pseudo tokens were also successfully used for specifying the translation style (Sennrich et al., 2016) and for domain adaptation (Tars and Fishel, 2018). Alternatively, pseudo-tokens can be used as prefix constraints in the beginning of the (generated) target sentence (Takeno et al., 2017).

The disadvantage of pseudo-tokens is that they only encode one piece of information, and their influence on the produced NMT output is limited, especially in cases where the differentiating power of the additional meta-information is small, e.g. when the meta-information encodes domains/topics which are similar. In such cases it is advisable to use factored machine translation and encode the extra meta-information as an additional factor for each source word (García-Martínez et al., 2016; Wilken and Matusov, 2019). In this way, the meta-information will have a stronger influence, since the NMT encoder would then be able to learn for which words the meta-information factor is more important than for the other words. For some types of meta-information, like speaker gender, the factor (e.g. male/female gender) can be assigned to the relevant words only (e.g. personal pronouns and verbs whose translation may be different depending on the speaker/author gender). All other words in this case can be assigned a third,

"neutral" value.

When the meta-information about a sentence or document is automatically predicted with a certain probability, it is advisable to directly include this probability into the NMT training. Thus, in case of genre prediction, assuming 20 different genres, the additional input can be a 20-dimensional vector with probabilities for each genre given the input source sentence or document. This dense representation can then be associated with a genre embedding and included in the NMT architecture in a variety of ways, e.g. via a separate attention component to the genre/topic embedding. A stronger influence of meta-information on the decoder can be achieved by concatenating each current state with the genre/topic embedding before the next decoder state is predicted. Details can be found e.g. in (Chen et al., 2016).

At inference time, we assume that the extra meta-data is provided by the user/customer or is automatically generated by an upstream component (such as speaker gender classifier or a topic classifier). At training time, the meta-information can also be already available (e.g. domain of a document or a whole collection of documents, language or language variety, or even style). This is especially true of recent customer-specific data, since a lot of companies, language service providers in particular, have started to pay attention to consistent storage of meta-data that accompanies their translation content. Other types of meta-information can be directly computed for each pair of parallel sentences in the training data, like the length ratio between the source and the target sentence which can be used to classify translations into short, medium-length, and long (see Section 4). Or, it can be derived using regular expressions or more complex tools such as syntactic parsers and part-of-speech taggers. A "garbage" class can be assigned to sentences which do not match any of the regular expressions. See Section 3.1 for more details.

For more complex types of meta-information that is not available for a given set of parallel training sentence pairs, a classifier can be trained that predicts this information either on the sentence-level or document-level. To reduce error propagation, the vector of posterior probabilities for all the predicted classes can be directly used in NMT as opposed to the first-best predicted label. For example, the genre and topic of a document can be predicted automatically with a trained classifier, but also e.g. its dialect or language variety. External monolingual labeled data can be used to select the label set and train the classifier. Usually, the classifier is trained for the source language so that it can be applied both at training time and at inference time as described above. More elaborate approaches such as the work of Zeng et al. (2018) jointly model NMT with monolingual attention-based classification tasks (in this particular case, domain classification).

3 Types of Customization

3.1 Style

The style or tone of a translation is very important for its acceptance. Thus, it is not appropriate to use an informal style in legal documents, etc. At the same time, a formal, polite style can not be used in translations of movie dialogs, chat messages, and other cases with colloquial language.

A single NMT system can be trained to support multiple styles. In what style the translation is generated depends on the additional input (selector) from the user, also called side constraints (Sennrich et al., 2016; Feely et al., 2019). In our experiments with English as the source language, we differentiated in particular between a formal style that uses a polite version of the second-person pronoun "you" (which is different from the informal pronoun in many languages such as German, Russian, French, Greek, etc.). The parallel training data was partitioned into 3 classes based on whether the formal or informal version of the pronoun was used in the target language sentence, or none at all. For corpora where document identity was available

System	BLEU [%]
AppTek baseline	27.9
AppTek style token informal	28.7
AppTek style token formal	26.7
On-line G 2020-06-18	21.8
On-line B 2020-06-18	27.3

Table 1: BLEU scores in % on an English-to-Greek subtitle test set of 50K running words, 5.5K sentences (held-out for the AppTek systems).

source	I am at your service.
formal	Ich stehe **Ihnen** zu Diensten.
informal	Ich stehe zu **deinen** Diensten.
source	I see you all are interested in media and subtitling.
formal	Ich sehe, **Sie** alle interessieren sich für Medien und Untertitelung.
informal	Ich sehe, **ihr seid** alle an Medien und Untertitelung interessiert.
source	Please hold the balls in your hands.
formal	Bitte **halten Sie** die **Bälle** in **Ihren** Händen.
informal	Bitte **halte** die **Eier** in **deinen** Händen.

Table 2: Translation examples for English-to-German NMT with formal vs. informal meta-information provided as pseudo-tokens. All of the NMT-generated translations are correct for these examples.

for each sentence, we assigned the whole document to the formal/informal class if the majority of its target sentences contained the formal/informal pronoun. This is a simple, yet effective rule-based approach; for a more sophisticated method, cf. (Niu and Carpuat, 2019).

We experimented with two language pairs: English-to-German and English-to-Greek. In both cases we used state-of-the-art NMT systems trained with Transformer architectures using millions of sentence pairs. For English-to-Greek, the MT quality as measured with the BLEU score (Papineni et al., 2002) on a held-out test set of movie subtitles (Table 1) shows that our systems compare favorably to two major online translation providers. The style information was provided to the system as a pseudo-token (one of 3) both at training and at inference time. At inference time, we always used either the formal or the informal pseudo-token for all sentences in the test set. Since the test set mostly includes popular movies with informal style, the improvement in BLEU when using the informal style token was expected.

We let a professional Greek-native translator check the output of the baseline system that does not use style tokens, compared to the systems that use the formal or informal style token. This was done on a subset of a held-out subtitle file that contains informal dialogs. Whereas no quantitative evaluation was conducted, the translator noted a generally good quality of all outputs. She found that the grammatical part of style adaptation, i.e. the correct second-person pronouns, seemed to work, with the formal version using mostly the formal form, correctly per the style chosen, despite the informal material it was applied on. She also noted that "the vocabulary choices in the MT output depending on the style chosen were fascinating". This underlines the other interesting aspect of style transfer: although not explicitly modelled when partitioning the training data, the vocabulary choice for the informal vs. formal style seems to correlate with the usage of the second-person pronouns.

Similar findings were made for English-to-German. Examples of formal vs. informal style are given in Table 2. Note that both singular and plural second-person pronouns (including

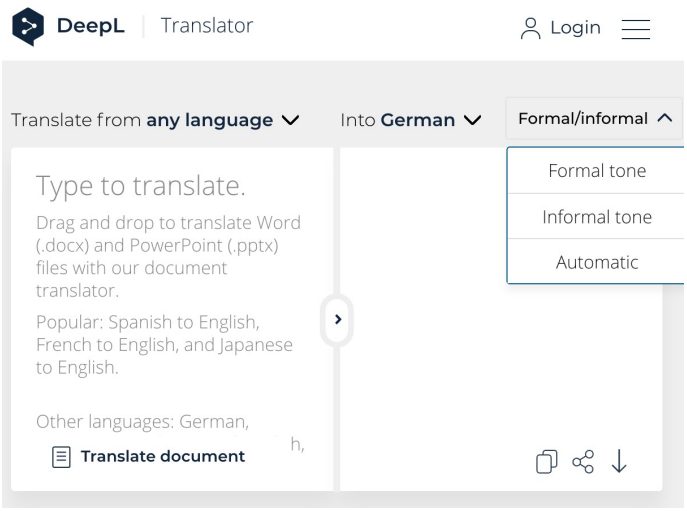

Figure 1: An example of a style switch menu in the on-line UI of the MT provider deepl.com (screenshot from August 28, 2020).

possessive ones) are translated correctly w.r.t. the requested style (`Ihnen` vs. `deinen`, `Sie` vs. `ihr`), correct auxiliary verb forms corresponding to these pronouns are used. The sentence structure is in some cases significantly different. The last anecdotal example in Table 2 where in case of the informal style the English word "balls" is translated in its profane meaning into "Eier" shows that style adaptation and transfer is also about lexical choice and meaning disambiguation. In particular, it is also possible to introduce additional constraints on the style, e.g. informal without obscene language, informal with obscene language, etc.

In practical applications of post-editing by professional translators the information about style can be provided for the whole document in advance before NMT is used to translate the document. However, in many cases style can change within a given document. For example, subtitles of a given film can include formal dialogs using the polite form of "you" as well as informal dialogs (or dialogs with a child) using the informal second-person pronouns. A button in the user interface can be implemented to help the post-editor instantly switch the translation of a single sentence to a different style when he or she notices such style changes. Of course, multiple translations for different styles would have to be pre-fetched in the background in order for this to work seamlessly.

Buttons or menu items for style switches in on-line translation tools for free text translation have started to appear already, as illustrated in Figure 1.

3.2 Domain, Genre and Topic

Domain, genre, and topic are almost synonyms in the sense that they refer to sometimes minuscule, sometimes large differences in content, combined with stylistic differences which are much harder to grasp or explain than the formal/informal style differences. Yet any information about such "world context" that goes beyond the context of the given and surrounding sentences is very important for correct translation, both by humans and by machines. For some genres it is all about correct terminology translation, whereas for others the differences are more subtle.

Usually, no fixed definitions or taxonomy of domains and genres are available. Nevertheless, the sources of monolingual and bilingual data often give a hint at the genre and domain. Yet in many cases especially the parallel data is crawled from multiple sources, often individ-

source	You need an **apple product** to obtain the best shape.
prose	Du brauchst ein **Apfelprodukt**, um die beste Form zu bekommen.
manuals	Sie brauchen ein **Apple-Produkt**, um die beste Form zu erhalten.
source	The bushing 20 is inserted in a **hole** 21 of the cover 12 of the **base** 11.
news texts	Die Buchse 20 wird in ein **Loch** 21 des Deckels 12 der **Basis** 11 eingesetzt.
patents	Die Buchse 20 ist in eine **Bohrung** 21 des Deckels 12 des **Sockels** 11 eingesetzt.
source	He came here to look for **food**.
documentary	Er kam her, um nach **Nahrung** zu suchen.
talks	Er kam her, um nach **Essen** zu suchen.

Table 3: Examples of translations by the same NMT system which change depending on additional meta-information about the input genre (English-to-German).

ual sentence pairs are taken from an unknown bilingual document or are even extracted from non-parallel, comparable corpora. Furthermore, genre can slightly vary even within a single document, or can be a new genre that has traces of the previously observed genres. Also, we would like that at inference time the genre/domain is either provided by the user, or automatically predicted for a given input sentence or document, or is not provided (and then the NMT system falls back to genre-agnostic translation). Most closely related work to our approach is by Kobus et al. (2017).

In our experiments, we decided to focus on genres. Some of them were defined in international MT research projects like GALE (Olive et al., 2011): newswire text, web (blog) text, broadcast news and conversations. We identified further genres based on the available English data. These include chat messages and comments, e-commerce product descriptions, customer product reviews, subtitles (film dialogs), documentary subtitles, emails, government texts, legal texts, software and hardware manuals, marketing material, military-related texts, non-fiction books, fiction (prose), poetry, patents, religious texts, educational (school) material, scientific texts including research papers, as well as parliamentary speeches and public talks.

We sampled 10M English sentences per genre and trained a bidirectional 1-layer LSTM classifier to predict the genre labels on a sentence level. The classifier obtained an accuracy of 77% on a held-out set of 6000 sentences that contained 250 sentences of each genre.

As mentioned in Section 2, the best way to integrate the genre information would be to change the architecture to include the predicted genre distribution as an embedding vector. In preliminary experiments, however, we converted this prediction into a single label using a heuristic – if a single label was predicted with a probability of more than 0.5, we assigned this label to a given training sentence pair. In cases when none of the labels had such a high probability, we assigned a "no genre" label. We then used this label as a pseudo-token similarly to the style pseudo-tokens described in Section 3.1.

We trained an English-to-German system with genre pseudo-tokens and first verified that its quality as measured with BLEU on multiple test sets with different domains did not significantly degrade as compared to a baseline system that does not use any pseudo-tokens. For this sanity check, we prepended each sentence with the "no genre" pseudo-token. Then, we manually checked the system performance on a number of examples.

Generally, the effect of using just the pseudo token was minimal - the translation in many cases remained the same. That is why in our future work we would like to explore a stronger signal from the predicted genre distribution. However, if there was significant change in the output, it was always in the right direction for our examples, as can be observed in Table 3. In some cases, though, a more fine-grained distinction between genres may be desirable, leading to prediction of a topic distribution/profile of a given sentence or document. For instance, to

System	BLEU [%]
baseline	35.5
concatenate in training	36.3
concatenate in training + inference	37.0
on-line G 2020-06-18	33.2
on-line D 2020-08-12	34.9

Table 4: Translation results for MT with extended context on the English-to-German subtitle test set of 2378 sentences, 18K running words.

disambiguate the translation of "Apple" it is not enough to know whether the system deals with prose or marketing content, since in almost all of the defined genres both the fruit and the brand meaning can occur with high frequency. Topic modeling usually requires unsupervised clustering methods to obtain the right number of topics with as little overlap in their distributed representations as possible. In some applications like e-commerce, however, fine-grained topic taxonomies are already defined (e.g. for product categories and sub-categories) and can be used directly as supervising labels (Chen et al., 2016).

3.3 Extended Context

Topic modeling flows into the research on extended context for NMT and is related to document-level translation. Recently, there have been advances in this area, showing that additional context in the form of encoded previous and subsequent sentences from the same document is beneficial for improved MT quality (Werlen et al., 2018; Kim et al., 2019). In particular it can help with pronoun resolution (Müller et al., 2018). We argue that it is also possible to train a single NMT system that can either consume the additional context or can translate a single sentence without it, depending on the user request.

Table 4 summarizes the results of our experiments for English-to-German. In all cases we follow the simple concatenation approaches of Tiedemann and Scherrer (2017) and Junczys-Dowmunt (2019). We concatenate subsequent sentences appearing consecutively in the same document (subtitles for a single film) if the resulting sequence does not exceed a certain number of tokens (50). Multiple sentences on source and target side are concatenated using a special symbol `@sep@` so that the NMT system learns to generate such separator symbols. The concatenated data is added to the original training data; thus some of the sentences appear in the training data twice: on their own and concatenated with surrounding sentences. The test data is augmented in a similar way, except every sentence is translated exactly once.

AppTek's baseline state-of-the-art English-to-German system was trained on ca. 20M sentence pairs, including subtitle data. As can be inferred from Table 4, its performance on a held-out subtitle test set is better in terms of BLEU than when translating with two major on-line MT services. After augmentation via concatenation of some of the sentences which had document information associated with them, the total number of lines in the training data increased to 39M.

We observed significant increases in BLEU from doing the concatenation in training only, which shows that the proposed method does not harm the baseline translation quality. When short sentences are concatenated at inference time, the translation quality increases further. A detailed analysis of sentences of different lengths showed that in particular translations of very short segments benefited from the context of the previous and next sentences. The absolute BLEU improvement for sentences of length one (individual words) was 11% absolute, and for sentences of length from 2 to 9 words it was 2% absolute. But even for long sentences, a marginal BLEU score improvement was observed.

source	I found a watch and returned it to the owner.
pseudo-token male	Naš**el** jsem hodinky a vrátil je majiteli.
pseudo-token female	Naš**la** jsem hodinky a vrátil**a** je majiteli.

Table 5: An example of two translations into Czech of the same sentence by a single NMT system, the first time with the gender meta-information "male", the second time with the gender meta-information "female".

For real-life use, this means that an MT application can be programmed to use the additional document-level context (or simply put, the context of the surrounding sentences) on demand, when the context of a given sentence is not enough as determined by some objective criterion, the simplest of which can be input sentence length.

3.4 Speaker/Author Gender

In some languages, the morphological realizations of certain parts-of-speech depend on the gender of the speaker/author. Examples include past-tense singular verb forms in Russian, Czech, etc. When translating from languages such as English where this is mostly not the case, the NMT system chooses one of the gender-specific forms. With the absence of supporting gender-relevant context (e.g. "she said" vs. "he said"), it makes its decision mostly based on the examples that were observed in training. Usually, but not always, the training data is biased towards male word forms. Biased or not, however, an incorrect word form in the automatically generated translation is annoying and yet hard to fix; it can appear again and again throughout a given text or speech that, for instance, is a first-person narrative with many forms starting with the pronoun "I".

To explicitly use the information about the speaker or the author, we again propose to partition the training data into "male", "female", and "neutral" sentence pairs depending whether or not the corresponding male or female word forms are used throughout the target sentence (almost) exclusively. We realize that such a method requires many heuristics, but in the absence of data labeled with speaker gender that was used in related research of Vanmassenhove et al. (2018) it is difficult to come up with a better solution.

So far, we conducted only preliminary experiments for English-to-Czech, using 3 types of pseudo-tokens as described above. Table 5 shows an example where the correct gender forms are used in the Czech translation when the information about gender is provided to the system. This is a step in the right direction. We envision that especially for applications involving speech translation, speaker gender can be automatically predicted with high confidence and passed on to NMT for use as an additional signal. Also, in personalized translation applications, the correct gender can be set by the app user, and then her/his texts and messages would be translated from English using the right gender form of a given target language.

3.5 Length

In some applications it is desirable to control the length of MT output, as measured in words or characters, while minimizing any information loss. For instance, subtitle templates are usually created in the source language and have a fixed number of subtitles with a fixed duration of their appearance on the screen. Thus, a translation of a sentence in a subtitle that is significantly longer than the original sentence can only be inserted without changing the template by using more than the allowed number of lines per subtitle (usually two). This means that a faster reading speed is necessary to finish reading the text before the subtitle disappears, and should be avoided as much as possible. That is why shorter translations (from English) are preferred.

Another application is translation of user interface elements/menus in a software compo-

source	¿No te vas a sentir incómodo?
baseline	You're not gonna feel uncomfortable?
shortened	Won't you be uneasy?
source	Se llevan muy bien y la verdad es que me da mucha pena.
baseline	They get along very well, and the truth is, I feel very sorry for them.
shortened	They get along very well and I'm really sorry.
source	De ninguna manera me sentiré incómodo ni tengo problema en verla.
baseline	There's no way I'm gonna feel uncomfortable and I don't have a problem seeing her.
shortened	There's no way I'll feel uncomfortable or have a problem seeing her.

Table 6: Examples of translations shortened by N-best list rescoring aimed at penalizing long translations (Spanish-to-English subtitles).

nent. There, the maximum length may be technically limited by the width of the menu or a text field.

A number of research publications appeared recently which target length control, starting with the seminal work on length control in encoder-decoder architectures by Kikuchi et al. (2016). In (Lakew et al., 2019), the pseudo-tokens for short, medium, and long translations are assigned at training time. These labels are derived from the length ratios between each training source sentence and its target language translation. At inference time, the user provides the desired label, e.g. requesting a short translation. An approach with length constraints learned end-to-end in an unsupervised way is presented by (Niehues, 2020).

Another method that we tested tailored specifically to subtitle translation is to re-score the N-best output of the NMT system using a linear combination of the original NMT model score and a score derived from the the length of an N-best list hypothesis and the duration of the subtitle in which the source sentence, and thus also its translation, is to appear on the screen.

Since for judging the translation quality in cases of e.g. shortened MT output it is not reasonable to use the original reference translations created without such length constraints for computation of automatic MT error measures, we only conducted a small-scale manual evaluation of the resulting output.

We translated the content of 64 subtitles from Spanish to English with AppTek's state-of-the-art NMT system, performing N-best list rescoring aimed at penalizing all translations with a reading speed of 17 chars or more per second[1]. As a result, the average reading speed of the file reduced from 19.3 to 17.23 chars/s and the number of frames with a reading speed of more than 20 chars/s dropped from 33 to 13.

A professional translator noted that the shorter automatic translation versions "are mostly great, exactly what a subtitler would do". In very few cases, they do change the meaning, which is not acceptable, but can usually be fixed by quick post-editing.

Table 6 shows examples of translations shortened with the above approach, for which the meaning of the translation did not change.

3.6 Language Variety and Multilinguality

Multilingual NMT systems have shown to be effective in using parallel training data from high-resource language pairs to improve the quality of translation from or to a low-resource language (Firat et al., 2016; Johnson et al., 2017).

In case of multiple target languages, it is often sufficient to use a pseudo-token at the beginning of the source sentence that signals to what language it should be translated. With this

[1]The reading speed is defined as the subtitle length in characters (e.g. a maximum of 2 lines with a maximum of 42 characters per line) divided by the subtitle duration (usually 2-5 seconds).

simple approach, already an acceptable level of MT quality can be reached. Thus, a multilingual system can be viewed also a customization of a single NMT system with meta-information about the target language.

At AppTek, we use the multilingual approach also for language varieties or dialects, following also the work of Lakew et al. (2018). The main challenge is how to partition the training data: in most cases no reliable information about the used dialect is available, and automatic dialect prediction is a hard task. Following a pragmatic approach for English-to-Spanish translation of movie subtitles, we labeled those film subtitles in the training data as European Spanish which contained words and phrases used only in Spain. The rest was labeled as Latin American Spanish. We then trained a multilingual system with the two labels. Our customers can choose the language variety via an API parameter and obtain a possibly different translation for a given sentence using the same system. More details can be found in Matusov et al. (2019).

In case of multilingual, dialectal, or even mixed-language input, it is possible to train an NMT model which is sensitive to the meta-information about the input language or dialect. Again, in practical applications, such as computer-assisted translation from Arabic to English, a general translation can be generated prior to post-editing (assuming e.g. Modern Standard Arabic or MSA), together with translations for (a subset of) the Arabic dialects. Then, the professional translator can change the MT in the post-editing window when she or he notices that the language switched from MSA to a dialect. This can happen in particular when someone's dialectal speech is quoted in a news article written in MSA.

At the same time, for such multilingual or multi-dialect many-to-one systems it is advisable to use a "garbage" label which is associated randomly with a subset of the training data in any language or dialect. Providing this label may help when the dialect or language of the input is not known, or it is a mixed-language input. For instance, AppTek's multilingual NMT system that can translate from any of 12 Slavic languages into English is also able to translate mixed-language sentences like the following one which is a mix of Ukrainian and Russian (typical for messages and speech of a significant part of the population of Ukraine). Хлопці были у меня дома, но про дівчин они ничего не пліткували is correctly translated as "The boys were at my house, but they didn't say anything about the girls." (Ukrainian words in the otherwise Russian sentence are Хлопці, дівчин, and пліткували).

3.7 Glossaries

Terminology glossary entries or translation memory matches can accompany each translation request to an NMT system so that for any matched entry the translation from the glossary is forced to be used in-context in the MT system output. This so called glossary transfer or override is another user-specific customization of a given NMT system and can be implemented in professional post-editing UIs by e.g. giving the user the possibility to upload a glossary prior to populating the output window with the automatic translation. In other cases the glossary can be automatically created in a computer-assisted translation environment by memorizing past user translation corrections and choices.

In its simplest form, glossary transfer "as is", i.e. the exact copy of the target side of the glossary entry, is implemented using placeholder tokens. In training, a source word or phrase is replaced by such a placeholder token; the same token replaces the (consecutive sequence of) words in the target sentence which are word-aligned to this particular source word or phrase. If there are multiple replacements within a given sentence pair, different placeholder tokens are used. Thus, a system learns to translate (and thus also correctly position, if reordering is involved) a given placeholder token to itself in all cases.

At inference time, a matched glossary entry in the source sentence is replaced with such a placeholder token in preprocessing, and then the same token in the generated translation is

replaced with the target side of the corresponding glossary entry in postprocessing. The obvious disadvantage of this approach is that the context in the form of the glossary entry itself is lost during translation, since it is generalized to the placeholder token.

More complex algorithms involve encoding of the desired target translation in the source sentence using special markers (Dinu et al., 2019). Other methods try to use constrained decoding (Hasler et al., 2018) or NMT-internal attention mechanisms to override the translation of the next word if the current focus of the attention is on the corresponding matched source glossary entry (Dahlmann et al., 2017). With such approaches it is not guaranteed that the desired translation from the glossary will be used, but at the same time, it is possible that the system will learn that the glossary translation has to be used in a morphological form that is different from the (base) form present in the glossary because of the surrounding context.

To illustrate the basic approach and the challenges that the more advanced approaches can rarely master, we present two examples. In the first one, the sentence `Jack, when are you going back to Vienna?` is correctly translated into German as `Jack, wann fahren Sie zurück nach Wien?`. However, this translation is not correct if Vienna is referring to a city in the United States. Here, the simple approach of the "as is" glossary override via placeholder tokens can already enforce a glossary entry `Vienna` → `Vienna`. In the second example for English-to-Russian translation, the sentence `The Hatter put the Dormouse's head in a teapot and winked to the March Hare` from *Alice in Wonderland* by Lewis Carol can be translated with the help of the following glossary[2]:

```
Doormouse == Sonya
March Hare == Martovskiy Zayats
Hatter == Shlyapnik
```

However, the Russian translations of these fictional characters are given in the nominative case, whereas in a translation of the sentence some of them must be used in other cases with different suffixes/endings: "Shlyapnik polozhil golovu Son**i** v chainik i podmignul Martovsk**omu** Zay**tsu**". The changed suffixes are marked in bold. To the best of our knowledge, state-of-the-art glossary transfer methods for NMT are not able to satisfactorily address this task of glossary override for morphologically rich languages, which opens up possibilities for future work.

4 Conclusion

In this paper, we provided an overview of different customization opportunities so that a single neural machine translation system can be trained to accept additional meta-information as input and thus produce different translations of a given sentence based on the additional metadata. We showed how meta-information about style, genre, topic, and speaker/author gender can be obtained from customer databases or derived automatically, and then used in training and at inference to produce better, in-context translations with correct style, grammar, and correct word sense disambiguation. We discussed how extra context in the form of surrounding sentences from the same document can be "turned on" to improve the translation of a given sentence. Furthermore, we showed that translation length can be effectively controlled if necessary without significant information loss. We also showed how customization works in the context of multilinguality, language varieties and dialects, and even mixed-language input. Finally, we elaborated on the practical applications of single customizable NMT systems in several usage scenarios, with focus on user interfaces for efficient MT post-editing.

Of course, it is possible to combine all or some of the different types of metadata inputs described in this paper in a single NMT system. Our future plans are to train such a system and successfully use it for AppTek's customers.

[2]Transliteration of Russian is used here for better understanding.

References

Chen, W., Matusov, E., Khadivi, S., and Peter, J.-T. (2016). Guided alignment training for topic-aware neural machine translation. *AMTA 2016, Vol.*, page 121.

Dahlmann, L., Matusov, E., Petrushkov, P., and Khadivi, S. (2017). Neural machine translation leveraging phrase-based models in a hybrid search. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1411–1420, Copenhagen, Denmark. Association for Computational Linguistics.

Dinu, G., Mathur, P., Federico, M., and Al-Onaizan, Y. (2019). Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Feely, W., Hasler, E., and de Gispert, A. (2019). Controlling Japanese honorifics in English-to-Japanese neural machine translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 45–53, Hong Kong, China. Association for Computational Linguistics.

Firat, O., Cho, K., and Bengio, Y. (2016). Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875.

García-Martínez, M., Barrault, L., and Bougares, F. (2016). Factored neural machine translation architectures. In *International Workshop on Spoken Language Translation (IWSLT'16)*.

Ha, T.-L., Niehues, J., and Waibel, A. (2016). Toward multilingual neural machine translation with universal encoder and decoder. In *International Workshop on Spoken Language Translation (IWSLT'16)*.

Hasler, E., de Gispert, A., Iglesias, G., and Byrne, B. (2018). Neural machine translation decoding with terminology constraints. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 506–512, New Orleans, Louisiana. Association for Computational Linguistics.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Junczys-Dowmunt, M. (2019). Microsoft Translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233.

Kikuchi, Y., Neubig, G., Sasano, R., Takamura, H., and Okumura, M. (2016). Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338, Austin, Texas. Association for Computational Linguistics.

Kim, Y., Tran, D. T., and Ney, H. (2019). When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 24–34.

Kobus, C., Crego, J., and Senellart, J. (2017). Domain control for neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 372–378, Varna, Bulgaria. INCOMA Ltd.

Lakew, S. M., Di Gangi, M. A., and Federico, M. (2019). Controlling the output length of neural machine translation. In *16th International Workshop on Spoken Language Translation*.

Lakew, S. M., Erofeeva, A., and Federico, M. (2018). Neural machine translation into language varieties. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 156–164, Brussels, Belgium. Association for Computational Linguistics.

Matusov, E., Wilken, P., and Georgakopoulou, Y. (2019). Customizing neural machine translation for subtitling. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 82–93, Florence, Italy. Association for Computational Linguistics.

Müller, M., Gonzales, A. R., Voita, E., and Sennrich, R. (2018). A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72.

Niehues, J. (2020). Machine translation with unsupervised length-constraints. *arXiv preprint arXiv:2004.03176*.

Niu, X. and Carpuat, M. (2019). Controlling neural machine translation formality with synthetic supervision. *arXiv preprint arXiv:1911.08706*.

Olive, J., Christianson, C., and McCary, J. (2011). *Handbook of natural language processing and machine translation: DARPA global autonomous language exploitation*. Springer Science & Business Media.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA.

Sennrich, R., Haddow, B., and Birch, A. (2016). 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*, pages 35–40.

Takeno, S., Nagata, M., and Yamamoto, K. (2017). Controlling target features in neural machine translation via prefix constraints. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 55–63, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Tars, S. and Fishel, M. (2018). Multi-domain neural machine translation. *arXiv preprint arXiv:1805.02282*.

Tiedemann, J. and Scherrer, Y. (2017). Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92.

Vanmassenhove, E., Hardmeier, C., and Way, A. (2018). Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008.

Werlen, L. M., Ram, D., Pappas, N., and Henderson, J. (2018). Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954.

Wilken, P. and Matusov, E. (2019). Novel applications of factored neural machine translation. *arXiv preprint arXiv:1910.03912*.

Zeng, J., Su, J., Wen, H., Liu, Y., Xie, J., Yin, Y., and Zhao, J. (2018). Multi-domain neural machine translation with word-level domain context discrimination. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 447–457, Brussels, Belgium. Association for Computational Linguistics.