Multitask Models for Controlling
the Complexity of Neural Machine Translation

Sweta Agrawal
Department of Computer Science
University of Maryland
sweagraw@cs.umd.edu

Marine Carpuat
Department of Computer Science
University of Maryland
marine@cs.umd.edu

Abstract

We introduce a machine translation task where the output is aimed at audiences of different levels of target language proficiency. We collect a novel dataset of news articles available in English and Spanish and written for diverse reading grade levels. We leverage this dataset to train multitask sequence to sequence models that translate Spanish into English targeted at an easier reading grade level than the original Spanish. We show that multitask models outperform pipeline approaches that translate and simplify text independently. ¹

1 Introduction

Generating text at the right level of complexity is important to make machine translation (MT) more accessible to non-native speakers, language learners (Petersen and Ostendorf, 2007; Allen, 2009) or people who suffer from language impairments (Carroll et al., 1999; Canning et al., 2000; Inui et al., 2003). Simplification has been used to improve MT by restructuring complex sentences into shorter and simpler segments that are easier to translate (Gerber and Hovy, 1998; ˇStajner and Popovic, 2016; Hasler et al., 2017). Closest to our goal, Marchisio et al. (2019) address the task of producing either simple or complex translations of the same input, using automatic readability scoring of parallel corpora. Our work shares their goal of controlling translation complexity, but considers a broader range of reading grade levels and simplification operations grounded in professionally edited text simplification corpora.

We collect examples of Spanish sentences paired with several English translations that span a range of complexity levels from the Newsela website, which provides professionally edited simplifications and translations. While Newsela dataset has been used to build English text simplification systems (Xu et al., 2016; Zhang and Lapata, 2017; Scarton and Specia, 2018; Nishihara et al., 2019; Zhong et al., 2020) and Spanish simplification systems (Stajner et al., 2018), we exploit the document level alignment between English and Spanish articles to construct evaluation and training samples for complexity controlled MT. By contrast with MT parallel corpora, the English and Spanish translations at different grade levels are only comparable. We adopt a multitask approach that trains a single encoder-decoder model to perform the two distinct tasks of machine translation and text simplification and evaluate it on Spanish-English complexity controlled MT. Our empirical study shows that multitask models produce better and simpler translations than pipelines of independent translation and simplification models. Scripts to replicate our model configurations and our cross-lingual segment aligner are available at https://github.com/sweta20/ComplexityControlledMT.

2 A Multitask Approach to Complexity Controlled MT

We define complexity controlled MT as a task that takes two inputs: an input language segment $s_i$ and a target complexity $c$ representing the desired reading grade level of the output. The goal is to generate a translation $s_o$ in the output language with complexity $c$.

¹This paper is an abridged version of our work (Agrawal and Carpuat, 2019).

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/.
We model $P(s_o|s_i, c; \theta)$ as a neural encoder-decoder with attention (Bahdanau et al., 2015). Target complexity $c$ is incorporated as a special token appended to the beginning of the input sequence, which acts as a side constraint. Our multitask training configuration lets us exploit different types of training examples to train shared encoder-decoder parameters $\theta$. We use the following samples/tasks:

- Complexity controlled MT samples $(s_i, c_o, s_o)$: These are the closest samples to the task at hand, but are hard to obtain. They are used to defined the complexity-controlled MT loss, $L_{CMT} = \sum (s_i, c_o, s_o) \log P(s_o|s_i, c_o; \theta)$.
- MT samples $(s_i, s_o)$: These are sentence pairs drawn from parallel corpora. They are available in large quantities for many language pairs (Tiedemann, 2012) and are used to define the MT loss, $L_{MT} = \sum (s_i, s_o) \log P(s_o|s_i; \theta)$.
- Text simplification samples in the MT target language $(s_o, c_o', s_o')$ where $s_o'$ is a simplified version of complexity $c_o'$ for input $s_o$, which are likely to be available in much smaller quantities than MT samples and are used to define the simplification loss, $L_{Simplify} = \sum (s_o, c_o', s_o') \log P(s_o'|s_o, c_o'; \theta)$.

The multitask loss is simply obtained by summing the losses from individual tasks: $L_{CMT} + L_{MT} + L_{Simplify}$.

### 3 Evaluation and Discussion

We extract segment pairs from the Newsela corpus which consists of English articles in their original form, 4 or 5 different versions re-written by professionals to suit different grade levels as well as optional translations of original and/or simplified English articles into Spanish. English and Spanish segments are aligned across complexity levels to create a bilingual dataset for training and evaluating complexity controlled MT systems. We report results using both translation and simplification metrics: BLEU (Papineni et al., 2002), SARI (Xu et al., 2016) and the Pearson Correlation between the complexity of NMT outputs and reference translations (Heilman et al., 2008), where the reading grade level complexity of MT outputs and reference translations is estimated using the Automatic Readability Index (ARI).

We contrast the multitask system with pipeline based approaches, where translation and simplification are treated as independent consecutive steps. In the first pipeline setup, the output from the translation model is fed as input to an English simplification model while in the second, the output from the Spanish simplification model is fed as input to an translation model. Table 1 shows that compared to pipeline models, multitask models generate translations that better match human references according to BLEU. SARI suggests that multitask translations are simpler than baseline translations, and their resulting complexity correlates better with reference grade levels according to PCC. The “All tasks” model highlights the strengths of the multitask approach: combining training samples from many tasks yields improvements over the “Translate and Simplify” multitask model which is trained on the exact same data as the pipelines. However, even without additional training data, the multitask “Translate and Simplify” model improves over baselines mainly by simplifying the output more, which suggests that the simplification component of the multitask model benefits from the additional MT training data. Table 2 illustrates simplification operations observed for a fixed grade 12 Spanish input into English with target grade levels ranging from 9 to 3. For lower grade levels such as 7 and 5, paraphrasing (e.g. “inaugurar”

---

Table 1: Compared to pipeline models, multitask models produce complexity controlled translations that better match human references (BLEU), that are simpler (SARI), and whose resulting complexity correlates better with the target grade level (PCC).

<table>
<thead>
<tr>
<th>Complexity cont. MT</th>
<th>BLEU</th>
<th>SARI</th>
<th>PCC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pipeline Baselines</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Translate then Simplify</td>
<td>21.98</td>
<td>30.4</td>
<td>0.436</td>
</tr>
<tr>
<td>Simplify then Translate</td>
<td>17.09</td>
<td>37.4</td>
<td>0.275</td>
</tr>
<tr>
<td><strong>Multitask Models</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Translate and Simplify</td>
<td>22.51</td>
<td>44.8</td>
<td>0.572</td>
</tr>
<tr>
<td>All Tasks</td>
<td><strong>22.75</strong></td>
<td><strong>45.0</strong></td>
<td><strong>0.608</strong></td>
</tr>
</tbody>
</table>

---

2https://github.com/mmautner/readability
is translated as “set to open”) and sentence splitting is observed. For the simplest grade level, the model deletes additional content such as “authoritations” and “historical”.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>Ahora el museo Mauritshuis está por inaugurar una exposición dedicada a los autorretratos del siglo XVII, que destaca las similitudes y diferencias entre las fotos modernas y las obras de arte históricas.</td>
</tr>
<tr>
<td>9</td>
<td>Now the museum Mauritshois is launching an exhibition dedicated to the 18th century authoritations, highlighting the similarities and differences between modern photos and historical artworks.</td>
</tr>
<tr>
<td>7</td>
<td>The museum is now set to open an exhibition dedicated to the 18th century authoritations. It highlights the similarities and differences between modern photos and historical artworks.</td>
</tr>
<tr>
<td>5</td>
<td>The museum is now set to open an exhibit dedicated to the 18th century authoritations. It shows the similarities and differences between modern photos and historical artworks.</td>
</tr>
</tbody>
</table>

Table 2: Example of multi-task model outputs when translating grade 12 Spanish into increasingly simpler English: the multi-task model performs a wide range of simplification operations including sentence splitting, paraphrasing and deletion.

However, even when simplifying translations, multitask models are not yet able to exactly match the desired complexity level, and the gap between the complexity achieved and the target complexity increases with the amount of simplification required. Our datasets and models thus provide a foundation to investigate strategies for a tighter control on output complexity in future work using training objectives that explicitly addresses this gap or via modelling the type of lexical and syntactic operations performed when simplifying to a grade level.

References


