Sentence Compression for Arbitrary Languages via Multilingual Pivoting

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
2018 Conference on Empirical Methods in Natural Language Processing

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Sentence Compression for Arbitrary Languages via Multilingual Pivoting

Jonathan Mallinson, Rico Sennrich and Mirella Lapata
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB
J.Mallinson@ed.ac.uk, {rsennnic,mlap}@inf.ed.ac.uk

Abstract

In this paper we advocate the use of bilingual corpora which are abundantly available for training sentence compression models. Our approach borrows much of its machinery from neural machine translation and leverages bilingual pivoting: compressions are obtained by translating a source string into a foreign language and then back-translating it into the source while controlling the translation length. Our model can be trained for any language as long as a bilingual corpus is available and performs arbitrary rewrites without access to compression specific data. We release\footnote{Publicly available for download at https://github.com/Jmallins/MOSS} MOSS, a new parallel Multilingual Compression dataset for English, German, and French which can be used to evaluate compression models across languages and genres.

1 Introduction

Sentence compression aims to produce a summary of a single sentence that retains the most important information while preserving its fluency. The task has attracted much attention due to its potential for applications such as text summarization (Jing, 2000; Madnani et al., 2007; Woodsend and Lapata, 2010; Berg-Kirkpatrick et al., 2011), subtitle generation (Vandeeginste and Pan, 2004; Luoto-lahti and Ginter, 2015), and the display of text on small-screens (Corston-Oliver, 2001).

The bulk of research on sentence compression has focused on a simplification of the task involving exclusively word deletion (Knight and Marcu, 2002; Riezler et al., 2003; Turner and Charniak, 2005; McDonald, 2006; Clarke and Lapata, 2008; Cohn and Lapata, 2009), whereas a few approaches view sentence compression as a more general text rewriting problem (Galley and McKeown, 2007; Woodsend and Lapata, 2010; Cohn and Lapata, 2013). Irrespective of how the compression task is formulated, most previous work relies on syntactic information such as parse trees to help decide what to delete from a sentence or which rules to learn in order to rewrite a sentence using less words. More recently, there has been much interest in applying neural network models to natural language generation tasks, including sentence compression (Rush et al., 2015; Filippova et al., 2015; Chopra et al., 2016; Kikuchi et al., 2016). Filippova et al. (2015) focus on deletion-based sentence compression which they model as a sequence labeling problem using a recurrent neural network with long short-term memory units (LSTM; Hochreiter and Schmidhuber 1997). Rush et al. (2015) capture the full gamut of rewrite operations drawing insights from encoder-decoder models recently proposed for machine translation (Bahdanau et al., 2015).

Neural network-based approaches are data-driven, relying on the ability of recurrent architectures to learn continuous features without recourse to preprocessing tools or syntactic information (e.g., part-of-speech tags, parse trees). In order to achieve good performance, they require large amounts of training data, in the region of millions of long-short sentence pairs.\footnote{Rush et al. (2015) use approximately four million training instances and Filippova et al. (2015) two million.} Existing compression datasets are several orders of magnitude smaller. For example, the Ziff-Davis corpus (Knight and Marcu, 2002) contains 1,067 sentences and originated from a collection of news articles on computer products. Clarke and Lapata (2008) create two manual corpora sampled from written (1,433 sentences) and spoken sources (1,370 sentences). Cohn and Lapata (2013) elicit manual compressions for 625 sentences taken from newspaper articles. More recently, Toutanova et al. (2016) crowdsource a larger corpus which contains manual compressions for single \emph{and} multiple sentences (about 26,000 pairs of source and compressed texts).
Since large scale compression datasets do not occur naturally, they must be somehow approximated, e.g., by pairing headlines with the first sentence of a news article (Filippova and Altun, 2013; Rush et al., 2015). As a result, the training corpus construction process must be repeated and reconfigured for new languages and domains (e.g., many headline-first sentence pairs are spurious and need to be filtered using language and domain specific heuristics). And although it may be easy to automatically obtain large scale training data in the news domain, it is not clear how such data can be sourced for many other genres with different writing conventions.

Our work addresses the paucity of data for sentence compression models. We argue that multilingual corpora are a rich source for learning a variety of rewrite rules across languages and that existing neural machine translation (NMT) models (Sutskever et al. 2014; Bahdanau et al. 2015) can be easily adapted to the compression task through bilingual pivoting (Mallinson et al., 2017) coupled with methods which decode the output sequence to a desired length (e.g., subject to language and genre requirements). We obtain compressions by translating a source string into a foreign language and then back-translating it into the source while controlling the translation length (Kikuchi et al., 2016). Our model can be trained for any language as long as a bilingual corpus is available, and can perform arbitrary rewrites while taking advantage of multiple pivots if these exist. We also demonstrate that models trained on multilingual data perform well out-of-domain.

Although our approach does not employ compression corpora for training, for evaluation purposes, we create MOSS, a new Multilingual Compression dataset for English, French, and German. MOSS is a parallel corpus containing documents from the European parliament proceedings, TED talks, news commentaries, and the EU bookshop. Each document is written in English, French, and German, and compressed by native speakers of the respective language who process a document at a time. We obtain five compressions per document leading to 2,000 long-short sentence pairs per language. Like previous related resources (Clarke and Lapata, 2008; Cohn and Lapata, 2013; de Loupy et al., 2010) our corpus is curated manually, however it differs from Toutanova et al. (2016) in that it contains compressions for individual sentences, not documents.

There has been relatively little interest in compressing languages other than English. A few models have been proposed for Japanese (Hori and Furui, 2004; Hirao et al., 2009; Harashima and Kurohashi, 2012), including a neural network model (Hasegawa et al., 2017) which repurposes Filippova and Altun’s (2013) data construction method for Japanese. There is a compression corpus available for French (de Loupy et al., 2010), however, we are not aware of any modeling work on this language. Overall, there are no standardized datasets in languages other than English, either for training or testing.

Our contributions in this work are three-fold: a novel application of bilingual pivoting to sentence compression; corroborated by empirical results showing that our model scales across languages and text genres without additional supervision over and above what is available in the bilingual parallel data; and the release of a multilingual, multi-reference compression corpus which can be effectively used to gain insight in the compression task and facilitate further research in compression modeling.

2 Pivot-based Neural Compression

In our pivot-based sentence compression model an input sequence is first translated into a foreign language, and then back into the source language. Unlike previous paraphrasing pivoting models (Mallinson et al., 2017), we parameterize our translation models with a length feature, which allows us to produce compressed output. We define two models, performing compression in one step or alternatively in two steps which affords more flexibility in model output.

2.1 NMT Background

In the neural encoder-decoder framework for MT (Bahdanau et al., 2015; Sutskever et al., 2014), an encoder takes in a source $X = (x_1, ..., x_{T_x})$ of length $T_x$ and the decoder generates a target sequence $(y_1, ..., y_{T_y})$ of length $T_y$. Let $h_i$ be the hidden state of the source symbol at position $i$, obtained by concatenating the forward and backward encoder RNN hidden states, $h_i = [\overrightarrow{h_i}; \overleftarrow{h_i}]$. We deviate from previous work (Bahdanau et al., 2015; Sutskever et al., 2014) in that we initialize the decoder with the average of the hidden states, following Sennrich et al. (2017):

$$s_0 = \tanh(W_{init} \sum_{i=1}^{T_x} h_i / T_x)$$

(1)
where $W_{\text{init}}$ is a learnt parameter. Our decoder is a conditional recurrent neural network, specifically a gated recurrent unit (GRU, Cho et al., 2014) with attention, which we denote as $cGRU_{\text{att}}$. $cGRU_{\text{att}}$ takes as input the previous hidden state $s_{j-1}$, the source annotations $C = h_1, \ldots, h_T$, and the previously decoded symbol $y_{j-1}$ in order to update its hidden state $s_j$, which is used to decode symbol $y_j$ at position $j$:

$$s_j = cGRU_{\text{att}}(s_{j-1}, y_{j-1}, C)$$

(2)

cGRU_{\text{att}} consists of three components. The first combines the previously decoded symbol $y_{j-1}$ and the previous hidden state $s_{j-1}$ to generate an intermediate representation $s'_j$. The attention mechanism, $ATT$, inputs the entire context set $C$ along with intermediate hidden state $s'_j$ in order to compute the context vector $c_j$:

$$c_j = ATT(C, s'_j) = \sum_{i} \alpha_{ij} h_i$$

(3)

where $\alpha_{ij}$ is the normalized alignment weight between the source symbol at position $i$ and the target symbol at position $j$, and $f$ is a feedforward neural network.

Finally, we generate $s_j$, the hidden state of $cGRU_{\text{att}}$, by using the intermediate representation $s'_j$ and the context vector $c_j$. Given $s_j, y_{j-1}$, and $c_j$ the output probability $p(y_j | s_j, y_{j-1}, c_j)$ is computed using a feedforward neural network with a softmax activation. We define the probability of sequence $y$ as:

$$P(y | x; \theta) = \prod_{j=1}^{T_y} p(y_j | s_j, y_{j-1}, c_j)$$

(6)

### 2.2 Length Control

To be able to produce compressed sentences, we parameterize our model with a length vector which allows to control the output length. Our approach is similar to the LenInit model of Kikuchi et al. (2016), however we use a GRU instead of an LSTM. The hidden state of the decoder consists of the average of the encoder’s hidden states but also a length vector $LV$, a learnt parameter, which is scaled by the desired target length $T'_y$. We therefore rewrite Equation (1) as follows:

$$s'_0 = \tanh \left( W_{\text{init}} \left[ \frac{\sum_{i=1}^{T_x} h_i}{T_x}; LV \cdot T'_y \right] \right)$$

(7)

As such we now define our model as:

$$P(y | x; T'_y; \theta)$$

(8)

During training, the target length is set to $T'_y = T_y$. However, at test time, the target length generally varies according to the domain, genre, and language at hand. We determine the target length experimentally based on a small validation set.

### 2.3 Pivoting

Pivoting is often used in machine translation to overcome the shortage of parallel data, i.e., when there is no translation path from the source language to the target by taking advantage of paths through an intermediate language. The idea dates back at least to Kay (1997), who observed that ambiguities in translating from one language onto another may be resolved if a translation into some third language is available, and has met with success in phrase-based SMT (Wu and Wang, 2007; Utiyama and Isahara, 2007) and more recently in neural MT systems (Firat et al., 2016).

We use pivoting to provide a path from a source English sentence, via an intermediate foreign language, to English in a compressed form. We propose to extend Mallinson et al.’s (2017) approach to multi-pivoting, where a sentence $x$ is translated to $K$-best foreign pivots, $F_k = \{f_1, \ldots, f_K\}$. The probability of generating compression $y = y_1 \ldots y_{T_y}$ is decomposed as:

$$P(y | x) = \sum_{f \in F_k} P(y | f; \theta) \cdot p(f | x; \theta)$$

(9)

which we approximate as the tokenwise weighted average of the pivots:

$$P(y | x) \approx \prod_{f \in F_k} \sum_{j=1}^{T_x} P(y | y_{<j}, f) P(f | x)$$

(10)

where $y_{<j} = y_1 \ldots y_j$. To ensure a probability distribution, we normalize the $K$-best list $F_k$, such that the translation probabilities sum to one. We use beam search to decode tokens by conditioning on multiple pivoting sentences. The results with the best decoding scores are considered candidate compressions.

To ensure the model produces compressed output, we extend the pivoting approach in two ways. In single step compression, one of the translation
models is parameterized with length information:

\[ P(y|x, T_y) \approx \sum_f P(y|f, T_y; \theta) \cdot P(f|x; \theta) \]

In dual-step compression, we parameterize both translation models with length information:

\[ P(y|x, T_y', T_{y'}) \approx \sum_f P(y|f, T_y'; \theta') \cdot P(f|x, T_{y'}; \theta) \]

We find that dual-compression performs better when the system is expected to drastically compress the source sentence (e.g., in a headline generation task). Imposing a high compression ratio from the start tends to produce unintelligible text. The model attempts to reduce the length of the source at all costs, even at the expense of being semantically faithful to the input. Performing two moderate compressions in succession reduces both length and content conservatively and as a result produces more meaningful text.

In Figure 1 we illustrate how the pivot-based model sketched above can successfully control the output of the generated compressions. We show the output of a single-step compression model on three languages initialized with varying compression rates\(^3\) (see Section 4 for details on how the models were trained and tested). The compression rate (CR) is used to determine length parameter of Equation (8):

\[ T_y' = T_x \cdot CR \quad (11) \]

The figure shows how the output length varies compared to a vanilla encoder-decoder system which uses pivoting to backtranslate the source language (Mallinson et al., 2017). We can see that the majority of sentences are generated with length close to the desired compression rate.

### 3 The MOSS Dataset

For evaluation purposes, we created a multilingual sentence compression corpus in English, German, and French. The corpus was collated from existing document and sentence aligned multilingual datasets which vary both in terms of topic and genre. We sampled five documents each from:

1. Europarl, the European Parliament Proceedings Parallel Corpus (Koehn, 2005), has been used extensively in machine translation research; it contains the minutes of the European parliament and is a spoken corpus of formulaic nature; speakers take part in debating various issues concerning EU policy (e.g., taxation, environment).

2. The TED parallel Corpus (Cettolo et al., 2012) contains transcripts in multiple languages of short talks devoted to spreading powerful ideas on a variety of topics ranging from science to business and global issues.

3. The EU bookshop corpus (Skadiņš et al., 2014) contains publications from European institutions covering a variety of topics such as refugees, gender equality, and travel.

4. The News Commentary Parallel Corpus contains articles downloaded from Project Syndicate, an international media organization that publishes commentary on global topics (e.g., economics, world affairs).

We obtained compressions using the Crowdflower platform. Crowdworkers were given instructions that explained the task and defined sentence compression with the aid of examples. They
Table 1: Examples of crowdsourced compressions (in italics) from the MOSS corpus. Sentences shown (in order of appearance) from Europarl, TED, and News Commentary corpora.

<table>
<thead>
<tr>
<th>English</th>
<th>French</th>
<th>German</th>
</tr>
</thead>
<tbody>
<tr>
<td>On the very day that the earthquake struck, the European Council asked the High Representative and the Commission to mobilise all appropriate assistance. Assistance was mobilized on the very day of the earthquake.</td>
<td>Le jour même du tremblement de terre, le Conseil européen a demandé à la haute représentante et à la Commission de mobiliser toute l’aide appropriée.</td>
<td>Am gleichen Tag, an dem das Erdbeben ausbrach, ersuchte der Europäische Rat die Hohe Vertreterin und die Kommission um die Mobilisierung aller angemessenen Hilfe. Europa erbrachte Hilfe noch am selben Tag.</td>
</tr>
<tr>
<td>We’re at a tipping point in human history, a species poised between gaining the stars and losing the planet we call home. We’re at tipping point in human history, poised between gaining the stars and losing the Earth.</td>
<td>L’histoire humaine est à un tournant. Notre espèce hésite à toucher les étoiles ou à perdre la planète qui est la sienne.</td>
<td>Wir stehen vor einem historischen Wendepunkt: zwischen dem Griff nach den Sternen und dem Verlust unseres Heimatplaneten. Wir sind vor einem historischen Wendepunkt: zwischen dem Griff nach Sternen und Verlust unseres Planeten.</td>
</tr>
</tbody>
</table>

Table 2 presents various statistics on our corpus. As can be seen, Europarl contains the longest sentences across languages (see column SL), TED contains the shortest sentences, while the other two corpora are somewhere in-between. We also observe that crowdworkers compress the least when it comes to TED (see column CR), which is not surprising given the brevity of the utterances. Overall, French speakers seem more conservative when shortening sentences compared to English and German. In general, compression rates are genre dependent, they range from 0.58 (for English Europarl) to 0.84 (for German TED). We also examined the degree to which crowdworkers paraphrase the source sentence using Translation Edit Rate (TER; Snover et al., 2006), a measure commonly used to automatically evaluate the quality of machine translation output. We used TER to compute the (average) number of edits required to change a long sentence to shorter output. We also report the number of edits by type, i.e., the number of insertions, substitutions, deletions, and shifts needed (on average) to convert long to short sentences. We observe that crowdworkers perform a fair amount of rewriting across corpora and languages. The most frequent rewrite operations are deletions followed by substitutions, shifts, and insertions.

4 Experimental Setup

Neural Machine Translation Training Nematus (Sennrich et al., 2017) was used as the machine translation system for all our experiments. We generally used the default settings and training procedures as specified within Nematus. All networks have a hidden layer size of 1,000, and an embedding layer size of 512. In addition, layer normalization (Ba et al., 2016) was used. During training, we used ADAM (Kingma and Ba, 2014), a minibatch size of 80, and the training set was reshuffled between epochs. We also employed early stopping.
We experimented with various model variants using one or multiple pivots. The compression rate (see Equation (8)) was tuned experimentally on the validation set which consists of one document from each domain (20 source sentences; 100 compression-pairs). Compression rates varied from 0.55 to 0.85 and were broadly comparable to those shown in Table 2.

Table 2: MOSS statistics across corpora and languages: length of source (SL) and target sentence (TL), compression rate (CR), TER scores, and number of insertions (Ins), deletions (Del), substitutions (Sub), and shifts (Shft).

<table>
<thead>
<tr>
<th>Language</th>
<th>SL</th>
<th>TL</th>
<th>CR</th>
<th>Ins</th>
<th>Del</th>
<th>Sub</th>
<th>Shft</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>English</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EU Par</td>
<td>27.29</td>
<td>15.89</td>
<td>0.58</td>
<td>0.47</td>
<td>0.07</td>
<td>11.29</td>
<td>1.37</td>
</tr>
<tr>
<td>TED</td>
<td>10.64</td>
<td>7.70</td>
<td>0.72</td>
<td>0.40</td>
<td>0.11</td>
<td>2.89</td>
<td>1.33</td>
</tr>
<tr>
<td>News</td>
<td>20.32</td>
<td>15.25</td>
<td>0.75</td>
<td>0.42</td>
<td>0.21</td>
<td>5.58</td>
<td>2.64</td>
</tr>
<tr>
<td>Books</td>
<td>20.52</td>
<td>15.23</td>
<td>0.78</td>
<td>0.38</td>
<td>0.14</td>
<td>5.43</td>
<td>1.86</td>
</tr>
<tr>
<td><strong>All</strong></td>
<td>19.69</td>
<td>13.51</td>
<td>0.71</td>
<td>0.42</td>
<td>0.13</td>
<td>6.30</td>
<td>1.85</td>
</tr>
<tr>
<td><strong>French</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EU Par</td>
<td>29.40</td>
<td>23.00</td>
<td>0.78</td>
<td>0.29</td>
<td>0.06</td>
<td>6.39</td>
<td>1.89</td>
</tr>
<tr>
<td>TED</td>
<td>6.16</td>
<td>4.88</td>
<td>0.79</td>
<td>0.46</td>
<td>0.03</td>
<td>1.63</td>
<td>1.35</td>
</tr>
<tr>
<td>News</td>
<td>27.52</td>
<td>21.03</td>
<td>0.76</td>
<td>0.37</td>
<td>0.11</td>
<td>6.56</td>
<td>2.92</td>
</tr>
<tr>
<td>Books</td>
<td>21.97</td>
<td>17.93</td>
<td>0.82</td>
<td>0.24</td>
<td>0.08</td>
<td>4.12</td>
<td>1.25</td>
</tr>
<tr>
<td><strong>All</strong></td>
<td>21.26</td>
<td>16.71</td>
<td>0.79</td>
<td>0.34</td>
<td>0.07</td>
<td>4.66</td>
<td>1.85</td>
</tr>
<tr>
<td><strong>German</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EU Par</td>
<td>24.78</td>
<td>15.52</td>
<td>0.63</td>
<td>0.41</td>
<td>0.02</td>
<td>9.28</td>
<td>1.08</td>
</tr>
<tr>
<td>TED</td>
<td>5.44</td>
<td>4.54</td>
<td>0.84</td>
<td>0.25</td>
<td>0.02</td>
<td>0.84</td>
<td>0.49</td>
</tr>
<tr>
<td>News</td>
<td>23.35</td>
<td>14.68</td>
<td>0.63</td>
<td>0.47</td>
<td>0.08</td>
<td>9.42</td>
<td>2.11</td>
</tr>
<tr>
<td>Books</td>
<td>19.35</td>
<td>14.37</td>
<td>0.74</td>
<td>0.37</td>
<td>0.09</td>
<td>5.07</td>
<td>1.94</td>
</tr>
<tr>
<td><strong>All</strong></td>
<td>18.23</td>
<td>12.28</td>
<td>0.71</td>
<td>0.38</td>
<td>0.05</td>
<td>6.15</td>
<td>1.41</td>
</tr>
</tbody>
</table>

Comparison Systems We compared our model against ABS, a sequence-to-sequence attention-based model, developed by Rush et al. (2015). This model was trained on a monolingual dataset extracted from the Annotated English Gigaword corpus (Napoles et al., 2011). The dataset consists of approximately 4 million pairs of the first sentence from each source document and its headline. We also trained LenInit (Kikuchi et al., 2016) on the same corpus which is conceptually similar to ABS but additionally controls the output length using a length embedding vector (as described in Section 2.2). Unfortunately, we could not train these models for French or German, since there are no monolingual sentence compression datasets available at a similar scale. An obvious workaround is to translate Gigaword to French and German and then train compression models on the translated data. As the quality of the translation is relatively poor, we also translated German or French into English, compressed it with ABS and LenInit trained on the Gigaword corpus, and then translated the compressions back to French or German. Finally, we include a prefix (Pfix) baseline which does not perform any rewriting but simply truncates the source sentence so that it matches the compression ratio of the validation set.

5 Results

MOSS Evaluation We assessed model performance using three automatic metrics which represent different aspects of the compression task and have been found to correlate well with human judgments (Toutanova et al., 2016; Clarke and Lapata, 2006). These include a recall metric based on skip bi-grams, any pair of words in a sequence allowing for gaps of size four (RS-R); a recall metric based on bi-grams of dependency tree triples (D2-R); and bi-gram ROUGE (R2-F1). We used the Stanford neural network parser (Chen and Manning, 2014) to obtain dependency triples.

Table 3(a) reports results on English with a monolingual model which controls the output length (L) and uses either a single pivot (SP; K = 1) or multiple pivots (MP; K = 10). We experimented with French (fr) or German (de) as pivot languages. All pivot-based models perform compression in a single step (see Section 2.3). Dual-step compression was not available in our model.

<table>
<thead>
<tr>
<th>Language</th>
<th>SL</th>
<th>TL</th>
<th>CR</th>
<th>Ins</th>
<th>Del</th>
<th>Sub</th>
<th>Shft</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>English</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EU Par</td>
<td>27.29</td>
<td>15.89</td>
<td>0.58</td>
<td>0.47</td>
<td>0.07</td>
<td>11.29</td>
<td>1.37</td>
</tr>
<tr>
<td>TED</td>
<td>10.64</td>
<td>7.70</td>
<td>0.72</td>
<td>0.40</td>
<td>0.11</td>
<td>2.89</td>
<td>1.33</td>
</tr>
<tr>
<td>News</td>
<td>20.32</td>
<td>15.25</td>
<td>0.75</td>
<td>0.42</td>
<td>0.21</td>
<td>5.58</td>
<td>2.64</td>
</tr>
<tr>
<td>Books</td>
<td>20.52</td>
<td>15.23</td>
<td>0.78</td>
<td>0.38</td>
<td>0.14</td>
<td>5.43</td>
<td>1.86</td>
</tr>
<tr>
<td><strong>All</strong></td>
<td>19.69</td>
<td>13.51</td>
<td>0.71</td>
<td>0.42</td>
<td>0.13</td>
<td>6.30</td>
<td>1.85</td>
</tr>
</tbody>
</table>

5 We used our own implementation of ABS and LenInit which on DUC-2004 obtained ROUGE scores similar to those published in Rush et al. (2015) and Kikuchi et al. (2016).

6 We add a begin-of-sentence marker at the start of the candidate and reference sentences.
sion obtained inferior results which we omit for the sake of brevity. As can be seen, models which use a single pivot are better than those using multiple ones (German is better than French; see SP_{de} vs SP_{fr}). More pivots might introduce noise at the expense of translation quality.

Overall, pivot-based models outperform ABS and LenInit. This is perhaps to be expected since these models are tested on out of domain data with different vocabulary and writing conventions; MOSS does not contain any newspaper articles. Unfortunately, it is not possible to train ABS and LenInt on in-domain data as compression data only exists for the headlines-first sentences pairs. As an upper bound, we also report how well humans agree with each other, treating one (randomly selected) reference as system output and computing how it agrees with the rest (row Gold in Table 3). All models lag significantly behind human performance on this task.

Table 3:Automatic evaluation on MOSS; S/MP: single/multiple pivot models; L: length parameter; pivot languages: English (en), French (fr), German (de); ABS (Rush et al., 2015) and LenInit (Kikuchi et al., 2016) are sequence-to-sequence models trained on Gigaword; Gold is inter-annotator agreement.

<table>
<thead>
<tr>
<th>English</th>
<th>RS-R</th>
<th>D2-R</th>
<th>R2-F1</th>
<th>French</th>
<th>RS-R</th>
<th>D2-R</th>
<th>R2-F1</th>
<th>German</th>
<th>RS-R</th>
<th>D2-R</th>
<th>R2-F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pfix</td>
<td>45.38</td>
<td>47.57</td>
<td>53.67</td>
<td>Pfix</td>
<td>60.33</td>
<td>62.44</td>
<td>53.37</td>
<td>Pfix</td>
<td>56.28</td>
<td>50.78</td>
<td>45.84</td>
</tr>
<tr>
<td>ABS</td>
<td>18.29</td>
<td>23.55</td>
<td>15.60</td>
<td>ABS</td>
<td>13.84</td>
<td>18.00</td>
<td>9.74</td>
<td>ABS</td>
<td>5.72</td>
<td>12.95</td>
<td>5.21</td>
</tr>
<tr>
<td>LenInit</td>
<td>17.90</td>
<td>19.64</td>
<td>11.18</td>
<td>LenInit</td>
<td>9.91</td>
<td>14.52</td>
<td>8.08</td>
<td>LenInit</td>
<td>4.91</td>
<td>11.77</td>
<td>2.87</td>
</tr>
<tr>
<td>SP_{de}</td>
<td>34.60</td>
<td>37.97</td>
<td>22.67</td>
<td>SP_{en}</td>
<td>16.39</td>
<td>22.08</td>
<td>13.17</td>
<td>SP_{en}</td>
<td>9.43</td>
<td>14.78</td>
<td>6.79</td>
</tr>
<tr>
<td>SP_{fr}</td>
<td>27.42</td>
<td>32.34</td>
<td>19.29</td>
<td>LenInit</td>
<td>20.08</td>
<td>24.41</td>
<td>13.06</td>
<td>LenInit</td>
<td>13.19</td>
<td>18.67</td>
<td>7.65</td>
</tr>
<tr>
<td>MP_{de}</td>
<td>28.71</td>
<td>34.70</td>
<td>19.06</td>
<td>SP_{en}</td>
<td>31.55</td>
<td>37.88</td>
<td>26.59</td>
<td>MP_{en}</td>
<td>38.19</td>
<td>38.54</td>
<td>31.15</td>
</tr>
<tr>
<td>MP_{fr}</td>
<td>20.74</td>
<td>27.50</td>
<td>13.89</td>
<td>MP_{en}</td>
<td>31.55</td>
<td>37.88</td>
<td>26.59</td>
<td>MP_{en}</td>
<td>23.62</td>
<td>29.13</td>
<td>17.36</td>
</tr>
<tr>
<td>Gold</td>
<td>76.60</td>
<td>71.68</td>
<td>42.89</td>
<td>Gold</td>
<td>74.42</td>
<td>80.00</td>
<td>52.13</td>
<td>Gold</td>
<td>76.01</td>
<td>77.48</td>
<td>48.36</td>
</tr>
</tbody>
</table>

Table 4: System output for the example source sentences in Table 1.
We also elicited human judgments through the Crowdflower platform. We asked crowdworkers to rate the grammaticality of the target compressions and whether they preserved the most important information from the source. In both cases, they used a five-point rating scale where a high number indicates better performance. We randomly selected 25 sentences from each corpus from the test portion of MOSS, i.e., 100 long-short sentence pairs per language. We compared compressions generated by our model (SP\_L), with ABS models for the three languages, the prefix baseline, and (randomly selected) gold-standard reference (Ref) compressions from MOSS. All systems used the length parameter to allow comparisons with approximately the same compression rates. We collected five ratings per compression. Our results are summarized in Table 5. We show mean ratings for grammaticality (Gram), importance (Imp) and their combination (column Avg). Across languages our model (SP\_L) significantly \((p < 0.05)\) outperforms comparison systems (Pfix, ABS) on both dimensions of grammaticality and importance (significance tests were performed using a student \(t\)-test). All systems are significantly worse \((p < 0.05)\) than the human reference compressions.

Finally, in Table 6 we analyze the output of our best model (SP\_L) using the same statistics we applied to the human compressions (see Table 2). As can be seen, the model generally compresses more aggressively and applies more edits than the crowdworkers (both compression rates and TER scores are higher for all three languages). Although the rate of deletions is similar to humans, insertions, substitutions and shifts happen to a greater extent for our model, indicating that it performs a good amount of paraphrasing.

### DUC-2004 Evaluation

Besides MOSS, we evaluated our model on the benchmark DUC-2004 task-1 dataset. In this task, the aim is to create a very short summary (75 bytes) for a document. The evaluation set consists of 500 source documents (from the New York Times and Associated Press Wire services) each paired with four human-written (reference) summaries. We follow previous work (Rush et al., 2015; Chopra et al., 2016) in compressing the first sentence of the document and presenting this as the summary. To make the evaluation unbiased to length, the output of all systems is cut-off after 75-characters and no bonus is given for shorter summaries.

Our results are shown in Table 7. To compare with existing methods, we also report ROUGE (Lin, 2004) unigram and bigram overlap (Lin, 2004) and the longest common subsequence (ROUGE-L).\(^9\) We employed a dual step compression model (see Section 2) as preliminary experiments showed that it was superior to single-stage variants. We compared single and multiple pivot models against existing ABS and ABS+ (Rush et al., 2015), two encoder-decoder models trained on the English Gigaword. ABS+ applies minimum error rate (MERT) training as a copy-
Source King Norodom Sihanouk has declined requests to chair a summit of Cambodia’s top political leaders, saying the meeting would not bring any progress in deadlocked negotiations to form a government.

SP_L,de King Norodom Sihanouk has refused to chair Cambodia summit.

Gold Sihanouk refuses to chair Cambodian political summit at home or abroad.

Source Cambodia’s ruling party responded Tuesday to criticisms of its leader in the U.S. Congress with a lengthy defense of strongman Hun Sen’s human rights record.

SP_L,de Cambodia’s ruling party responded Tuesday to criticism of its leader in the US.

Gold Cambodian party defends leader Hun Sen against criticism of U.S. House.

Source The Swiss government has ordered no investigation of possible bank accounts belonging to former Chilean dictator Augusto Pinochet, a spokesman said Wednesday.

SP_L,de Swiss government ordered no inquiry into possible bank accounts of former Chilean dictator Augusto.

Gold Switzerland joins charges against Pinochet but avoids bank probe.

Table 8: System output for DUC-2004.

<table>
<thead>
<tr>
<th>Models</th>
<th>Gram</th>
<th>Imp</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pfix</td>
<td>3.03</td>
<td>2.93</td>
<td>2.98</td>
</tr>
<tr>
<td>SP_L,de</td>
<td>3.37</td>
<td>3.22</td>
<td>3.29</td>
</tr>
<tr>
<td>Topiary</td>
<td>3.05</td>
<td>3.15</td>
<td>3.10</td>
</tr>
<tr>
<td>ABS+</td>
<td>3.67</td>
<td>3.23</td>
<td>3.45</td>
</tr>
<tr>
<td>LenEmb</td>
<td>3.14</td>
<td>3.08</td>
<td>3.11</td>
</tr>
<tr>
<td>Ref</td>
<td>3.62</td>
<td>3.27</td>
<td>3.45</td>
</tr>
</tbody>
</table>

Table 9: Mean ratings elicited by humans on DUC-2004; Avg is the average rating of grammaticality and importance.

Using automatic metrics we see that our model generally performs worse compared to these systems and that German is the best pivot for English. Although the objective of this paper is not to obtain state-of-the-art scores on this evaluation set, it is interesting to see that our model is able to compress out-of-domain. We do not have access to headline-first sentence pairs, while all comparison systems do. We also elicited human judgments on the compressions of 100 lead sentences whose documents were randomly selected from the DUC-2004 test set. We compared the prefix baseline, our model (SP_L,de), ABS+ (Rush et al., 2015), LenEmb (Kikuchi et al., 2016), Topiary (Zajic et al., 2004), and a randomly selected reference. Topiary came top in almost all measures in the DUC-2004 evaluation; it first compresses the lead sentence using linguistically motivated heuristics and then enhances it with topic keywords. Crowdworkers rated grammaticality and importance, using a five-point scale; we collected five ratings per compression.

As shown in Table 9 ABS+ has the lead with our system following suit. In terms of grammaticality, ABS+ and SP_L,de are not significantly different from the gold standard or from each other (Pfix, Topiary, and LenEmb are significantly worse than Gold; \( p < 0.05 \)). In terms of importance, pairwise differences between systems and the gold standard are not significant. Overall, we observe that SP_L,de performs comparably to ABS+ even though it was not trained on any compression specific data. Inspection of system output reveals that our model performs more paraphrasing than comparison systems (a conclusion also confirmed by the statistics in Table 6).

6 Conclusions

In this paper we have shown that multilingual corpora can be used to bootstrap compression models across languages and text genres. Our approach adapts existing neural machine translation machinery to the compression task coupled with methods which decode the output to a desired length. An interesting direction for future work would be to train our model using reinforcement learning (Ranzato et al., 2016; Zhang and Lapata, 2017) in order to control the compression output more directly. Moreover, although we do not use any direct supervision in our experiments, it would be interesting to incorporate it as a means of domain adaptation (Cheng et al., 2016).

Acknowledgments The authors gratefully acknowledge the support of the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1; Mallinson) and the European Research Council (award number 681760; Lapata).
References


