Sentence Compression for Arbitrary Languages via Multilingual Pivoting

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, {rsennric,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[1] MOSS, a new parallel **M**ultilingual C**o**mpre**ss**ion 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 (Vandeghinste and Pan, 2004; Luotolahti 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.[2] 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 *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 approx-

[1] Publicly available for download at https://github.com/Jmallins/MOSS

[2] Rush et al. (2015) use approximately four million training instances and Filippova et al. (2015) two million.

imated, 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 **M**ultilingual C**o**mpre**ss**ion 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):

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

where W_{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 $cGRU_{att}$. $cGRU_{att}$ takes as input the previous hidden state s_{j-1}, the source annotations $C = h_1, ..., h_{T_x}$, 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:

$$s_j = cGRU_{att}(s_{j-1}, y_{j-1}, C) \tag{2}$$

$cGRU_{att}$ 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:

$$c_j = ATT(C, s'_j) = \sum_i^{T_x} \alpha_{ij} h_i \tag{3}$$

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{T_x} exp(e_{kj})} \tag{4}$$

$$e_{ij} = f(s'_j, h_i) \tag{5}$$

Where α_{ij} is the normalized alignment weight between the source symbol at position i and the target symbol at position j, and f is a feedfoward neural network.

Finally, we generate s_j, the hidden state of $cGRU_{att}$, 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:

$$P(y|x;\theta) = \prod_{j=1}^{T_y} p(y_j|s_j, y_{j-1}, c_j) \tag{6}$$

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:

$$s'_0 = tanh\left(W_{init}\left[\frac{\sum_{i=1}^{T_x} h_i}{T_x}; LV \cdot T_{y'}\right]\right) \tag{7}$$

As such we now define our model as:

$$P(y|x, T_{y'}; \theta) \tag{8}$$

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, genres, 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, $F_x = \{f_1, ..., f_K\}$. The probability of generating compression $y = y_1 ... y_{T_y}$ is decomposed as:

$$P(y|x) = \sum_f^{F_x} P(y|f; \overrightarrow{\theta}) \cdot P(f|x; \overleftarrow{\theta}) \tag{9}$$

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

$$P(y|x) \approx \prod_{j=1}^{T_y} \sum_f^{F_x} P(y_j|y_{<j}, f) P(f|x) \tag{10}$$

Where $y_{<j} = y_1, ... y_j$. To ensure a probability distribution, we normalize the K-best list F_x, 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:

$$P(y|x, T_{y'}) \approx \sum_f^{F} P(y|f, T_{y'}; \overrightarrow{\theta}) \cdot P(f|x; \overleftarrow{\theta})$$

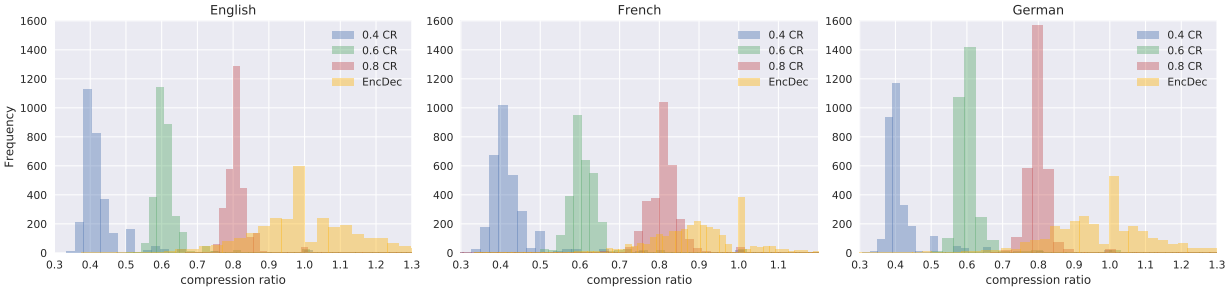

Figure 1: Histograms of output lengths at three compression rates (CR) compared to a vanilla encoder-decoder system which does not manipulate output length. German is used as pivot for English, and English as pivot for French and German.

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

$$P(y|x,T_{y'},T_{y''}) \approx \sum_{f}^{F} P(y|f,T_{y'};\overrightarrow{\theta}) \cdot P(f|x,T_{y''};\overleftarrow{\theta})$$

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:

$$T_{y'} = T_x \cdot CR \tag{11}$$

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 were asked to compress while preserving the most important information, ensuring the sentences remained grammatical and meaning preserving. Annotators were encouraged to use any rewriting operations that seemed appropriate, e.g., to delete words, add new words, substitute them, or reorder

[3]The term refers to the percentage of words retained from the source sentence in the compression.

English	French	German
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.*	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. *Le Conseil européen a demandé à la haute représentante et à la Commission de mobiliser l'aide.*	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.*
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.*	L'histoire humaine est à un tournant. Notre espèce hésite à toucher les étoiles ou à perdre la planète qui est la sienne. *L'humanité est à un tourt. Notre espèce a envie des étoiles ou à perdre sa planète.*	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.*
Surveys undertaken by the World Bank in developing countries show that when poor people are asked to name the three most important concerns they face good health is always mentioned. *World Bank surveys in developing countries show poor people always name good health as an important concern.*	Les enquêtes menées par la Banque mondiale dans les pays en développement montrent que, quand on demande aux populations pauvres de nommer les trois défis les plus importants qu'ils rencontrent, leur "bonne santé" fait toujours partie de cette liste. *Quand on demande aux populations pauvres de nommer les trois défis les plus importants qu'ils rencontrent, leur "bonne santé" fait toujours partie de la liste.*	Umfragen der Weltbank in Entwicklungsländern zeigen, wenn man Arme nach den drei wichtigsten Anliegen fragt, die sie beschäftigen, wird "Gesundheit" immer genannt. *Umfragen in Entwicklungsländern zeigen, dass bei Armen das wichtigste Anliegen Gesundheit ist.*

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.

them. Annotation proceeded on a document-by-document basis, line-by-line. Crowdworkers compressed the first twenty lines of each document and we elicited five compression per document. Example compressions are shown in Table 1.

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 0.64 (for English Europarl) to 0.85 (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 ratio 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 used up to four encoder-decoder NMT models in our experiments (BLEU scores[4] shown in parentheses): English→French (27.03), French→English (29.14), English→German (29.34), and German→English (26.60). German training/test data was taken from the WMT16

[4]BLEU scores were calculated using mteval-v13a.pl

English	SL	TL	CR	TER	Ins	Del	Sub	Shft
EUPar	27.29	17.48	0.64	0.45	0.11	10.66	1.72	0.45
TED	10.64	8.12	0.76	0.34	0.02	2.57	1.02	0.15
News	19.17	14.22	0.74	0.38	0.14	5.39	1.91	0.43
Books	20.52	16.12	0.78	0.32	0.11	4.50	1.54	0.38
All	19.41	13.99	0.73	0.37	0.10	5.78	1.55	0.35

French	SL	TL	CR	TER	Ins	Del	Sub	Shft
EUPar	29.40	23.48	0.79	0.43	0.83	7.04	2.90	0.38
TED	6.16	5.11	0.83	0.44	0.03	1.35	1.33	0.04
News	27.52	21.95	0.79	0.37	0.14	6.37	3.06	0.50
Books	22.32	18.48	0.83	0.36	0.52	4.21	1.79	0.20
All	21.35	17.26	0.81	0.40	0.38	4.74	2.27	0.28

German	SL	TL	CR	TER	Ins	Del	Sub	Shft
EUPar	24.53	16.87	0.69	0.38	0.10	8.70	1.14	0.18
TED	5.36	4.55	0.85	0.24	0.02	0.76	0.53	0.10
News	23.48	16.49	0.70	0.45	0.13	8.39	2.15	0.47
Books	19.83	14.97	0.75	0.50	0.52	5.66	2.89	0.34
All	18.30	13.22	0.75	0.39	0.19	5.88	1.68	0.27

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).

shared task and French from the WMT14 shared task. The training data was 4.2 million and 39 million sentence pairs for EN-DE, and EN-FR, respectively. We also used back-translated monolingual training data, from the news domain, (Sennrich et al., 2016a) in training for the German systems. The data was pre-processed using standard scripts found in MOSES (Koehn et al., 2007). Rare words were split into sub-word units, using byte pair encoding (BPE; Sennrich et al. 2016b). The BPE operations are shared between language directions.

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.

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).[5] 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[6] (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 model which controls the output length ($\mathcal{L}$) 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 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

[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

English	RS-R	D2-R	R2-F1
Pfix	45.38	47.57	33.67
ABS	18.29	23.55	15.60
LenInit	17.90	19.64	11.18
$SP_{\mathcal{L},de}$	**34.60**	**37.97**	**22.67**
$SP_{\mathcal{L},fr}$	27.42	32.34	19.29
$MP_{\mathcal{L},de}$	28.71	34.70	19.06
$MP_{\mathcal{L},fr}$	20.74	27.50	13.89
Gold	76.60	71.68	42.89

(a)

French	RS-R	D2-R	R2-F1
Pfix	60.33	62.44	53.37
ABS	13.84	18.00	9.74
ABS_{en}	16.39	22.08	13.17
LenInit	9.91	14.52	8.08
$LenInit_{en}$	20.08	24.41	13.06
$SP_{\mathcal{L},en}$	**43.38**	**46.17**	**35.07**
$MP_{\mathcal{L},en}$	31.55	37.88	26.59
Gold	74.42	80.00	52.13

(b)

German	RS-R	D2-R	R2-F1
Pfix	56.28	50.78	45.84
ABS	5.72	12.95	5.21
ABS_{en}	9.43	14.78	6.79
LenInit	4.91	11.77	2.87
$LenInit_{en}$	13.19	18.67	7.65
$SP_{\mathcal{L},en}$	**38.19**	**38.54**	**31.15**
$MP_{\mathcal{L},en}$	23.62	29.13	17.36
Gold	76.01	77.48	48.36

(c)

Table 3: Automatic evaluation on MOSS; S/MP: single/multiple pivot models; $\mathcal{L}$: 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.

	English	French	German
ABS	Europe urged to help quake victims.	Le Conseil Européen demande une aide pour les victimes du tremblement de terre.	Europäischer Rat sucht Hilfen für Quiz-Opfer.
SP	The European Council called on the High Representative and the Commission to mobilise all appropriate assistance.	Le Conseil Européen a demandé au Haut Représentant et à la Commission de mobiliser l'assistance.	Am selben Tag forderte der Europäische Rat die Hohe Vertreterin und die Kommission auf, jede Hilfe.
ABS	Advance for Sunday July a new look at the world.	Un tournant pour le tournant.	Die Stars der Stars und die Stars.
SP	We are at a turning point in human history and losing the planet we call home.	L'histoire de l'humanité est à la croisée des chemins et de l'histoire.	Zwischen dem Griff der Sterne und dem Verlust unseres Planeten stehen wir vor.
ABS	Poor people ask to name the three most important concerns.	Les enquêtes de la Banque mondiale révèlent que la santé fait toujours partie de la liste.	Weltbank-Umfragen zeigen arme Menschen in Entwicklungsländern.
SP	Polls conducted by the World Bank show that when poor people are asked to mention the three main concerns.	Les enquêtes menées par la Banque mondiale dans les pays en développement montrent que, lorsqu'on demande aux pauvres de nommer les trois plus grands éfis.	Wenn man die Armen nach den drei Hauptanliegen fragt, werden sie gefordert.

Table 4: System output for the example source sentences in Table 1.

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.

Tables 3(b) and 3(c) report results on French and German, respectively. For these languages, we obtained best results with English as pivot, using a single-step compression model. ABS and LenInit perform poorly when trained directly on translations of Gigaword into French and German; their performance improves considerably when they are trained on the Gigaword and used to compress English translations of French or German (ABS_{en}, $LenInit_{en}$). Again, we observe that our models ($SP_{\mathcal{L},en}$, $MP_{\mathcal{L},en}$) outperform the comparison systems across all metrics and that using a single pivot yields better compressions. Example compressions are given in Table 4 where we show output produced by ABS and SP for each language (see the supplementary material for more examples). Finally, notice that automatic scores for the prefix baseline across languages are misleadingly high, since it simply repeats the source sentence up to a fixed length without performing any rewriting.

	English			French			German		
Models	Imp	Gram	Avg	Imp	Gram	Avg	Imp	Gram	Avg
Pfix	2.72	2.98	2.85	2.73	2.89	2.80	3.17	2.96	3.06
LenInit	2.51	3.0	2.75	1.82	2.62	2.22	2.10	3.25	2.67
$SP_{\mathcal{L}}$	**3.27**	**3.69**	**3.48**	**3.48**	**3.60**	**3.54**	**3.30**	**3.87**	**3.59**
Ref	3.47	3.80	3.63	4.05	4.14	4.10	3.97	4.26	4.10

Table 5: Mean ratings elicited by humans on MOSS; Avg is the average rating of grammaticality and importance.

	SL	TL	CR	TER	Ins	Del	Sub	Shft
English	19.41	12.31	0.63	0.65	0.10	6.68	2.14	0.44
French	21.35	14.98	0.70	0.67	0.29	5.71	3.36	0.61
German	18.30	12.51	0.68	0.67	0.16	6.38	2.94	0.50

Table 6: Statistics of model output ($SP_{\mathcal{L}}$) on MOSS (aggregated across domains): length of source (SL) and target (TL), compression rate (CR), TER scores, and number of insertions (Ins), deletions (Del), substitutions (Sub), and shifts (Shft).

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 ($SP_{\mathcal{L}}$), 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 ($SP_{\mathcal{L}}$) using the same statistics we applied to the human compressions (see Table 2). As can be seen, the model generally compressess 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 insertions and deletions is similar to humans, 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.

Models	RS-R	D2-R	R2-F1	R1-R	R2-R	RL-R
Pfix	15.25	15.59	5.38	20.42	5.86	18.07
$SP_{\mathcal{L},de}$	**12.93**	**13.89**	**4.97**	**20.70**	**5.35**	**18.35**
$SP_{\mathcal{L},fr}$	12.06	12.18	4.42	19.77	4.75	17.40
$MP_{\mathcal{L},fr}$	10.38	11.85	3.70	18.67	4.03	16.20
$MP_{\mathcal{L},de}$	11.06	13.26	4.30	19.10	4.69	16.84
Gold	16.41	18.12	7.72	26.95	7.72	22.79
ABS[7] (Rush et al., 2015)				26.55	7.06	22.05
ABS+ (Rush et al., 2015)				28.18	8.49	23.81
RAS (Chopra et al., 2016)				28.97	8.26	24.06
LenInit [8] (Kikuchi et al., 2016)				25.87	8.27	23.24
LenEmb (Kikuchi et al., 2016)				26.73	8.40	23.88

Table 7: DUC-2004 results (75 char length cap); results for comparison systems are taken from their respective papers.

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 copying mechanism. LenEmb and LenInit include a length parameter (Kikuchi et al., 2016), whereas RAS uses a specialized recurrent neural network architecture (Elman, 1990). We also report how well DUC-2004 abstractors agree with each other (row Gold in Table 7). Example compressions are given in Table 8, where we show output produced by SP_en and a corresponding human reference (see the supplementary material for further examples).

Using automatic metrics we see that our model generally performs worse compared to these sys-

[7] Our ABS implementation obtains R1-R 25.03, R2-R 8.40, and RL-R: 22.35

[8] Our LenInit implementation obtains R1-R 29.26, R2-R 9.56, and RL-R 25.70

[9] We used ROUGE version 1.5.5 with the original DUC-2004 ROUGE parameters.

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**$_{\mathcal{L},de}$: 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**$_{\mathcal{L},de}$: 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**$_{\mathcal{L},de}$: 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.

tems 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 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$_{\mathcal{L},de}$), 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$_{\mathcal{L},de}$ 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$_{\mathcal{L},de}$ 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).

Models	Gram	Imp	Avg
Pfix	3.03	2.93	2.98
SP$_{\mathcal{L},de}$	3.37	3.22	3.29
Topiary	3.05	3.15	3.10
ABS+	**3.67**	**3.23**	**3.45**
LenEmb	3.14	3.08	3.09
Ref	3.62	3.27	3.45

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

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

Lei Jimmy Ba, Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*, San Diego, California.

Taylor Berg-Kirkpatrick, Dan Gillick, and Dan Klein. 2011. Jointly learning to extract and compress. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 481–490, Portland, Oregon, USA.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 740–750, Doha, Qatar.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1965–1974, Berlin, Germany.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, Doha, Qatar.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, San Diego, California.

James Clarke and Mirella Lapata. 2006. Models for sentence compression: A comparison across domains, training requirements and evaluation measures. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 377–384.

James Clarke and Mirella Lapata. 2008. Global inference for sentence compression: An integer linear programming approach. *Journal of Artificial Intelligence Research*, 31:273–381.

Trevor Cohn and Mirella Lapata. 2009. Sentence compression as tree transduction. *Journal of Artificial Intelligence Research*, 34:637–674.

Trevor Cohn and Mirella Lapata. 2013. An abstractive approach to sentence compression. *ACM Transactions on Intelligent Systems and Technology*, 4(3):1–35.

Simon Corston-Oliver. 2001. Text Compaction for Display on Very Small Screens. In *Proceedings of the NAACL Workshop on Automatic Summarization*, pages 89–98, Pittsburgh, Pennsylvania.

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

Katja Filippova, Enrique Alfonseca, Carlos A. Colmenares, Lukasz Kaiser, and Oriol Vinyals. 2015. Sentence compression by deletion with LSTMs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 360–368, Lisbon, Portugal.

Katja Filippova and Yasemin Altun. 2013. Overcoming the lack of parallel data in sentence compression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1481–1491, Seattle, Washington, USA.

Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277. Association for Computational Linguistics.

Michel Galley and Kathleen McKeown. 2007. Lexicalized Markov grammars for sentence compression. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 180–187, Rochester, New York.

Jun Harashima and Sadao Kurohashi. 2012. Flexible Japanese sentence compression by relaxing unit constraints. In *Proceedings of COLING 2012*, pages 1097–1112, Mumbai, India.

Shun Hasegawa, Yuta Kikuchi, Hiroya Takamura, and Manabu Okumura. 2017. Japanese sentence compression with a large training dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 281–286, Vancouver, Canada.

Tsutomu Hirao, Jun Suzuki, and Hideki Isozaki. 2009. A syntax-free approach to japanese sentence compression. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 826–833, Suntec, Singapore.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Chiori Hori and Sadaoki Furui. 2004. Speech summarization: an approach through word extraction and a method for evaluation. *IEICE Transactions on Information and Systems*, E87-D(1):15–25.

Hongyan Jing. 2000. Sentence Reduction for Automatic Text Summarization. In *Proceedings of the 6th ANLP*, pages 310–315, Seattle,WA.

Martin Kay. 1997. The proper place of men and machines in language translation. *Machine Translation*, 12(1-2):3–23.

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 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.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Kevin Knight and Daniel Marcu. 2002. Summarization beyond sentence extraction: a probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1):91–107.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, pages 70–86, Phuket, Thailand.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain.

Claude de Loupy, Marie Guégan, Christelle Ayache, Somara Seng, and Juan-Manuel Torres Moreno. 2010. A french human reference corpus for multi-document summarization and sentence compression. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Juhani Luotolahti and Filip Ginter. 2015. Sentence compression for automatic subtitling. In *Proceedings of the 20th Nordic Conference for Computational Linguistics*, pages 135–143, Vilnius, Lithuania.

Nitin Madnani, David Zajic, Bonnie Dorr, Necip Fazil Ayan, and Jimmy Lin. 2007. Mutiple alternative sentence compressions for automatic text summarization. In *Proceedings of the 2007 Document Understanding Conference*, Rochester, NY, USA.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893, Valencia, Spain.

Ryan McDonald. 2006. Discriminative sentence compression with soft syntactic constraints. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 297–304, Trento, Italy.

Courtney Napoles, Chris Callison-Burch, Juri Ganitkevitch, and Benjamin Van Durme. 2011. Paraphrastic sentence compression with a character-based metric: Tightening without deletion. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 84–90, Portland, Oregon.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *Proceedings of the International Conference on Learning Representations*, San Juan, Puerto Rico.

Stefan Riezler, Tracy H. King, Richard Crouch, and Annie Zaenen. 2003. Statistical sentence condensation using ambiguity packing and stochastic disambiguation methods for lexical-functional grammar. In *Proceedings of the HLT/NAACL*, pages 118–125, Edmonton, Canada.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany.

Raivis Skadiņš, Jörg Tiedemann, Roberts Rozis, and Daiga Deksne. 2014. Billions of parallel words for

free: Building and using the EU bookshop corpus. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation of the Americas*, pages 223–231, Cambridge, Massachusetts.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Kristina Toutanova, Chris Brockett, Ke M. Tran, and Saleema Amershi. 2016. A dataset and evaluation metrics for abstractive compression of sentences and short paragraphs. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 340–350, Austin, Texas.

Jenine Turner and Eugene Charniak. 2005. Supervised and unsupervised learning for sentence compression. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 290–297, Ann Arbor, Michigan.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491, Rochester, New York.

Vincent Vandeghinste and Yi Pan. 2004. Sentence compression for automated subtitling: A hybrid approach. In *Proceedings of the ACL Workshop on Text Summarization*, pages 89–95, Barcelona, Spain.

Kristian Woodsend and Mirella Lapata. 2010. Automatic generation of story highlights. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 565–574, Uppsala, Sweden.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21(3):165–181.

David Zajic, Bonnie Dorr, and Richard Schwartz. 2004. Bbn/umd at duc-2004: Topiary. In *In Proceedings of the NAACL Workshop on Document Understanding*, pages 112–119, Boston, MA.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Copenhagen, Denmark.