Comparison of Data Selection Techniques for the Translation of Video Lectures

Joern Wuebker, Hermann Ney, Martínez-Villaronga Adrià, Adrià Giménez, Alfons Juan, Christophe Servan, Marc Dymetman, Shachar Mirkin

Comparison of Data Selection Techniques for the Translation of Video Lectures

Joern Wuebker[1] wuebker@cs.rwth-aachen.de
Hermann Ney[1,2] ney@cs.rwth-aachen.de
[1]RWTH Aachen University, Aachen, Germany
[2]Univ. Paris-Sud, France and LIMSI/CNRS, Orsay, France

Adrià Martínez-Villaronga amartinez1@dsic.upv.es
Adrià Giménez agimenez@dsic.upv.es
Alfons Juan ajuan@dsic.upv.es
Universitat Politècnica de València, València, Spain

Christophe Servan christophe.servan@xrce.xerox.com
Marc Dymetman marc.dymetman@xrce.xerox.com
Shachar Mirkin shachar.mirkin@xrce.xerox.com
Xerox Research Centre Europe, Meylan, France

Abstract

For the task of online translation of scientific video lectures, using huge models is not possible. In order to get smaller and efficient models, we perform data selection. In this paper, we perform a qualitative and quantitative comparison of several data selection techniques, based on cross-entropy and infrequent n-gram criteria. In terms of BLEU, a combination of translation and language model cross-entropy achieves the most stable results. As another important criterion for measuring translation quality in our application, we identify the number of out-of-vocabulary words. Here, infrequent n-gram recovery shows superior performance. Finally, we combine the two selection techniques in order to benefit from both their strengths.

1 Introduction

With the continuous growth of available bitexts and research advances of the underlying technology, statistical machine translation (SMT) has become popular for many real world tasks. The most common approach is still the phrase-based paradigm (Koehn et al., 2003), that provides an efficient framework with good translation quality for many language pairs.

This work focuses on the application of SMT to the task of translating scientific video lectures. Online scientific video lectures are becoming increasingly popular, e.g. in the context of massive open online courses (MOOCs). Being able to provide high quality automatic translations for this kind of technical talks could, e.g., prove beneficial to education at universities,

sharing technical knowledge and connecting researchers around the world.

With today's large amounts of available data for SMT training, selecting the most valuable portions can be crucial to obtain good performance. First, for the practical task of online translation, using huge models is inefficient and can render real-time translation impossible, especially on mobile devices. The use of smaller training data leads to faster and more space-efficient translation systems. Secondly, selecting the data that is most relevant to the domain at hand, e.g. scientific lectures, can have a significant impact on translation quality. This is why we look for approaches that get or adapt small and efficient models. The task of adapting a translation system to perform well on a specific type of language is called *domain adaptation* and will be discussed in Section 2. One of the prominent branches of domain adaptation research is data selection.

In this work, we perform a qualitative and quantitative comparison of several data selection techniques based on two oppositional criteria, cross-entropy and infrequent n-gram recovery. While the cross-entropy criterion selects sentences that are most similar to a given domain, infrequent n-gram recovery puts the emphasis onto adding new information to the translation system. Our results show that in terms of BLEU, a combination of translation and language model cross-entropy achieves the most stable results.

However, for the task of translating scientific lectures, the number of out-of-vocabulary (OOV) words is also an important criterion to evaluate translation quality. Although in our experiments OOV words make up only a small portion of the data and thus have no visible effect on BLEU, we show examples where it does impact the translation quality as perceived by humans. With respect to the OOV rate, infrequent n-gram recovery has strong advantages over cross-entropy based methods.

Finally, we propose to combine the data selected with two different approaches in order to benefit from both new information provided by infrequent n-gram recovery and domain-related distribution.

The paper is organized as follows. We will discuss previous research on domain adaptation in Section 2 and provide a detailed description of the data selection techniques in Section 3. Section 4 gives an account of the statistical translation system used in our experiments. Finally, the experimental setup and results are discussed in Section 5 and we conclude with Section 6.

2 Domain Adaptation

Domain adaptation can be performed in different ways: using lightly-supervised approaches, model combination/update or data selection.

2.1 Lightly-supervised approaches

A common way to adapt a statistical machine translation model is to use lightly-supervised approaches. These approaches aim to self-enhance the translation model. This was first proposed by Ueffing (2006) and refined by Ueffing et al. (2007). The main idea is to filter the translations with the translated test data. This process involves a confidence measure in order to select the

most reliable data to train a small additional phrase table (PT). The generic and the new phrase tables are used jointly for translation, which can be seen as a mixture model with one specific PT built for each test set.

The lightly-supervised training approach proposed by Schwenk (2008) does not adapt the model to the test data, but it proposes to add large amounts of monolingual training data translated using a completely new model. Lambert et al. (2011) enhanced this approach by using the translations of monolingual data in the target language.

2.2 Model combination and update

One way to adapt MT models is to combine translation models. Models can be combined using the mixture-model approach, a log-linear combination or through incremental learning approaches.

To obtain a mixture of domain-specific models trained on several different domain-specific corpora, they can be combined using a log-linear or a linear approach (Foster and Kuhn, 2007; Civera and Juan, 2007). The standard log-linear model may be used to combine some domain-specific models (Koehn and Schroeder, 2007). In the same way target language models may be combined using a log-linear or a linear combination (Schwenk and Koehn, 2008). Sennrich et al. (2013) proposed to combine different specific parts of the phrase-table during translation leading to a multi-domain adaptation approach.

Niehues and Waibel (2012) compared several incremental approaches, namely the back-off, the factored, the log-linear and the fill-up (Bisazza et al., 2011) techniques. These approaches aim at adapting an MT system towards a target domain using small amounts of parallel in-domain data. The main outcome of this paper is that all the approaches successfully improve the generic model and none of them is better than the others. The performances of the approaches mainly depend on their match to the specific data.

2.3 Data selection

The main idea of data selection is to try to take advantage of a generic corpus by picking out a subset of training data that is most relevant to the domain of interest.

Two main approaches are used to perform domain adaptation. On one hand, such approaches use information retrieval techniques and similarity scores. On the other hand, language models are used associated to perplexity and cross-entropy.

Intuitively, seeking the data closest to the test set is related to information retrieval techniques. Lü et al. (2007) present this approach using the standard measure $TF.IDF$ (Term Frequency – Inverse Document Frequency) to measure the similarity between the test sentences and the training sentences. This approach is based on a bag-of-words scheme.

The second approach, based on language models (LMs), was originally proposed by Gao and Zhang (2002). Here, the generic corpus is scored against an LM trained on a seed of domain-specific data, and the cross-entropy is computed for each sentence. Then, the same generic corpus is scored against an LM trained on a random sample taken from itself. Now,

sentences of the generic corpus are sorted regarding the computation of the difference between domain-specific score and generic score. At last, the best amount of the sorted data has to be determined. This best point is found by minimizing the perplexity of a development set on growing percentages of the sorted corpus.

Moore and Lewis (2010) reported that the perplexity decreases when less, but more appropriate data is used. Recent works expand this approach to bitexts (Axelrod et al., 2011; Mansour et al., 2011).

Approaches like corpus weighting (Shah et al., 2010) or sentence weighting (Matsoukas et al., 2009; Mansour and Ney, 2012) are not suitable to our translation task because these approaches can produce huge models by considering the whole data.

3 Cross-entropy based Data Selection versus Infrequent n-gram Recovery

In this section we detail the different approaches experimented with for data selection. On one hand we process the data selection for both LM and translation model (TM) using cross-entropy. On the other hand, the infrequent n-gram recovery (Gascó et al., 2012), is explored.

3.1 Language Model Cross-entropy

The LM cross-entropy difference can be used for both monolingual data selection for LM training as described by Moore and Lewis (2010), or bilingual selection for translation model training (Axelrod et al., 2011).

Given an in-domain corpus I and an out-of-domain or general-domain corpus O, first we generate a random subset $\hat{O} \subseteq O$ of approximately the same size as I, and train the LMs LM_I and $LM_{\hat{O}}$ using the corresponding training data. Afterwards, each sentence $o \in O$ is scored according to:

$$H_{LM_I}(o) - H_{LM_{\hat{O}}}(o) \tag{1}$$

where H is the length-normalised LM cross-entropy, which is defined by:

$$H_{LM}(x) = -\sum_{i=1}^{|x|} \frac{1}{|x|} \log p_{LM}(x_i|x_{i-1}) \tag{2}$$

for an LM with a 2-gram order. $|x|$ denotes the number of tokens in sentence $x = x_1, x_2, \ldots, x_{|x|}$. It is computed analogously for higher order LMs.

This idea was adapted by Axelrod et al. (2011) for bilingual data selection for the purpose of translation model training. In this case, we have both source and target in-domain corpora I_{src} and I_{trg}, and correspondingly, out-of-domain corpora O_{src} and O_{trg}, with random subsets $\hat{O}_{src} \subseteq O_{src}$ and $\hat{O}_{trg} \subseteq O_{trg}$. We score each sentence pair (s, t) by the sum of the cross-entropy differences on both source and target side:

$$\bar{H}_{LM}(s,t) = H_{LM_{I_{src}}}(s) - H_{LM_{\hat{O}_{src}}}(s) + H_{LM_{I_{trg}}}(t) - H_{LM_{\hat{O}_{trg}}}(t) \tag{3}$$

Note that since the scores in Equation 3 are computed for the source and target separately, any target sentence t' whose cross-entropy score is similar to that of t can exchange t and have a similar score assigned to it by this method. As a result, poorly aligned data can not be detected by LM cross-entropy scoring only.

3.2 Translation Model Cross-entropy

The IBM-Model 1 (M1) (Brown et al., 1993) is a model used in state-of-the-art SMT systems for a variety of applications. In this work, we apply M1 scores to achieve adaptation to some domain specific data. Mansour et al. (2011) extend the formulation by Axelrod et al. (2011), which is described in Equation (3), by adding the M1 cross-entropy score to the LM cross-entropy score. The M1 cross-entropy for a sentence pair $(s,t) = \left(\left(s_1,...,s_{|s|}\right),\left(t_1,...,t_{|t|}\right)\right)$ is defined as:

$$\bar{H}_{M1}(s,t) = H_{M1_I}(t|s) - H_{M1_{\hat{O}}}(t|s) + H_{M1_I}(s|t) - H_{M1_{\hat{O}}}(s|t) \tag{4}$$

where

$$H_{M1}(t|s) = -\sum_{i=1}^{|t|}\frac{1}{|t|}\log\left(\frac{1}{|s|}\sum_{j=1}^{|s|}p_{M1}(t_i|s_j)\right) \tag{5}$$

The cross-entropy of the inverse M1 model $H_{M1}(s|t)$ is calculated by switching s and t in Equation (5).

This metric has several advantages:

- both standard and inverse direction M1 is used, which leads to a more balanced scoring
- it uses cross-entropy *difference* which has a better correlation with the sample's similarity to a specific domain than simple cross-entropy (cf. (Moore and Lewis, 2010))
- M1 is a translation model and thus can capture the translation quality of a given sentence pair.

We use a linear interpolation of LM and M1 cross-entropy scores for data selection, which Mansour et al. (2011) have shown to perform best. Such a combination is similar to an SMT system decoder score, where one combines several model scores including an LM and a TM. The score of the interpolated metric is defined by:

$$\alpha \cdot \bar{H}_{LM}(s,t) + (1-\alpha)\cdot\bar{H}_{M1}(s,t) \tag{6}$$

In our experiments, the value of α is set to $\alpha = 0.8$, which has proven to perform well on previous tasks. In the following sections, we will refer to the interpolated metric defined by Equation 6 as the selection based on translation model (TM) cross-entropy.

3.3 Infrequent n-gram Recovery

The performance of phrase-based machine translation systems relies on the quality of the phrases extracted from the training samples. Unfortunately, training corpora typically yield

sparse phrases. This means that those word alignments that appear rarely in the training corpus cannot be accurately computed and consequently the phrases cannot be properly extracted.

The goal of infrequent n-gram recovery, introduced by Gascó et al. (2012), is to increase the informativeness of the training set by adding sentences that provide information not seen in the given training corpus. The sentences selected from a generic parallel corpus (from here on, referred to as pool) must contain *infrequent n-grams*, i.e. n-grams that appear less than a given threshold τ in the training corpus, referred to as infrequency threshold. If the source language sentences to be translated are known beforehand, the set of infrequent n-grams can be reduced to the ones present in those sentences.

An infrequency score is defined for the sentences, so that they can be sorted to select the most informative ones. Let $\mathcal{X}$ be the set of n-grams that appear in the sentences to be translated and $\mathbf{w}$ one of them; $C(\mathbf{w})$ the counts of $\mathbf{w}$ in the source language training set; and $N_{\mathbf{f}}(\mathbf{w})$ the counts of $\mathbf{w}$ in $\mathbf{f}$, where $\mathbf{f}$ is the sentence from the pool to be scored. The infrequency score of $\mathbf{f}$ is defined as follows:

$$i(\mathbf{f}) = \sum_{\mathbf{w} \in \mathcal{X}} \frac{\min(1, N_{\mathbf{f}}(\mathbf{w})) \max(0, \tau - C(\mathbf{w}))}{Z(\mathbf{f}, |\mathbf{w}|)} \tag{7}$$

In Equation 7, in order to avoid assigning a high score to noisy sentences with many occurrences of the same infrequent n-gram, only one occurrence of each n-gram is taken into account when computing the score. Additionally, a normalization constant $Z(\mathbf{f}, |\mathbf{w}|)$ is included in the equation, which will be set to 1 if no normalization is used, or to the number of n-grams of order $|\mathbf{w}|$ in $\mathbf{f}$, i.e. $|\mathbf{f}| - |\mathbf{w}| + 1$, otherwise.

Each time a sentence is selected, the scores of the remaining sentences are updated in order to avoid the selection of too many sentences with the same infrequent n-gram. However, since rescoring the whole pool would incur a very high computational cost, a suboptimal search strategy is followed. The search is constrained to the set of the one million highest scoring sentences.

In the experiments performed in this work, we will consider n-grams up to order 3 and an infrequency threshold of $\tau = 25$, values that have proven to perform well in similar previous tasks. Note that, as mentioned, this selection technique depends on the sentences to be translated which, for these experiments, are the source sentences from the test set.

3.4 Comparison

In this paper, cross-entropy based and infrequent n-grams based approaches are compared. But some adjustments need to be made in order to compare them.

Different from the cross-entropy based methods described in Sections 3.1 and 3.2, the selection based on infrequent n-grams uses knowledge of the actual development and test sets (described in Section 5). For a fair comparison, we want to see if the cross-entropy based technique can also benefit from this additional knowledge. To that end, we exchanged the in-domain training corpus I with a concatenation of development and test set to perform a cross-

entropy based selection from the out-of-domain data. Here, we denote the development set as D and the test set as T. We only use the source side of the data, which renders IBM-Model 1 cross-entropy unusable. Thus, we use only language model cross-entropy, modifying Equation 3 by dropping the terms based on the target language. We will refer to this technique as *test cross-entropy*:

$$\bar{H}_{LM}(s,t) = H_{LM_{D_{src}+T_{src}}}(s) - H_{LM_{\hat{O}_{src}}}(s) \tag{8}$$

It should also be noted that the criteria applied for data selection are oppositional between the cross-entropy and the infrequent n-gram approaches. The cross-entropy based methods select the sentences that are most *similar* to a given in-domain data set. The goal here is to use the most domain-relevant data. On the other hand, infrequent n-gram recovery selects those sentences that are most *different* to the data that is already given, trying to provide the translation system with new information. Therefore, it seems natural to combine the two orthogonal techniques.

Combined method: We performed additional experiments, where part of the data is selected based on infrequent n-gram recovery and part is selected with TM model cross-entropy. This way, we hope to benefit from the new information introduced by the first while reinforcing a domain-specific distribution at the same time. In practice we start with the maximum amount of data selected by infrequent n-gram recovery. On top of this, we now add increasing amounts of data selected by TM model cross-entropy, until the full general domain data has been added.

4 Statistical Translation System

We use the standard phrase-based translation decoder from the open source toolkit *Jane* (Wuebker et al., 2012) for all translation experiments. The translation process is framed as a log-linear combination of models, which is a generalization of the source-channel paradigm introduced by Brown et al. (1993). The decoder searches for the best translation $\hat{e}_1^{\hat{I}}$ as defined by the models $h_m(e_1^I, s_1^K, f_1^J)$. It can be written as (Och and Ney, 2004)

$$\hat{e}_1^{\hat{I}} = \arg\max_{I, e_1^I} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, s_1^K, f_1^J) \right\}, \tag{9}$$

where $f_1^J = f_1 \dots f_J$ is the source sentence, $e_1^I = e_1 \dots e_I$ the target sentence and $s_1^K = s_1 \dots s_K$ their phrase segmentation and alignment.

The feature functions h_m include translation channel models in both directions, lexical smoothing models in both directions, an n-gram language model, phrase and word penalty, a jump-distance-based distortion model, a hierarchical orientation model (Galley and Manning, 2008) and an n-gram cluster language model (Wuebker et al., 2013). The log-linear feature weights λ_m are optimized on a development data set with minimum error rate training (MERT) (Och, 2003). As optimization criterion we use BLEU (Papineni et al., 2001).

5 Experiments

In this section we describe the different experiments we made in order to compare between the approaches.

5.1 The VideoLectures.NET Repository

VideoLectures.NET[1] is a free and open access repository of video lectures mostly filmed by people from the Jožef Stefan Institute (JSI, Slovenia) at major conferences, summer schools, workshops and science promotional events from many fields of science. VideoLectures.NET has so far published more than $15K$ lectures, all of them recorded with high-quality, homogeneous standards. VideoLectures.NET is a major player in the diffusion of the open-source Matterhorn platform[2].

VideoLectures.NET has been adopted as the main target repository in the transLectures[3] project. The main objective of transLectures is to develop innovative, cost-effective solutions for producing accurate transcriptions and translations of lectures published on Matterhorn-related repositories. For system development and evaluation purposes, about 27 English lectures (20 hours) from VideoLectures.NET were manually transcribed and translated into several languages. In particular, 23 of these 27 lectures (16 hours) were translated into French by professional translators.

5.2 Data

Our experiments are performed on the task of translating manually transcribed English video lectures into French. In addition to around 5000 sentence pairs from VideoLectures.NET, we use the parallel TED talk data provided for the shared translation task of the *International Workshop on Spoken Language Translation*[4] as in-domain data. The general domain data consists of several corpora. The COSMAT scientific thesis abstracts (Lambert et al., 2012) and the news-commentary-v8 corpus, provided by the *ACL 2013 8th Workshop on Statistical Machine Translation*[5] (WMT), are directly added to the baseline without instance selection due to their small size. The large corpora on which data selection is performed, are the Europarl-v7 corpus (also provided by WMT), the JRC-Acquis corpus (Steinberger et al., 2006) and the Open Subtitles corpus[6] (Tiedemann, 2012). Data statistics for the complete in-domain and out-of-domain data are given in Table 1. For the development and test sets we selected four video lectures each, that were manually transcribed and professionally translated, resulting in a total of 1013 and 1360 sentences for development and test, respectively.

In addition to the target side of the bilingual data, we leverage large amounts of monolingual resources for language model training. These include the Common Crawl Corpus, the 10^9 French-English corpus, the UN corpus and the News Crawl articles, available from the WMT

[1] http://videolectures.net
[2] http://opencast.org/matterhorn
[3] http://translectures.eu
[4] http://www.iwslt2013.org
[5] http://www.statmt.org/wmt13
[6] http://www.opensubtitles.org

		English	French
in-domain	Sentences	159K	
	Running Words	3.1M	3.3M
	Vocabulary	49K	63K
out-of-domain	Sentences	13.9M	
	Running Words	175M	179M
	Vocabulary	648K	617K

Table 1: Data statistics for the bilingual training data. 'Vocabulary' denotes the number of distinct words (i.e. unigrams) that appear in the data.

website. In addition, we use the LDC French Gigaword corpus.[7] Altogether, the language models are trained on 3.1 billion running words.

5.3 Experimental Setup

The baseline system is trained only on the VideoLectures.NET, TED, COSMAT and news-commentary corpora. For all other systems this data is also used, but is extended by the selected sentences. In all experiments we use the concatenation of the Europarl, JRC and Open Subtitles data as the pool for data selection. As in-domain data I for the LM and TM cross-entropy based selection, we concatenate the VideoLectures.NET and TED corpora. For the *test cross-entropy* technique (cf. Section 3.4), these are replaced by the concatenation of the development and test sets, which is denoted by $D_{src} + T_{src}$ in Eq. 8. Infrequent n-gram recovery is performed separately for the development and test set. To compare the effectiveness of the different approaches, we select increasing amounts of data with each technique, starting with 250K source words and going up to the full data. The selected data is then added to the baseline system and the models are retrained. However, with infrequent n-gram recovery, the maximum number of selected source words is 5M. Then, the infrequency threshold is reached and the technique does not select any more sentences.

For the combined method (cf. Section 3.4), we use this maximum amount of data selected with infrequent n-gram recovery and gradually add additional portions by TM cross-entropy selection.

The language models are kept fixed throughout all experiments. We use a 4-gram standard language model and a 7-gram cluster language model. All results are arithmetic means over three independent MERT runs to control for optimizer stability and are reported in BLEU.

5.4 Results

The BLEU scores we obtained on the test set with the different data selection techniques are plotted in Figure 1. The baseline system without any of the additional data already reaches 33.56% BLEU, while the system using all data yields 33.96% BLEU. Using the development and test data for cross-entropy based selection (*test cross-entropy*) is clearly not a good idea. The small amount of training data for the language models that are used to compute the cross-

[7] http://catalog.ldc.upenn.edu/LDC2009T28

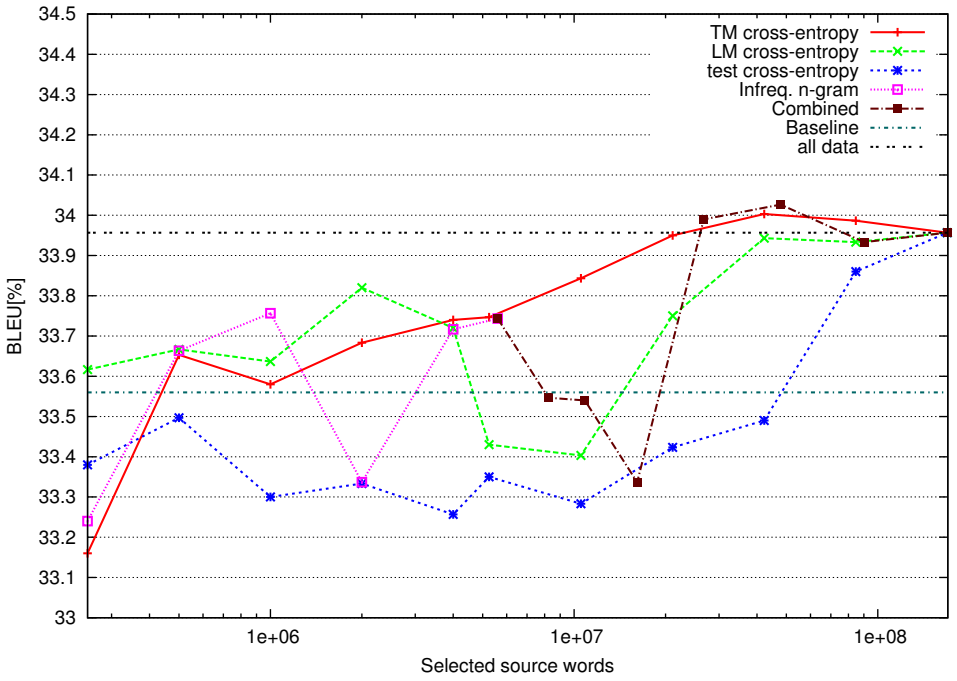

Figure 1: BLEU scores for the different data selection techniques. The x-axis denotes the number of selected source words on a logarithmic scale. The infrequent n-gram recovery selects a maximum of 5M source words, after which the infrequency threshold is reached for all n-grams. The combined method adds additional sentences selected with TM cross-entropy on top of these 5M words.

entropy seems to result in very unreliable estimations for the quality of the data. Further, we can assume that source-only cross-entropy is less stable than complementing it with the target side. However, as it directly makes use of the test set for data selection, it is quicker to recover OOVs (cf. Fig. 2). Regarding the remaining techniques, it is hard to draw a clear conclusion. Due to the small impact of the additional data, all observed values are very close to each other. Altogether, TM cross-entropy seems to yield the most stable results, where translation quality increases with the data size. Both LM and TM cross-entropy based selection reach the same BLEU level as the system using the full data with only $\frac{1}{4}$ of the data. TM cross-entropy has a slight advantage here, reaching 34.00% BLEU. The best result with selecting only 1M sourced words (0.6% of the full out-of-domain data) is achieved by the infrequent n-gram recovery. However, we observe a drop at the next data point, suggesting that the subsequently selected 1M words perturb the domain-specific distribution, resulting in a lower score.

As was mentioned, the infrequent n-gram recovery selects a maximum of 5M words, after which the infrequency threshold is reached for all n-grams. In order to combine this method with cross-entropy based selection, we kept this maximum selection fixed and gradually added increasing amounts of data selected with the TM cross-entropy criterion. Again, adding only a little data yields a decreasing BLEU score. However, after adding an additional $\frac{1}{4}$ of the full data, we reach a score of 34.03% BLEU, which is on the same level as the TM cross-entropy selection alone.

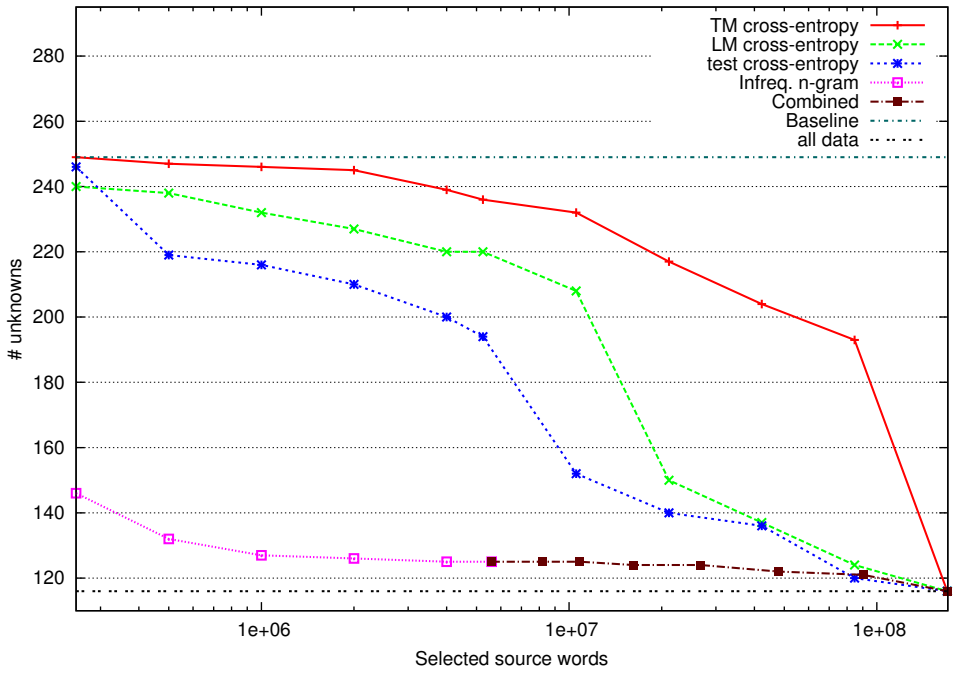

Figure 2: Number of words unknown to the translation model for the different data selection techniques. The x-axis denotes the number of selected source words on a logarithmic scale.

Another relevant measure for the user-oriented translation of technical talks is the the number of words (i.e. unigrams) that are unknown to the translation decoder, which are left untranslated. Figure 2 displays the number of these out-of-vocabulary words for each selection technique. On this criterion, infrequent n-gram selection is clearly superior to the cross-entropy based techniques. After adding only an additional 500k source words (0.3% of the full out-of-domain data), the number of unknown words is reduced by 47% from 249 to 132. Using all data yields a total of 116 unknown tokens in the test set. From the cross-entropy based methods, selection based on the test set has the best recovery of unknown words, followed by LM cross-entropy scoring. The combined method obviously benefits from the strong performance of the infrequent n-gram recovery, but can hardly add any additional words to its vocabulary.

To illustrate the importance of translating unknown words, we have selected two example sentences from the VideoLectures.NET test set and compared their translations with TM cross-entropy selection and infrequent n-gram selection in Figure 3. In both cases, 1M words were selected from the out-of-domain data. In the first example, the English word *re-sell* is left untranslated by the system trained with cross-entropy selection, but correctly translated with infrequent n-gram selection. In the second example, *commercialise* is left untranslated by the first and correctly translated by the latter. Here, the translation does also affect the surrounding words, so that the verb *get* is translated to *aller*, which was simply dropped with the TM cross-entropy method.

source	[...] let's say that you add some software to an image and you want to **re-sell** it, [...]
reference	[...] disons qu'on ajoute un logiciel à une image que l'on veut **revendre**, [...]
TM cross-entropy	[...] disons que vous ajoutez des logiciels à une image et vous voulez **re-sell**, [...]
infrequent n-gram	[...] disons que vous ajoutez des logiciels à une image et vous voulez le **revendre**, [...]
source	[...] here's a great technology, we can **commercialise** it quickly and ***get*** to an exit.
reference	[...] voilà une technologie intéressante, on peut la **commercialiser** rapidement et ***parvenir*** à une sortie.
TM cross-entropy	[...] voici une grande technologie , nous pouvons **commercialise** rapidement et à une sortie.
infrequent n-gram	[...] voici une grande technologie , on peut **commercialiser** rapidement et ***aller*** vers une sortie.

Figure 3: Example sentences from the VideoLectures.NET test set. 1M source words were selected by both the TM cross-entropy and the infrequent n-gram methods.

6 Conclusion

For the task of translating online scientific video lectures efficient and compact systems are essential, as they may need to be applied in real-time or on mobile devices. Selecting only the most relevant parts of the training data reduces both model size and time and memory requirements and in previous work has also improved translation quality. Therefore, we compared several data selection techniques based on cross-entropy and infrequent n-gram recovery criteria for the translation of English-French video lectures.

As infrequent n-gram recovery uses knowledge of the test set, we also experimented with cross-entropy selection based on the test corpus for a fair comparison. However, in terms of BLEU this method did not prove to be competitive with the standard cross-entropy based approaches. Among the cross-entropy based methods, TM cross-entropy yielded the most stable results, reaching the same performance as using the entire data by selecting a quarter of it. However, it has limited capabilities of adding new words to the vocabulary. With respect to the number of unknown words, infrequent n-gram recovery clearly outperforms the cross-entropy based methods, which can be expected given its design. We illustrated the importance of recovering out-of-vocabulary words for the domain of video lectures on two example sentences. Finally, by combining the two approaches, we achieve the best results both in terms of BLEU and OOV rate.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 287755 (transLectures), and the Spanish MINECO Active2Trans (TIN2012-31723) research project.

References

Axelrod, A., He, X., and Gao, J. (2011). Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Bisazza, A., Ruiz, N., and Federico, M. (2011). Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation. In *International Workshop on Spoken Language Translation (IWSLT)*, pages 136–143, San Francisco, CA.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R. L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Civera, J. and Juan, A. (2007). Domain adaptation in statistical machine translation with mixture modelling. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Foster, G. and Kuhn, R. (2007). Mixture-model adaptation for smt. In *Second Workshop on Statistical Machine Translation*, pages 128–135, Prague, Czech Republic.

Galley, M. and Manning, C. D. (2008). A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 848–856, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gao, J. and Zhang, M. (2002). Improving language model size reduction using better pruning criteria. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 176–182, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gascó, G., Rocha, M.-A., and Sanchis-Trilles, G. (2012). Does more data always yield better translations? In *13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 152–1611, Avigvnon, France.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical Phrase-Based Translation. In *Proceedings of the 2003 Meeting of the North American chapter of the Association for Computational Linguistics (NAACL-03)*, pages 127–133, Edmonton, Alberta.

Koehn, P. and Schroeder, J. (2007). Experiments in domain adaptation for statistical machine translation. In *Second Workshop on Statistical Machine Translation*, pages 224–227, Prague, Czech Republic.

Lambert, P., Schwenk, H., Servan, C., and Sadaf, A.-R. (2011). Investigations on translation model adaptation using monolingual data. In *6th Workshop on Statistical Machine Translation*, pages 284–293, Edinburgh, UK.

Lambert, P., Senellart, J., Romary, L., Schwenk, H., Zipser, F., Lopez, P., and Blain, F. (2012). Collaborative machine translation service for scientific texts. In *13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 11–15, Avignon, France.

Lü, Y., Huang, J., and Liu, Q. (2007). Improving statistical machine translation performance by training data selection and optimization. In *45th Annual Meeting of the Association of Computational Linguistics*, pages 343–350, Prague, Czech Republic.

Mansour, S. and Ney, H. (2012). A simple and effective weighted phrase extraction for machine translation adaptation. In *International Workshop on Spoken Language Translation*, pages 193–200, Hong Kong.

Mansour, S., Wuebker, J., and Ney, H. (2011). Combining translation and language model scoring for domain-specific data filtering. In *International Workshop on Spoken Language Translation*, pages 222–229, San Francisco, California, USA.

Matsoukas, S., Rosti, A.-V. I., and Zhang, B. (2009). Discriminative corpus weight estimation for machine translation. In *Conference on Empirical Methods in Natural Language Processing*, pages 708–717, Singapore.

Moore, R. C. and Lewis, W. (2010). Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. Association for Computational Linguistics.

Niehues, J. and Waibel, A. (2012). Detailed analysis of different strategies for phrase table adaptation in SMT. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA)*.

Och, F. J. (2003). Minimum Error Rate Training for Statistical Machine Translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan.

Och, F. J. and Ney, H. (2004). The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). Bleu: a Method for Automatic Evaluation of Machine Translation. IBM Research Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598.

Schwenk, H. (2008). Investigations on large-scale lightly-supervised training for statistical machine translation. In *International Workshop on Spoken Language Translation*, pages 182–189, Hawaii, USA.

Schwenk, H. and Koehn, P. (2008). Large and diverse language models for statistical machine translation. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP)*.

Sennrich, R., Schwenk, H., and Aransa, W. (2013). A multi-domain translation model framework for statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 832–840, Sofia, Bulgaria.

Shah, K., Barrault, L., and Schwenk, H. (2010). Translation model adaptation by resampling. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 392–399, Stroudsburg, PA, USA. Association for Computational Linguistics.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, A., Erjavec, T., and Tufiş, D. (2006). The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 2142–2147.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *8th International Conference on Language Resources and Evaluation (LREC 2012)*.

Ueffing, N. (2006). Using monolingual source-language data to improve MT performance. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pages 174–181.

Ueffing, N., Haffari, G., and Sarkar, A. (2007). Transductive learning for statistical machine translation. In *45th Annual Meeting of the Association of Computational Linguistics*, pages 25–32, Prague, Czech Republic.

Wuebker, J., Huck, M., Peitz, S., Nuhn, M., Freitag, M., Peter, J.-T., Mansour, S., and Ney, H. (2012). Jane 2: Open source phrase-based and hierarchical statistical machine translation. In *International Conference on Computational Linguistics*, pages 483–491, Mumbai, India.

Wuebker, J., Peitz, S., Rietig, F., and Ney, H. (2013). Improving statistical machine translation with word class models. In *Conference on Empirical Methods in Natural Language Processing*, pages 1377–1381, Seattle, USA.