The NICT Translation System for IWSLT 2012

Andrew Finch† *Ohnmar Htun*‡ *Eiichiro Sumita*†

† Multilingual Translation Group
MASTAR Project
National Institute of Information and
Communications Technology
Kyoto, Japan
andrew.finch,eiichiro.sumita@nict.go.jp

‡ Dept. of Management and
Information System Science
Nagaoka University of Technology
Nagaoka, Japan
s097001@stn.nagaokaut.ac.jp

Abstract

This paper describes NICT's participation in the IWSLT 2012 evaluation campaign for the TED speech translation Russian-English shared-task. Our approach was based on a phrase-based statistical machine translation system that was augmented by using transliteration mining techniques.

The basic premise behind our approach was to try to use sub-word-level alignments to guide the word-level alignment process used to learn the phrase-table. We did this by first mining a corpus of Russian-English transliterations pairs and cognates from a set of interlanguage link titles from Wikipedia. This corpus was then used to build a many-to-many nonparametric Bayesian bilingual alignment model that could be used to identify the occurrence of transliterations and cognates in the training corpus itself. Alignment counts for these mined pairs were increased in the training corpus to increase the likelihood that these pairs would align in training. Our experiments on the test sets from the 2010 and 2011 shared tasks, showed that an improvement in BLEU score can be gained in translation performance by encouraging the alignment of cognates and transliterations during word alignment.

1. Introduction

In the IWSLT 2012 evaluation campaign [1], the NICT team participated in TED [2] speech translation shared-task for Russian-English. This paper describes the machine translation approach adopted for this campaign.

Our overall approach was to take a phrase-based statistical machine translation decoder and increase its performance by improving the word alignment. Typically only word co-occurrence statistics are used in determining the word-to-word alignments during training, however certain classes of words can offer additional features that can be used to assist in the prediction of their alignment: these words are transliterations and cognates. Transliterations are words that have been borrowed from another language; loan words imported into the language while preserving their phonetics as far as possible. So for example, the Italian name 'Donatello' would be transcribed into the Cyrillic alphabet as 'Донателло' (DONATELLO). The upper case form in parentheses is a romanized form of the preceding Russian character sequence, which in this case is exactly the same as the original English word, but in general this is not necessarily the case.

Cognates are words that share a common etymological origin, for example the word 'milk' in English is a cognate of the German word 'milch' and the Russian word 'молоко' (MOLOKO). Transliterations are derived directly from the word in the language from which they are being borrowed, and cognates are both derived from their common root. Our hypothesis is that these relationships can be modeled and thereby detected in bilingual data. Our approach is to model both cases using a generative model, under the assumption that there exists some generative process that can reliably assign a higher generation probability to cognates and transliterations than a model designed to explain random pairs of words. Furthermore, we assume that if two words are assigned a relatively high probability from such a model, then they are likely to be aligned in the data. This assumption is not true in general due to the existence of false cognates; words may appear to be cognates, when in fact there is no genetic relationship between them. Nonetheless, we anticipate that pathological occurrences of this kind will be rare, and that relying on the assumptions mentioned earlier will result an overall benefit.

Due to an unfortunate error in the processing of the phrase-tables of our systems for the final submission to the shared task, the official scores for our system are several BLEU points below what could be expected of the system had there been no error, we therefore do not report the official results for our system on the 2012 test data, but instead rely on experiments based on systems trained on the 2012 training set, and tested on the 2010 and 2011 test sets.

The overall layout of our paper is as follows. In the next section we describe the underlying phrase-based statistical

machine translation system that forms the basis of all of the systems reported in this paper. In the following section we describe the techniques we used to incorporate information from sub-word alignments into the word alignment process. Then we present our experiments comparing our system to a baseline system. Finally we conclude and offer some directions for future research.

2. The Base System

2.1. Decoder

The decoder used in these experiments is an in-house phrase-based statistical machine translation decoder OCTAVIAN than can operate in a similar manner to the publicly available MOSES decoder [3]. The base decoder used a standard set of features that were integrated into a log-linear model using independent exponential weights for each feature. These features consisted of: a language mode; five translation model features; a word penalty; and a lexicalized re-ordering model with monotone, discontinuous, swap features for the current and previous phrase-pairs.

Based on a set of pilot experiments we decoded with a maximum distance of 5 on the distances phrases could be moved in the re-ordering process during decoding.

2.2. Pre-processing

The English data was tokenized by applying a number of regular expressions to separate punctuation, and split contractions such as "it's" and "hasn't" into two separate tokens. We also removed all case information from the English text to help to minimize issues of data sparseness in the models of the translation system. All punctuation was left in both source and target. We took the decision to generate target punctuation directly using the process of translation, rather than as a punctuation restoration step in post processing based on experiments carried out for the 2010 IWSLT shared evaluation [4].

2.3. Post-processing

The output of the translation system was subject to the following post-processing steps which were carried out in the order in which that are listed.

1. Out of vocabulary words (OOVs) were passed through the translation process unchanged, some of these OOVs were Russian and some English. We took the decision to delete only those OOVs containing cyrillic characters not included in the ASCII character set and leave words containing only ASCII characters in the output.

2. The output was de-tokenized using a set of heuristics implemented as regular expressions designed to undo the process of English tokenization. Punctuation was attached to neighboring words and tokens that form split contractions were combined into a single token.

3. The output was re-cased using the re-casing tool supplied with the MOSES [3] toolkit. We trained the re-casing tool on untokenized text from the TED talk training data.

2.4. Training

2.4.1. Data

We trained out translation and language models using only the in-domain TED data supplied for the task. This data consisted of approximately 120k bilingual sentence pairs containing about 2.4 million words of English, and 2 million words of Russian. In addition to this data, we used approximately 600,000 bilingual article title pairs extracted from the interlanguage links of the most recent dump of the Russian Wikipedia database. In the remainder of this section we describe the details of the process of building the machine translation engine used in our experiments. A description of the training and application of the transliteration mining component of our system follows in the next section.

2.4.2. Language Model

The language models were built using the SRI language modeling toolkit [5]. A 5-gram model was built for decoding the development and test data for evaluation, and a 3-gram model was built on the same data for efficient tuning. Pilot experiments indicated that using a lower order language model for tuning did not significantly affect the translation quality of the systems produced by the MERT process. The language models were smoothed using modified Knesser-Ney smoothing.

2.4.3. Translation Model

The translation model for the base system was built in the standard manner using a 2-step process. First the training data was word-aligned using GIZA++. Second, the grow-diag-final-and phrase-extraction heuristics from the MOSES [3, 6] machine translation toolkit were used to extract a set of bilingual phrase-pairs using the alignment produced by GIZA++. However before training the proposed system, mined single-word transliteration/cognate pairs were added to the training data set. In doing this, these word pairs are guaranteed to align, increasing their alignment counts thereby encouraging their alignment where they occur together in the remainder of the corpus. Pilot experiments were run on development data to assess the effect of adding these transliteration/cognate pairs multiple times to the data. We found that adding the pairs a single time was the most effective strategy.

2.4.4. Parameter Tuning

To tune the values for the log-linear weights in our system, we used the standard minimum error-rate training procedure

(MERT) [7]. The weights for the models were tuned using the development data supplied for the task.

3. Using Sub-word Alignment

3.1. Motivation

The use of transliterations to aid the alignment process was first proposed by [8], and has been shown to improve word alignment quality in [9]. The idea is based on the simple principle that for transliterations and cognates there exist similarities at the substring level due to the relationships these words possess, these relationships can be discovered by bilingual alignment at the grapheme level, and may be used as additional alignment evidence during a word alignment process. However this promising idea has received little attention in the literature. Our system is based on a two step process: first a bilingual alignment model is built from noisy data using a transliteration mining process; in the second step the training corpus itself is mined for transliterations/cognates using the model built from the first step. We describe these two steps in more detail in the next two subsections.

3.2. Transliteration Mining

3.2.1. Corpus

To train the mining system we extracted 629,021 bilingual Russian-English interlanguage link title pairs from the most recent (July 2012) Wikipedia database dump. From this data we selected only the single word pairs for training, leaving a corpus of 145,817 noisy word pairs. We expected (based on our experience building transliteration generation models on these languages) that the amount of clean data in this corpus would be sufficient for training the transliteration component of our generative model since the grapheme vocabulary sizes for both languages are not large, and the alignments are often reasonably direct (as can be seen in the set of examples given below). 98,902 pairs were automatically extracted from this corpus as transliteration/cognate pairs.

3.2.2. Methodology

The mining model we used was based on the research of [10] which in turn draws on the work of [11] and [12].

The mining system is capable of simultaneously modeling and clustering the data. It does this by means of a single generative model that is composed of two sub-models: the first models the transliterations/cognates; the second models the noise. The generative story for this model is as follows:

1. Choose whether to generate noise (with probability λ), or a transliteration/cognate pair (probability $1 - \lambda$);
2. Generate the noise pair, or the transliteration pair with the respective sub-model.

The noise and transliteration/cognate sub-models are both unigram joint source-channel models [13]: the joint probability of generating a bilingual word pair is given by the product of the probabilities of a sequence steps each involving the generation of a bilingual grapheme sequence pair. The difference between these models being the types of grapheme sequence pair they are allowed to generate.

As in [10], we have extended the nonparametric Bayesian alignment model of [12] to include null alignments to either single characters or sequences of graphemes up to a maximum specified length. The alignment model is symmetrical with respect to the source and target languages and therefore these null alignments can be to either source or target grapheme sequences, and their probabilities are learned during training in the same manner as the other parameters in the model.

The difference between the noise and transliteration/cognate sub-models was that the noise sub-model was restricted to generate using only null alignments. In other words, the noise sub-model generates the source and target sequences independently. Constraining the noise model in this way allows it to distribute more of its probability mass onto those model parameters that are useful for explaining data where there is no relationship between source and target. The transliteration/cognate sub-model on the other hand is able to learn the many-to-many grapheme substitution operations useful in modeling pairs that can be generated by bilingual grapheme sequence substitution. During the sampling process, both models compete to explain the word pairs in the corpus, thereby naturally clustering them into two sets while learning.

Our Bayesian alignment model is able to perform many-to-many alignment without the overfitting problems commonly encountered when using maximum likelihood training. In the experiments reported here, we arbitrarily limit the maximum source and target sequence lengths to 3 graphemes on each side. This was done to speed up the training process, but was not strictly necessary.

The aligner was trained using block Gibbs sampling using the efficient forward-filter backward-sample dynamic programming approach set out in [14]. The initial alignments were chosen randomly using an initial backward sampling pass with a uniform distribution on the arcs in the alignment graph. The prior probability of the pairs being noise (λ) was set to 0.5 in the first iteration. During the training λ was updated whenever the class (transliteration/cognate or noise) of a bilingual word pair was changed in the sampling process. λ was calculated based on a simple frequency count of the classes assigned to all the word pairs while sampling.

3.3. Mining the Training Set

In order to discover alignments of transliteration/cognate pairs in the training data we again applied a mining approach. We aligned each Russian word to each English word in the same sentence of the training corpus, and then used the approach of [15] to determine whether these pairs were transliterations/cognates. In principle it would be possible to apply

the approach described in the previous section here, however, we chose not to attempt this due to the considerably larger amount of noise in this data, and also because of the size of this corpus. For full details of this method the reader is referred to [15], but in brief the technique mines data by first aligning it using an alignment model similar to the transliteration sub-model described in the previous section. Then features extracted from the alignment are combined with features derived from the characteristics of the word pairs (for example their relative lengths); these features are then used to classify the data. The advantages of this approach over the method described in the previous section are firstly that it utilizes a model already trained on relatively clean data, and so will not be affected by the noise in the corpus being mined, and secondly no iterative learning is required; the process is effectively the same as the backward sampling step and can proceed very rapidly given an already trained model. The mining process yielded a sequence of word pairs that the system considered to be likely candidates for transliterations/cognates. This sequence of pairs was added to the training data used to build the translation model, in doing so these word pairs were forced to align to each other and the counts for their alignments were increased thereby encouraging their alignments in the remainder of the corpus. We ran pilot experiments to determine the effect of increasing the counts further by adding the mined pairs multiple times to the corpus, and although the performance seemed reasonably insensitive to the number of copies of the data we used, the experiments with a single copy of the data gave the highest scores. In future research we would seek to either soften this parameter and then optimize it on the data set (in a similar manner to [11]), or ideally remove it altogether by integrating the mining and alignment processes.

3.4. Examples

Some typical examples of mined transliteration/cognate pairs are given in Table 3.4. Notice that in many of the examples (for example Соционика/Socionics) most of the mapping is possible with simple grapheme-to-grapheme substitutions. In this example, a transformation of the word ending (ика↦ics) is also required. This transformation is quite common in the corpus and the aligner learned this as a model parameter. Furthermore, the grapheme sequence pair was used as a single step in aligning both this word pair and others with analogous endings in the corpus. The mining process was able to learn to be robust to small variations in the data. For example in the pair Посткапитализм/Post-capitalism a hyphen is present on the English side, but not on the Russian side. The aligner learned to delete hyphens in the data by aligning them to null, thereby learning to model its asymmetrical usage in the data.

Russian	**Engish**
Космополитизм (KOSMOPOLITIZM)	Cosmopolitanism
Посткапитализм (POSTKAPITALIZM)	Post-capitalism
Соционика (SOCIONIKA)	Socionics
Физика (FIZIKA)	Physics
Механика MEHANIKA	Mechanics
Парапсихология (PARAPSIHOLOGIJA)	Parapsychology
Хронология (HRONOLOGIJA)	Chronology
Спагетти (SPAGETTI)	Spaghetti
Париж (PARIZH)	Paris

Table 1: Examples of transliteration/cognate pairs discovered by mining Wikipedia interlanguage link titles.

3.5. Experiments

We evaluated the effectiveness of our approach using the the supplied training, development and IWSLT2010 and IWSLT2011 test data sets. The baseline model was trained identically, but without using the mined data. The results are shown in Figure 3.5. Our results show a modest but consistent improvement in translation performance on both test sets, motivating further development of this approach. We analyzed the results to investigate the impact of the approach on the number of OOVs in the test data. Surprisingly on both IWSLT2010 and IWSLT2011 test sets our approach gave rise to a 0.2% increase in number of OOVs. This may indicate our approach is succeeding by improving the overall word alignment, rather than by improving the translation of words with cognates and transliterations in the target language.

Model	**IWSLT2010**	**IWSLT2011**
Baseline	16.23	18.08
Proposed	16.77	18.53

Table 2: The effect on BLEU score of using sub-word alignments to assist word alignment.

4. Conclusions

This paper described NICT's system for the IWSLT 2012 evaluation campaign for the TED speech translation Russian-English shared-task. Our approach was based on a fairly typical phrase-based statistical machine translation system that was augmented using a transliteration mining approach designed to exploit the alignments between transliterations and

cognates to improve the word alignment. Our experimental results on the IWSLT2010 and IWSLT2011 test sets gave improvements of approximately 0.5 BLEU percentage points.

In future work we would like to explore integrate the transliteration/cognate mining techniques more tightly into the word alignment process. We believe it should be possible to simultaneously word align while mining the corpus for sub-word alignments, within a single nonparametric Bayesian alignment process.

5. References

[1] M. Federico, M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker, "Overview of the IWSLT 2012 evaluation campaign," in *Proc. of the International Workshop on Spoken Language Translation*, Hong Kong, HK, December 2012.

[2] M. Cettolo, C. Girardi, and M. Federico, "Wit^3: Web inventory of transcribed and translated talks," in *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, May 2012, pp. 261–268.

[3] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowa, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: open source toolkit for statistical machine translation," in *ACL 2007: proceedings of demo and poster sessions*, Prague, Czeck Republic, June 2007, pp. 177–180.

[4] C.-L. Goh, T. Watanabe, M. Paul, A. Finch, and E. Sumita, "The NICT Translation System for IWSLT 2010," in *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, M. Federico, I. Lane, M. Paul, and F. Yvon, Eds., 2010, pp. 139–146.

[5] A. Stolcke, "Srilm - an extensible language model toolkit," 1999. [Online]. Available: http://www.speech.sri.com/projects/srilm

[6] P. Koehn, "Pharaoh: a beam search decoder for phrase-based statistical machine translation models," in *Machine translation: from real users to research: 6th conference of AMTA*, Washington, DC, 2004, pp. 115–124.

[7] F. J. Och, "Minimum error rate training for statistical machine translation," in *Proceedings of the ACL*, 2003.

[8] U. Hermjakob, "Improved word alignment with statistics and linguistic heuristics," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, August 2009, pp. 229–237. [Online]. Available: http://www.aclweb.org/anthology/D/D09/D09-1024

[9] H. Sajjad, A. Fraser, and H. Schmid, "An algorithm for unsupervised transliteration mining with an application to word alignment," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, ser. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 430–439. [Online]. Available: http://dl.acm.org/citation.cfm?id=2002472.2002527

[10] O. Htun, A. Finch, E. Sumita, and Y. Mikami, "Improving transliteration mining by integrating expert knowledge with statistical approaches," *International Journal of Computer Applications*, vol. 58, November 2012.

[11] H. Sajjad, A. Fraser, and H. Schmid, "A statistical model for unsupervised and semi-supervised transliteration mining," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Jeju Island, Korea: Association for Computational Linguistics, July 2012, pp. 469–477. [Online]. Available: http://www.aclweb.org/anthology/P12-1049

[12] A. Finch and E. Sumita, "A Bayesian Model of Bilingual Segmentation for Transliteration," in *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, M. Federico, I. Lane, M. Paul, and F. Yvon, Eds., 2010, pp. 259–266.

[13] H. Li, M. Zhang, and J. Su, "A joint source-channel model for machine transliteration," in *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 2004, p. 159.

[14] D. Mochihashi, T. Yamada, and N. Ueda, "Bayesian unsupervised word segmentation with nested pitman-yor language modeling," in *ACL-IJCNLP '09: 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: Volume 1*. Morristown, NJ, USA: Association for Computational Linguistics, 2009, pp. 100–108.

[15] T. Fukunishi, A. Finch, S. Yamamoto, and E. Sumita, "Using features from a bilingual alignment model in transliteration mining," in *Proceedings of the 3rd Named Entities Workshop (NEWS 2011)*, 2011, pp. 49–57.