NICT@WMT09: Model Adaptation and Transliteration for Spanish-English SMT

Michael Paul, Andrew Finch and Eiichiro Sumita
Language Translation Group
MASTAR Project
National Institute of Information and Communications Technology
Michael.Paul@nict.go.jp

Abstract

This paper describes the NICT statistical machine translation (SMT) system used for the WMT 2009 Shared Task (WMT09) evaluation. We participated in the *Spanish-English* translation task. The focus of this year's participation was to investigate model adaptation and transliteration techniques in order to improve the translation quality of the baseline phrase-based SMT system.

1 Introduction

This paper describes the NICT statistical machine translation (SMT) system used for the shared task of the Fourth Workshop on Statistical Machine Translation. We participated in the *Spanish-English* translation task under the *Constrained Condition*. For the training of the SMT engines, we used two parallel Spanish-English corpora provided by the organizers: the *Europarl* (*EP*) corpus (Koehn, 2005), which consists of 1.4M parallel sentences extracted from the proceedings of the European Parliament, and the *News Commentary* (*NC*) corpus (Callison-Burch et al., 2008), which consists of 74K parallel sentences taken from major news outlets like *BBC*, *Der Spiegel*, and *Le Monde*.

In order to adapt SMT systems to a specific domain, recent research focuses on model adaptation techniques that adjust their parameters based on information about the evaluation domain (Foster and Kuhn, 2007; Finch and Sumita, 2008a). Statistical models can be trained on *in-domain* and *out-of-domain* data sets and combined at run-time using probabilistic weighting between domain-specific statistical models. As the official WMT09 evaluation testset consists of documents taken from the news domain, we applied statistical model adaptation techniques to combine *translation models* (tm), *language models* (lm) and *distortion models* (dm) trained on (a) the in-domain *NC* corpus and (b) the out-of-domain *EP* corpus (cf. Section 2).

One major problem in the given translation task was the large amount of *out-of-vocabulary* (OOV) words, i.e., source language words that do not occur in the training corpus. For unknown words, no translation entry is available in the statistical translation model (*phrase-table*). As a result, these OOV words cannot be translated. Dealing with languages with a rich morphology like *Spanish* and having a limited amount of bilingual resources make this problem even more severe.

There have been several efforts in dealing with OOV words to improve translation quality. In addition to parallel text corpora, external bilingual dictionaries can be exploited to reduce the OOV problem (Okuma et al., 2007). However, these approaches depend on the coverage of the utilized external dictionaries.

Data sparseness problems due to inflectional variations were previously addressed by applying word transformations using stemming or lemmatization (Popovic and Ney, 2005; Gupta and Federico, 2006). A tight integration of morpho-syntactic information into the translation model was proposed by (Koehn and Hoang, 2007) where lemma and morphological information are translated separately, and this information is combined on the output side to generate the translation. However, these approaches still suffer from the data sparseness problem, since lemmata and inflectional forms never seen in the training corpus cannot be translated.

In order to generate translations for unknown words, previous approaches focused on *transliteration* methods, where a sequence of characters is mapped from one writing system into another. For example, in order to translate names and technical terms, (Knight and Graehl, 1997) introduced a probabilistic model that replaces Japanese

Proceedings of the 4th EACL Workshop on Statistical Machine Translation , pages 105–109,
Athens, Greece, 30 March – 31 March 2009.

katakana[1] words with phonetically equivalent English words. More recently, (Finch and Sumita, 2008b) proposed a transliteration method that is based directly on techniques developed for phrase-based SMT, and transforms a character sequence from one language into another in a subword-level, character-based manner. We extend this approach by exploiting the phrase-table of the baseline SMT system to train a phrase-based transliteration model that generates English translations of Spanish OOV words as described in Section 3. The effects of the proposed techniques are investigated in detail in Section 4.

2 Model Adaptation

Phrase-based statistical machine translation engines use multiple statistical models to generate a translation hypothesis in which (1) the *translation model* ensures that the source phrases and the selected target phrases are appropriate translations of each other, (2) the *language model* ensures that the target language is fluent, (3) the *distortion model* controls the reordering of the input sentence, and (4) the *word penalty* ensures that the translations do not become too long or too short. During decoding, all model scores are weighted and combined to find the most likely translation hypothesis for a given input sentence (Koehn et al., 2007).

In order to adapt SMT systems to a specific domain, separate statistical models can be trained on parallel text corpora taken from the respective domain (*in-domain*) and additional *out-of-domain* language resources. The models are then combined using mixture modeling (Hastie et al., 2001), i.e., each model is weighted according to its fit with in-domain development data sets and the linear combination of the respective scores is used to find the best translation hypothesis during the decoding of unseen input sentences.

In this paper, the above model adaptation technique is applied to combine the *NC* and the *EP* language resources provided by the organizers for the *Spanish-English* translation task. As the WMT09 evaluation testset consists of documents taken from the news domain, we used the *NC* corpus to train the in-domain models and the *EP* corpus to train the out-of-domain component models. Using mixture modeling, the above mentioned statistical models are combined where each component model is optimized separately. Weight optimization is carried out using a simple grid-search method. At each point on the grid of weight parameter values, the translation quality of the combined weighted component models is evaluated for development data sets taken from (a) the *NC* corpus and (b) from the *EP* corpus.

[1] A special syllabary alphabet used to write down foreign names or loan words.

3 Transliteration

Source language input words that cannot be translated by the standard phrase-based SMT models are either left untranslated or simply removed from the translation output. Common examples are named entities such as *personal names* or *technical terms*, but also include content words like *common nouns* or *verbs* that are not covered by the training data. Such unknown occurrences could benefit from being transliterated into the MT system's output during translation of orthographically related languages like Spanish and English.

In this paper, we apply a phrase-based transliteration approach similar to the one proposed in (Finch and Sumita, 2008b). The transliteration method is based directly on techniques developed for phrase-based SMT and treats the task of transforming a character sequence from one language into another as a character-level translation process. In contrast to (Finch and Sumita, 2008b) where external dictionaries and inter-language links in Wikipedia[2] are utilized, the transliteration training examples used for the experiments in Section 4 are extracted directly from the phrase-table of the baseline SMT systems trained on the provided data sets. For each phrase-table entry, corresponding word pairs are identified according to a string similarity measure based on the *edit-distance* (Wagner, 1974) that is defined as the sum of the costs of *insertion*, *deletion*, and *substitution* operations required to map one character sequence into the other and can be calculated by a *dynamic programming* technique (Cormen et al., 1989). In order to reduce noise in the training data, only word pairs whose word length and similarity are above a pre-defined threshold are utilized for the training of the transliteration model.

The obtained transliteration model is applied as a post-process filter to the SMT decoding process, i.e.. all source language words that could not be translated using the SMT engine are replaced with the corresponding transliterated word forms in order to obtain the final translation output.

[2] http://www.wikipedia.org

4 Experiments

The effects of *model adaptation* and *transliteration* techniques were evaluated using the Spanish-English language resources summarized in Table 1. In addition, the characteristics of this year's testset are given in Table 2. The sentence length is given as the average number of words per sentence. The OOV word figures give the percentage of words in the evaluation data set that do not appear in the *NC*/*EP* training data. In order to get an idea how difficult the translation task may be, we also calculated the language perplexity of the respective evaluation data sets according to 5-gram target language models trained on the *NC*/*EP* data sets.

Concerning the development sets, the *news-dev2009* data taken from the same news sources as the evaluation set of the shared task was used for the tuning of the SMT engines, and the *devtest2006* data taken from the *EP* corpus was used for system parameter optimization. For the evaluation of the proposed methods, we used the testsets of the Second Workshop on SMT (*nc-test2007* for *NC* and *test2007* for *EP*). All data sets were case-sensitive with punctuation marks tokenized.

The numbers in Table 1 indicate that the characteristics of this year's testset differ largely from testsets of previous evaluation campaigns. The *NC* devset (2,438/1,378 OOVs) contains twice as many untranslatable Spanish words as the *NC* evalset (1,168/73 OOVs) and the *EP* devset (912/63 OOVs). In addition, the high language perplexity figures for this year's testset show that the translation quality output for both baseline systems is expected to be much lower than those for the *EP* evaluation data sets. In this paper, translation quality is evaluated according to (1) the *BLEU* metrics which calculates the *geometric mean of n-gram precision* by the system output with respect to reference translations (Papineni et al., 2002), and (2) the *METEOR* metrics that calculates unigram overlaps between translations (Banerjee and Lavie, 2005). Scores of both metrics range between 0 (worst) and 1 (best) and are displayed in percent figures.

4.1 Baseline

Our baseline system is a fairly typical phrase-based machine translation system (Finch and Sumita, 2008a) built within the framework of a feature-based exponential model containing the following features:

Table 1: Language Resources

Corpus			Train	Dev	Eval
NC	Spanish	sentences	74K	2,001	2,007
		words	2,048K	49,116	56,081
		vocab	61K	9,047	8,638
		length	27.6	24.5	27.9
		OOV (%)	–	5.2 / 2.9	1.4 / 0.9
	English	sentences	74K	2,001	2,007
		words	1,795K	46,524	49,693
		vocab	47K	8,110	7,541
		length	24.2	23.2	24.8
		OOV (%)	–	5.2 / 2.9	1.2 / 0.9
		perplexity	–	349 / 381	348 / 458
EP	Spanish	sentences	1,404K	1,861	2,000
		words	41,003K	50,216	61,293
		vocab	170K	7,422	8,251
		length	29.2	27.0	30.6
		OOV (%)	–	2.4 / 0.1	2.4 / 0.2
	English	sentences	1,404K	1,861	2,000
		words	39,354K	48,663	59,145
		vocab	121K	5,869	6,428
		length	28.0	26.1	29.6
		OOV (%)	–	1.8 / 0.1	1.9 / 0.1
		perplexity	–	210 / 72	305 / 125

Table 2: Testset 2009

Corpus			Test
NC	Spanish	sentences	3,027
		words	80,591
		vocab	12,616
		length	26.6

- Source-target phrase translation probability
- Inverse phrase translation probability
- Source-target lexical weighting probability
- Inverse lexical weighting probability
- Phrase penalty
- Language model probability
- Lexical reordering probability
- Simple distance-based distortion model
- Word penalty

For the training of the statistical models, standard word alignment (GIZA++ (Och and Ney, 2003)) and language modeling (SRILM (Stolcke, 2002)) tools were used. We used 5-gram language models trained with modified Knesser-Ney smoothing. The language models were trained on the target side of the provided training corpora. Minimum error rate training (MERT) with respect to BLEU score was used to tune the decoder's parameters, and performed using the technique proposed in (Och, 2003). For the translation, the in-house multi-stack phrase-based decoder **CleopATRa** was used.

The automatic evaluation scores of the baseline systems trained on (a) only the *NC* corpus and (b) only on the *EP* corpus are summarized in Table 3.

Table 3: Baseline Performance

	NC Eval		*EP* Eval	
	BLEU	METEOR	BLEU	METEOR
baseline	17.56	40.52	33.00	56.50

4.2 Effects of Model Adaptation

In order to investigate the effect of model adaptation, each model component was optimized separately using the method described in Section 2. Table 4 summarizes the automatic evaluation results for various model combinations. The combination of *NC* and *EP* models using equal weights achieves only a slight improvement for the *NC* task (BLEU: +0.4%, METEOR: +0.4%), but a large improvement for the *EP* task (BLEU: +1.0%, METEOR: +1.7%). Weight optimization further improves all translation tasks where the highest evaluation scores are achieved when the optimized weights for all statistical models are used. In total, model adaptation gains 1.1% and 1.3% in BLEU and 0.8% and 1.8% in METEOR for the *NC* and *EP* translation tasks, respectively.

Table 4: Effects of Model Adaptation

weight optimization	*NC* Eval		*EP* Eval	
	BLEU	METEOR	BLEU	METEOR
–	17.92	40.72	34.00	58.20
tm	18.13	40.95	34.05	58.23
tm+lm	18.25	41.23	34.12	58.22
tm+dm	18.36	41.06	34.24	58.34
tm+lm+dm	**18.65**	**41.35**	**34.35**	**58.36**

4.3 Effects of Transliteration

In order to investigate the effects of transliteration, we trained three different transliteration using the phrase-table of the baseline systems trained on (a) only the *NC* corpus, (b) only the *EP* corpus, and (c) on the merged corpus (*NC+EP*). The performance of these phrase-based transliteration models is evaluated for 2000 randomly selected transliteration examples. Table 5 summarizes the haracter-based automatic evaluation scores for the *word error rate* (WER) metrics, i.e., the edit distance between the system output and the closest reference translation (Niessen et al., 2000), as well as the BLEU and METEOR metrics. The best performance is achieved when training examples from both domains are exploit to transliterate unknown Spanish words into English. Therefore, the *NC+EP* transliteration model was applied to the translation outputs of all mixture models described in Section 4.2.

The effects of the transliteration post-process are summarized in Table 6. Transliteration consistently improves the translation quality of all mixture models, although the gains obtained for the *NC* task (BLEU: +1.3%, METEOR: +1.3%) are much larger than those for the *EP* task (BLEU: +0.1%, METEOR: +0.2%) which is due to the larger amount of untranslatable words in the *NC* evaluation data set.

Table 5: Transliteration Performance

Training Data	character-based		
	WER	BLEU	METEOR
NC	13.10	83.62	86.74
EP	11.76	85.93	87.89
NC+EP	**11.72**	**86.08**	**87.89**

Table 6: Effects of Transliteration

weight optimization	*NC* Eval		*EP* Eval	
	BLEU	METEOR	BLEU	METEOR
tm	19.14	42.39	34.11	58.46
tm+lm	19.46	42.65	34.16	58.44
tm+dm	19.77	42.35	34.38	58.57
tm+lm+dm	**19.95**	**42.64**	**34.48**	**58.60**

4.4 WMT09 Testset Results

Based on the automatic evaluation results presented in the previous sections, we selected the SMT engine based on the *tm+lm+dm* weights optimized on the *NC* devset as the primary run for our testset run submission. All other model weight combinations were submitted as contrastive runs. The BLEU scores of these runs are listed in Table 7 and confirm the results obtained for the above experiments, i.e., the best performing system is the one based on the mixture models using separately optimized weights in combination with the transliteration of untranslatable Spanish words using the phrase-based transliteration model trained on all available language resources.

Table 7: Testset 2009 Performance

weight optimization	*NC* Eval BLEU	*EP* Eval BLEU
tm	21.07	20.81
tm+lm	20.95	20.59
tm+dm	21.45	21.32
tm+lm+dm	**21.67***	21.27

5 Conclusion

The work for this year's shared task focused on the task of effectively utilizing out-of-domain language resources and handling OOV words to improve translation quality. Overall our experiments show that the incorporation of mixture models and phrase-based transliteration techniques largely out-performed standard phrase-based SMT engines gaining a total of 2.4% in BLEU and 2.1% in METEOR for the news domain.

References

S. Banerjee and A. Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT*, pages 65–72, Ann Arbor, Michigan.

C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder. 2008. Further Meta-Evaluation of Machine Translation. In *Proceedings of the 3rd Workshop on SMT*, pages 70–106, Columbus, Ohio.

H. Cormen, C. Leiserson, and L. Rivest. 1989. *Introduction to Algorithms*. MIT Press.

A. Finch and E. Sumita. 2008a. Dynamic Model Interpolation for Statistical Machine Translation. In *Proceedings of the 3rd Workshop on SMT*, pages 208–215, Columbus, Ohio.

A. Finch and E. Sumita. 2008b. Phrase-based Machine Transliteration. In *Proceedings of the IJCNLP*, pages 13–18, Hyderabad, India.

G. Foster and R. Kuhn. 2007. Mixture-Model Adaptation for SMT. In *Proceedings of the 2nd Workshop on SMT*, pages 128–135, Prague, Czech Republic.

D. Gupta and M. Federico. 2006. Exploiting Word Transformation in SMT from Spanish to English. In *Proceedings of the EAMT*, pages 75–80, Oslo, Norway.

T. Hastie, R. Tibshirani, and J. Friedman. 2001. *The Elements of Statistical Learning*. Springer, New York.

K. Knight and J. Graehl. 1997. Machine Transliteration. In *Proceedings of the 35th ACL*, pages 128–135, Madrid, Spain.

P. Koehn and H. Hoang. 2007. Factored Translation Models. In *Proceedings of the EMNLP-CoNLL*, pages 868–876, Prague, Czech Republic.

P. Koehn, F.J. Och, and D. Marcu. 2007. Statistical Phrase-Based Translation. In *Proceedings of the HLT-NAACL*, pages 127–133, Edmonton, Canada.

P. Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the MT Summit X*, pages 79–86, Phuket, Thailand.

S. Niessen, F.J. Och, G. Leusch, and H. Ney. 2000. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proc. of the 2nd LREC*, pages 39–45, Athens, Greece.

F.J. Och and H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

F.J. Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st ACL*, pages 160–167, Sapporo, Japan.

H. Okuma, H. Yamamoto, and E. Sumita. 2007. Introducing Translation Dictionary into phrase-based SMT. In *Proceedings of MT Summit XI*, pages 361–368, Copenhagen, Denmark.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th ACL*, pages 311–318, Philadelphia, USA.

M. Popovic and H. Ney. 2005. Exploiting Phrasal Lexica and Additional Morpho-synatctic Language Resources for SMT with Scarce Training Data. In *Proceedings of the EAMT*, pages 212–218, Budapest, Hungary.

A. Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of ICSLP*, pages 901–904, Denver.

R.W. Wagner. 1974. The string-to-string correction problem. *Journal of the ACM*, 21(1):169–173.