On the Importance of Pivot Language Selection for Statistical Machine Translation

Michael Paul∗†**, Hirofumi Yamamoto**†‡**, Eiichiro Sumita**† **and Satoshi Nakamura**†
† NICT, Hikaridai 2-2-2, Keihanna Science City, 619-0288 Kyoto, Japan
‡ Kinki University School of Science and Engineering, Higashi-Osaka City, 577-8502, Japan
Michael.Paul@nict.go.jp

Abstract

Recent research on multilingual statistical machine translation focuses on the usage of *pivot languages* in order to overcome resource limitations for certain language pairs. Due to the richness of available language resources, *English* is in general the pivot language of choice. In this paper, we investigate the appropriateness of languages other than English as pivot languages. Experimental results using state-of-the-art statistical machine translation techniques to translate between twelve languages revealed that the translation quality of 61 out of 110 language pairs improved when a non-English pivot language was chosen.

1 Introduction

The translation quality of state-of-the-art, phrase-based statistical machine translation (SMT) approaches heavily depends on the amount of bilingual language resources available to train the statistical models. For frequently used language pairs like *French-English* or *Chinese-English*, large-sized text data sets are readily available. There exist several data collection initiatives like the *Linguistic Data Consortium*[1], the *European Language Resource Association*[2], or the *GSK*[3], amassing and distributing large amounts of textual data. However, for less frequently used language pairs, e.g., most of the Asian languages, only a limited amount of bilingual resources are available, if at all.

In order to overcome such language resource limitations, recent research on multilingual SMT focuses on the usage of *pivot languages*. Instead of a direct translation between two languages where only a limited amount of bilingual resources is available, the *pivot translation* approach makes use of a third language that is more appropriate due to the availability of more bilingual corpora and/or its relatedness towards either the source or the target language. Several pivot translation techniques like *cascading*, *phrase-table combination*, or *pseudo corpus generation* have already been proposed (cf. Section 2).

However, for most recent research efforts, *English* is the pivot language of choice due to the richness of available language resources. For example, the Europarl corpus is exploited in (Utiyama and Isahara, 2007) for comparing pivot translation approaches between *French*, *German* and *Spanish* via *English*. Other research efforts tried to exploit the closeness between specific language pairs to generate high-quality translation hypotheses in the first step to minimize the pivot detoriation effects, e.g., for *Catalan*-to-*English* translations via *Spanish* (Gispert and Marino, 2006).

This paper investigates the appropriateness of languages other than English as pivot languages to support future research on machine translation between under-resourced language pairs. Pivot translation experiments using state-of-the-art SMT techniques are carried out to translate between twelve of the major world languages covering Indo-European as well as Asian languages and the effects of selecting a non-*English* language as the pivot language are discussed in Section 3.

2 Pivot Translation

Pivot translation is a translation from a source language (SRC) to a target language (TRG) through an intermediate *pivot* (or *bridging*) *language* (PVT). Within the SMT framework, the following coupling strategies have already been investigated:

1. *cascading of two translation systems* where the first MT engine translates the source language input into the pivot language and the second MT engine takes the obtained pivot language output as its input and translates it into the target language.
2. *pseudo corpus* approach that (i) creates a "noisy" SRC-TRG parallel corpus by translating the pivot language parts of the SRC-PVT and PVT-TRG training resources into the target language using an SMT engine trained on the PVT-TRG and PVT-SRC language resources, respectively, and (ii) directly translates the source language input into the target language using a single SMT engine that is trained on the obtained SRC-TRG language resources (Gispert and Marino, 2006).
3. *phrase-table composition* in which the translation models of the SRC-PVT and PVT-TRG translation engines are combined to a new SRC-TRG phrase-table by merging SRC-PVT and PVT-TRG phrase-table entries with identical pivot language phrases and mul-

[1]LDC: http://www.ldc.upenn.edu
[2]ELRA: http://www.elra.info
[3]GSK: http://www.gsk.or.jp/catalog.html

Proceedings of NAACL HLT 2009: Short Papers, pages 221–224,
Boulder, Colorado, June 2009.

tiplying posterior probabilities (Utiyama and Isahara, 2007; Wu and Wang, 2007).

4. *bridging at translation time* where the coupling is integrated into the SMT decoding process by modeling the pivot text as a hidden variable and assuming independence between source and target sentences (Bertoldi et al., 2008).

3 Pivot Language Selection

The effects of using different pivot languages are investigated using the multilingual *Basic Travel Expressions Corpus* (BTEC), which is a collection of sentences that bilingual travel experts consider useful for people going to or coming from another country. For the pivot translation experiments, we selected twelve of the major world languages covered by BTEC, favoring languages that are actively being researched on, i.e.,*Chinese* (zh), *English* (en), *French* (fr), *German* (de), *Hindi* (hi), *Indonesian* (id), *Japanese* (ja), *Korean* (ko), *Malay* (ms), *Spanish* (es), *Thai* (th), and *Vietnamese* (vi). These languages differ largely in *word order* (SVO, SOV), *segmentation unit* (phrase, word, none), and *degree of inflection* (high, moderate, light). All data sets were case-sensitive with punctuation marks preserved.

However, in a real-world application, identical language resources covering three or more languages are not necessarily to be expected. In order to avoid a trilingual scenario for the pivot translation experiments described in this paper, the 160k sentence-aligned BTEC corpus was randomly split into two subsets of 80k sentences each, whereby the first set of sentence pairs was used to train the source-to-pivot translation models ($80k^{sp}$) and the second subset of sentence pairs was used to train the pivot-to-target translation models ($80k^{pt}$). Table 1 summarizes the characteristics of the BTEC corpus data sets used for the training (*train*) of the SMT models, the tuning of model weights (*dev*), and the evaluation of translation quality (*eval*). Besides the number of sentences (*sen*) and the vocabulary (*voc*), the sentence length (*len*) is also given, as the average number of words per sentence.

For the training of the SMT models, standard word alignment (Och and Ney, 2003) and language modeling (Stolcke, 2002) tools were used. Minimum error rate training (MERT) was used to tune the decoder's parameters, and performed on the *dev* set using the technique proposed in (Och and Ney, 2003). For the translation, an in-house multi-stack phrase-based decoder comparable to MOSES was used. For the evaluation of translation quality, we applied standard automatic evaluation metrics, i.e., BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005). For the experimental results in this paper, the given scores are calculated as the average of the respective BLEU and METEOR scores obtained for each system output and are listed as percent figures.

Table 1: Language Resources

BTEC Corpus		train 80k^{sp}	train 80k^{pt}	dev set	eval set
# of sen		80,000	80,000	1,000	1,000
en	voc	12,264	11,047	1,262	1,292
	len	7.8	7.2	7.1	7.2
de	voc	19,593	17,324	1,486	1,491
	len	7.4	6.8	6.7	6.8
es	voc	16,317	14,807	1,486	1,511
	len	7.6	7.1	7.0	7.2
fr	voc	15,319	13,663	1,455	1,466
	len	7.8	7.3	7.1	7.3
hi	voc	26,096	19,906	1,558	1,588
	len	8.1	7.6	7.4	7.5
id	voc	14,585	13,224	1,433	1,394
	len	7.0	6.5	6.3	6.4
ja	voc	13,868	12,517	1,407	1,408
	len	8.8	8.2	8.1	8.2
ko	voc	13,546	12,281	1,366	1,365
	len	8.3	7.8	7.7	7.8
ms	voc	15,113	13,616	1,459	1,438
	len	7.1	6.6	6.4	6.5
th	voc	6,103	5,603	1,081	1,053
	len	8.1	7.4	7.3	7.4
vi	voc	7,980	7,335	1,245	1,267
	len	9.4	8.7	8.5	8.6
zh	voc	11,084	10,159	1,312	1,301
	len	7.1	6.6	6.4	6.5

In order to get an idea of how difficult the translation task for the different languages is supposed to be, the automatic evaluation scores for the direct translation approach using the $80k^{sp}$ language resources are summarized in Section 3.1. The effects of the pivot language selection are discussed in Section 3.2 using the pivot translation method of *cascading two SMT systems*. In addition, the dependency between selecting the optimal pivot language for a given language pair and the amount of available training resources are described in Section 3.3.

3.1 Direct Translation Results

The automatic evaluation scores for all source and target language pair combinations of the direct translation approach are given in Table 2. For each target language, the highest evaluation scores are marked in boldface and the lowest scores are marked in typewriter mode.

The highest translation quality was achieved for the *Japanese* $\Leftrightarrow$ *Korean*, *Indonesian* $\Leftrightarrow$ *Malay*, and *Spanish* $\Leftrightarrow$ *English* translation tasks. In addition, relatively high evaluation scores were achieved for *Japanese* $\Leftrightarrow$ *Chinese* and for translations from *English* into *German*, *French*, *Hindi*, *Thai*, and *Vietnamese*. On the other hand, the most difficult translation tasks were those having *Korean* or *Chinese* as the source language.

3.2 Pivot Translation Results

The automatic evaluation scores for all pivot translation combinations are summarized in Table 3 whereby for each source-target language pair, the results of the experiments using (i) *English* (*en*) and (ii) the best performing language (*best*) as the pivot language are listed.

Comparing the results of the pivot translation experiments towards the direct translation results, we can see

Table 2: Translation Quality of Direct Translation Approach

SRC	de	en	es	fr	hi	id	ja	ko	ms	th	vi	zh
de	–	74.24	56.22	49.78	63.25	69.31	54.09	50.88	69.33	66.83	67.17	51.59
en	**63.31**	–	**64.30**	**56.10**	**66.43**	73.46	55.64	54.15	73.66	**70.57**	**72.64**	53.18
es	58.98	**76.43**	–	53.53	63.60	70.46	55.37	51.41	70.46	67.69	69.15	52.03
fr	55.45	72.24	57.25	–	61.70	68.58	55.17	52.15	68.72	65.03	65.97	52.83
hi	52.89	67.82	50.69	45.53	–	68.65	52.94	50.93	68.14	66.44	66.88	51.31
id	52.75	67.58	52.06	46.00	62.43	–	55.52	52.90	**88.69**	67.20	68.01	52.77
ja	35.43	51.65	37.82	32.70	46.94	52.90	–	**78.73**	53.26	54.14	51.45	**67.83**
ko	32.65	50.12	36.97	31.62	44.67	53.51	**78.88**	–	51.75	52.35	51.34	63.19
ms	53.16	68.17	53.06	45.30	63.36	**91.12**	54.88	52.18	–	67.79	67.93	53.23
th	49.66	64.53	50.16	42.70	59.40	66.58	53.82	50.81	65.76	–	65.90	52.22
vi	52.59	69.16	53.17	45.60	61.19	68.39	52.95	50.68	69.44	67.64	–	51.29
zh	34.18	49.79	37.13	31.16	44.33	52.72	65.64	62.23	52.46	51.88	51.09	–

Table 4: Pivot Language Selection

PVT	usage (%)	
en	49	(44.5)
ms	16	(14.5)
id	16	(14.5)
ja	14	(12.7)

PVT	usage (%)	
ko	12	(10.9)
zh	2	(1.8)
es	1	(0.9)

that in general the pivot translation approach performs worse than the direct translation approach due to the effect of error chaining, i.e., translation errors of the SRC-PVT engine cause a degradation in translation quality of the PVT-TRG system output. However, for language pairs like *Korean* ⇔ *German*, *Japanese* ⇔ *Indonesian* and *German/Spanish* ⇔ *Korean*, the best pivot translation system outperforms the direct translation approach slightly. This phenomenon is caused mainly by the high SRC-PVT (*PVT-TRG*) translation quality in combination with a better *PVT-TRG* (SRC-PVT) performance compared to the direct *SRC-TRG* system output results.

Besides the automatic evaluation scores, Table 3 lists also the optimal pivot language for each source-target language pair in boldface. The experimental results show that *English* is indeed the best pivot language when translating between languages, like *German*, *Spanish*, *French*, *Hindi*, *Thai*, and *Vietnamese*, whose direct translation performance from/into *English* is high. For these six languages, all language pair combinations achieved the highest scores using the *English* pivot translation approach. In contrast, *English* is the pivot language of choice for only 16.2% (11 out of 68) of the language pairs when translating from/into *Japanese*, *Korean*, *Indonesian*, or *Malay*. In the remaining cases, the language with the highest direct translation scores is in general selected as the optimal pivot language, i.e., *Japanese* for *Korean*, *Malay* for *Indonesian* and vice versa. For *Chinese*, the choice of the optimal pivot language varies largely depending on the language direction. However, the selection of the optimal pivot language is not symmetric for 34.5% of the language pairs, i.e., a different optimal pivot language was obtained for the SRC-TRG compared to the TRG-SRC translation task. This indicates that the choice of the optimal pivot language depends on the relatedness of the SRC and PVT languages as well as the relatedness of the PVT and TRG languages.

The distribution of the optimal pivot language selection for all language pairs is given in Table 4. The figures show that the *English* pivot approach still achieves the highest scores for the majority of the examined language pairs. However, in 55.5% (61 out of 110) of the cases, a non-English pivot language, mainly *Malay*, *Indonesian*, *Japanese*, or *Korean*, is to be preferred.

Table 5: Pivot Selection Shifts for 10k vs. 80k Training Data

10k PVT	80k PVT	pivot translation language pair
ko	**en**	ja-**fr**, ja-**de**, ja-**vi**
ko		zh-**fr**, zh-**es**, zh-**hi**
ja		ko-**vi**, zh-**vi**, zh-**th**
ms		id-**fr**
ja	**es**	ko-**hi**
ja	**id**	ko-**ms**,th-**zh**
en		es-**ms**,hi-**zh**,hi-**ja**

10k PVT	80k PVT	pivot translation language pair
ko	**ms**	ja-**id**
en		vi-**zh**
es		fr-**zh**
en	**ja**	fr-**ko**, hi-**ko**, vi-**ko**
id		zh-**ms**
ms	**ko**	id-**ja**
es		fr-**ja**
en		de-**ja**,es-**ja**

3.3 Training Data Size Dependency

In order to investigate the dependency between selecting the optimal pivot language for a given language pair and the amount of available training resources, we repeated the pivot translation experiments described in Section 3.2 for statistical models trained on subsets of 10k sentences randomly extracted from the $80k^{sp}$ and the $80k^{pt}$ corpora, respectively.

The results showed that 75.5% of the pivot language selections are identical for small (10k) and large (80k) training data sets. For the remaining 27 out of 110 translation tasks, Table 5 lists how the optimal pivot language selection changed. In the case of small training data sets, the pivot language is closely related (in terms of high direct translation quality) to the source language. However, for larger training data sets, the focus shifts towards closely related target languages. Therefore, the higher the translation quality of the pivot translation task is, the more dependend the selection of the optimal pivot language is on the system performance of the PVT-TRG task.

4 Conclusion

In this paper, the effects of using non-English pivot languages for translations between twelve major world languages were compared to the standard English pivot translation approach. The experimental results revealed that *English* was indeed more frequently (45.5% out of

Table 3: Translation Quality of Pivot Translation Approach

SRC	PVT	de	es	fr	hi	id	ja	ko	ms	th	vi	zh
de	en	–	54.69	47.01	60.48	66.42	52.53	51.10	66.47	65.06	66.08	50.46
	best	–	(en) 54.69	(en) 47.01	(en) 60.48	(**ms**) **66.92**	(**ko**) **52.67**	(en) 51.10	(en) 66.47	(en) 65.06	(en) 66.08	(en) 50.46
es	en	55.37	–	48.75	60.24	68.10	52.68	51.80	67.54	65.59	66.99	51.08
	best	(en) 55.37	–	(en) 48.75	(en) 60.24	(**ms**) **69.29**	(**ko**) **53.10**	(en) 51.80	(**id**) **68.37**	(en) 65.59	(en) 66.99	(en) 51.08
fr	en	52.03	53.88	–	58.27	65.59	52.51	51.19	65.43	62.47	64.34	50.12
	best	(en) 52.03	(en) 53.88	–	(en) 58.27	(**ms**) **67.25**	(**ko**) **53.06**	(**ja**) **51.81**	(en) 65.43	(en) 62.47	(en) 64.34	(**ms**) **50.35**
hi	en	48.56	48.69	41.71	–	63.01	50.21	48.96	63.13	62.08	62.48	48.12
	best	(en) 48.56	(en) 48.69	(en) 41.71	–	(**ms**) **65.43**	(**id**) **51.09**	(**ja**) **49.06**	(**id**) **65.54**	(en) 62.08	(en) 62.48	(**id**) **48.71**
id	en	48.97	49.48	42.56	57.41	–	51.30	50.19	72.94	62.40	64.60	49.45
	best	(**ms**) **49.19**	(**ms**) **50.16**	(en) 42.56	(**ms**) **60.30**	–	(**ko**) **54.12**	(**ja**) **51.54**	(en) 72.94	(**ms**) **64.51**	(**ms**) **65.51**	(**ms**) **51.82**
ja	en	33.43	36.61	31.20	44.27	52.31	–	56.34	51.34	52.57	50.97	52.85
	best	(en) 33.43	(**ko**) **36.88**	(en) 31.20	(**ko**) **44.96**	(**ms**) **53.13**	–	(**zh**) **60.99**	(**ko**) **51.37**	(**ko**) **52.65**	(en) 50.97	(**ko**) **62.65**
ko	en	31.52	34.50	29.01	43.23	50.70	54.43	–	49.83	50.74	49.97	51.66
	best	(**ja**) **33.23**	(**ja**) **36.18**	(**ja**) **31.20**	(**es**) **44.24**	(**ja**) **52.21**	(**zh**) **60.10**	–	(**id**) **51.79**	(**ja**) **51.98**	(en) 49.97	(**ja**) **62.74**
ms	en	49.64	49.71	42.39	57.85	73.25	51.01	49.52	–	62.64	64.09	49.22
	best	(**id**) **51.14**	(**id**) **50.95**	(**id**) **43.87**	(**id**) **60.76**	(en) 73.25	(**id**) **54.56**	(**ja**) **50.94**	–	(**id**) **65.99**	(**id**) **66.97**	(**id**) **52.46**
th	en	46.57	46.61	39.83	55.41	61.88	50.54	48.75	61.09	–	61.50	47.75
	best	(en) 46.57	(en) 46.61	(en) 39.83	(en) 55.41	(**ms**) **63.22**	(**ko**) **51.37**	(**ja**) **50.39**	(**id**) **62.36**	–	(en) 61.50	(**id**) **48.72**
vi	en	49.87	50.17	43.04	57.42	64.94	50.68	49.45	64.60	62.50	–	48.12
	best	(en) 49.87	(en) 50.17	(en) 43.04	(en) 57.42	(**ms**) **67.14**	(**ko**) **51.86**	(**ja**) **49.48**	(**id**) **66.57**	(en) 62.50	–	(**ms**) **48.86**
zh	en	32.26	35.29	28.35	43.20	50.11	53.27	52.53	49.20	51.54	49.92	–
	best	(en) 32.26	(en) 35.29	(en) 28.35	(en) 43.20	(**ms**) **52.18**	(**ko**) **61.96**	(**ja**) **60.64**	(**ja**) **49.71**	(en) 51.54	(en) 49.92	–

110 language pairs) selected as the best pivot language over any other examined language. However, its usage is limited to translations between Indo-European languages and some Asian languages like *Thai* or *Vietnamese*. Otherwise, the *English* pivot approach is largely outperformed by using Asian languages as the pivot languages, especially *Japanese*, *Malay*, *Indonesian*, or *Korean*.

The analysis of the results revealed that the selection of the optimal pivot language largely depends on the SRC-PVT and PVT-TRG translation performance, i.e., for small training corpora, the relationship between source/pivot languages seems to be more important, whereas the selection criteria moves towards the relationship between pivot/target languages for larger amounts of training data and thus for MT engines of higher translation quality.

In order to explore the question of pivot selection further and arrive at firmer conclusions, future work will have to investigate in detail what kind of features are important in selecting a pivot language for a given language pair. Besides the translation quality of SMT engines, automatic metrics to measure the relatedness of a language pair should also be taken into account to find optimal pivot languages. For example, (Birch et al., 2008) proposes features like *amount of reordering*, *the morphological complexity of the target language*, and *historical relatedness of the two languages* as strong predictors for the variability of SMT system performance.

In addition, concerning the question of how the pivot language selection criteria depends on the choice of the pivot translation method, future work will also have to investigate the effects of pivot language selection for the other pivot translation approaches described in Section 2.

Based on these findings, we plan to determine the contribution of different language characteristics on the system performance automatically to obtain useful indicators that could be used to train statistical classification models to predict the best pivot language for a new language pair and improve the usability of machine translation between under-resourced languages further.

Acknowledgment

This work is partly supported by the Grant-in-Aid for Scientific Research (C) Number 19500137 and "Construction of speech translation foundation aiming to overcome the barrier between Asian languages", the Special Coordination Funds for Promoting Science and Technology of the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

S. Banerjee and A. Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation. In *Proc. of the ACL*, pages 65–72, Ann Arbor, US.

N. Bertoldi, M. Barbaiani, M. Federico, and R. Cattoni. 2008. Phrase-Based SMT with Pivot Languages. In *Proc. of the IWSLT*, pages 143–149, Hawaii, US.

A. Birch, M. Osborne, and P. Koehn. 2008. Predicting Success in MT. In *Proc. of the EMNLP*, pages 744–753, Hawaii, US.

A. Gispert and J. Marino. 2006. Catalan-English SMT without Parallel Corpus: Bridging through Spanish. In *Proc. of 5th LREC*, pages 65–68, Genoa, Italy.

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

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

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

M. Utiyama and H. Isahara. 2007. A comparison of pivot methods for phrase-based SMT. In *Proc. of HLT*, pages 484–491, New York, US.

H. Wu and H. Wang. 2007. Pivot Language Approach for Phrase-Based SMT. In *Proc. of ACL*, pages 856–863, Prague, Czech Republic.