How to Choose the Best Pivot Language for Automatic Translation of Low-Resource Languages

MICHAEL PAUL, ANDREW FINCH, and EIICHRIO SUMITA, National Institute of Information and Communications Technology

Recent research on multilingual statistical machine translation focuses on the usage of pivot languages in order to overcome language resource limitations for certain language pairs. Due to the richness of available language resources, English is, in general, the pivot language of choice. However, factors like language relatedness can also effect the choice of the pivot language for a given language pair, especially for Asian languages, where language resources are currently quite limited. In this article, we provide new insights into what factors make a pivot language effective and investigate the impact of these factors on the overall pivot translation performance for translation between 22 Indo-European and Asian languages. Experimental results using state-of-the-art statistical machine translation techniques revealed that the translation quality of 54.8% of the language pairs improved when a non-English pivot language was chosen. Moreover, 81.0% of system performance variations can be explained by a combination of factors such as language family, vocabulary, sentence length, language perplexity, translation model entropy, reordering, monotonicity, and engine performance.

Categories and Subject Descriptors: I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Machine translation*

General Terms: Languages, Performance, Measurement

Additional Key Words and Phrases: Machine translation, pivot language selection, translation quality indicators, Asian languages

ACM Reference Format:
Paul, M., Finch, A., and Sumita, E. 2013. How to choose the best pivot language for automatic translation of low-resource languages. *ACM Trans. Asian Lang. Inform. Process.* 12, 4, Article 14 (October 2013), 17 pages.
DOI:http://dx.doi.org/10.1145/2505126

1. INTRODUCTION

The quality of statistical machine translation (SMT) approaches heavily depends on the amount and coverage of bilingual language resources available for training the statistical models. There exist several data collection initiatives[1] amassing and distributing large amounts of textual data. For frequently used language pairs like French-English, large text datasets are readily available. However, for most of the other language pairs, only a limited amount of bilingual resources are available, if any at all.

[1] LDC (http://www.ldc.upenn.edu), ELRA (http://www.elra.info), GSK (http://www.gsk.or.jp/index_e.html), etc.

M. Paul is currently affiliated with ATR-Trek Co., Ltd, Nishinakajima 6-1-1, 532-0011 Osaka, Japan.
Authors' addresses: M. Paul (corresponding author), A. Finch, and E. Sumita, National Institute of Information and Communications Technology, Hikaridai 3-5, 619-0289 Kyoto, Japan; email: mihyaeru.pauru@gmail.com.

DOI:http://dx.doi.org/10.1145/2505126

In order to overcome language resource limitations, recent research on SMT has focused on the usage of *pivot languages* [Bertoldi et al. 2008; de Gispert and Marino 2006; Utiyama and Isahara 2007; Wu and Wang 2007]. 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 facilitates the use of larger amounts of bilingual data for training. In a first step, the source language input is translated into the pivot language using statistical translation models trained on the *source-pivot* language resources. In the second step, the translation in the pivot language is translated into the target language using a second translation engine trained on the *pivot-target* language resources. Although the pivot translation approach does enable translation between languages where no bilingual resources exist at all, the drawback of this translation method is that the translation quality may deteriorate in the two-step process, that is, small translation errors during the first step may lead to severe errors in the target language output.

In previous research on pivot translation, the pivot language was typically selected based on two criteria: (1) the availability of bilingual language resources and (2) the language relatedness between source and pivot languages.

In most recent research, English has been the pivot language of choice due to the richness of available language resources. For example, Utiyama and Isahara [2007] exploited the Europarl[2] corpus for comparing pivot translation approaches between French, German, and Spanish via English, and the IWSLT evaluation campaign [Paul 2008] featured a pivot translation task for Chinese-Spanish translation via English. In addition, several research efforts tried to exploit the closeness between specific language pairs to achieve high-quality translation hypotheses in the first step to minimize the deterioration effect in the pivot approach. For example, de Gispert and Marino [2006] proposed a method for translating Catalan-English via Spanish, and Babych et al. [2005] translated Ukrainian-English via Russian. Moreover, Cohn and Lapata [2007] exploited multiple translations of the same source phrase to obtain more reliable translation frequency estimates from small datasets and showed that using more than one pivot improve the overall system performance. Leusch et al. [2010] generate intermediate translations in several pivot languages and using system combination techniques to output a consensus translation.

However, the preceding criteria might not be sufficient for choosing the best pivot language, especially for Asian languages. With the exception of Chinese, only a few parallel text corpora for Asian languages and English are publicly available. Moreover, language families in Asia cover a large number of different languages and are more linguistically diverse than Indo-European language families. Recent research on pivot translation from or into Asian languages has shown that the usage of non-English pivot languages can improve translation quality for certain language pairs [Paul et al. 2009].

Concerning the contribution of aspects of different language pairs on the quality of machine translation, Birch et al. [2008] identified three features (morphological complexity, amount of reordering, historical relatedness) for predicting the success of MT in translations between the official languages of the European Union. Moreover, Koehn et al. [2009] investigated an additional feature (translation model complexity) using the JRC-Aquis corpus covering not only Indo-European languages but also one Semitic and three Finno-Ugric languages. Specia et al. [2011] investigated the applicability of quality estimation indicators (complexity, fluency, named entities) in predicting the adequacy of translations on the sentence level for Arabic-English.

[2] `www.statmt.org/europarl`

This article differs from previous research in the following aspects: (1) it focuses on the framework of pivot translation, where a target language translation of a source language input is obtained through an intermediate pivot language; (2) it investigates what factors make a pivot language effective; and (3) it analyzes what impact these factors have on the overall translation quality of language pairs, not only including Indo-European languages, but also a large variety of Asian languages. Pivot-based SMT experiments translating between 22 Indo-European and Asian languages are carried out and analyzed in Section 2 to provide new insights into how much language differences affect the translation performance of pivot translation approaches. In Section 3, eight factors (language family, vocabulary, sentence length, language perplexity, translation model entropy, reordering, monotonicity, engine performance) are investigated to determine the significance of each factor in predicting translation quality using linear regression analysis.

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*. 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*. (a) Creates a “noisy” SRC-TRG parallel corpus by translating the pivot language parts of the SRC-PVT training resources into the target language using an SMT engine trained on the PVT-TRG language resources; and (b) 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 [de Gispert and Marino 2006].

(3) *Phrase-Table Composition*. 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 multiplying posterior probabilities [Utiyama and Isahara 2007; Wu and Wang 2007].

(4) *Bridging at Translation Time*. 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 language sentences [Bertoldi et al. 2008].

(5) *Multi-Pivot Translation*. Intermediate translations into several pivot languages are used to generate a final translation by probabilistic combination of translation models or system combination techniques [Cohn and Lapata 2007; Leusch et al. 2010].

However, as the scope of this article is not to improve pivot translation methods but to investigate the effects of the pivot language selection for statistical machine translation involving low-resource languages, the method of cascading two translation systems is adopted in the pivot translation experiments reported in this article.

Pivot translation using the cascading approach requires two MT engines, where the first engine translates the source language input into the pivot language and the second engine takes the obtained pivot language output as its input and translates it into the target language. Given *N* languages, a total of *2*N*(N-1)* SMT engines have to be built in order to cover all *N*(N-1)*(N-2)* SRC-PVT-TRG language pair combinations.

2.1. Language Resources

The effects of pivot language selection on MT quality 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 [Kikui et al. 2006]. The sentence-aligned corpus consists of 160K sentences pairs[3] covering 22 Indo-European and Asian languages which belong to a variety of language families, including Germanic (DA, DE, EN, NL), Romance (ES, FR, IT, PT, PTB), Slavic (PL, RU), Indo-Iranian (HI), Afro-Asiatic (AR), Austronesian (ID, MS, TL), Taid-Kadai (TH), Austro-Asiatic (VI), Sino-Tibetan (ZH, ZHT), Japanese (JA), and Korean (KO) languages.

The corpus statistics are summarized in Table I, where *Voc* specifies the vocabulary size, *Len* the average sentence length, and *OOV* the percentage of unknown words[4] in the respective datasets. These languages differ largely in word order (i.e., *order*: subject-object-verb (*SOV*), subject-verb-object (SVO), verb-subject-object (*VSO*), no dominant word-order[5] (*mixed*)), segmentation unit (i.e., *unit*: phrase, word, none), and degree of inflection (i.e., *inflection*: high, moderate, light). Very similar characteristics can be seen for Indo-European languages and for certain subsets of Asian languages (JA, KO; ID, MS).

In addition, Indo-European languages have, in general, a higher degree of inflection compared to Asian languages. Concerning word segmentation for languages that do not use a white space to separate word/phrase tokens, the corpora were preprocessed using language-specific word segmentation tools, that is, CHASEN[6] for Japanese, ICTCLAS[7] for Chinese, WORDCUT[8] for Thai, and an inhouse segmenter for Korean. For all other languages, simple tokenization tools were applied. All datasets were processed in a case-sensitive manner with punctuation marks preserved.

The language resources were randomly split into three subsets for the evaluation of translation quality (*eval*, 1,000 sentences); the tuning of the SMT model weights (*dev*, 1,000 sentences), and the training of the statistical models (*train*). 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, the *train* corpus was randomly split into two subsets of 80K sentences each, whereby the first set of sentence pairs was used to train the SRC-PVT translation models and the second subset of sentence pairs was used to train the PVT-TRG translation models. In total, 924 SMT translation engines were built to cover all 9,240 language-pair combinations.

The SMT model training as well as the evaluation of the MT results were carried out in a case-sensitive fashion with punctuation marks reserved. For the training of the SMT models, standard word alignment [Och and Ney 2003] and language modeling [Stolcke 2002] tools were used. For the translation, a multistack phrase-based decoder [Finch et al. 2007] built within the framework of a feature-based exponential model containing a standard set of features[9] was used. Minimum error rate training (MERT)

[3] The BTEC corpus was created by translating the original English sentences into the respective languages.

[4] Words of the evaluation dataset that are not occuring in the training datasets.

[5] World Atlas of Language Structres: `http://wals.info/feature/81A`.

[6] `http://chasen-legacy.sourceforge.jp`

[7] `https://github.com/pierrchen/ictclas_plus`

[8] `http://www.sourceforge.net/projects/thaiwordseg`

[9] The feature set included phrase translation and inverse phrase table probabilities, lexical weighting and inverse inverse lexical weighting probabilities, phrase penalty, 5-gram language model probability, lexical reordering probability, simple distance-based distortion model, and word penalty.

Table I. Language Resources (BTEC_{160K})

(Indo-European Languages)

Language		Voc	Len	OOV	Order	Unit	Inflection
Danish	DA	26.5k	7.2	1.0	SVO	word	high
German	DE	25.7k	7.1	1.1	mixed	word	high
English	EN	15.4k	7.5	0.4	SVO	word	moderate
Spanish	ES	20.8k	7.4	0.8	SVO	word	high
French	FR	19.3k	7.6	0.7	SVO	word	high
Hindi	HI	33.6k	7.8	3.8	SOV	word	high
Italian	IT	23.8k	6.7	0.9	SVO	word	high
Dutch	NL	22.3k	7.2	1.0	mixed	word	high
Polish	PL	36.4k	6.5	1.1	SVO	word	high
Portuguese	PT	20.8k	7.0	1.0	SVO	word	high
Brazilian Portuguese	PTB	20.5k	7.0	1.0	SVO	word	high
Russian	RU	36.2k	6.4	2.3	SVO	word	high

(Asian Languages)

Language		Voc	Len	OOV	Order	Unit	Inflection
Arabic	AR	47.8k	6.4	2.1	VSO	word	high
Indonesian	ID	18.6k	6.8	0.8	SVO	word	high
Japanese	JA	17.2k	8.5	0.5	SOV	none	moderate
Korean	KO	17.2k	8.1	0.8	SOV	phrase	moderate
Malay	MS	19.3k	6.8	0.8	SVO	word	high
Thai	TH	7.4k	7.8	0.4	SVO	none	light
Tagalog	TL	28.7k	7.4	0.7	VSO	word	high
Vietnamese	VI	9.9k	9.0	0.2	SVO	phrase	light
Chinese	ZH	13.3k	6.8	0.5	SVO	none	light
Taiwanese	ZHT	39.5k	5.9	0.6	SVO	none	light

was used to tune the decoder's parameters, and was performed on the *dev* set using the technique proposed by Och and Ney [2003].

For the translation quality evaluation, we applied the standard automatic metric BLEU [Papineni et al. 2002], which calculates the geometric mean of n-gram precision of the system output with respect to reference translations multiplied by a brevity penalty to prevent very short candidates from receiving too high a score. Scores range between 0% (worst) and 100% (best). For the experiments reported in this article, single translation references were used.

2.2. Language Diversity

In order to get an idea of how diverse the investigated languages are, we calculated the language perplexity of the target language evaluation datasets according to a standard

Table II. Language Perplexity (BTEC_{160K})

(Indo-European Languages)

Language	Perplexity	Total Entropy
DA	13.9	26007.0
DE	17.2	27820.2
EN	13.3	26781.7
ES	13.2	26501.7
FR	11.8	25906.5
HI	17.6	30902.4
IT	16.9	26169.1
NL	15.5	27026.8
PL	19.5	25749.1
PT	15.6	26249.3
PTB	15.2	26028.8
RU	18.1	25275.6

(Asian Languages)

Language	Perplexity	Total Entropy
AR	19.6	26450.7
ID	15.8	25565.6
JA	10.9	28012.9
KO	11.4	27102.7
MS	15.7	25848.6
TH	14.7	28462.5
TL	16.3	28775.8
VI	11.9	30627.6
ZH	17.2	26755.5
ZHT	20.6	24483.5

5-gram language model trained on the respective training datasets. Table II lists the *language perplexity* and the *total entropy*, that is, the entropy multiplied by the number of words of the evaluation dataset. The total entropy figures represent the entropy of the whole corpus, and the numbers indicate that Hindi and Vietnamese are supposed to be the most difficult languages, followed by Tagalog, Thai, and Japanese. In general, the total entropy figures of Indo-European languages are much lower than those of Asian languages.

In order to get an idea of how difficult the translation task for the different languages is supposed to be, we calculated the BLEU scores for all the language-pair combinations of the direct translation approach using the SRC-TRG engines trained on the full corpus. The obtained results are summarized in Table III.

For each source (target) language, the language pair achieving the highest evaluation scores are highlighted using black (white) scores in boldface (italic), respectively.[10] The highest evaluation scores were achieved for closely related language pairs, such as Portuguese ⇔ Brazilian Portuguese, Indonesian ⇔ Malay, English ⇔ Spanish, and Japanese ⇔ Korean. The lowest translation quality were obtained when translating from Chinese, Japanese, or Korean into any of the not closely related languages and vice versa.

The results show the large diversity between the investigated language pairs. In general, the evaluation scores for Indo-European-only language pairs are much higher than those for language pairs involving Asian languages. Interestingly, not all language pairs having English as the source language always achieved the highest scores, especially when translating into Asian languages. Similarly, the quality of English translations depends largely on the respective source language.

This indicates that a deterioration in translation quality is to be expected when English is used as the pivot language compared to other pivot languages, where higher evaluation scores for the direct translation from/into the pivot language were obtained.

[10] Due to differences in word units and reference translations, the BLEU scores are not directly comparable across different target languages.

Table III. Direct Translation Quality (BTEC$_{160K}$, BLEU%)

		(Indo-European Languages)												(Asian Languages)										
	TRG → ↓ SRC	DA	DE	EN	ES	FR	HI	IT	NL	PL	PT	PTB	RU	AR	ID	JA	KO	MS	TH	TL	VI	ZH	ZHT
(Indo-European Languages)	DA	–	62.7	**72.6**	67.9	67.7	53.7	63.6	67.1	59.4	66.0	66.4	60.3	58.8	61.4	42.0	40.6	62.1	60.5	56.8	60.8	39.1	50.3
	DE	66.1	–	**71.4**	66.7	65.3	55.0	63.1	68.8	58.9	65.0	65.1	58.5	58.2	60.7	43.9	40.9	61.2	59.5	58.0	58.8	36.9	51.5
	EN	*72.6*	*67.3*	–	*75.0*	*73.8*	*57.9*	*70.0*	*72.8*	*64.2*	72.2	72.6	*63.3*	*64.2*	66.1	45.2	42.9	65.4	*63.7*	*61.2*	*64.0*	41.4	53.0
	ES	66.1	63.4	**74.4**	–	69.4	55.3	66.9	67.6	59.7	70.9	71.0	59.9	60.5	62.0	43.2	40.8	61.8	59.6	58.1	61.2	37.9	49.0
	FR	65.2	59.4	**69.0**	66.9	–	52.8	61.7	64.7	57.1	64.3	65.0	57.0	56.5	59.5	44.8	43.1	59.8	56.1	55.5	56.7	41.0	49.5
	HI	60.9	59.6	**64.4**	61.5	61.9	–	59.8	61.4	54.9	59.6	60.5	55.7	56.5	61.0	42.3	40.0	59.8	59.1	55.5	59.4	39.1	50.0
	IT	64.7	62.7	**71.0**	69.1	67.8	54.0	–	65.4	58.8	66.1	66.3	57.1	58.8	60.6	43.2	40.9	60.9	58.6	56.1	60.4	39.4	50.6
	NL	67.4	66.5	**74.1**	68.8	68.8	55.3	64.4	–	58.9	65.0	66.1	59.5	58.1	61.8	42.0	40.9	61.9	59.9	56.8	60.0	39.5	51.4
	PL	63.6	60.4	**68.1**	64.2	61.5	53.9	62.4	63.3	–	63.8	63.4	60.4	57.4	61.0	44.3	41.7	61.5	57.2	55.8	57.9	38.4	50.9
	PT	66.9	63.0	74.1	72.1	69.4	53.0	66.4	66.5	59.7	–	*92.8*	60.4	60.3	61.9	43.0	40.4	61.3	60.1	55.8	61.3	39.8	50.3
	PTB	66.8	62.7	73.6	72.7	69.8	54.4	65.7	66.5	60.8	*92.5*	–	60.4	60.8	61.6	42.3	41.5	61.3	60.0	56.2	60.4	38.5	49.0
	RU	61.1	57.7	**65.0**	63.2	61.2	52.8	59.4	60.7	58.2	61.2	61.5	–	56.2	58.6	41.5	40.3	58.0	58.5	54.1	58.2	37.9	48.8
(Asian Languages)	AR	64.4	59.8	**70.4**	66.2	65.6	53.5	61.4	63.3	57.4	63.7	64.6	58.0	–	61.7	41.4	38.6	60.1	57.3	54.2	59.7	37.8	50.0
	ID	59.4	58.1	64.3	61.7	60.8	53.4	58.0	59.4	56.7	60.1	58.7	56.4	56.7	–	42.7	40.3	*85.1*	58.7	56.1	59.1	40.0	49.1
	JA	40.5	38.6	44.9	43.7	45.8	32.7	39.0	40.3	37.4	41.3	40.6	37.8	36.3	38.7	–	*72.7*	39.9	43.2	39.9	40.7	59.0	39.9
	KO	36.5	35.0	40.4	39.8	42.4	32.2	35.7	35.5	33.8	39.3	39.2	33.9	32.9	37.3	*71.7*	–	39.0	40.8	36.4	35.8	53.6	38.1
	MS	60.6	57.2	65.4	62.2	60.9	53.8	59.2	61.1	56.6	59.5	59.5	55.7	56.8	*88.1*	43.5	41.2	–	60.1	55.4	60.6	39.3	50.1
	TH	58.2	54.5	**60.4**	56.8	56.6	49.1	55.8	56.3	50.7	55.2	55.7	54.1	51.6	56.3	41.4	40.4	56.6	–	52.9	57.6	37.3	46.7
	TL	60.6	59.1	**67.6**	62.3	62.1	52.7	58.0	60.3	55.6	61.2	61.3	55.4	54.3	59.9	44.7	41.4	58.9	57.7	–	58.3	40.1	49.0
	VI	59.7	58.3	**66.2**	63.1	61.6	52.1	58.7	59.7	54.3	60.3	60.2	55.6	55.8	59.3	41.6	39.4	61.5	60.5	55.5	–	39.0	49.2
	ZH	37.0	37.1	41.3	41.8	42.4	32.6	36.7	39.2	36.1	39.4	38.8	35.7	33.0	38.2	**56.2**	52.2	38.6	40.4	36.1	38.4	–	*53.3*
	ZHT	51.4	50.7	56.5	53.0	53.1	45.9	51.0	53.0	49.2	51.5	52.0	49.6	49.9	53.2	44.5	44.6	52.0	51.9	48.9	50.0	*59.3*	–

2.3. Pivot Language Selection

Figure 1 summarizes the BLEU score ranges ([MIN:MAX]) for all the pivot translation experiments obtained for the given pivot language in terms of a box-and-whisker diagram. Each box part goes from the first to the third quartiles, and the dot in the box represents the mean score of the respective BLEU score distribution. The results show a large variation in BLEU scores for all pivot languages, indicating that there is not a single "best" pivot language, but the quality of a given pivot translation task largely depends on the respective source and target languages. For Indo-European pivot languages, the best language combination scores are, in general, much higher than the ones obtained for Asian pivot languages.

Table IV lists the highest BLEU scores for the pivot translation experiments obtained for all language-pair combinations. The pivot languages achieving the highest scores (*oracle pivot*) for translating the source language into the target language are given in parentheses. Non-English oracle pivot languages are highlighted in boldface. The figures show that the English pivot approach still achieves the highest scores for the majority of the examined language pairs. However, in 54.8% (230 out of 420) of the cases, a non-English pivot language (mainly PT, PTB, MS, ID, JA, KO) is preferable.

In addition, the experimental results show that the selection of the best pivot language is not symmetric for 21.4% (90 out of 420) of the investigated language pairs.

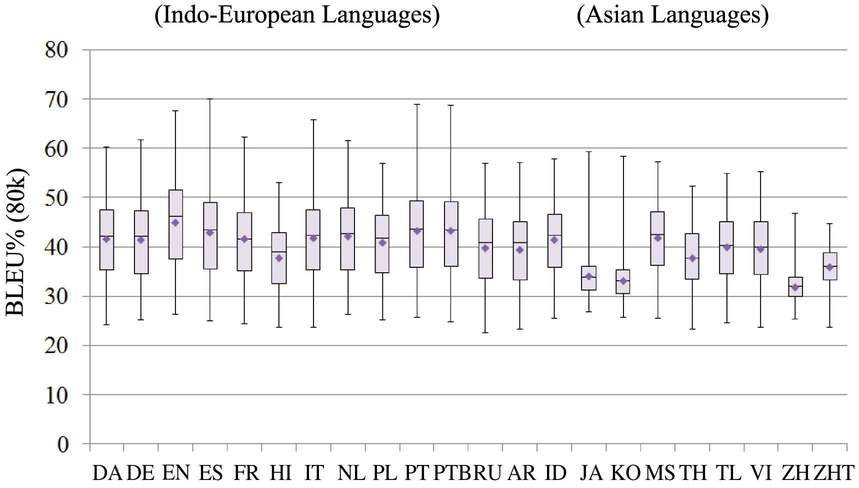

Fig. 1. Pivot language dependency.

For languages that are closely related, such as Portuguese versus Brazilian Portuguese and Malay versus Indonesian, the related language should be chosen as the pivot language when either translating from or into the respective language for 88.7% (71 out of 80) and 85.0% (68 out of 80) of the pivot translation experiments. Moreover, Japanese is the dominant pivot language when translating from Korean into another language (95.0%, 19 out of 20) , but not for the translation into Korean (30.0%, 6 out of 20). These results suggest that in general pivot languages closely related to the source language have a larger impact on the overall pivot translation quality than pivot languages related to the target language.

Interestingly, for Indo-European-only language pairs, only Indo-European languages are the oracle pivot language, the majority of which is English. In addition, Spanish is the pivot language of choice when translating from English into another Indo-European language, and the Dutch pivot achieved the highest BLEU scores for Germanic-only language pairs. On the other hand, when translating between Asian languages, only 65.6% (59 out of 90) of the oracle pivot languages are Asian languages.

In order to investigate the dependency of pivot language selection and language families further, Table V summarizes the BLEU scores of pivot translations between only (a) non-English Indo-European and (b) Asian language pairs. The results of the Indo-European-only language pairs in the table on the left confirm the findings of Table IV. Portuguese and Brazilian Portuguese are still the dominant pivot languages for non-English Indo-European language pairs. An increase of Spanish (Dutch) oracle pivot language pairs can be seen for the translation between only Romance (Germanic) languages, respectively. Similarly, Malay and Indonesian are the dominant pivot languages, followed by Japanese and Korean, for Asian-only language pairs, most of which achieve BLEU scores that are only slightly lower than the ones for the English oracle pivot language experiments reported in Table IV.

Table VI summarizes the proportion of the experiments in which the respective pivot language achieved the highest evaluation score for the pivot translation experiments summarized in Table IV (all language pairs) and Table V (non-English Indo-European language pairs, Asian language pairs). The results show that English is indeed the pivot language of choice for the majority of the investigated translation directions, but for almost half of the language pairs, a non-English pivot language is preferable.

Table IV. Oracle Pivot Translation Quality (BTEC$_{80K}$, BLEU%)

		(Indo-European Languages)												(Asian Languages)									
	TRG → ↓ SRC	DA	DE	EN	ES	FR	HI	IT	NL	PL	PT	PTB	RU	AR	ID	JA	KO	MS	TH	TL	VI	ZH	ZHT
(Indo-European Languages)	DA	–	53.9 (en)	60.3 **(nl)**	59.1 (en)	57.6 (en)	45.3 (en)	53.4 (en)	57.6 (en)	49.8 (en)	57.8 **(ptb)**	57.8 (en)	49.5 (en)	48.8 (en)	52.5 **(ms)**	37.5 **(ko)**	36.9 (en)	51.9 **(id)**	51.6 (en)	47.7 (en)	52.6 (en)	34.2 (en)	39.9 (en)
	DE	57.2 (en)	–	61.3 **(nl)**	59.3 (en)	57.3 (en)	45.6 (en)	53.6 (en)	58.5 (en)	49.7 (en)	59.2 **(ptb)**	58.3 **(pt)**	49.1 (en)	47.8 (en)	52.1 **(ms)**	37.8 (en)	36.8 (en)	51.5 (en)	51.8 (en)	48.3 (en)	52.2 (en)	33.3 (en)	41.1 (en)
	EN	59.8 **(es)**	55.5 **(nl)**	–	62.7 **(pt)**	60.7 **(es)**	45.8 **(es)**	56.9 **(es)**	60.1 **(es)**	49.9 **(es)**	65.7 **(ptb)**	65.5 **(pt)**	50.7 **(es)**	50.1 **(es)**	57.2 **(ms)**	39.4 **(ko)**	38.0 **(ja)**	56.8 **(id)**	51.3 **(es)**	49.4 **(es)**	53.6 **(es)**	33.6 **(es)**	40.4 **(es)**
	ES	59.0 (en)	54.4 (en)	63.3 **(pt)**	–	59.4 (en)	45.6 (en)	55.7 (en)	58.6 (en)	51.7 (en)	64.7 **(ptb)**	64.6 **(pt)**	50.5 (en)	50.1 (en)	55.3 **(ms)**	38.5 **(ko)**	37.7 (en)	54.4 **(id)**	52.5 (en)	49.6 (en)	54.0 (en)	34.1 (en)	40.4 (en)
	FR	56.4 (en)	50.9 (en)	58.8 **(es)**	58.2 (en)	–	43.2 (en)	52.4 (en)	54.8 (en)	47.3 (en)	58.7 **(ptb)**	57.9 **(pt)**	47.3 (en)	48.2 (en)	52.5 **(ms)**	37.8 (en)	37.6 **(ja)**	51.1 **(id)**	49.5 (en)	46.6 (en)	50.5 (en)	33.4 **(es)**	39.9 (en)
	HI	50.3 (en)	47.4 (en)	50.5 **(ptb)**	51.8 (en)	50.8 (en)	–	47.9 (en)	50.2 (en)	44.4 (en)	51.5 **(ptb)**	51.6 **(pt)**	44.6 (en)	44.7 (en)	50.3 **(ms)**	35.7 **(ko)**	34.6 (en)	50.8 **(id)**	48.1 (en)	43.6 (en)	48.2 (en)	30.8 (en)	36.9 (en)
	IT	56.7 (en)	52.8 (en)	60.6 **(pt)**	59.5 (en)	58.1 (en)	44.8 (en)	–	55.7 (en)	48.8 (en)	60.5 **(ptb)**	60.2 **(pt)**	48.1 (en)	47.1 (en)	52.5 **(ms)**	38.1 (en)	36.8 (en)	52.1 **(id)**	50.6 (en)	47.3 (en)	51.6 (en)	32.3 **(es)**	40.5 (en)
	NL	60.3 (en)	55.8 (en)	60.9 **(es)**	61.5 (en)	59.6 (en)	46.3 (en)	55.0 (en)	–	51.1 (en)	60.0 **(ptb)**	59.5 (en)	50.3 (en)	49.7 (en)	52.6 **(ms)**	37.7 (en)	36.8 (en)	51.9 (en)	52.0 (en)	49.0 (en)	53.3 (en)	33.3 (en)	39.9 (en)
	PL	54.7 (en)	51.1 (en)	56.1 **(ptb)**	56.2 (en)	54.0 (en)	44.2 (en)	51.2 (en)	53.5 (en)	–	56.1 **(ptb)**	56.6 **(pt)**	48.7 (en)	46.4 (en)	52.3 **(ms)**	37.4 **(ko)**	37.6 (en)	51.6 **(id)**	50.1 (en)	47.3 (en)	50.5 (en)	32.7 (en)	39.7 (en)
	PT	60.6 **(ptb)**	55.8 **(ptb)**	68.7 **(ptb)**	67.0 **(ptb)**	63.6 **(ptb)**	47.3 **(ptb)**	58.8 **(ptb)**	60.1 **(ptb)**	51.8 **(ptb)**	–	67.8 **(es)**	52.2 **(ptb)**	52.4 **(ptb)**	54.8 **(ms)**	38.1 **(ko)**	37.3 (en)	53.6 **(id)**	53.5 **(ptb)**	50.1 **(ptb)**	54.8 **(ptb)**	34.1 **(ptb)**	42.6 **(ptb)**
	PTB	60.4 **(pt)**	56.5 **(pt)**	68.9 **(pt)**	66.9 **(pt)**	62.8 **(pt)**	47.9 **(pt)**	59.1 **(pt)**	60.0 **(pt)**	52.8 **(pt)**	70.0 **(es)**	–	51.5 **(pt)**	52.2 **(pt)**	54.9 **(pt)**	38.7 **(ko)**	37.2 (en)	54.2 **(pt)**	52.9 **(pt)**	50.5 **(pt)**	54.8 **(pt)**	34.8 **(pt)**	42.2 **(pt)**
	RU	51.6 (en)	47.6 (en)	53.6 **(ptb)**	53.8 (en)	51.5 (en)	42.2 (en)	47.5 (en)	51.2 (en)	46.8 (en)	53.2 **(ptb)**	53.8 **(pt)**	–	44.8 (en)	50.3 **(ms)**	36.7 (en)	35.8 (en)	50.3 **(id)**	47.4 (en)	44.2 (en)	49.1 (en)	32.0 (en)	37.0 (en)
(Asian Languages)	AR	54.7 (en)	51.3 (en)	56.5 **(pt)**	57.1 (en)	56.1 (en)	44.9 (en)	51.7 (en)	54.4 (en)	47.2 (en)	55.7 **(ptb)**	55.6 **(pt)**	47.9 (en)	–	52.0 **(ms)**	36.4 (en)	36.1 (en)	52.0 (en)	49.2 (en)	45.6 (en)	51.8 (en)	32.4 (en)	38.7 (en)
	ID	52.0 **(ms)**	48.5 **(ms)**	56.7 **(ms)**	54.0 **(ms)**	51.9 **(ms)**	46.1 **(ms)**	49.2 **(ms)**	51.7 **(ms)**	48.4 **(ms)**	51.3 **(ptb)**	51.3 **(pt)**	46.9 **(ms)**	47.8 **(ms)**	–	39.1 **(ms)**	37.5 **(ja)**	59.6 (en)	51.7 **(ms)**	47.8 **(ms)**	52.7 **(ms)**	34.6 **(ms)**	41.5 **(ms)**
	JA	33.5 (en)	31.9 (en)	38.8 **(ko)**	37.9 **(ko)**	38.6 (en)	29.3 **(ko)**	33.0 (en)	34.1 (en)	31.1 (en)	35.8 **(ptb)**	36.3 **(pt)**	30.7 (en)	29.8 **(ko)**	35.5 **(ko)**	–	46.7 **(zh)**	33.9 **(id)**	37.9 **(ko)**	33.7 **(ko)**	35.5 **(ko)**	46.9 **(ko)**	33.1 **(ko)**
	KO	33.2 **(ja)**	31.8 **(ja)**	38.7 **(ja)**	37.1 **(ja)**	38.8 **(ja)**	28.8 **(ja)**	32.4 **(ja)**	32.7 **(ja)**	30.7 **(ja)**	34.5 **(ja)**	36.3 **(ja)**	29.5 **(ja)**	29.7 **(ja)**	35.2 **(ja)**	45.8 **(zh)**	–	34.2 **(id)**	38.1 **(ja)**	32.4 **(ja)**	33.7 **(ja)**	47.2 **(ja)**	32.9 **(ja)**
	MS	53.1 **(id)**	50.5 **(id)**	57.8 **(id)**	55.1 **(id)**	53.8 **(id)**	47.3 **(id)**	49.5 **(id)**	53.0 **(id)**	49.3 **(id)**	52.3 **(id)**	53.2 **(id)**	48.7 **(id)**	48.5 **(id)**	60.2 (en)	39.8 **(id)**	37.0 **(id)**	–	53.2 **(id)**	48.7 **(id)**	53.4 **(id)**	35.1 **(id)**	42.9 **(id)**
	TH	49.5 (en)	45.0 (en)	50.1 **(ptb)**	49.2 (en)	48.5 (en)	40.6 (en)	45.1 (en)	47.9 (en)	43.2 (en)	50.1 **(ptb)**	49.8 **(pt)**	40.8 (en)	41.4 (en)	48.5 **(ms)**	36.1 **(ko)**	36.8 **(ja)**	47.9 **(id)**	–	41.7 (en)	47.5 (en)	31.4 **(id)**	37.0 (en)
	TL	53.6 (en)	50.2 (en)	54.5 **(pt)**	55.6 (en)	53.7 (en)	43.7 (en)	49.7 (en)	51.8 (en)	47.0 (en)	53.5 **(ptb)**	53.1 **(pt)**	46.0 (en)	45.4 (en)	52.5 **(ms)**	37.5 **(ko)**	36.7 (en)	50.8 **(id)**	50.2 (en)	–	51.3 (en)	33.0 (en)	39.3 (en)
	VI	53.2 (en)	48.8 (en)	53.7 **(pt)**	53.8 (en)	52.6 (en)	42.8 (en)	48.8 (en)	51.8 (en)	46.4 (en)	52.2 **(ptb)**	53.3 **(pt)**	45.4 (en)	46.0 (en)	53.0 **(ms)**	36.8 **(ko)**	35.4 **(ja)**	52.8 **(id)**	49.3 (en)	45.2 (en)	–	31.6 **(ms)**	38.3 (en)
	ZH	31.7 (en)	31.2 **(nl)**	35.4 **(zht)**	35.8 (en)	34.9 (en)	27.9 **(ja)**	31.5 (en)	32.1 (en)	29.1 (en)	33.1 **(ptb)**	33.4 **(ja)**	27.7 **(ms)**	27.0 **(nl)**	34.3 **(ms)**	47.4 **(ko)**	47.6 **(ja)**	32.3 **(id)**	36.3 (en)	30.9 (en)	33.2 (en)	–	33.8 **(ja)**
	ZHT	44.1 (en)	41.4 (en)	44.5 **(pt)**	45.4 (en)	44.5 (en)	36.8 (en)	40.8 (en)	43.5 (en)	39.4 (en)	44.6 **(ptb)**	44.3 **(pt)**	39.5 (en)	38.2 (en)	44.5 **(ms)**	40.8 **(zh)**	38.5 **(zh)**	44.0 **(id)**	43.0 (en)	39.4 (en)	42.8 (en)	35.0 **(ja)**	–

Table V. Changes in Pivot Selection for Non-English Language Pairs (BTEC$_{80K}$, BLEU%)

(Indo-European Languages)

TRG → ↓ SRC	DA	DE	ES	FR	HI	IT	NL	PL	PT	PTB	RU
DA	–	51.7 **(nl)**	56.0 **(nl)**	56.2 **(es)**	43.1 **(es)**	50.6 **(es)**	55.2 **(es)**	46.7 **(pt)**	57.8 (ptb)	57.6 **(pt)**	47.5 **(es)**
DE	55.4 **(nl)**	–	57.1 **(ptb)**	55.4 **(ptb)**	43.3 **(nl)**	51.9 **(ptb)**	54.8 **(es)**	47.3 **(nl)**	59.2 (ptb)	58.3 (pt)	47.8 **(nl)**
ES	57.0 **(pt)**	52.9 **(pt)**	–	58.4 **(pt)**	43.9 **(pt)**	54.7 **(pt)**	56.0 **(ptb)**	48.8 **(pt)**	64.7 (ptb)	64.6 (pt)	48.4 **(ptb)**
FR	53.2 **(pt)**	50.1 **(nl)**	57.2 **(pt)**	–	41.6 **(es)**	50.7 **(es)**	53.4 **(es)**	45.0 **(es)**	58.7 (ptb)	57.9 (pt)	45.8 **(es)**
HI	47.3 **(ptb)**	45.6 **(nl)**	49.6 **(ptb)**	48.4 **(ptb)**	–	45.2 **(ptb)**	47.6 **(de)**	41.4 **(es)**	51.5 (ptb)	51.6 (pt)	41.7 **(es)**
IT	53.8 **(pt)**	50.4 **(nl)**	58.5 **(pt)**	56.4 **(ptb)**	42.4 **(pt)**	–	53.8 **(es)**	46.8 **(pt)**	60.5 (ptb)	60.2 (pt)	47.3 **(es)**
NL	55.5 **(es)**	51.6 **(da)**	57.7 **(ptb)**	56.7 **(es)**	43.9 **(es)**	52.0 **(es)**	–	47.7 **(pt)**	60.0 (ptb)	59.5 **(pt)**	47.9 **(es)**
PL	51.8 **(pt)**	47.9 **(pt)**	53.9 **(ptb)**	51.8 **(pt)**	41.9 **(pt)**	49.6 **(ptb)**	51.3 **(es)**	–	56.1 (ptb)	56.6 (pt)	45.6 **(es)**
PT	60.6 (ptb)	55.8 (ptb)	67.0 (ptb)	63.6 (ptb)	47.3 (ptb)	58.8 (ptb)	60.1 (ptb)	51.8 (ptb)	–	67.8 (es)	52.2 (ptb)
PTB	60.4 (pt)	56.5 (pt)	66.9 (pt)	62.8 (pt)	47.9 (pt)	59.1 (pt)	60.0 (pt)	52.8 (pt)	70.0 (es)	–	51.5 (pt)
RU	50.0 **(pt)**	46.8 **(nl)**	52.5 **(pt)**	50.6 **(pt)**	40.7 **(es)**	46.9 **(ptb)**	49.7 **(es)**	44.2 **(pt)**	53.2 (ptb)	53.8 (pt)	–

(Asian Languages)

TRG → ↓ SRC	AR	ID	JA	KO	MS	TH	TL	VI	ZH	ZHT
AR	–	52.0 (ms)	35.6 **(id)**	33.9 **(id)**	52.0 **(id)**	46.1 **(ms)**	41.3 **(id)**	46.7 **(ms)**	31.1 **(id)**	36.4 **(id)**
ID	47.8 (ms)	–	39.1 (ms)	37.5 (ja)	54.9 **(vi)**	51.7 (ms)	47.8 (ms)	52.7 (ms)	34.6 (ms)	41.5 (ms)
JA	29.8 (ko)	35.5 (ko)	–	46.7 (zh)	33.9 (id)	37.9 (ko)	33.7 (ko)	35.5 (ko)	46.9 (ko)	33.1 (ko)
KO	29.7 (ja)	35.2 (ja)	45.8 (zh)	–	34.2 (id)	38.1 (ja)	32.4 (ja)	33.7 (ja)	47.2 (ja)	32.9 (ja)
MS	48.5 (id)	53.8 **(ar)**	39.8 (id)	37.0 (id)	–	53.2 (id)	48.7 (id)	53.4 (id)	35.1 (id)	42.9 (id)
TH	39.4 **(ms)**	48.5 (ms)	36.1 (ko)	36.8 (ja)	47.9 (id)	–	40.5 **(id)**	44.3 **(ms)**	31.4 (id)	34.7 **(ms)**
TL	40.8 **(id)**	52.5 (ms)	37.5 (ko)	36.7 **(ja)**	50.8 (id)	46.5 (ms)	–	47.0 **(ms)**	32.3 **(id)**	36.5 **(ms)**
VI	42.5 **(ms)**	53.0 (ms)	36.8 (ko)	35.4 (ja)	52.8 (id)	48.6 **(ms)**	43.6 **(ms)**	–	31.6 (ms)	37.0 **(ms)**
ZH	26.9 **(zht)**	34.3 (ms)	47.4 (ko)	47.6 (ja)	32.3 (id)	35.9 **(ja)**	30.8 **(ko)**	32.6 **(zht)**	–	33.8 (ja)
ZHT	35.9 **(id)**	44.5 (ms)	40.8 (zh)	38.5 (zh)	44.0 (id)	40.6 **(id)**	36.8 **(id)**	40.5 **(ms)**	35.0 (ja)	–

In order to investigate how much of an improvement in pivot translation performance can be achieved by using non-English pivot languages instead of an English pivot, we calculated the difference in BLEU scores for all 188 non-English language pairs, where the non-English pivot language improved translation quality. Table VII summarizes the average, minimal, and maximal gains in BLEU scores for the respective pivot language translation experiments. The pivot languages are sorted according to the highest average increase in translation performance, and the amount of improved language pairs are given in parentheses. In total, an average gain of 2.2 BLEU points was obtained for the investigated language pairs. The highest gains (13.3/11.4 BLEU points) were achieved for the Japanese/Korean pivots when translating Korean/Japanese into Chinese, respectively.

If we had to select a single pivot languages for all translation directions, however, English seems to be the best choice. Figure 2 lists the average BLEU score differences of the respective non-English pivot towards the English pivot translation tasks.

2.4. Training Data Size Dependency

In order to investigate the dependency between the best pivot language selection and the amount of available training resources, we repeated the pivot translation experiments described in the previous sections for SMT models trained on 10K sentence subsets (BTEC$_{10k}$) randomly extracted from the BTEC$_{80k}$ corpora.

The results showed that 86.4% of the pivot language selections are identical for the small (10K) and large (80K) training data conditions. For the remaining 63 out of

Table VI. Oracle Pivot Language Distribution (BTEC_{80K})

(All Languages)

PVT	usage (%)	
EN	232	(50.2)
PT	40	(8.7)
PTB	38	(8.2)
ID	37	(8.0)
MS	36	(7.8)
JA	29	(6.3)
KO	21	(4.5)
ES	19	(4.1)
NL	5	(1.1)
ZH	4	(0.9)
ZHT	1	(0.2)

(Indo-European)

PVT	usage (%)	
PT	40	(36.3)
PTB	32	(29.1)
ES	26	(23.7)
NL	10	(9.1)
DE	1	(0.9)
DA	1	(0.9)

(Asian)

PVT	usage (%)	
ID	28	(31.1)
MS	27	(30.0)
JA	15	(16.6)
KO	12	(13.3)
ZH	4	(4.4)
ZHT	2	(2.2)
VI	1	(1.1)
AR	1	(1.1)

Table VII. Gain of Non-English Pivot (BTEC_{80K})

PVT	(oracle)	Gain in BLEU% (80K)		
		avg	min	max
ZH	(4)	4.7	3.2	6.1
JA	(27)	2.5	0.1	13.3
ID	(35)	2.4	0.6	5.4
PT	(31)	2.3	0.3	4.6
PTB	(32)	2.1	0.3	4.9
KO	(19)	1.9	0.1	11.4
MS	(34)	1.8	0.1	3.9
ES	(4)	0.8	0.1	2.4
NL	(2)	0.6	0.5	0.8

462 translation tasks, Table VIII lists how the oracle pivot language selection changed. In the case of the small training datasets, the pivot language is closely related (in terms of direct translation quality) to the source language. However, for larger training datasets, the focus shifts towards closely related target languages (marked in boldface) for the majority (37 out of 63) of the investigated language pairs that are listed in the left part of Table VIII. Therefore, in general, the higher the translation quality of the pivot translation task, the more dependent the selection of the best pivot language is on the system performance of the PVT-TRG task. Moreover, for 18 out of 63 translation tasks, the pivot language changed to English even for tasks where the 10K oracle pivot is closely related to either the source or the target language. The remaining eight translation tasks where the oracle pivot selection depends on the training data size translated mainly from or into Chinese and consist of the more difficult translation tasks investigated in this article. This indicates that languages closely related to either

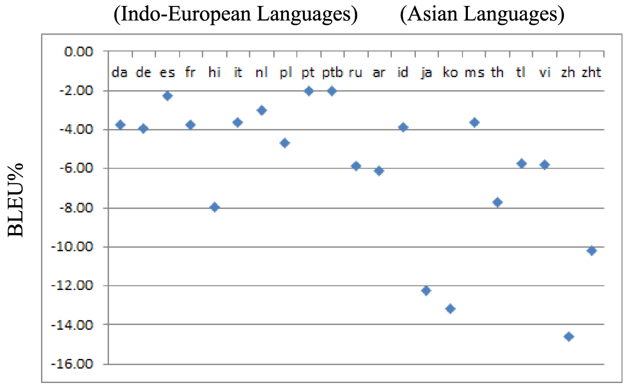

Fig. 2. BLEU score differences between non-English and English pivot.

Table VIII. Oracle Pivot Selection Changes

BTEC$_{10K}$ PVT	BTEC$_{80K}$ PVT	Language Pair
EN	**ID** (11)	DA-**MS**, ES-**MS**, FR-**MS**, IT-**MS**, PL-**MS**, RU-**MS**, TL-**MS**
JA		KO-**MS**, ZH-**MS**
KO		JA-**MS**
PTB		PT-**MS**
KO	**EN** (9)	JA-**DA**, JA-**DE**, JA-**FR**, JA-**IT**, JA-**NL**, JA-**PL**, JA-**RU**, ZH-**ES**, ZH-**IT**
EN	**KO** (8)	DA-**JA**, ES-**JA**, HI-**JA**
FR		PL-**JA**
ID		VI-**JA**
PT		PTB-**JA**, TL-**JA**
PTB		PT-**JA**
EN	**MS** (3)	DA-**ID**
JA		ZH-**ID**
PTB		PT-**ID**
en	**JA** (3)	FR-**KO**, VI-**KO**
ms		ID-**KO**
JA	**PTB** (2)	ZH-**PT**
MS		ID-**PT**
KO	**PT**	JA-**PTB**
ES	EN (18)	RU-IT
FR		IT-JA
ID		TH-ZHT
JA		ZH-TH, ZH-VI
NL		DE-JA
PT		DA-PTB, NL-PTB, FR-JA, NL-JA
PTB		ES-IT, FR-IT, AR-JA, ZHT-IT
ZH		ZHT-TH, ZHT-VI
ZHT		ZH-FR, ZH-TL
EN	ES (2)	FR-ZH
FR		IT-ZH
EN	ID (2)	MS-JA
JA		TH-ZH
KO	JA (2)	ZH-HI, ZH-ZHT
EN	NL (2)	ZH-DE
KO		ZH-AR

the source or the target language are to be preferred as pivot languages for language pairs of low translation quality which augurs well for data availability.

3. INDICATORS OF PIVOT TRANSLATION QUALITY

The diversity of the best pivot languages reported in the last section give rise to the question of what makes a language an effective pivot language for a given language pair.

We investigated the following eight factors (comprised of a total of 45 distinct features) based on the language resources and SMT engines (SRC-PVT, PVT-TRG) used for the pivot translation experiments described in Section 2. The number given in parentheses after each factor indicates the total number of features of the respective factor.

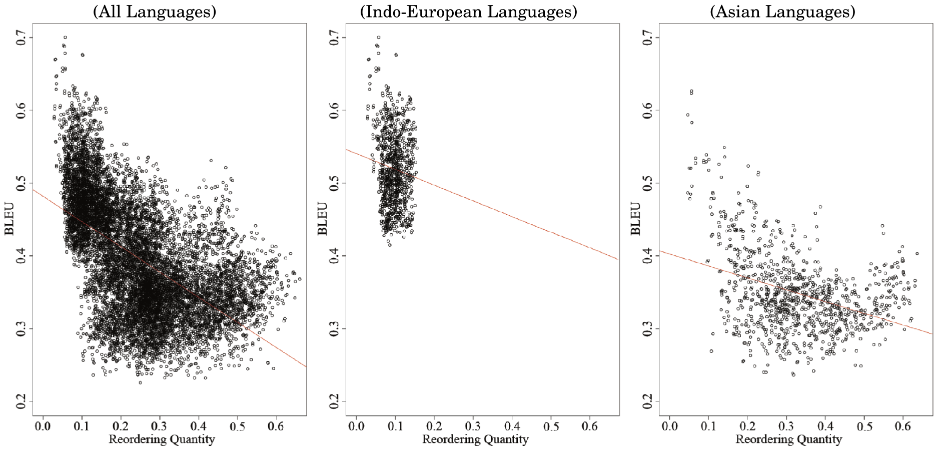

Fig. 3. Linear regression example (reordering quantity).

For SMT engine-related features, both translation directions (SRC-PVT, PVT-TRG) are taken into account.

- *Language Family (2)*. A binary feature verifying whether or not the source and target languages of the SMT engines belong to the same family (as defined in Section 2.1).
- *Vocabulary (15)*. The training data vocabulary size of source and target languages, the ratio of source and target vocabulary sizes, and the overlap between source and target vocabulary.
- *Sentence Length (12)*. The average sentence length (computed in terms of words) of source and target training sets and the ratio of source and target sentence length.
- *Reordering (6)*. The amount and span of word order differences (reordering) in the training data and the *reordering quantity* score, as proposed in Birch et al. [2008].
- *Language Perplexity (4)*. The perplexity of the utilized language models measured on the *dev/eval* datasets.
- *Translation Model Entropy (2)*. The amount of uncertainty involved in choosing candidate translation phrases, as proposed in Koehn et al. [2009].
- *Engine Performance (2)*. The BLEU scores of the respective SMT engines used for the pivot translation experiments.
- *Monotonicity (2)*. The BLEU score difference of a given SMT engine for decoding with and without a reordering model.

The impact of these factors in isolation on the translation performance is measured using linear regression, which models the relationship between a response variable and one or more explanatory variables. Datasets are modeled using linear functions, and unknown model parameters are estimated from the data. In this article, the response variable is defined by the BLEU metric (measuring the pivot translation performance), and the explanatory variables are given by the feature values obtained for each of the respective language pair combinations.

Figure 3 gives an example for a simple linear regression using the *reordering quantity* feature as the explanatory variable for (a) all language pairs, (b) Indo-European languages only, and (c) Asian languages only. The closely grouped plot of the Indo-European languages indicates that word-order differences are quite limited. In contrast, the Asian language plot is quite scattered, and therefore more errors are to be

Table IX. Impact on Translation Performance

Explanatory Variable	R^2 All	R^2 Indo-European	R^2 Asian
all factors	0.8102	0.9106	0.5880
engine performance	0.7438	0.7906	0.5151
translation model entropy	0.4604	0.3669	0.1661
reordering	0.4383	0.4593	0.1806
vocabulary	0.3112	0.3867	0.2389
monotonicity	0.2682	0.0149	0.1323
sentence length	0.1717	0.6052	0.0724
language family	0.1204	0.1280	0.0982
language perplexity	0.0826	0.1100	0.0337

expected for the translation between these languages. Taking into account translations between Indo-European and Asian languages, translation errors due to word-order differences are even more severe, as illustrated in the all-language plot.

The "goodness of fit" of the explanatory variable(s) is calculated using the R^2 coefficient of determination, which is a statistical measure of how well the regression line approximates the real data points. An R^2 of 1.0 indicates that the regression line perfectly fits the data. For the reordering quantity factor, for example, we obtain an R^2 of 0.2385 for all language pairs, which indicates that 23.85% of the differences in translation performance can be explained by this factor.

3.1. Predictive Power of Single Factors

Table IX summarizes the R^2 scores of the multiple linear regression analysis of the respective investigated factors, that is, all features of a given factor are combined and treated as multiple explanatory variables. In total, 81% of the system performance variations can be explained when all investigated factors are taken into account. For Indo-European language pairs, the impact is even larger (91%). However, for Asian language pairs, the investigated factors have much less correlation (an R^2 of 0.5888) with the overall pivot translation translation quality, indicating the difficulty of selecting an appropriate pivot language for translation tasks, including Asian languages.

The impact of each factor on the translation performance is also given in Table IX. The results show that *engine performance* is the most correlated factor, followed by *translation model entropy* and *reordering* when all language combinations are taken into account. *Language family* and *language perplexity* seem to have the least impact on translation performance. However, when applying linear regression on language subsets (only Indo-European vs. only Asian languages), the impact of the factors largely differs. Similar for all language pairs, the *engine performance* factor is most relevant for both Indo-European and Asian language subsets.

For pivot translations between Indo-European languages, *sentence length*, *reordering*, and *vocabulary* are more predictive than the *translation model entropy* factor. Moreover, the *monotonicity* factor obtains the lowest R^2 score, indicating that word-order differences between Indo-European languages occur mainly on a phrase level (*local reordering*) and that only minor gains can be achieved when reordering successive phrases. The high R^2 score for *sentence length* also suggests that the ratio of

Table X. Factor Contribution

Explanatory Variable	R^2 All	R^2 Indo-European	R^2 Asian
all factors	0.8102	0.9106	0.5880
w/o engine performance	0.5621	0.8755	0.3683
w/o language perplexity	0.7734	0.8895	0.5488
w/o sentence length	0.7856	0.8989	0.5501
w/o reordering	0.7958	0.8999	0.5712
w/o vocabulary	0.7961	0.8766	0.5669
w/o translation model entropy	0.8004	0.9024	0.5748
w/o monotonicity	0.8026	0.9024	0.5768
w/o language family	0.8035	0.9022	0.5793

sentence length is an important feature when selecting an appropriate pivot language for closely related languages.

On the other hand, looking at the Asian language pair regression results, the lower R^2 scores underline the large diversity between the Asian languages. Relatively high R^2 scores for *reordering* and *monotonicity* are obtained for Asian languages, indicating that structural differences between the pivot language and the source/target language largely affect the overall pivot translation quality.

3.2. Contribution of Single Factors

Besides the predictive power of each factor, we calculated the R^2 scores of all the factors besides one (*leave-one-out*) in order to investigate the contribution of each factor to the multiple linear regression analysis. In general, the smaller the R^2 score after omitting a given factor, the larger the contribution of this factor to the explanation of the overall translation performance is supposed to be.

The results summarized in Table X show that the largest contribution for all language pairs is obtained for the *engine performance* factor, followed by *language perplexity* and *sentence length*. Interestingly, the *vocabulary* factor contributes as much as the *engine performance* factor for Indo-European languages, but not for Asian languages. This confirms that morphological similarities between highly inflected languages are important for identifying an appropriate pivot language. Moreover, for Indo-European-only and Asian-only language pairs, the omission of any of these factors led to lower R^2 scores, but the difference to the complete factor set is much smaller. This shows the importance of all the investigated features for the task of pivot language selection, especially if largely diverse languages are to be taken into account.

3.3. Translation Direction Dependency

In order to investigate whether the selection of a pivot language depends more on its relationship to the source language or the target language, we carried out a linear regression analysis based on all factors using (a) only source language-related features (*SRC-PVT only*) and (b) only target language-related features (*PVT-TRG only*). The results are summarized in Table XI.

The source language features seem to be more predictive than the target language features. However, for more coherent language pairs, like in the case of Indo-European

Table XI. Source vs. Target Dependency

Explanatory Variable	R^2 All	 Indo-European	 Asian
all factors	0.8102	0.9106	0.5880
SRC-PVT only	0.4923	0.3125	0.2805
PVT-TRG only	0.4732	0.6505	0.2986

languages, the impact on how much language diversity affects pivot translation performance shifts towards the target language-related features. Moreover, limiting the features to either only the source or only the target features leads to a large decrease in the R^2 scores for all language datasets, underlining the importance of both source language-related and target language-related feature sets for identifying an appropriate pivot language for a given language pair.

4. CONCLUSION

In this article, the effects of using non-English pivot languages for translations between 22 Indo-European and Asian languages were compared to the standard English pivot translation approach. The experimental results revealed that English is the best pivot for the majority of the investigated languages, but for 54.8% of language pairs, a non-English pivot language is preferable. On average, a gain of 2.2 BLEU points can be obtained by using non-English pivot languages instead of an English pivot.

In addition, the choice of the best pivot is not symmetric for 21.4% language pairs. Interestingly, for Indo-European-only language pairs, only Indo-European languages are the oracle pivot language, whereas only 65.6% of the oracle pivot languages are Asian languages when translating between Asian languages.

In order to get an idea of what makes a language an effective pivot language for a given language pair, we investigated the impact of eight translation quality indicators. A linear regression analysis showed that 81% of the variation in translation performance differences can be explained by a combination of these factors. The most informative factor in identifying the best pivot language is *engine performance*, that is, the translation quality of the SMT engines used to translate (a) the source input into the pivot language and (b) the pivot language MT output into the target language. In addition, the highest correlation of the investigated factors to pivot translation performance was obtained when both source language-related and target language-related features were combined. The importance of source versus target language features largely depends on the diversity of the investigated language pairs, that is, source language features are preferable for heterogeneous language pairs, whereas the focus shifts towards target language-related features for more coherent language pairs. In addition, the differentiation between Indo-European and Asian languages revealed that the task of identifying a pivot language for new language pairs largely depends on the availability of structurally similar languages.

As future work, we are planning to investigate the importance of the factors analyzed in Section 3 in the selection of pivot languages for new language pairs by applying a machine learning approach, such as support vector machines (SVM) to train discriminative models for the task of predicting a pivot language that achieves the highest translation performance for a given translation task.

In addition, we would like to study the effects of pivot language selection on pivot translation methods other than the *cascading* method utilized here. Although such methods are computationally more expensive, we expect these to improve the overall pivot translation quality further.

REFERENCES

Bogdan Babych, Anthony Hartley, and Serge Sharoff. 2005. Translating from under-resourced languages: Comparing direct transfer against pivot translation. In *Proceedings of the Machine Translation Summit XI*. International Association for Machine Translation, 29–35.

Nicola Bertoldi, Madalina Barbaiani, Marcello Federico, and Roldano Cattoni. 2008. Phrase-based statistical machine translation with pivot languages. In *Proceedings of the 5th International Workshop on Spoken Language Translation (IWSLT)*. 143–149.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2008. Predicting success in machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 745–754.

Trevor Cohn and Mirella Lapata. 2007. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, 728–735.

Adria de Gispert and Jose B. Marino. 2006. Catalan-English statistical machine translation without parallel corpus: Bridging through Spanish. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*. 65–68.

Andrew Finch, Etienne Denoual, Hideo Okuma, Michael Paul, Hirofumi Yamamoto, Keiji Yasuda, Ruiqiang Zhang, and Eiichiro Sumita. 2007. The NICT/ATR speech translation system for IWSLT 2007. In *Proceedings of the 4th International Workshop on Spoken Language Translation (IWSLT)*. 103–110.

Genichiro Kikui, Seiichi Yamamoto, Toshiyuki Takezawa, and Eiichiro Sumita. 2006. Comparative study on corpora for speech translation. *IEEE Trans. Audio Speech Lang. 14*, 5, 1674–1682.

Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 2009. 462 Machine translation systems for Europe. In *Proceedings of the Machine Translation Summit XII*. International Association for Machine Translation, 65–72.

Gregor Leusch, Aurélien Max, Josep Maria Crego, and Hermann Ney. 2010. Multi-pivot translation by system combination. In *Proceedings of the 7th International Workshop on Spoken Language Translation (IWSLT)*. 299–306.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics 29*, 1, 19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Association for Computational Linguistics (ACL)*. 311–318.

Michael Paul. 2008. Overview of the IWSLT 2008 evaluation campaign. In *Proceedings of the 5th International Workshop on Spoken Language Translation (IWSLT)*. 1–17.

Michael Paul, Hirofumi Yamamoto, Eiichiro Sumita, and Satoshi Nakamura. 2009. On the importance of pivot language selection for statistical machine translation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL/HLT)*. Association for Computational Linguistics, 221–224.

Lucia Specia, Najeh Hajlaoui, Catalina Hallet, and Wiler Aziz. 2011. Predicting machine translation adequacy. In *Proceedings of the Machine Translation Summit XIII*. International Association for Machine Translation, 513–520.

Andreas Stolcke. 2002. SRILM - An extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*. 901–904.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Proceedings of the Human Language Technologies (HLT)*. 484–491.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. In *Proceedings of the 45th Association for Computational Linguistics (ACL)*. 856–863.

Received January 2013; revised May 2013; accepted July 2013