Translation Quality Indicators for Pivot-based Statistical MT

Michael Paul and Eiichiro Sumita
National Institute of Information and Communications Technology
MASTAR Project
Kyoto, 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. This paper provides new insights into what factors make a good pivot language and investigates the impact of these factors on the overall pivot translation performance. Pivot-based SMT experiments translating between 22 Indo-European and Asian languages were used to analyze the impact of eight factors (*language family*, *vocabulary*, *sentence length*, *language perplexity*, *translation model entropy*, *reordering*, *monotonicity*, *engine performance*) on pivot translation performance. The results showed that 81% of system performance variations can be explained by these factors.

1 Introduction

The translation quality of statistical machine translation (SMT) approaches heavily depends on the amount and coverage of bilingual language resources available to train 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 data sets are readily available. However, for less frequently used language pairs only a limited amount of bilingual resources are available, if any at all.

In order to overcome language resource limitations, recent research on SMT focuses on the usage of *pivot languages* (de Gispert and Marino, 2006; Utiyama and Isahara, 2007; Wu and Wang, 2007; Bertoldi et al., 2008). 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. For most recent research efforts, *English* is the pivot language of choice due to the richness of available language resources. However, recent research on pivot translation has shown that the usage of non-English pivot languages can improve translation quality for certain language pairs (Paul et al., 2009; Leusch et al., 2010).

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 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.

This paper differs from previous research in the following aspects: we focus on the framework of *pivot translation*, where a target language translation of a source language input is obtained through an intermediate (*pivot*) language, investigate what factors make a good pivot language and what impact these factors have on the overall translation quality of language pairs not only including Indo-Euopean languages, but also a large variety of Asian languages. In Section 2, we report on pivot-based SMT experiments translating between 22 Indo-European as well as Asian languages in order to provide new insights into how much language diversity affects the translation performance of pivot translation approaches. In Section 3, eight factors (*language family*, *vocabulary*, *sentence length*, *language perplexity*,

[1] LDC: http://www.ldc.upenn.edu, ELRA: http://www.elra.info

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 811–818,
Chiang Mai, Thailand, November 8 – 13, 2011.

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, various coupling strategies like *cascading*, *phrase-table composition*, and *pseudo-corpus generation* have been proposed. For the experiments reported in this paper, we utilized the *cascading* approach because it is computational less expensive, but still performs comparably well compared to the other, more sophisticated pivot translation approaches. Pivot translation using the *cascading* approach requires two translation 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.

The importance of translation quality factors in pivot translation 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 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), *Semitic* (ar), *Austronesian* (id,ms,tl), *Tai* (th), *Mon-khmer* (vi), and *Sinitic* (zh,zht) languages. The corpus statistics are summarized in Table 1, where *Voc* specifies the vocabulary size and *Len* the average sentence length of the respective data sets. These languages differ largely in word order (*Order*: subject-object-verb (*SOV*), subject-verb-object (SVO), verb-subject-object (*VSO*)), segmentation unit (*Unit*: phrase, word, none), and degree of inflection (*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, the corpora were preprocessed using language-specific word-segmentation tools for languages that do not use white-space to separate word/phrase tokens (ja,ko,th,zh,zht). For all other languages, simple tokenization tools were applied. All data sets were case-sensitive with punctuation marks preserved.

Table 1: Language Resources

(European Languages)

Language		Voc	Len	Order	Unit	Inflection
Danish	da	26.5k	7.2	SVO	word	high
German	de	25.7k	7.1	SVO	word	high
English	en	15.4k	7.5	SVO	word	moderate
Spanish	es	20.8k	7.4	SVO	word	high
French	fr	19.3k	7.6	SVO	word	high
Hindi	hi	33.6k	7.8	SOV	word	high
Italian	it	23.8k	6.7	SVO	word	high
Dutch	nl	22.3k	7.2	SVO	word	high
Polish	pl	36.4k	6.5	SVO	word	high
Portuguese	pt	20.8k	7.0	SVO	word	high
Brazilian Portuguese	ptb	20.5k	7.0	SVO	word	high
Russian	ru	36.2k	6.4	SVO	word	high

(Asian Languages)

Language		Voc	Len	Order	Unit	Inflection
Arabic	ar	47.8k	6.4	VSO	word	high
Indonesian	id	18.6k	6.8	SVO	word	high
Japanese	ja	17.2k	8.5	SOV	none	moderate
Korean	ko	17.2k	8.1	SOV	phrase	moderate
Malay	ms	19.3k	6.8	SVO	word	high
Thai	th	7.4k	7.8	SVO	none	light
Tagalog	tl	28.7k	7.4	VSO	word	high
Vietnamese	vi	9.9k	9.0	SVO	phrase	light
Chinese	zh	13.3k	6.8	SVO	none	light
Taiwanese	zht	39.5k	5.9	SVO	none	light

The language resources were randomly split into three subsets for the evaluation of translation quality (*eval*, 1000 sentences), the tuning of the SMT model weights (*dev*, 1000 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.

For the training of the SMT models, standard

Table 3: Oracle Pivot Translation Quality (BLEU)

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

Table 2: Pivot Language Dependency

(European Languages)

PVT	BLEU (%) *low*		*high*
da	24.2	∼	60.2
de	25.1	∼	61.8
en	26.2	∼	67.7
es	24.9	∼	70.0
fr	24.5	∼	62.3
hi	23.7	∼	53.1
it	23.7	∼	65.8
nl	26.2	∼	61.6
pl	25.2	∼	56.8
pt	25.8	∼	68.9
ptb	24.9	∼	68.8
ru	22.6	∼	56.9

(Asian Languages)

PVT	BLEU (%) *low*		*high*
ar	23.3	∼	57.1
id	25.6	∼	57.9
ja	26.8	∼	59.4
ko	25.7	∼	58.3
ms	25.5	∼	57.3
th	23.3	∼	52.3
zht	23.7	∼	44.7
vi	23.6	∼	55.3

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 was 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 was used. For the evaluation of translation quality, we applied the standard automatic evaluation metric BLEU which calculates the geometric mean of n-gram precision by 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 1 (best) (Papineni et al., 2002). For our experiments, single translation references were used.

Table 2 summarizes the BLEU score ranges of all pivot translation experiments obtained for a given pivot language. The results show a large variation in BLEU scores for all pivot languages indicating that the best pivot choice largely depends on the respective source and target language. For European pivot languages, the best language combination scores are in general much higher than the ones obtained for Asian pivot languages.

Table 3 lists the highest BLEU scores of the pivot translation experiments obtained for all language pair combinations. The pivot language achieving

Table 4: Changes in Pivot Selection for Non-English European and Asian Language Pairs (BLEU)

(Non-English European Language Pairs)

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

(Asian Language Pairs)

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

the highest scores (*oracle pivot*) for translating the source (S) language into the target (T) language are given in parantheses. 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 49.8% (230 out of 462) of the cases, a non-English pivot language, mainly *Portuguese*, *Brazilian Portuguese*, *Malay*, *Indonesian*, *Japanese*, *Korean*, is preferable. For languages that are closely related like Portuguese vs. Brazilian Portuguese and Malay vs. 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 privot translation experiments, respectively. Moreover, *Japanese* is the dominant pivot language when translating from Korean into an other 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 overal pivot translation quality than pivot languages related to the target language.

Interestingly, for European-only language pairs, only 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 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, 65.6% (59 out of 90) of the oracle pivot languages are Asian languages. The Spanish (Chinese) oracle pivot languages for translations between Portuguese and Brazilian Portuguese (Japanese and Korean) also stresses the importance of language relatedness.

In order to investigate the dependency of pivot language selection and language families further, Table 4 summarizes the BLEU scores of pivot translations between only (a) non-English European and (b) Asian language pairs. The results of the European-only language pairs in the table on the left confirm the findings of Table 3. *Portuguese* and *Brazilian Portuguese* are still the dominant pivot languages for non-English 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 3.

Table 5 summarizes the percentages for the language pairs where the respective pivot language achieved the highest automatic evaluation score for the pivot translation experiments summarized in Table 3 (all language pairs) and Table 4 (non-English 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.

In order to investigate how much improvement

Table 5: Oracle Pivot Language Distribution

(All Language Pairs)

PVT	usage (%)		PVT	usage (%)	
en	232	(50.2)	ko	21	(4.5)
pt	40	(8.7)	es	19	(4.1)
ptb	38	(8.2)	nl	5	(1.1)
id	37	(8.0)	zh	4	(0.9)
ms	36	(7.8)	zht	1	(0.2)
ja	29	(6.3)			

(Non-English European Language Pairs)

PVT	usage (%)		PVT	usage (%)	
pt	40	(36.3)	nl	10	(9.1)
ptb	32	(29.1)	de	1	(0.9)
es	26	(23.7)	da	1	(0.9)

(Asian Language Pairs)

PVT	usage (%)		PVT	usage (%)	
id	28	(31.1)	zh	4	(4.4)
ms	27	(30.0)	zht	2	(2.2)
ja	15	(16.6)	vi	1	(1.1)
ko	12	(13.3)	ar	1	(1.1)

Table 6: Gain of non-English Pivot Languages

PVT	(oracle)	Gain in BLEU (%) 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

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 6 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 parantheses. In total, an average gain of 2.2 BLEU points were obtained for the investigated language pairs. The highest gains (13.4/11.4 BLEU points) were achieved for the Japanese/Korean pivots when translating Korean/Japanese into Chinese, respectively.

3 Indicators of Pivot Translation Quality

The diversity of the pivot language selection reported in the last section rises the question of what makes a language a good 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 where the total number of features of each factor is given in brackets. For SMT-engine-related features, both translation directions (SRC-PVT, PVT-TRG) are taken into account.

- *language family* [2]: a binary feature verifying whether the source and target languages of the SMT engines belong to the same family or not.
- *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 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]: perplexity of the utilized language models measured on the *dev*/*eval* data sets.
- *translation model entropy* [2]: 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 engine 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 the above 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. Data sets are modeled using linear functions and unknown model parameters are estimated from the data. In this paper, 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 1 gives an example for a simple linear regression using the *reordering quantity* feature as the explanatory variable for (a) all language pairs, (b) European languages only, and (c) Asian languages only. The “goodness of

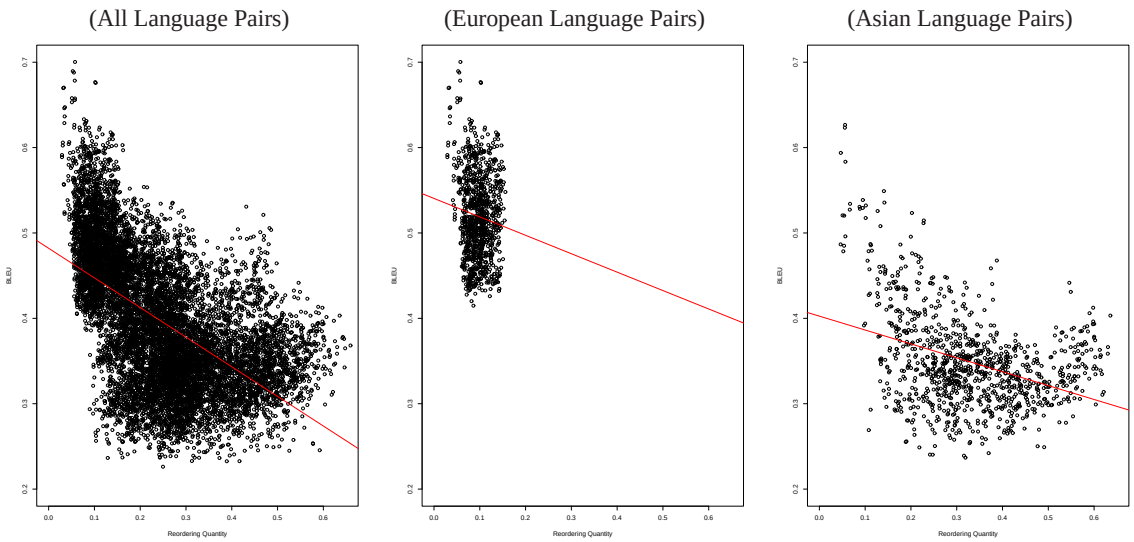

Figure 1: Linear Regression Example (*Reordering Quantity*)

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 *translation model entropy* factor, for example, we obtain an R^2 of 0.4604 for all language pairs, which indicates that 46.04% of the differences in translation performance can be explained by this factor.

3.1 Predictive Power of Single Factors

Table 7 summarizes the R^2 scores of the multiple linear regression analysis of the respective investigated factors, i.e. 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 European language pairs, the impact is even larger (91%). However, for Asian language pairs, the investigated factors have much less correlation (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 7. 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* seems to have the least impact on translation performance. However, when applying linear regression on language subsets (only European vs. only Asian languages), the impact of factors largely differs. Similar to all language pairs, the *engine performance* factor is most relevant for both European and Asian language subsets.

Table 7: Impact on Translation Performance

Explanatory Variable	R^2 All	R^2 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

For pivot translations between 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 European languages occur mainly on the 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 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 lan-

Table 8: Factor Contribution

Explanatory Variable	R^2 All	European	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

guages, indicating that structural differences between the pivot language and the source/target language largely affects the overal 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 on the explanation of the overall translation performance is supposed to be.

The results summarized in Table 8 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 European languages, but not for Asian languages. This confirms that morphological similarities between highly inflected languages are important to identify an appropriate pivot language. Moreover, for European-only and Asian-only language pairs, the omission of any of these factors led to lower R^2 scores, but the difference towards 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 languages of large diversity 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 towards 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 9.

Table 9: Source vs. Target Language Dependency

Explanatory Variable	R^2 All	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

In order to distinguish between languages of large diversity, 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 European languages, the impact on how much language diversity affects pivot translation performance shifts towards target-language-related features. However, the restriction to either the source or the target features leads to a large decrease in the R^2 scores for all language data sets, underlining the importance of both source-language-related and target-language-related feature sets to identify an appropriate pivot language for a given language pair.

4 Conclusion

We investigated the impact of eight translation quality indicators for the task of pivot translation between 22 languages covering a large diversity of language characteristics. A linear regression analysis showed that 81% of the variation in translation performance differences can be explained by the combination of these factors. The most informative factor in identifying the best pivot language is *engine performance*, i.e., the translation quality of the SMT engines used to translate (a) the source input into the intermediate language and (b) the intermediate language MT output into the target language. In addition, the highest correlation of the investigated factors towards pivot translation performance was obtained when both source-language-related and target-language-related features were combined. The importance of source vs. target language features largely depends on the diversity of the investigated language pairs, i.e., source language features are preferable for heterogenous language pairs whereas the focus shifts towards target-language-related features for more coherent language pairs. In addition, the differentiation between European and Asian languages revealed that the task of identifying a pivot lan-

guage 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 algorithm like *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.

References

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)*, pages 143–149, Hawaii, USA.

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)*, pages 745–754, Honolulu, Hawaii.

Adria de Gispert and Jose B. Marino. 2006. Catalan-english statistical machine translation without parallel corpus: bridging through spanish. In *Proceedings of 5th International Conference on Language Resources and Evaluation (LREC)*, pages 65–68, Genoa, Italy.

Genichiro Kikui, Seiichi Yamamoto, Toshiyuki Takezawa, and Eiichiro Sumita. 2006. Comparative study on corpora for speech translation. *IEEE Transactions on Audio, Speech and Language*, 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*, Ottawa, Canada.

Gregor Leusch, Aurélien Max, Josep Maria Crego, and Hermann Ney. 2010. Multi-Pivot Translation by System Combination. In *Proceedings of 7th International Workshop on Spoken Language Translation (IWSLT)*, pages 299–306, Paris, France.

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 Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, USA.

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)*, pages 221–224, Boulder, USA.

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

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

Hua Wu and Haifeng Wang. 2007. Pivot Language Approach for Phrase-Based Statistical Machine Translation. In *Proceedings of 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 856–863, Prague, Czech Republic.