462 Machine Translation Systems for Europe

Philipp Koehn
School of Informatics
University of Edinburgh
pkoehn@inf.ed.ac.uk

Alexandra Birch
School of Informatics
University of Edinburgh
a.birch@sms.ed.ac.uk

Ralf Steinberger
Joint Research Centre
European Commission
Ralf.Steinberger@jrc.it

Abstract

We built 462 machine translation systems for all language pairs of the Acquis Communautaire corpus. We report and analyse the performance of these system, and compare them against pivot translation and a number of system combination methods (multi-pivot, multi-source) that are possible due to the available systems.

1 Introduction

While its many languages pose a challenge to the economic and cultural integration of Europe, it also provides an excellent test bed for machine translation research. The official European languages come from a variety of language families and vary along many linguistic dimensions (morphology, word order, etc.). Some are closely related (such as Portuguese and Spanish), while some are very distant (such as Finnish and German). The data comes from seven language families, two of which are not Indo-European as shown in Table 1.

In this paper, we will describe how the JRC-Acquis corpus was used to build machine translation systems for 462 language pairs. This allows us to analyse the challenges of the different language pairs by carrying out a regression study to determine the main factors for differences in performance.

We also compare the direct translation systems against pivot translation through English and French. Surprisingly, translation performance is often better when pivoting through English, while it decreases for any other pivot language.

The availability of translation systems for so many language pairs also allows us to employ a system combination method to combine systems in a novel way. We report on multi-pivot and multi-source translation, which leads to gains of in the area of 0.5-1 %BLEU and 2-5 %BLEU, respectively.

Indo-European			
Germanic		Slavic	
Swedish	sv	Polish	pl
German	de	Slovak	sk
Dutch	nl	Czech	cs
Danish	da	Slovene	sl
English	en	Bulgarian	bg
Romance		Baltic	
French	fr	Lithuanian	lt
Portuguese	pt	Latvian	lv
Italian	it	Greek	
Spanish	es	Greek	el
Romanian	ro		
Non Indo-European			
Finno-Ugric		Semitic	
Finnish	fi	Maltese	mt
Estonian	et		
Hungarian	hu		

Table 1: Acquis languages in their language families

2 Acquisition of the Corpus

The corpus used to develop the 462 MT systems is the JRC-Acquis (Steinberger et al., 2006), a multilingual parallel corpus consisting of altogether over 1 billion words (almost 50 million words per language; see Table 2). To our knowledge, it is the largest parallel corpus in so many languages. Apart from its size, the most special and useful feature of the JRC-Acquis is the fact that includes a number of under-resourced languages and language pairs.

The JRC-Acquis is to a large extent based on the Acquis Communautaire, which is the body of common rights and obligations which have been adopted by all European Union (EU) Member States. For the texts to become legally binding in the EU Member States, they had to be translated into the 23 official EU languages. The Irish version (the 23rd official EU language), however, is not yet available.

As text types, the corpus contains documents on political objectives, treaties, declarations, resolutions, agreements, EU legislation, and more. It is thus mostly of a legal nature, but as the law and the agreements cover most domains of life, it does contain vocabulary from a wide range of subject fields,

Language ISO code	N° of texts	Text body			Signatures	Annexes	Total N° words (text + signatures + annexes):
		Total N° words	Total N° characters	Average n° words	Total N° words	Total N° words	
bg	10956	15208341	98288769	1388.13	2052773	12885853	30146967
cs	21438	22843279	148972981	1065.55	7225300	16763733	46832312
da	23624	31459627	213468135	1331.68	2629786	16855213	50944626
de	23541	32059892	232748675	1361.87	2542149	16327611	50929652
el	23184	36453749	**239583543**	1572.37	2973574	16459680	55887003
en	23545	34588383	210692059	1469.03	3198766	17750761	55537910
es	23573	38926161	238016756	1651.3	**3490204**	19716243	**62132608**
et	23541	24621625	192700704	1045.9	1336051	14995748	40953424
fi	23284	24883012	212178964	1068.67	2677798	12547171	40107981
fr	**23627**	**39100499**	234758290	1654.91	3021013	**19978920**	62100432
hu	22801	28602380	213804614	1254.44	2529488	15056496	46188364
it	23472	35764670	230677013	1523.72	3120797	18331535	57217002
lt	23379	26937773	199438258	1152.22	2436585	15018484	44392842
lv	22906	27592514	196452051	1204.6	1673124	15437969	44703607
mt	10545	20926909	128906748	**1984.53**	1336042	15620611	37883562
nl	23564	35265161	231963539	1496.57	3039580	18467115	56771856
pl	23478	29713003	214464026	1265.57	2513141	17027393	49253537
pt	23505	37221668	227499418	1583.56	3034308	19350227	59606203
ro	6573	9186947	60537301	1397.68	514296	11185842	20887085
sk	21943	26792637	179920434	1221.01	3227852	16190546	46211035
sl	20642	27702305	178651767	1342.04	3103193	16837717	47643215
sv	20243	29433037	199004401	1453.99	2575771	14965384	46974192
Total	**463364**	**635283572**	**4282728446**	**1385.88**	**60251591**	**357770252**	**1053305415**

Table 2: Size of the JRC-Acquis Communautaire corpus

including human and veterinary medicine, the environment, fishery and agriculture, banking and commerce, transport, energy, science, social and religious issues, geography and more.

The corpus was compiled by crawling documents from the EU's Eur-Lex website[1] and by then selecting those documents that existed in at least ten languages, of which at least three had to be languages from the states that joined the EU in 2004

Each JRC-Acquis document has been manually classified according to the multilingual EUROVOC thesaurus[2], which distinguishes over 6,000 subject domain classes.

3 Data Preparation

Training a statistical machine translation system requires a sentence-aligned parallel corpus to build the model, as well as tuning and test sets to optimize and assess its performance.

3.1 Training Data

The JRC-Acquis corpus provides already the data in the form required for training a statistical machine translation system, and very little additional processing is needed.

It is hard to quantify how much training data is needed to achieve a minimum level of performance. This depends on the expansiveness of the domain and the language pair. Typically, tens of millions of words give decent performance: For instance, systems trained on the 30–40 million word Europarl corpus are competitive with commercial systems, typically better on this domain and even close in performance when translating related material such as news stories (Callison-Burch et al., 2008).

The JRC-Acquis corpus is large enough to expect decent translation performance within its domain, but on the other hand, the domain is also very specific. Translation models trained on such legal texts do not necessarily perform well on other domains.

3.2 Tuning and Test Sets

Since we develop machine translation systems for 462 language pairs, we wanted to have a common tuning and testing environment. Hence, we extracted from part of the corpus subset where sentences are aligned one-to-one across all languages.

First, we identified all documents that exist for all languages. This is a set of 5383 documents. From

[1] http://eur-lex.europa.eu/

[2] http://europa.eu/eurovoc/

	Target Language																					
	en	bg	de	cs	da	el	es	et	fi	fr	hu	it	lt	lv	mt	nl	pl	pt	ro	sk	sl	sv
en	–	40.5	46.8	52.6	50.0	41.0	55.2	34.8	38.6	50.1	37.2	50.4	39.6	43.4	39.8	52.3	49.2	55.0	49.0	44.7	50.7	52.0
bg	61.3	–	38.7	39.4	39.6	34.5	46.9	25.5	26.7	42.4	22.0	43.5	29.3	29.1	25.9	44.9	35.1	45.9	36.8	34.1	34.1	39.9
de	53.6	26.3	–	35.4	43.1	32.8	47.1	26.7	29.5	39.4	27.6	42.7	27.6	30.3	19.8	50.2	30.2	44.1	30.7	29.4	31.4	41.2
cs	58.4	32.0	42.6	–	43.6	34.6	48.9	30.7	30.5	41.6	27.4	44.3	34.5	35.8	26.3	46.5	39.2	45.7	36.5	43.6	41.3	42.9
da	57.6	28.7	44.1	35.7	–	34.3	47.5	27.8	31.6	41.3	24.2	43.8	29.7	32.9	21.1	48.5	34.3	45.4	33.9	33.0	36.2	47.2
el	59.5	32.4	43.1	37.7	44.5	–	54.0	26.5	29.0	48.3	23.7	49.6	29.0	32.6	23.8	48.9	34.2	52.5	37.2	33.1	36.3	43.3
es	60.0	31.1	42.7	37.5	44.4	39.4	–	25.4	28.5	51.3	24.0	51.7	26.8	30.5	24.6	48.8	33.9	57.3	38.1	31.7	33.9	43.7
et	52.0	24.6	37.3	35.2	37.8	28.2	40.4	–	37.7	33.4	30.9	37.0	35.0	36.9	20.5	41.3	32.0	37.8	28.0	30.6	32.9	37.3
fi	49.3	23.2	36.0	32.0	37.9	27.2	39.7	34.9	–	29.5	27.2	36.6	30.5	32.5	19.4	40.6	28.8	37.5	26.5	27.3	28.2	37.6
fr	64.0	34.5	45.1	39.5	47.4	42.8	60.9	26.7	30.0	–	25.5	56.1	28.3	31.9	25.3	51.6	35.7	61.0	43.8	33.1	35.6	45.8
hu	48.0	24.7	34.3	30.0	33.0	25.5	34.1	29.6	29.4	30.7	–	33.5	29.6	31.9	18.1	36.1	29.8	34.2	25.7	25.6	28.2	30.5
it	61.0	32.1	44.3	38.9	45.8	40.6	57.2	25.0	29.7	52.7	24.2	–	29.4	32.6	24.6	50.5	35.2	56.5	39.3	32.5	34.7	44.3
lt	51.8	27.6	33.9	37.0	36.8	26.5	41.0	34.2	32.0	34.4	28.5	36.8	–	40.1	22.2	38.1	31.6	31.6	29.3	31.8	35.3	35.3
lv	54.0	29.1	35.0	37.8	38.5	29.7	42.7	34.2	32.4	35.6	29.3	38.9	38.4	–	23.3	41.5	34.4	39.6	31.0	33.3	37.1	38.0
mt	72.1	32.2	37.2	37.9	38.9	33.7	48.7	26.9	25.8	42.4	22.4	43.7	30.2	33.2	–	44.0	37.1	45.9	38.9	35.8	40.0	41.6
nl	56.9	29.3	46.9	37.0	45.4	35.3	49.7	27.5	29.8	43.4	25.3	44.5	28.6	31.7	22.0	–	32.0	47.7	33.0	30.1	34.6	43.6
pl	60.8	31.5	40.2	44.2	42.1	34.2	46.2	29.2	29.0	40.0	24.5	43.2	33.2	35.6	27.9	44.8	–	44.1	38.2	38.2	39.8	42.1
pt	60.7	31.4	42.9	38.4	42.8	40.2	60.7	26.4	29.2	53.2	23.8	52.8	28.0	31.5	24.8	49.3	34.5	–	39.4	32.1	34.4	43.9
ro	60.8	33.1	38.5	37.8	40.3	35.6	50.4	24.6	26.2	46.5	25.0	44.8	28.4	29.9	28.7	43.0	35.8	48.5	–	31.5	35.1	39.4
sk	60.8	32.6	39.4	48.1	41.0	33.3	46.2	29.8	28.4	39.4	27.4	41.8	33.8	36.7	28.5	44.4	39.0	43.3	35.3	–	42.6	41.8
sl	61.0	33.1	37.9	43.5	42.6	34.0	47.0	31.1	28.8	38.2	25.7	42.3	34.6	37.3	30.0	45.9	38.2	44.1	35.8	38.9	–	42.7
sv	58.5	26.9	41.0	35.6	46.6	33.3	46.6	27.4	30.9	38.9	22.7	42.0	28.2	31.0	23.7	45.6	32.2	44.2	32.7	31.3	33.5	–

Table 3: Translation performance as measured in %BLEU for all 462 language pairs

these, we selected a subset of 270 documents to extract tuning and test sets.

We rely on the word alignment provided along with the JRC-Acquis corpus to match up the sentences. There are several strategies to match up sentences across all languages in a multi-lingual corpus: (1) We extract those sentences that are aligned 1-1 across all languages. (2) We allow many-to-many alignments between sentences and extract minimal sets of sentences for each language that are aligned between each other but not other sentences. (3) We choose one language as pivot language and find matches in all the other languages based on the alignment to the pivot languages.

While we would have preferred one of the first two methods, they were not practical. Extracting only 1-1 sentence alignment yielded mostly only very short sentences, and extracting sets of sentences under transitive closure of the sentence alignment very often matched up entire documents. But either too short or too long sentences do not serve well as tuning and test sets.

So, only the last option was practical, and we selected English as pivot language. This gave us a set of 12,322 sentences aligned across all 22 languages of the corpus. We split this set into three parts, a tuning set for parameter optimization, a development test set for experimentation and a final test set to report translation performance.

Since these sets contain many short and a few very long sentences, we reduced the tuning set further, by requiring that all sentences are between 8 and 60 words long. This left us with a tuning set of 1944 sentences per language.

3.3 Training

For the development of the translation system, we used the defaults of the Moses toolkit (Koehn et al., 2007) with the following additional settings: maximum sentence length 80 words, bi-directional msd reordering model, 5-gram language model.

4 Performance

A thorough evaluation of the translation quality of translation systems for 462 different language pairs would be a daunting task, so we rely on automatic metrics. The most commonly used metric in statistical machine translation is the BLEU score (Papineni et al., 2002). Table 3 shows the scores for all the 462 translation systems.

Performance varies widely for the different language pairs. For instance, French–English translation (64.0) is better than Bulgarian–Hungarian (24.7).

French Input	French–English MT System	English Reference Translation
LE CONSEIL DE LA COMMUNAUTÉ ÉCONOMIQUE EUROPÉENNE,	The Council of the European Economic Community,	THE COUNCIL OF THE EUROPEAN ECONOMIC COMMUNITY,
considérant que l'instauration d'une politique commune des transports comporte entre autres l'établissement de règles communes applicables aux transports internationaux de marchandises par route, exécutés au départ ou à destination du territoire d'un état membre, ou traversant le territoire d'un ou plusieurs états membres;	Whereas the establishment of a common transport policy entails, inter alia, laying down common rules applicable to the international carriage of goods by road, to or from the territory of a Member State or passing across the territory of one or more Member States;	Whereas the adoption of a common transport policy involves inter alia laying down common rules for the international carriage of goods by road to or from the territory of a Member State or passing across the territory of one or more Member States;
Les transports faisant l'objet de l'annexe II ne devront plus être soumis à un régime de contingentement. Ils pourront cependant demeurer sujets à autorisation pour autant qu'aucune restriction quantitative n'en résulte ; chaque état membre devra en pareil cas veiller à ce qu'une décision intervienne dans les cinq jours suivant l'introduction de la demande d'autorisation.	The carriage listed in Annex II shall not be subject to a quota system. They may, however, remain subject to authorisation provided that any quantitative restrictions arises; each Member State may in such cases ensure that a decision is taken within five days of submission of the application for authorisation.	The types of carriage listed in Annex II shall no longer be subject to a quota system. They may, however, remain subject to authorisation provided no quantitative restriction is involved ; in such case Member States shall ensure that decisions on applications for authorisation are given within five days of receipt.

Figure 1: System output for French–English on the beginning of the test set used in the evaluation.

Compared to BLEU scores for other training scenarios and test sets, these numbers are fairly high, indicating that the systems work very well on the domain of European law. European law is a very well-defined domain that does not allow a lot of variation in translation, so it is possible for a statistical system to pick up on the correct words and phrase to use. See also Figure 1 for sample output of the French–English system.

To get a better sense of the translation performance, we wanted to compare the translation systems against a translation system trained on the Europarl corpus. On the news set of the 2008 ACL Workshop on SMT, the Acquis system achieved a score of 11.6, while the Europarl system scored 15.7, for German–English.

5 Analysis

The Acquis corpus comprises of a very large number and variety of language pairs. The breadth of data conditions make this corpus ideal for performing experiments which investigate language pair characteristics and the effect they have on translation. This allows us to provide a wide perspective on the challenges facing machine translation and provide strong motivation for further research on important factors.

5.1 Factors

In this paper we extend and enhance previous research (Birch et al., 2008) by using a much larger number of language pairs and by investigating a new factor, translation model entropy, which captures the amount of uncertainty present when choosing candidate translation phrases. We have also included corpus size as a factor as the amount of Acquis data per language pair can vary by a factor of four. The following characteristic form part of our analysis:

Morphological Complexity The morphological complexity of the language pairs is an important factor influencing translation performance. A simple method of measuring this complexity is to use vocabulary size. Vocabulary size is strongly influenced by the number of word forms for number, case, tense etc. and it is also affected by the number of agglutinations in the language.

Reordering We measure word order differences between languages by assuming that reordering is a binary process between two blocks that are adjacent in the source and whose order is reversed in the target. Word alignments are extracted using GIZA++ and then merged using the grow-final-diag algorithm. Reorderings are then extracted using the shift-reduce algorithm (Galley and Manning, 2008). These reorderings are used to extract a sentence level metric, RQuantity (Birch et al., 2008), which is the sum of the widths of all the reorderings on the source side, normalized by the length of the source sentence. This measure is averaged over a random sample of 2000 training sentences to get the corpus RQuantity.

Language Relatedness Languages which are closely related could share morphological forms which might be captured reasonably well in translation models. We include a measure of language relatedness to take this into account. Lexicostatistics provides a quantitative measure of language relatedness by comparing lists of lexical cognates. We use the data from Dyen et al. (1992) who developed a list of 200 meanings for 84 Indo-European languages. Non-Indo-European languages are assigned a minimal score.

Corpus Size The sizes of the parallel corpora varies considerably and we take this into account by using the number of sentence pairs used for training the systems as a factor.

These factors, together with translation model entropy, which is described in the following section, form the basis of our analysis of the Acquis corpus.

5.2 Translation Model Entropy

Translation model entropy captures the amount of uncertainty involved in choosing candidate translation phrases. Some language pairs can cause translation models to have higher entropy because there is no clear correlation between concepts in one language and the other. Translating from morphologically poor languages into richer languages could also lead to high entropy models, due to the lack of certainty as to which word form to choose. To the best of our knowledge, this important characteristic of translation has not been investigated until now.

The entropy of the translation model is calculated on the test sets. We perform a search through all possible segmentations of the source sentence. Each segmentation, or source phrase, has a set of possible translations in the phrase table T. The entropy H for a source phrase s is calculated as follows:

$$H(s) = -\sum_{t \in T} p(t|s) * log_2 p(t|s)$$

The search returns the set of segments which covers the source sentence with the lowest average entropy per word. Longer phrases tend to have lower entropy with fewer phrase table entries and more of the probability mass concentrated on fewer alternatives, and they will tend to be selected when present in the phrase table. This is similar to the actual translation process.

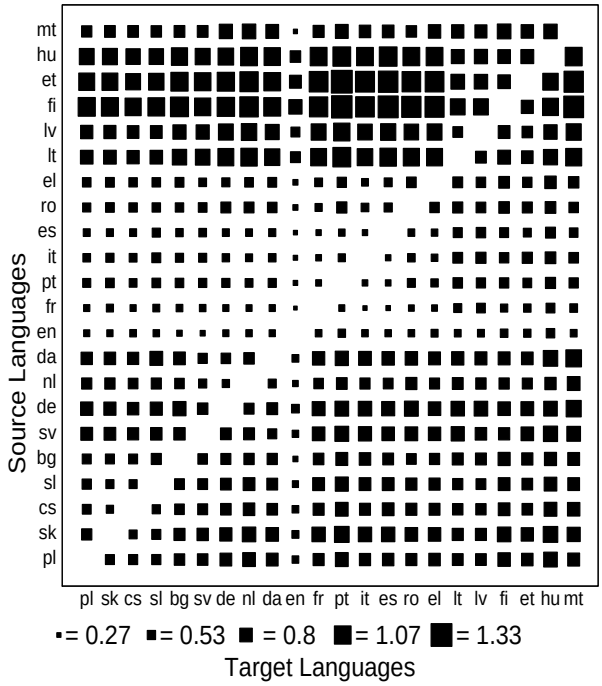

Figure 2: Matrix of Translation Model's Entropies

Figure 2 shows the average sentence entropy for the Acquis matrix. The matrix has a wide variety of entropy values for different language pairs from the lowest, fr-en with 0.22, to the highest, et-pt with 1.33. It seems that models of languages pairs with a Romance Language or English as the source generally have low entropy. The target language does not seem to affect entropy as much, except in the case of English where model entropy is particularly low. This confirms our intuition that translating from morphologically rich languages into poorer ones should lead to lower entropy as English is the language with the lowest morphological complexity and smallest vocabulary size. The models with the highest entropy seem to be those with very rich morphology in the source, which does not uphold our intuition that the poor-rich translation models would have a high entropy.

In order to better understand the entropy results we fit a number of simple linear regression models, with entropy as the independent variable. The results are shown in Table 4 where we present the R^2, which is the fraction of the variance explained by the model, or its goodness of fit and the significance of

Factor	R^2	Significance
Reordering Amnt	0.310	***
Source Vocab Size	0.285	***
Lang. Relatedness	0.123	***
Target Vocab Size	0.056	***
Source Corpus Size	0.003	

Table 4: Simple linear regression models showing correlation of entropy with other factors.

the correlation, where * means $p < 0.05$, ** means $p < 0.01$, and *** means $p < 0.001$. We can see that reordering amount is the most correlated factor. This is almost certainly not a causal relationship and it does not explain the entropy results. However, the fact that source vocabulary is more strongly correlated with entropy than target vocabulary size could explain the fact that entropy seems to depend more on the source language than on the target language. Finally we can see that entropy is not at all correlated with corpus size. Phrase table entropy cannot be defined simply in terms of other measures. It captures a new aspect of translation difficulty which is very important, as we shall see in the next section.

5.3 Individual Impact on Performance

In order to establish the relative impact of the different factors on translation performance, we fit a number of simple linear regression models. The results are shown in Table 5. Translation model entropy is the factor which best explains the variation in performance seen between language pairs. The amount of reordering accounts for a similar amount of variation as entropy, while language relatedness and target vocabulary size account for less than half of the variation. These findings support the results presented by Birch et al. (2008), showing that with a great number and variety of language pairs, reordering has an important effect on performance.

5.4 Combined Impact on Performance

Although simple regressions can show the impact of the different factors in isolation, we are also interested in how they interact. We fit a multiple regression model to the data where all explanatory variable vectors were normalized to be more comparable.

In Table 6 we can see the relative contribution of the different factors to the model, although the factors are correlated. This means that the magnitudes of the coefficients are unreliable as the explanatory power of one variable could be shifted to another correlated variable. The R^2 of the model is 0.745 which means that 74.5% of the variation in BLEU can be explained by these factors.

Factor	R^2	Significance
Entropy	0.276	***
Reordering Amnt	0.267	***
Lang. Relatedness	0.115	***
Target Vocab Size	0.101	***
Source Corpus Size	0.034	***
Target Corpus Size	0.034	***
Source Vocab Size	0.001	

Table 5: Simple linear regression models showing correlation of BLEU with explanatory factors. An R^2 of 0.276 implies that entropy explains 27.6% of the difference in performance.

Explanatory Variable	Coefficient	
Entropy	-5.147	***
Corpus Size	24.412	***
Target Vocab. Size	-21.759	***
Language Similarity	3.736	***
Reordering Amount	-11.215	***
Target Vocab. Size^2	6.885	***
Interaction: Corp.Size/L.Sim.	4.377	***
Interaction: Corp.Size/Reord.	-5.456	***
Interaction: Corp.Size/Entropy	2.449	*
Interaction: T.Vocab.Size/L.Sim.	-4.325	***
Interaction: T.Vocab.Size/Reord.	3.453	***

Table 6: The impact of the various explanatory features on the BLEU score via their coefficients in the minimal adequate model.

6 System Combination

Let us now look at some types of system combinations that we are able to explore using our matrix of translation systems. They are illustrated in Figure 3: pivot translation, multi-pivot translation, and multi-source translation.

6.1 Pivot Translation

Instead of building machine translation systems for each language pair, we may want to resort to a simpler strategy. We pick one language as the pivot, and only build systems translating into and out of this language. When translating a language pair not including the pivot, then we chain together the source–pivot system and the pivot–target system.

Recent work on pivot translation with statistical machine translation has investigated more sophisticated approaches, such as the merging of phrase tables (Wu and Wang, 2007), but simple chaining performs comparably well. Pivoting reduces the number of required systems to $2(n-1)$ instead of

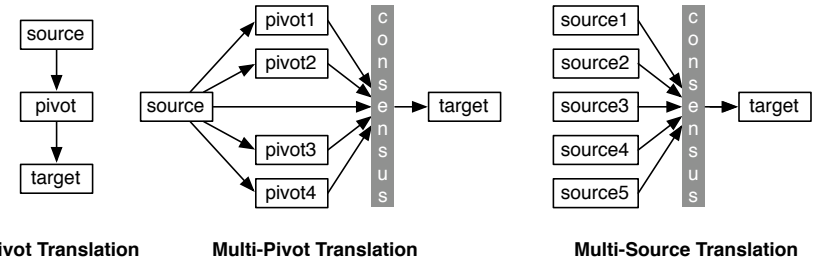

Figure 3: Three types of system combinations explored: (a) translating through a pivot language, (b) consensus of multiple pivot translations, (c) consensus of translating from multiple source languages.

BLEU Diff.	LPs via en	LPs via fr
$<$ -15	0 (0%)	2 (0%)
-15 to -10	0 (0%)	37 (8%)
-10 to -5	3 (0%)	126 (30%)
-5 to -2	16 (3%)	183 (43%)
-2 to 2	120 (28%)	71 (16%)
2 to 5	122 (29%)	1 (0%)
5 to 10	151 (35%)	0 (0%)
$\geq$ 10	8 (1%)	0 (0%)

Table 7: Pivot translation. Using English (en) as pivot mostly gains in BLEU over direct translation, while pivoting through French (fr) and other languages generally hurts.

$n(n-1)$, so in our case to 42 instead of 462.

We experimented with different pivot languages. Surprisingly, using English as a pivot *increases* translation performance more often than not. This is not the case for other languages. See Table 7 for summary statistics for English and French as pivot.

When using English as pivot, we find not much difference (BLEU diverges by up to 2 points) for about a third of language pairs, for another third there are significant gains (2-5 points) and for another third even larger gains (5-10 points). However, using French as pivot generally decreases performance, only for a sixth of language pairs there is not much difference.

English as pivot has also shown to be beneficial for Arabic–Chinese translation (Habash and Hu, 2009). We find it hard to claim that this is due to linguistic reasons, but rather an artifact of the data set we are using. It is likely that most of the text was originally authored in English.

6.2 Multi-Pivot Translation

While pivoting through any language but English does generally lead to worse translations, it does constitute an alternative translation path. A recent trend in statistical machine translation is to combine the output of different MT systems in form of a consensus translation. In multi-pivot translation, we combine the direct translation system with several pivot systems, a novel method.

Our system combination method is an adaption of Rosti et al. (2007). The multiple translations obtained from the different systems are compiled into a word lattice that is searched for the most likely translation, with the aid of a language model. The combination method is optimized, using the originating system of each competing output word and phrase as a feature.

Such multi-pivot system combination may be done for any language pair. We only did this for language pairs with English as target language, partly due to the large computational burden and partly because we wanted to compare this method against a strong baseline. Table 8 shows the performance of such multi-pivot systems with all possible source languages translated into English. We varied the number of added pivot system. We achieved relatively small gains (typically 0.5-1% BLEU), depending on the language pair and the number of pivot systems added to the direct translation baseline.

6.3 Multi-Source Translation

Since documents often have to be translated into multiple languages, one strategy to improve translation performance is to use already generated translations in some languages to translate into yet another. This is called multi-source translation.

Again, we use consensus translation methods - the same way as for multi-pivot translation. In our experimental set-up, we assume that we already have the documents in all the other 21 languages when translating them into the 22nd language. The baseline is the easiest source language for each target language. We then add additional source languages,

Source	Direct	3 Best	6 Best
bg	61.3	61.7 (+0.4%)	61.8 (+0.5%)
de	53.6	54.0 (+0.4%)	54.4 (+0.8%)
cs	58.4	59.1 (+0.7%)	59.2 (+0.8%)
da	57.6	58.0 (+0.4%)	57.9 (+0.3%)
el	59.5	60.0 (+0.5%)	60.2 (+0.7%)
es	60.0	60.2 (+0.2%)	
et	52.0	52.4 (+0.4%)	52.5 (+0.5%)
fi	49.3	50.1 (+0.8%)	50.2 (+0.9%)
fr	64.0	64.4 (+0.4%)	64.5 (+0.5%)
hu	48.0	48.5 (+0.5%)	
it	61.0	61.6 (+0.6%)	61.7 (+0.7%)
lt	51.8	52.3 (+0.5%)	52.2 (+0.4%)
lv	54.0	54.6 (+0.6%)	54.9 (+0.9%)
mt	72.1	72.2 (+0.1%)	72.3 (+0.2%)
nl	56.9	57.4 (+0.5%)	57.6 (+0.7%)
pl	60.8	61.1 (+0.3%)	61.3 (+0.5%)
pt	60.7	61.0 (+0.3%)	61.2 (+0.5%)
ro	60.8	61.6 (+0.8%)	61.9 (+1.1%)
sk	60.8	61.3 (+0.5%)	61.5 (+0.7%)
sl	61.0	61.0 (+0.0%)	61.2 (+0.2%)
sv	58.5	58.9 (+0.4%)	59.0 (+0.5%)

Table 8: Multi-Pivot: Improving direct translation by system combination with pivot translation (all translations into English)

starting with the next easiest, and so on.

Table 9 shows the results. With more source languages, translation performance improves. For instance, for Spanish the easiest source language is French with 60.9%BLEU. By combining the output from translating three source languages (French, Portuguese, Italian), we achieve 63.0%BLEU (+2.1). Improvements vary for different target languages, but they are typically in the range of 2–5%.

7 Conclusions

We built translation systems for the largest number of language pairs known to us using the JRC-Aquis corpus. We carried out a regression study to determine the main factors of translation difficulty, which explaine 74.5% of differences in scores. We also contrasted direct translation systems against pivot translation and improved them with multi-pivot and multi-source system combination methods.[3]

References

Birch, A., Osborne, M., and Koehn, P. (2008). Predicting success in machine translation. In *EMNLP*.

[3]This work was supported by the EuroMatrix/EuroMatrixPlus project funded by the European Commission (6/7th Framework Programme) and made use of the resources provided by the Edinburgh Compute and Data Facility (`http://www.ecdf.ed.ac.uk/`). The ECDF is partially supported by the eDIKT initiative (`http://www.edikt.org.uk/`).

Target	Best	3 Best	6 Best
en	72.1	73.3 (+1.2%)	74.5 (+2.4%)
bg	40.5	41.5 (+1.0%)	42.1 (+1.6%)
de	46.9	49.9 (+3.0%)	50.7 (+3.8%)
cs	52.6	53.8 (+1.2%)	54.5 (+1.9%)
da	50.0	51.9 (+1.9%)	52.8 (+2.8%)
el	42.8	45.7 (+2.9%)	46.5 (+3.7%)
es	60.9	63.0 (+2.1%)	63.7 (+2.8%)
et	34.9	40.4 (+5.5%)	41.9 (+7.0%)
fi	38.6	43.2 (+4.6%)	44.0 (+5.4%)
fr	53.2	63.7 (+10.5%)	66.2 (+13.0%)
hu	37.2	38.9 (+1.7%)	39.3 (+2.1%)
it	56.1	59.8 (+3.7%)	61.5 (+5.4%)
lt	39.6	43.0 (+3.4%)	43.4 (+3.8%)
lv	43.4	44.1 (+0.7%)	45.6 (+2.2%)
mt	39.8	39.9 (+0.1%)	
nl	52.3	54.5 (+2.2%)	55.5 (+3.2%)
pl	49.2	49.6 (+0.4%)	50.0 (+0.8%)
pt	61.0	61.2 (+0.2%)	62.9 (+1.9%)
ro	49.0	50.0 (+1.0%)	50.0 (+1.0%)
sk	44.7	46.8 (+2.1%)	47.3 (+2.6%)
sl	50.7	51.5 (+0.8%)	52.1 (+1.4%)
sv	52.0	52.5 (+0.5%)	52.7 (+0.7%)

Table 9: Multi-Source: Combining translations from different source languages

Callison-Burch, C., Fordyce, C. S., Koehn, P., Monz, C., and Schroeder, J. (2008). Further meta-evaluation of machine translation. In *3rd Workshop on SMT*, Columbus, Ohio.

Dyen, I., Kruskal, J., and Black, P. (1992). An Indoeuropean classification, a lexicostatistical experiment. *Transactions of the American Philosophical Society*, 82(5).

Galley, M. and Manning, C. D. (2008). A simple and effective hierarchical phrase reordering model. In *EMNLP*.

Habash, N. and Hu, J. (2009). Improving Arabic-Chinese statistical machine translation using English as pivot language. In *4th Workshop on SMT*, Athens, Greece.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C. J., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *ACL Demo and Poster Session*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *ACL*.

Rosti, A.-V. I., Xiang, B., Matsoukas, S., Schwartz, R., Ayan, N. F., and Dorr, B. J. (2007). Combining output from multiple machine translation systems. In *HLT-NAACL*.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., and Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *LREC*.

Wu, H. and Wang, H. (2007). Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21(3):165–182.