Abstract

We built 462 machine translation systems for all language pairs of the Acquis Communautaire corpus. We report and analyse the performance of these systems, 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.
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 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/

Table 2: Size of the JRC-Acquis Communautaire corpus

| Language ISO code | N° of texts | Total N° words | Total N° characters | Average n° words | Signatures | Annexes | Total N° words (text + signatures + annexes) |
|-------------------|------------|----------------|--------------------|-----------------|------------|--------|---------------------------------|
| bg                | 1056       | 15208341       | 98298769           | 1388.13         | 2052773   | 1698563 | 30145967                       |
| cs                | 2143       | 22843279       | 14897289           | 1065.55         | 2725400   | 1673772 | 48837212                       |
| da                | 2564       | 31496827       | 21346813           | 1311.68         | 2839786   | 1685621 | 45846262                       |
| de                | 2541       | 32059592       | 22374865           | 1361.87         | 2542149   | 1632711 | 50928662                       |
| el                | 2184       | 36453349       | 25935434           | 1672.37         | 2349252   | 1634967 | 55687003                       |
| en                | 2345       | 34588383       | 21069205           | 1469.03         | 3198766   | 1777061 | 55357108                       |
| es                | 2357       | 38926161       | 23081676           | 1651.3          | 3490204   | 1971623 | 62132686                       |
| et                | 2041       | 24621625       | 21570074           | 1054.9          | 1336031   | 1495978 | 40953242                       |
| fi                | 2384       | 24830102       | 21278964           | 1068.67         | 2677798   | 2547171 | 40107961                       |
| fr                | 2627       | 39100499       | 24735829           | 1654.91         | 3021013   | 1997892 | 62100432                       |
| hu                | 2586       | 25623280       | 21380414           | 254.44          | 2529448   | 1506496 | 41663864                       |
| it                | 2342       | 35784670       | 23067721           | 1623.72         | 3120979   | 1721353 | 57217002                       |
| it                | 2379       | 26937773       | 19543268           | 1152.22         | 2343685   | 1501848 | 44392842                       |
| lv                | 2296       | 27582145       | 21564205           | 1204.6          | 1671324   | 1547369 | 44703667                       |
| no                | 1054       | 20926909       | 14896074           | 1964.53         | 1336042   | 1562011 | 37885652                       |
| nl                | 2364       | 35285161       | 21963359           | 1496.57         | 3039580   | 1847515 | 56771865                       |
| pl                | 2378       | 29713003       | 21440626           | 1325.67         | 2531141   | 1762739 | 49263537                       |
| pt                | 2305       | 37221668       | 22749941           | 1583.56         | 3034368   | 1953027 | 59606253                       |
| ro                | 653        | 9186437        | 6373971            | 1397.68         | 3142584   | 1465862 | 40887085                       |
| sk                | 2143       | 25728337       | 17920434           | 1221.01         | 3227852   | 1619046 | 46213059                       |
| sl                | 2064       | 27702260       | 17851787           | 1342.04         | 3103195   | 1467817 | 47543215                       |
| sv                | 2024       | 29433097       | 19084301           | 1453.99         | 2575771   | 1495834 | 46764192                       |
| Total             | 453364     | 635283572      | 4267278446         | 1358.05         | 6029159   | 37770752 | 105330545                       |

\(^1\)http://eur-lex.europa.eu/
\(^2\)http://europa.eu/eurovoc/
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 language.

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).
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 vary 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:

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

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

| 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 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 | Significance |
|----------------------|-------------|-------------|
| 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/Record. | -5.456 | ***         |
| Interaction: Corp.Size/Entropy | 2.449 | *           |
| Interaction: T.Vocab.Size/L.Sim. | -4.325 | ***         |
| Interaction: T.Vocab.Size/Record. | 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.

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.

\[
\begin{array}{|c|c|c|}
\hline
\text{BLEU Diff.} & \text{LPs via en} & \text{LPs via fr} \\
\hline
< -15 & 0 (0\%) & 2 (0\%) \\
-15 to -10 & 0 (0\%) & 37 (8\%) \\
-10 to -5 & 3 (0\%) & 126 (30\%) \\
-5 to -2 & 16 (3\%) & 183 (43\%) \\
-2 to 2 & 120 (28\%) & 71 (16\%) \\
5 to 10 & 151 (35\%) & 0 (0\%) \\
\geq 10 & 8 (1\%) & 0 (0\%) \\
\hline
\end{array}
\]

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,
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 explain 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.  

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 (6th 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/).

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

| 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%) | 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%) | 48.6 (+0.5%) |
| it     | 61.0   | 61.6 (+0.6%) | 61.7 (+0.7%) |
| it     | 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 9: Multi-Source: Combining translations from different 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%) | 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%) | 48.6 (+0.5%) |
| it     | 61.0   | 61.6 (+0.6%) | 61.7 (+0.7%) |
| it     | 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 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., 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, 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.

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