Europarl: A Multilingual Corpus for Evaluation of Machine Translation

Philipp Koehn
koehn@isi.edu
Information Sciences Institute
University of Southern California
Marina del Rey, CA 90292, USA

December 28, 2002

Abstract

This paper describes the acquisition, preparation and properties of a corpus extracted from the Proceedings of the European Parliament. This corpus is available in 11 languages, consists of over 20 million words per language, and it is preprocessed for the use in statistical machine translation. We describe the methods we used for crawling, document alignment, and sentence alignment. We also present a common test set for machine translation and report the results of a number of basic statistical machine translation experiments.

1 Introduction

Over the last decade, there has been an increasing effort to build computer systems that learn from data how to translate human languages. Given a large collection of sentence pairs, these systems find correspondences and learn what happens when translating, say, French to English.

The primary training data for such statistical machine translation systems are parallel corpora, also called bitexts. In this report, we describe how we acquired such a corpus from a web source and describe a number of experiments of a current statistical machine translation systems that is trained on this data.

Various methods are used to establish correspondences between the given translations of the text: on document level, paragraph level, sentence level, all the way down to the word level. Having such correspondences (or alignments) has shown to be a useful resource for other natural language tasks, such as information retrieval, word sense disambiguation, anaphora resolution, induction of tools across languages and many more.

2 Acquisition of the Corpus

The acquisition of a parallel corpus is done in a number of steps:

- Obtain the raw data (e.g., by crawling from a web source)
- Extract and map parallel chunks of text (document alignment)
- Break the texts into sentences (sentence splitting)
- Standardize the format of the text (tokenization)
- Map sentences in one language to sentences in the other language (sentence alignment)

In the following, we will describe in detail the acquisition of the Europarl corpus of European Parliament Proceedings from the website of the European Parliament. These proceedings are published in all the 11 official languages of the European Union. This means that we can not only extract a conventional parallel corpus, but a multilingual corpus of 11 languages, or 10 parallel corpora any of the languages.

2.1 Crawling

The website of the European Parliament[1] provides the Proceedings of the European Parliament in form of HTML files. Each file contains the utterances of one speaker in turn. The URL for each file contains relevant information for identification, such as its language, the day and number of the thread of discussion and number of the utterance.

We crawl this web resource with a web spider by starting at an index page and following certain links based on inclusion and exclusion rules. Since the corpus is broken in many small parts, the crawling process is very time consuming. Per language, it took multiple days to obtain the roughly 80,000 files each. Although such crawls are slow, it is typically easier than contacting directly the technical staff of the website and negotiate a transfer procedure.

Usually, there are also copyright concerns, although less so for information from government sources. Fortunately, the European Parliament web site states: "Except where otherwise indicated, reproduction is authorized, provided that the source is acknowledged." Such liberal copyright policy can not necessarily expected. Often a longer legal process is required to get permission and access to the data.

Besides identifying sources for parallel corpora manually, it is also possible to mine the web for such data. Resnik [1999] proposes such a system, called STRAND.

[1] `http://www3.europarl.eu.int/omk/omnsapir.so/calendar?APP=CRE&LANGUE=EN`

2.2 Document Alignment

Each sitting of the European Parliament covers a number of topics. A first step is to identify the texts belonging to each topic, and matching these between languages. To obtain the maximum amount of data, we match these topics for each of the 55 possible language pairs.

Large data collections such as the Proceedings of the European Parliament are created over the period of many years, often with changing formatting standards and other sources of error. For instance, part of the "English" part of the proceedings contain actually French texts (21-24 May 1996) at the time of our crawl.

The extraction of relevant text from noisy HTML is a cumbersome enterprise that requires constant refinement and adaptation. We process the HTML data with a Perl program that uses pattern matching to detect and extract the identity of the speaker and her statements including paragraph markers.

There is work on automatically learning systems that extract structured information from web sources or other forms of unstructured data. This task is called wrapper induction. See for instance work by Muslea et al. [1999]. For a single data source, however, a manual approach is often more efficient, especially if it is noisy.

For each day, the data is stored in a file per language with some markup of meta information:

```
File: de-en/en/ep-00-01-17.txt (German-English, English half, 17 Jan 2000)
<CHAPTER ID=1>
Resumption of the session
<SPEAKER ID=1 NAME="President">
I declare resumed the session of the European Parliament ...
<P>
Although, as you will have seen, the dreaded 'millennium bug' ...
```

The parallel corpora is available on our website[2] in this format. It is also available in sentence aligned format, which we will describe below.

The document alignment is done without tokenization and sentence splitting. The motivation behind this is that these are error prone processes for which multiple standards could be applied, and we do not want to force any specific standard at this step.

[2] http://www.isi.edu/~koehn/publications/europarl/

2.3 Sentence Splitting and Tokenization

Sentence splitting and tokenization requires specialized tools for each language. Unfortunately, we do not have such tools available for all the languages under consideration.

One problem of sentence splitting is the ambiguity of the period "." as either a end of sentence marker, or as a marker for an abbreviation. For English, French and German, we acquired a list of known abbreviations that are typically followed by a period. Another heuristic is the case of the first word following a period ("ca. three thousand men"), which indicates an abbreviation opposed to end of sentence.

There has been extensive work on empirical methods to learn sentence breaking. See for instance the work on SATZ [Palmer and Hearst, 1997]. Various machine learning methods can be applied to this problem, such as decision trees [Riley, 1989] and maximum entropy [Reynar and Ratnaparkhi, 1997].

Issues with tokenization include the English merging of words such as in "can't" (which we transform to "can not"), or the separation of possessive markers ("the man's" becomes "the man 's"). We do not perform any specialized treatment for other languages than English at this point. A recent study of a number of tokenization problems was carried out by Mikheev [2002].

For training a statistical machine translation system, usually all words are lowercased to eliminate the differences between different spelling of words depending on their occurrence at the beginning of a sentence (*The*), in the middle (*the*), or in a headline (*THE*). A more sophisticated approach is true-casing, which allows the distinction names (*Mr. Black*) and regular words (*black*).

2.4 Sentence Alignment

Sentence alignment is usually a hard problem, but in our case it is simplified by the fact that the texts already available in paragraph aligned format. Each paragraph consists typically of only 2-5 sentences.

If the number of paragraphs of a speaker utterance differs in the two languages, we discard this data for quality reasons. The alignment of sentences in the corpus is done with an implementation the algorithm by Gale and Church [1993]. This algorithms tries to match sentences of similar length in sequence and merges sentences if necessary (e.g. two short sentences in one language to one long sentence in the other language), based on the number of words in the sentence. Since there are so few sentences per paragraph, alignment quality is very high.

There is considerable work on better sentence alignment algorithms. One obvious extension is to not only consider sentence length, but also potential word correspondences within sentence pairs. Work by Melamed [1999] is an example for such an approach.

The sentence aligned data is stored in a file per day per language, so that lines with the same line number in a file pair are mappings of each other. The markup from the document aligned files is stripped out.

2.5 Size of the Corpus

Table 1 contains a breakdown of the size of the corpus. The chapters in the corpus refer to different topics of discussion on each day. During each topic, different speakers take turns. Each speaker turn is separated into a number of paragraphs.

Language	Days	Chapters	Speaker Turns	Paragraphs	Sentence Pairs	Words
Danish	388	3,882	69,995	274,011	763,919	21,111,943
German	398	3,993	71,736	282,069	759,220	21,158,370
Greek	363	3,560	64,334	251,900	623,604	17,636,960
Spanish	388	3,894	70,412	275,787	746,274	20,803,150
Finnish	338	3,398	61,449	241,259	668,487	18,512,071
French	388	3,866	69,700	273,174	746,147	20,843,596
Italian	388	3,854	69,239	271,506	693,792	20,255,923
Dutch	388	3,865	69,478	272,346	743,880	20,480,082
Portuguese	388	3,867	69,696	273,211	726,880	20,546,564
Swedish	388	3,398	61,328	240,709	626,769	17,296,225

Table 1: Size of the Europarl corpus, for each language aligned to English. Number of chapters, speaker turns, paragraphs for the document aligned corpus. Number of sentences pairs and words for the sentence aligned corpus. Words includes separated punctuation.

The number of days varies due to availability of the data. Data is also lost during document alignment and then again during sentence alignment.

2.6 Extraction of a Common Test Set

To allow the comparison of machine translation system, it is necessary not only to define a common training set (as the Europarl corpus), but also a common test set. We suggest to reserve the last quarter if 2000 (November-December 2000) as test set, and to use the rest of the corpus as training data. This portion of the corpus comprises of over one million words in over 40,000 sentences.

Typically, a portion of this test set is chosen based on sentence length, for instance all 10-12 word sentences in the source language.

To be able to compare system performance on different language pairs, we also extracted the set of sentences that are aligned to each other across all 11 languages. Here is one of the sentences from this collection:

English: that is almost a personal record for me this autumn !

Danish: det er næsten en personlig rekord for mig dette efterår .

German: das ist für mich fast persönlicher rekord in diesem herbst .

Greek: .

Spanish: es la mejor marca que he alcanzado este otoño .

Finnish: se on melkein minun ennätykseni tänä syksynä !

French: c ' est pratiquement un record personnel pour moi , cet automne !

Italian: e ' quasi il mio record personale dell ' autunno .

Dutch: dit is haast een persoonlijk record deze herfst .

Portuguese: é quase o meu recorde pessoal deste semestre !

Swedish: det är nästan personligt rekord för mig denna höst !

Note that this data is also lowercased, which is not done for the released sentence aligned data.

3 Training a Statistical Machine Translation System

This section shortly describes how to train a IBM Model 4 statistical machine translation system and how to translate foreign texts with it. This are described by Brown et al. [1990]. It was recently reimplemented at the 1999 John Hopkins summer workshop [Al-Onaizan et al., 1999], and made freely available[3]. Although better SMT systems have been developed recently, it provides a useful starting point and benchmark for research in this field. Also note that many of the more sophisticated methods are based on intermediate stages of these models.

Fortunately, all the necessary tools are freely available for research purposes. Besides being an end to itself, the various models also provide useful information for related research goals. Especially the word aligned data and translation lexicons produced in the process has shown to be a promising resource for various endeavors.

The following main steps are required:

- Train the model with Giza
- Train a language model
- Translate new texts with a decoder

3.1 Training the SMT Model

3.2 Training the LM Model

3.3 Using the Decoder to Translate

[3]Giza++ is available at `http://www-i6.informatik.rwth-aachen.de/ och/software/GIZA++.html`

4 Performance of Statistical Machine Translation Systems

Given the parallel corpora, we can now train baseline SMT systems to get some measure of performance on this corpus. Fine-tuning of parameters and dedicated preprocessing for each languages will improve results.

From the multi-language test data we extracted only the sentences that have a length of 5 to 15 words per sentence in English and a length of 2 to 25 words in each of the other languages. This results in 1755 sentences for testing.

4.1 Results

We trained systems based on IBM Model 4 with the Giza toolkit for each language pair including English. We then tested the performance of these translation systems with the greedy decoder developed at ISI. The results are measured with the BLEU metric [Papinini et al., 2001], which measures similarity to the English reference translation. See Table 2 for details.

from↓ to→	Danish	German	Greek	English	Spanish	Finnish	French	Italian	Dutch	Portug.	Swedish
Danish	–	0.2271	0.1666	0.2903	0.2271	0.1855	0.2210	0.1971	0.2513	0.2006	0.3246
German	0.2546	–	0.1752	0.2534	0.2252	0.1752	0.2267	0.2008	0.2654	0.2121	0.2236
Greek	0.1992	0.1763	–	0.2597	0.2779	0.1581	0.2460	0.2513	0.1996	0.2573	0.1966
English	0.2561	0.2040	0.1973	–	0.2700	0.1773	0.2555	0.2322	0.2285	0.2430	0.2595
Spanish	0.2166	0.2008	0.2194	0.2812	–	0.1742	0.3389	0.3059	0.2044	0.3371	0.2006
Finnish	0.2126	0.1804	0.1483	0.2178	0.1827	–	0.1683	0.1687	0.1792	0.1661	0.2070
French	0.2307	0.2018	0.1995	0.2787	0.3452	0.1636	–	0.2626	0.2021	0.2890	0.2110
Italian	0.2126	0.1904	0.2143	0.2741	0.3234	0.1663	0.3235	–	0.2265	0.3120	0.2004
Dutch	0.2493	0.2436	0.1488	0.2635	0.2316	0.1507	0.2238	0.2016	–	0.2080	0.2137
Portug.	0.2170	0.1893	0.2112	0.2668	0.3572	0.1634	0.3184	0.2963	0.2308	–	0.2068
Swedish	0.3379	0.2129	0.1886	0.3137	0.2368	0.1835	0.2202	0.2003	0.2379	0.2177	–

Table 2: Performance of baseline statistical machine translation methods on the Europarl corpus, as measured by the BLEU score

The performance scores (the higher the better) reflect very nicely the relatedness of language pairs. Translation from Portuguese to Spanish is relatively easy (0.3572), while translating from Greek to Finnish is relatively hard (0.1581). Note that this is not the only explanation for translation difficulties, e.g. translating from German to English (0.2534) is easier than the other way around (0.2040).

In the following, we will only discuss experimental results when translating to English, which is the most common target language in the research literature.

4.2 Discussion

Consider the English translations of the greedy decoder for the sentence from Section 2.6:

Danish: it is almost a personal record for me this autumn .

German: it is for me , almost personal record in this autumn .

Greek: this is the personal my record this autumn .

Spanish: it is the best mark which have reached this autumn .

Finnish: it is almost me ennätykseni autumn !

French: it is almost record staff for me , this autumn .

Italian: it is almost my records personal autumn .

Dutch: this is almost a personal privilege this autumn .

Portuguese: it is almost record my personal this half !

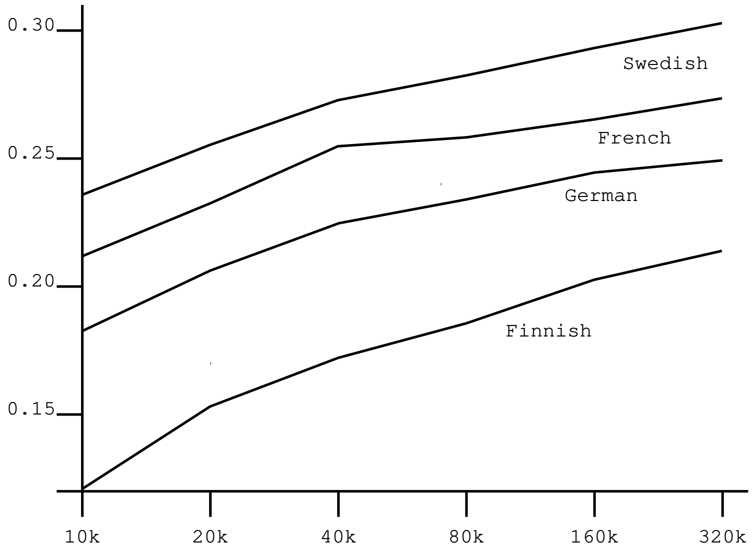
Swedish: it is almost record for me this personal autumn !

These translations illustrate a number of properties of the current state of statistical machine translation:

- The bad preprocessing of, for instance, the French "c'est" is not a big problem
- Morphological variations cause sparse data problems, such as displayed by the Finnish "ennätykseni" which has not been observed in this word form in the training data.
- An inherent problem of the IBM models is that a foreign word can not translate into multiple English words, as illustrated by the Portuguese "semestre" which may translate as "half year", but is only translated as "half".
- The output is often not grammatical – e.g. the translation from Spanish is missing the subject pronoun "I".
- The output may contain unrealistic bigrams, such as "personal autumn" for Swedish, since the trigram language model is not strong enough for rare words.
- There are a number of word choice errors, such as the translation of the French "personnel" as "staff" instead of "personal".
- Reordering seems often arbitrary, since it is largely driven by the trigram language model, leading to construction such as "record my personal" (from Portuguese).

4.3 Learning Curves

It is a truism in the field of empirical methods in natural language processing, that more data is better, and even more data is even better. We can demonstrate this effect for statistical machine translation systems as well. We trained the system with different training sizes, and compared performance on the common test set. Figure 1 shows the result of the experiments for a number of language pairs.



Training corpus size	Finnish-English	German-English	French-English	Swedish-English
10,000 sentence pairs	0.1215	0.1828	0.2115	0.2359
20,000 sentence pairs	0.1563	0.2071	0.2343	0.2550
40,000 sentence pairs	0.1729	0.2250	0.2550	0.2725
80,000 sentence pairs	0.1870	0.2336	0.2585	0.2853
160,000 sentence pairs	0.2029	0.2434	0.2665	0.2932
320,000 sentence pairs	0.2135	0.2494	0.2735	0.3035
all sentence pairs	0.2178	0.2534	0.2787	0.3137

Figure 1: Performance increases with more training data for translation and language model (number of all sentences used in training is roughly 500,000)

As the smallest training corpus size we used 10,000 sentence pairs for training the translation model as well as the language model. When doubling the training size, we can observe an almost log-linear improvement in translation quality, as measured by the BLEU score. This observation holds for both the easier and more difficult language pairs.

Language Model Size

How much is translation quality driven by the quality of the translation model versus the language model? The results in Table 3 offer a hint of an answer to this question. Here we compare again the performance difference for large training corpus versus a small training corpus for both translation model and language model training. As a third experiment, we check the performance for a system trained with a small corpus for the translation model and a large corpus for the language model.

Language Model	320,000	692,606	40,000
Translation Model	320,000	40,000	40,000
Swedish-English	0.3035	0.2867	0.2725
Spanish-English	0.2800	0.2565	0.2482
French-English	0.2735	0.2631	0.2550
Danish-English	0.2752	0.2573	0.2393
Dutch-English	0.2570	0.2366	0.2264
Portuguese-English	0.2648	0.2512	0.2375
Italian-English	0.2688	0.2531	0.2420
Greek-English	0.2564	0.2314	0.2259
German-English	0.2494	0.2337	0.2250
Finnish-English	0.2135	0.1845	0.1729
Average	0.2642	0.2454	0.2348

Table 3: Size of training corpus for translation model and language model

The results come out somewhat in the middle, depending on the language pair. This shows the fairly strong effects of a good language model. Also note, that since the language model is trained on a different (larger) corpus, the application of the Bayes rule $\text{argmax}_e p(e|f) = \text{argmax}_e p(f|e)p(e)$ is mathematically not correct. This does not seem to be very harmful and opens the doors to training language models on much larger corpora of widely available monolingual data.

4.4 Lexical Translation Accuracy

The more frequent a word occurs in the training corpus, the more likely it will be translated correctly by a statistical machine translation system. Koehn and Knight [2001] carried out a study to demonstrate this effect.

Figure 2 shows their results, which are obtained from a different training corpus and limited to nouns. Words that occurred only once in training were translated correctly less than 10%

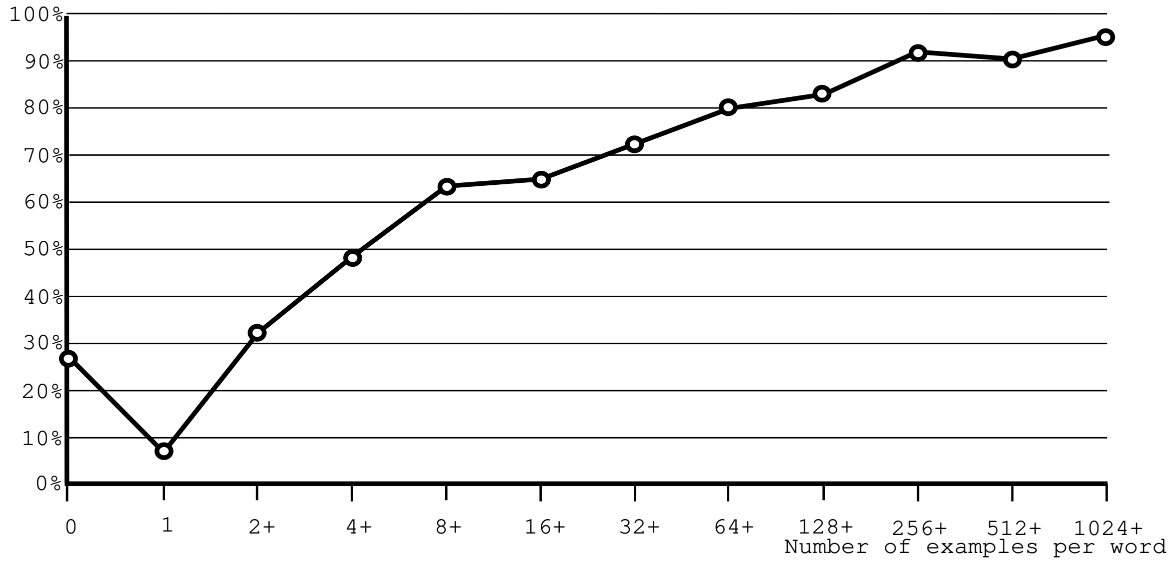


Figure 2: Frequency of a word in the training corpus vs. translation accuracy

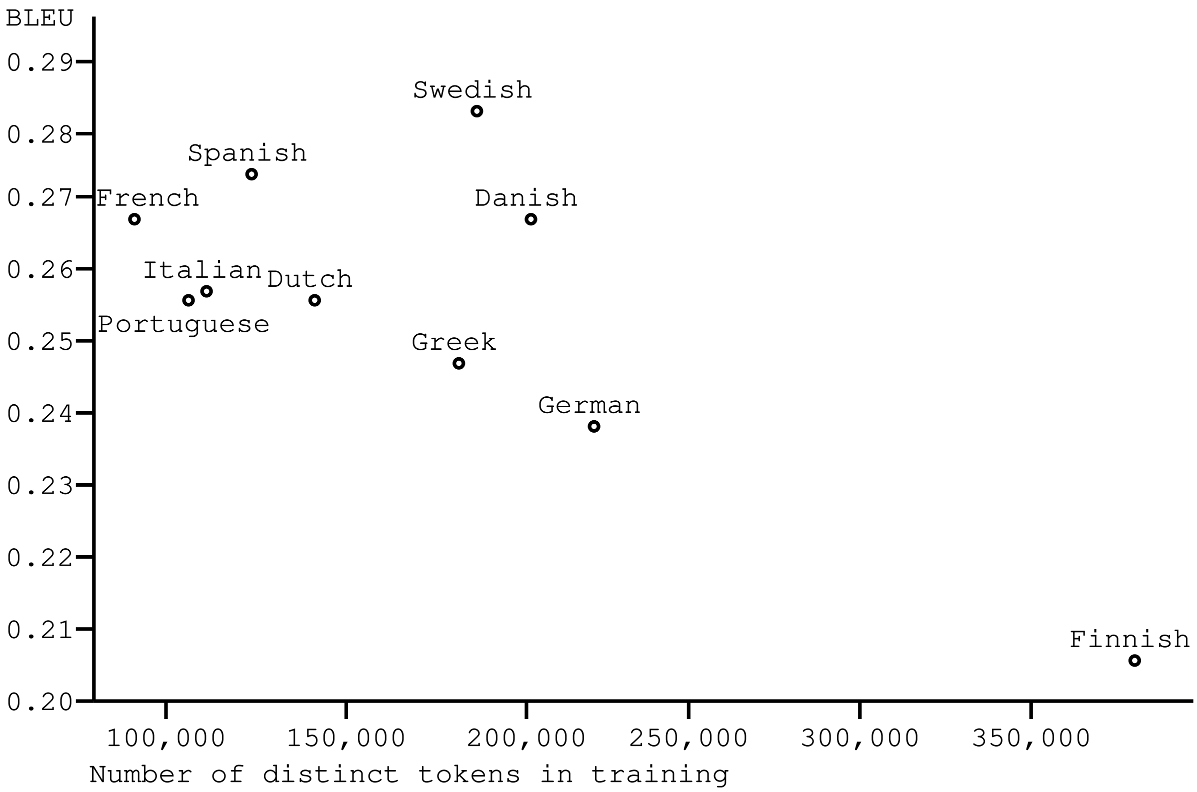
of the time, while translation accuracy is above 90% for the most frequent words. Words that occurred never in training are duplicated in verbatim in the output, which is often a good strategy for names.

4.5 Vocabulary Size

Why are some languages harder than others to translate? One obvious answer is that languages that are more related more easily translate into each other. This explains, for instance, why the performance for Swedish or French to English is so much higher than the performance for Finnish to English.

But also a rich morphology causes problems: It leads to a much higher number of word forms, which causes sparse data problems. As a general rule, rarely or never before seen words are much harder to translate that frequently seen ones (see also the next section).

In Figure 3 we plotted the number of distinct tokens in the training corpus against translation quality (as measured by the BLEU score). While the resulting points do not form a clear line, there is some correlation visible. For instance, the low translation quality for German, albeit closely related to English, may be due to some degree to its large vocabulary which stems from morphological variation and compounding of words.



Language	Vocabulary Size	BLEU
Danish	206,120	0.2686
German	219,401	0.2393
Greek	180,453	0.2477
Spanish	122,307	0.2069
Finnish	377,052	0.2069
French	92,731	0.2683
Italian	112,240	0.2573
Dutch	141,927	0.2563
Portuguese	107,361	0.2563
Swedish	187,379	0.2826

Figure 3: Vocabulary size vs. Performance: A larger number of distinct tokens in a language leads to worse translation quality

4.6 Morphology

Morphology plays a role in increasing the number of word forms on both the foreign and English side. If we observe in the parallel corpus the word *house*, we might learn something that is also useful for translation the plural form *houses*. However, these two words are treated as unrelated token in the standard model.

The problem increases when the foreign language is morphological much richer than the target language, e.g. when much larger number of word forms exist. This is the case for German, where even adjectives and determiners have different endings depending on case and gender. While this additional morphology matters, it usually has no impact on word-level translation. So, stripping out redundant morphology may be beneficial.

This can be done, if appropriate tools exist. To demonstrate this effect, we used the German morphological analyzer Morphix [Finkler and Neumann, 1998] and the POS tagger TnT [Brants, 2000] to prepare the German side of the training and test corpus. We collapsed the different word forms into equivalence classes, as suggested by Nießen and Ney [2001]: All adjective forms are stemmed into one base form. All verbs are reduced to the four classes present, past, past-participle, and infinitive. All nouns are reduced to either singular or plural.

The impact on performance is given in Table 4. In our experiments, the morphological reduction reduced lexicon size by 10 percent, but with no significant effect on performance. The performance, of course, also depends on the quality of the morphological analyzer.

Training corpus size	Raw corpus	Reduced Morphology
10,000 sentence pairs	0.1828	0.1758
20,000 sentence pairs	0.2071	0.2080
40,000 sentence pairs	0.2250	0.2228
80,000 sentence pairs	0.2330	0.2332
160,000 sentence pairs	0.2434	0.2415
320,000 sentence pairs	0.2494	0.2472
all sentence pairs	0.2534	0.2545

Table 4: Translation performance improves, if redundant morphology is stripped out (German-English)

Often, morphology is ambiguous and good morphological analyzers are not available for all languages. There has been some work in automatically learning morphology from monolingual data [Schone and Jurawsky, 2001; Yarowsky and Wicentowski, 2000], or projecting morphology across languages [Yarowsky et al., 2001]. These techniques may be useful for improving translation quality.

4.7 Noisy Training Data

Not all training data can be expected to be of high quality. If data is collected from different sources, preprocessing is much harder. Especially if the web is mined for parallel corpora, a lot of noise can be expected. Also, sentence alignment might be much harder for corpora that are not paragraph-aligned or have translation gaps.

Does the performance of a machine translation system degrade, when trained on noisy data? Wang [2002] addressed this question by artificially adding noise to a clean training corpus. Specifically, a certain percentage of sentence alignments are distorted to simulated misaligned training data.

His results suggest that the quality of the translation system only starts to significantly degrade, if half of the training data is distorted this way. In experiments on the Canadian Hansards[4] and the Europarl corpus, distortion of up to 25% of the training data has barely any impact on the translation quality, while 50% distorted training data reduces performance, as measured by the BLEU score, only by about 10%.

4.8 Out of Domain Testing

Machine translation systems are never general purpose – for optimal performance they have to be tuned for a specific domain, which may have a distinct topic or style. This means that a statistical MT system that is trained on the Europarl corpus may show poor performance in translation vastly different material.

To illustrate this, we considered another corpus – the German news corpus de-news, which is also freely available[5]. This corpus covers German news items translated into English. Although there is some topical overlap in political news, it also contains some sports and entertainment news.

First we tested a system trained solely on the Europarl corpus on a test set of the de-news corpus. We then built a second system trained on the de-news corpus, which is much smaller (roughly 1.5 million words, compared with the 20 million word Europarl corpus). We also built a third system trained on both corpora. The results are summarized in Table 5.

The results suggest that the much smaller in-domain de-news system trained on the same type of material fares much better (0.1143) than the bigger out-of-domain Europarl system (0.0835), when testing on de-news data. Note that the BLEU scores are not comparable to the Europarl BLEU scores from the preceding sections for a number of reasons (more difficult material, longer sentences, etc.).

---

[4] Available at `http://www.isi.edu/natural-language/download/hansard/`

[5] `http://www.isi.edu/~koehn/publications/de-news/`

Training corpus	Unique sentence pairs	Unique words	BLEU
de-news	65,078	63,066	0.1143
Europarl	525,994	218,861	0.0835
Europarl + de-news	586,502	245,343	0.1118
Europarl + de-news (10-fold)	586,502	245,343	0.1211

Table 5: Performance of systems trained on data from different domains when translating a de-news test set

Adding the Europarl data to the de-news data actually decreases system performance (to 0.1118), unless the de-news data is weighted much more heavily – with 10-fold weighting of this in-domain data, performance can be improved to 0.1211.

These results show that the performance of a statistical machine translation system decreases quite drastically, when simply trained on out-of-domain data. This can be alleviated by a added some in-domain data. These domain effects may also be addressed by language models trained in the target domain, or by providing additional domain specific dictionaries.

4.9 Other Statistical Machine Translation Systems

The performance of the publicly available IBM Model 4 system can be used as a baseline to measure the performance of other translation systems. The performance can be improved in many ways: better translation models, better language models, better decoders, better preprocessing, better tuning of the system to the test data, additional training data, additional linguistic resources, etc. A good evaluation of a different system should factor out the contribution of each of the changes.

Some translation systems may have difficulties with the vast amount of training data of the Europarl system. For instance, we compared the performance of a phrased-based model [Marcu and Wong, 2002] against IBM Model 4. However, at the time of this writing, this system was not yet able to cope with the large number of sentence pairs in the Europarl corpus. Therefore, we only trained the system on up to 40,000 sentence pairs.

We also trained an alignment template model [Och, 2002] on the same training data. The results are displayed in Table 6. On the same amount of training data, the phrase-based model shows superior performance than Model 4. The alignment template model comes out ahead of both.

Training corpus size	Al. Templ.	Phrase	Model 4
10,000 sentences	0.2152	0.1984	0.1828
20,000 sentences	0.2362	0.2175	0.2071
40,000 sentences	0.2557	0.2332	0.2250

Table 6: Performance of an Alignment Template model [Och, 2002] and a Phrase-Based Model [Marcu and Wong, 2002] compared with IBM Model 4 (German-English)

References

Al-Onaizan, Y., Curin, J., Jahr, M., Knight, K., Lafferty, J., Melamed, D., Och, F.-J., Purdy, D., Smith, N. A., and Yarowsky, D. (1999). Statistical machine translation. Technical report, John Hopkins University Summer Workshop `http://www.clsp.jhu.edu/ws99/projects/mt/`.

Banko, M. and Brill, E. (2001). Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*.

Brants, T. (2000). TnT - a statistical part-of-speech tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP*.

Brown, P., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Rossin, P. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16(2):76–85.

Finkler, W. and Neumann, G. (1998). Morphix. A fast realization of a classification-based approach to morphology. In *4. Österreichische Artificial-Intelligence-Tagung. Wiener Workshop - Wissensbasierte Sprachverarbeitung.*

Gale, W. and Church, K. (1993). A program for aligning sentences in bilingual corpora. *Computational Lingustics*, 19(1).

Koehn, P. and Knight, K. (2001). Knowledge sources for word-level translation models. In *Proceedings of EMNLP*.

Marcu, D. and Wong, W. (2002). A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP*.

Melamed, D. (1999). Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25(1):107–130.

Mikheev, A. (2002). Periods, capitalized words, etc. *Computational Linguistics*, 28(3):289–318.

Muslea, I., Minton, S., and Knoblock, C. (1999). A hierarchical approach to wrapper induction. In Etzioni, O., Müller, J. P., and Bradshaw, J. M., editors, *Proceedings of the Third International Conference on Autonomous Agents (Agents'99)*, pages 190–197, Seattle, WA, USA. ACM Press.

Nießen, S. and Ney, H. (2001). Toward hierarchical models for statistical machine translation of inflected languages. In *Workshop on Data-Driven Machine Translation at ACL 39*, pages 47–54.

Och, F. J. (2002). *Statistical Machine Translation: From Single-Word Models to Alignment Templates*. PhD thesis, RWTH Aachen, Germany.

Palmer, D. D. and Hearst, M. A. (1997). Adaptive mulilingual sentence boundary detection. *Computational Lingustics*, 23(2):241–267.

Papinini, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). BLEU: a method for automatic evaluation of machine translation. Technical Report RC22176(W0109-022), IBM Research Report.

Resnik, P. (1999). Mining the web for bilingual text. In *Proceedings of the International Conference of the Association of Computational Linguistics*.

Reynar, J. C. and Ratnaparkhi, A. (1997). A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the Firth Conference on Applied Natural Language Processing*, pages 803–806.

Riley, M. D. (1989). Some applications of tree-based modelling to speech and language. In *Proceedings of the DARPA Speech and Language Technology Workshop*, pages 339–352.

Schone, P. and Jurawsky, D. (2001). Knowledge-free induction of inflectional morphologies. In *Proceedings of NAACL*.

Wang, S. (2002). Machine translation on noisy training data. Master's thesis, University of Southern California.

Yarowsky, D., Ngai, G., and Wicentowski, R. (2001). Inducing multilingual text analysis tools via robust projection across corpora. In *Proceedings of HLT*.

Yarowsky, D. and Wicentowski, R. (2000). Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of ACL*.