Language Models for Machine Translation: Original vs. Translated Texts

Gennadi Lembersky and **Noam Ordan** and **Shuly Wintner**
Department of Computer Science, University of Haifa, 31905 Haifa, Israel
glembers@campus.haifa.ac.il, noam.ordan@gmail.com, shuly@cs.haifa.ac.il

Abstract

We investigate the differences between language models compiled from original target-language texts and those compiled from texts manually translated to the target language. Corroborating established observations of Translation Studies, we demonstrate that the latter are significantly better predictors of translated sentences than the former, and hence fit the reference set better. Furthermore, translated texts yield better language models for statistical machine translation than original texts.

1 Introduction

Statistical machine translation (MT) uses large target language models (LMs) to improve the fluency of generated texts, and it is commonly assumed that for constructing language models, "more data is better data" (Brants and Xu, 2009). Not all data, however, are created the same. In this work we explore the differences between LMs compiled from texts originally written in the target language and LMs compiled from *translated* texts.

The motivation for our work stems from much research in Translation Studies that suggests that original texts are significantly different from translated ones in various aspects (Gellerstam, 1986). Recently, corpus-based computational analysis corroborated this observation, and Kurokawa et al. (2009) apply it to statistical machine translation, showing that for an English-to-French MT system, a *translation* model trained on an English-translated-to-French parallel corpus is better than one trained on French-translated-to-English texts. Our research question is whether a *language model* compiled from translated texts may similarly improve the results of machine translation.

We test this hypothesis on several translation tasks, where the target language is always English. For each language pair we build two English language models from two types of corpora: texts originally written in English, and human translations from the source language into English. We show that for each language pair, the latter language model better fits a set of reference translations in terms of perplexity. We also demonstrate that the differences between the two LMs are not biased by content but rather reflect differences on abstract linguistic features.

Research in Translation Studies suggests that all translated texts, irrespective of source language, share some so-called *translation universals*. Consequently, translated texts from several languages to a single target language resemble each other along various axes. To test this hypothesis, we compile additional English LMs, this time using texts translated to English from languages *other* than the source. Again, we use perplexity to assess the fit of these LMs to reference sets of translated-to-English sentences. We show that these LMs depend on the source language and differ from each other. Whereas they outperform original-based LMs, LMs compiled from texts that were translated from the *source* language still fit the reference set best.

Finally, we train phrase-based MT systems (Koehn et al., 2003) for each language pair. We use four types of LMs: original; translated from

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 363–374,
Edinburgh, Scotland, UK, July 27–31, 2011.

the source language; translated from other languages; and a mixture of translations from several languages. We show that the translated-from-source-language LMs provide a significant improvement in the quality of the translation output over all other LMs, and that the mixture LMs always outperform the original LMs. This improvement persists even when the original LMs are up to ten times larger than the translated ones.

The main contributions of this work are therefore a computational corroboration of the hypotheses that

1. original and translated texts exhibit significant, measurable differences;
2. LMs compiled from translated texts better fit translated references than LMs compiled from original texts of the same (and much larger) size (and, to a lesser extent, LMs compiled from texts translated from languages other than the source language); and
3. MT systems that use LMs based on manually translated texts significantly outperform LMs based on originally written texts.

It is important to emphasize that translated texts abound: Many languages, especially low-resource ones, are more likely to have translated texts (religious scripts, educational materials, etc.) than original ones. Some numeric data are listed in Pym and Chrupała (2005). Furthermore, such data can be automatically identified (see Section 2). The practical impact of our work on MT is therefore potentially dramatic.

This paper is organized as follows: Section 2 provides background and describes related work. We explain our research methodology and resources in Section 3 and detail our experiments and results in Section 4. Section 5 discusses the results and their implications.

2 Background and Related Work

Numerous studies suggest that translated texts are different from original ones. Gellerstam (1986) compares texts written originally in Swedish and texts translated from English into Swedish. He notes that the differences between them do not indicate poor translation but rather a statistical phenomenon, which he terms *translationese*. He focuses mainly on lexical differences, for example less colloquialism in the translations, or foreign words used in the translations "with new shades of meaning taken from the English lexeme" (p.91). Only later studies consider grammatical differences (see, e.g., Santos (1995)). The features of translationese were theoretically organized under the terms *laws of translation* and *translation universals*.

Toury (1980, 1995) distinguishes between two laws: the *law of interference* and the *law of growing standardization*. The former pertains to the fingerprints of the source text that are left in the translation product. The latter pertains to the effort to standardize the translation product according to existing norms in the target language (and culture). Interestingly, these two laws are in fact reflected in the architecture of statistical machine translation: interference corresponds to the translation model and standardization to the language model.

The combined effect of these laws creates a hybrid text that partly corresponds to the source text and partly to texts written originally in the target language but in fact belongs to neither (Frawley, 1984). Baker (1993, 1995, 1996) suggests several candidates for translation universals, which are claimed to appear in any translated text, regardless of the source language. These include *simplification*, the tendency of translated texts to simplify the language, the message or both; and *explicitation*, their tendency to spell out implicit utterances that occur in the source text.

Baroni and Bernardini (2006) use machine learning techniques to distinguish between original and translated Italian texts, reporting 86.7% accuracy. They manage to abstract from content and perform the task using only morpho-syntactic cues. Ilisei et al. (2010) perform the same task for Spanish but enhance it theoretically in order to check the simplification hypothesis. The most informative features are lexical variety, sentence length and lexical density.

van Halteren (2008) focuses on six languages from Europarl (Koehn, 2005): Dutch, English, French, German, Italian and Spanish. For each

of these languages, a parallel six-lingual subcorpus is extracted, including an original text and its translations into the other five languages. The task is to identify the source language of translated texts, and the reported results are excellent. This finding is crucial: as Baker (1996) states, translations do resemble each other; however, in accordance with the law of interference, the study of van Halteren (2008) suggests that translation from different source languages constitute different sublanguages. As we show in Section 4.2, LMs based on translations from the source language outperform LMs compiled from non-source translations, in terms of both fitness to the reference set and improving MT.

Kurokawa et al. (2009) show that the direction of translation affects the performance of statistical MT. They train systems to translate between French and English (and vice versa) using a French-translated-to-English parallel corpus, and then an English-translated-to-French one. They find that in translating into French it is better to use the latter parallel corpus, and when translating into English it is better to use the former. Whereas they focus on the translation model, we focus on the language model in this work. We show that using a LM trained on a text translated from the source language of the MT system does indeed improve the results of the translation.

3 Methodology and Resources

3.1 Hypotheses

We investigate the following three hypotheses:

1. Translated texts differ from original texts;
2. Texts translated from one language differ from texts translated from other languages;
3. LMs compiled from manually translated texts are better for MT as measured using BLEU than LMs compiled from original texts.

We test our hypotheses by considering translations from several languages to English. For each language pair we create a reference set comprising several thousands of sentences written originally in the source language and manually translated to English. Section 3.4 provides details on the reference sets.

To investigate the first hypothesis, we train two LMs for each language pair, one created from original English texts and the other from texts translated into English. Then, we check which LM better fits the reference set.

Fitness of a LM to a set of sentences is measured in terms of *perplexity* (Jelinek et al., 1977; Bahl et al., 1983). Given a language model and a test (reference) set, perplexity measures the predictive power of the language model over the test set, by looking at the average probability the model assigns to the test data. Intuitively, a better model assigns higher probablity to the test data, and consequently has a *lower* perplexity; it is *less* surprised by the test data. Formally, the perplexity PP of a language model L on a test set $W = w_1\, w_2 \ldots w_N$ is the probability of W normalized by the number of words N Jurafsky and Martin (2008, page 96):

$$PP(L, W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P_L(w_i | w_1 \ldots w_{i-1})}} \quad (1)$$

For the second hypothesis, we extend the experiment to LMs created from texts translated from other languages to English. For example, we test how well a LM trained on French-to-English-translated texts fits the German-to-English reference set; and how well a LM trained on German-to-English-translated texts fits the French-to-English reference set.

Finally, for the third hypothesis, we use these LMs for statistical MT (SMT). For each language pair we build several SMT systems. All systems use a translation model extracted from a parallel corpus which is oblivious to the direction of the translation; and one of the above-mentioned LMs. Then, we compare the translation quality of these systems in terms of the BLEU metric (Papineni et al., 2002).

3.2 Language Models

In all the experiments, we use SRILM (Stolcke, 2002) to train 4-gram language models (with the default backoff model) from various corpora. Our main corpus is Europarl (Koehn, 2005), specifically portions collected over years 1996 to

1999 and 2001 to 2009. This is a large multilingual corpus, containing sentences translated from several European languages. However, it is organized as a collection of bilingual corpora rather than as a single multilingual one, and it is hard to identify sentences that are translated to several languages.

We therefore treat each bilingual sub-corpus in isolation; each such sub-corpus contains sentences translated from various languages. We rely on the **language** attribute of the **speaker** tag to identify the source language of sentences in the English part of the corpus. Since this tag is rarely used with English-language speakers, we also exploit the **ID** attribute of the **speaker** tag, which we match against the list of British members of the European parliament.

We focus on the following languages: German (DE), French (FR), Italian (IT), and Dutch (NL). For each of these languages, L, we consider the L-English Europarl sub-corpus. In each sub-corpus, we extract chunks of approximately 2.5 million *English* tokens translated from each of these source languages (T-L), as well as sentences written originally in English (O-EN). The mixture corpus (MIX), which is designed to represent "general" translated language, is constructed by randomly selecting sentences translated from any language (excluding original English sentences). Table 1 lists the number of sentences, number of tokens and average sentence length, for each sub-corpus and each original language.

In addition, we use the Hansard corpus, containing transcripts of the Canadian parliament from 1996–2007[1]. This is a bilingual French–English corpus comprising about 80% original English texts (EO) and about 20% texts translated from French (FO). We first separate original English from the original French and then, for each original language, we randomly extract portions of texts of different sizes: 1M, 5M and 10M tokens from the FO corpus and 1M, 5M, 10M, 25M, 50M and 100M tokens from the EO corpus; see Table 2.

[1]We are grateful to Cyril Goutte, George Foster and Pierre Isabelle for providing us with an annotated version of this corpus.

German–English			
Orig. Lang.	**Sent's**	**Tokens**	**Len**
MIX	82,700	2,325,261	28.1
O-EN	91,100	2,324,745	25.5
T-DE	87,900	2,322,973	26.4
T-FR	77,550	2,325,183	30.0
T-IT	65,199	2,325,996	35.7
T-NL	94,000	2,323,646	24.7
French–English			
Orig. Lang.	**Sent's**	**Tokens**	**Len**
MIX	90,700	2,546,274	28.1
O-EN	99,300	2,545,891	25.6
T-DE	94,900	2,546,124	26.8
T-FR	85,750	2,546,085	29.7
T-IT	72,008	2,546,984	35.4
T-NL	103,350	2,545,645	24.6
Italian–English			
Orig. Lang.	**Sent's**	**Tokens**	**Len**
MIX	87,040	2,534,793	29.1
O-EN	93,520	2,534,892	27.1
T-DE	90,550	2,534,867	28.0
T-FR	82,930	2,534,930	30.6
T-IT	69,270	2,535,225	36.6
T-NL	96,850	2,535,053	26.2
Dutch–English			
Orig. Lang.	**Sent's**	**Tokens**	**Len**
MIX	90,500	2,508,265	27.7
O-EN	97,000	2,475,652	25.5
T-DE	94,200	2,503,354	26.6
T-FR	86,600	2,523,055	29.1
T-IT	73,541	2,518,196	34.2
T-NL	101,950	2,513,769	24.7

Table 1: Europarl corpus statistics

To experiment with a non-European language (and a different genre) we choose Hebrew (HE). We use two English corpora: The *original* (O-EN) corpus comprises articles from the *International Herald Tribune*, downloaded over a period of seven months (from January to July 2009). The articles cover four topics: news (53.4%), business (20.9%), opinion (17.6%) and arts (8.1%). The *translated* (T-HE) corpus consists of articles collected from the Israeli newspaper *HaAretz* over the same period of time. *HaAretz* is published in Hebrew, but portions of

Original French			
Size	**Sent's**	**Tokens**	**Len**
1M	54,851	1,000,076	18.23
5M	276,187	5,009,157	18.14
10M	551,867	10,001,716	18.12
Original English			
Size	**Sent's**	**Tokens**	**Len**
1M	54,216	1,006,275	18.56
5M	268,806	5,006,482	18.62
10M	537,574	10,004,191	18.61
25M	1,344,580	25,001,555	18.59
50M	2,689,332	50,009,861	18.60
100M	5,376,886	100,016,704	18.60

Table 2: Hansard corpus statistics

it are translated to English. The O-corpus was downsized, so both corpora had approximately the same number of tokens in each topic. Table 3 lists basic statistics for these corpora.

Hebrew–English			
Orig. Lang.	**Sent's**	**Tokens**	**Len**
O-EN	135,228	3,561,559	26.3
T-HE	147,227	3,561,556	24.2

Table 3: Hebrew-to-English corpus statistics

3.3 SMT Training Data

To focus on the effect of the *language* model on translation quality, we design SMT training corpora to be oblivious to the direction of translation. Again, we use Europarl (January 2000 to September 2000) as the main source of our parallel corpora. We also use the Hansard corpus: We randomy extract 50,000 sentences from the original French sub-corpora and another 50,000 sentences from the original English sub-corpora. For Hebrew we use the Hebrew–English parallel corpus (Tsvetkov and Wintner, 2010) which contains sentences translated from Hebrew to English (54%) and from English to Hebrew (46%). The English-to-Hebrew part comprises many short sentences (approximately 6 tokens per sentence) taken from a movie subtitle database. This explains the small token to sentence ratio of this particular corpus. Table 4 lists some details on those corpora.

Lang's	**Side**	**Sent's**	**Tokens**	**Len**
DE-EN	DE	92,901	2,439,370	26.3
	EN	92,901	2,602,376	28.0
FR-EN	FR	93,162	2,610,551	28.0
	EN	93,162	2,869,328	30.8
IT-EN	IT	85,485	2,531,925	29.6
	EN	85,485	2,517,128	29.5
NL-EN	NL	84,811	2,327,601	27.4
	EN	84,811	2,303,846	27.2
Hansard	FR	100,000	2,167,546	21.7
	EN	100,000	1,844,415	18.4
HE-EN	HE	95,912	726,512	7.6
	EN	95,912	856,830	8.9

Table 4: SMT training data details

3.4 Reference Sets

The reference sets have two uses. First, they are used as the test sets in the experiments that measure the perplexity of the language models. Second, in the MT experiments we use them to randomly extract 1000 sentences for tuning and 1000 (different) sentences for evaluation.

For each language L we use the L-English subcorpus of Europarl (over the period of October to December 2000), containing only sentences originally produced in language L. The Hansard reference set is completely disjoint from the LM and SMT training sets and comprises only original French sentences. The Hebrew-to-English reference set is an independent (disjoint) part of the Hebrew-to-English parallel corpus. This set mostly comprises literary data (88.6%) and a small portion of news (11.4%). All sentences are originally written in Hebrew and are manually translated to English. See Table 5.

4 Experiments and Results

We detail in this section the experiments performed to test the three hypotheses: that translated texts can be distinguished from original ones, and provide better language models of other translated texts; that texts translated from other languages than the source are still better predictors of translations than original texts (Section 4.1); and that these differences are important for SMT (Section 4.2).

Lang's	Side	Sent's	Tokens	Len
DE-EN	DE	6,675	161,889	24.3
	EN	6,675	178,984	26.8
FR-EN	FR	8,494	260,198	30.6
	EN	8,494	271,536	32.0
IT-EN	IT	2,269	82,261	36.3
	EN	2,269	78,258	34.5
NL-EN	NL	4,593	114,272	24.9
	EN	4,593	105,083	22.9
Hansard	FR	8,926	193,840	21.72
	EN	8,926	163,448	18.3
HE-EN	HE	7,546	102,085	13.5
	EN	7,546	126,183	16.7

Table 5: Reference sets

4.1 Translated vs. Original texts

We train several 4-gram LMs for each Europarl sub-corpus, based on the corpora described in Section 3.2. For each language L, we train a LM based on texts translated from L, from languages other than L as well as texts originally written in English. The LMs are applied to the reference set of texts translated from L, and we compute the perplexity: the fitness of the LM to the reference set. Table 6 details the results, where for each sub-corpus and LM we list the number of unigrams in the test set, the number of out-of-vocabulary items (OOV) and the perplexity (PP). The lowest perplexity (reflecting the **best** fit) in each sub-corpus is typeset in boldface, and the highest (*worst* fit) is slanted.

These results overwhelmingly support our hypothesis. For each language L, the perplexity of the LM that was created from L translations is lowest, followed immediately by the MIX LM. Furthermore, the perplexity of the LM created from originally-English texts is highest in all experiments. In addition, the perplexity of LMs constructed from texts translated from languages other than L always lies between these two extremes: it is a better fit of the reference set than original texts, but not as good as texts translated from L (or mixture translations). This corroborates the hypothesis that translations form a language in itself, and translations from L_1 to L_2, form a sub-language, related to yet different from translations from other languages to L_2.

German to English translations			
Orig. Lang.	**Unigrams**	**OOV**	**PP**
MIX	32,238	961	83.45
O-EN	31,204	1161	*96.50*
T-DE	27,940	963	**77.77**
T-FR	29,405	1141	92.71
T-IT	28,586	1122	95.14
T-NL	28,074	1143	89.17
French to English translations			
Orig. Lang.	**Unigrams**	**OOV**	**PP**
MIX	33,444	1510	87.13
O-EN	32,576	1961	*105.93*
T-DE	28,935	2191	96.83
T-FR	30,609	1329	**82.23**
T-IT	29,633	1776	91.15
T-NL	29,221	2148	100.18
Italian to English translations			
Orig. Lang.	**Unigrams**	**OOV**	**PP**
MIX	33,353	462	90.71
O-EN	32,546	633	*107.45*
T-DE	28,835	628	100.46
T-FR	30,460	524	92.18
T-IT	29,466	470	**80.57**
T-NL	29,130	675	105.07
Dutch to English translations			
Orig. Lang.	**Unigrams**	**OOV**	**PP**
MIX	33,050	651	87.37
O-EN	32,064	771	*100.75*
T-DE	28,766	778	90.35
T-FR	30,502	775	96.38
T-IT	29,386	916	99.26
T-NL	29,178	560	**78.25**

Table 6: Fitness of various LMs to the reference set

A possible explanation for the different perplexity results between the LMs could be the specific contents of the corpora used to compile the LMs. To rule out this possibility and to further emphasize that the corpora are indeed *structurally* different, we conduct more experiments, in which we gradually abstract away from the domain- and content-specific features of the texts and emphasize their syntactic structure. We focus on German-to-English.

First, we remove all punctuation to eliminate

possible bias due to differences in punctuation conventions. Then, we use the Stanford Named Entity Recognizer (Finkel et al., 2005) to identify named entities, which we replace with a unique token ('NE'). Next, we replace all nouns with their POS tag; we use the Stanford POS Tagger (Toutanova and Manning, 2000). Finally, for full lexical abstraction, we replace all words with their POS tags.

At each step, we train six language models on O- and T-texts and apply them to the reference set (adapted to the same level of abstraction, of course). As the abstraction of the text increases, we also increase the order of the LMs: From 4-grams for text without punctuation and NE abstraction to 5-grams for noun abstraction to 8-grams for full POS abstraction. The results, which are depicted in Table 7, consistently show that the T-based LM is a better fit to the reference set, albeit to a lesser extent. While we do not show the details here, the same pattern is persistent in all the other Europarl languages we experiment with.

We repeat this experiment with the Hebrew-to-English reference set. We train two 4-gram LMs on the O-EN and T-HE corpora. We then apply the two LMs to the reference set and compute the perplexity. The results are presented in Table 8. Although the T-based LM has more OOVs, it is a better fit to the translated text than the O-based LM: Its perplexity is lower by 20.1%. Interestingly, the O-corpus LM has more unique unigrams than the T-corpus LM, supporting the claim of Al-Shabab (1996) that translated texts have lower type-to-token ratio.

We also conduct the above-mentioned abstraction experiments. The results, which are depicted in Table 9, consistently show that the T-based LM is a better fit to the reference set.

Clearly, then, translated LMs better fit the references than original ones, and the differences can be traced back not just to (trivial) specific lexical choice, but also to syntactic structure, as evidenced by the POS abstraction experiments. In fact, in order to retain the low perplexity level of translated texts, a LM based on original texts must be approximately ten times larger. We establish this by experimenting with the Hansard corpus. The results are persistent, but are omitted for lack of space.

No Punctuation			
Orig. Lang.	**OOVs**	**PP**	**PP diff.**
MIX	770	109.36	7.58%
O-EN	946	*127.03*	20.43%
T-DE	795	**101.07**	0.00%
T-FR	909	122.03	17.18%
T-IT	991	125.36	19.38%
T-NL	936	117.37	13.89%
NE Abstraction			
Orig. Lang.	**OOVs**	**PP**	**PP diff.**
MIX	643	99.13	6.99%
O-EN	772	*114.19*	19.26%
T-DE	661	**92.20**	0.00%
T-FR	752	110.22	16.35%
T-IT	823	112.72	18.21%
T-NL	771	105.81	12.86%
Noun Abstraction			
Orig. Lang.	**OOVs**	**PP**	**PP diff.**
MIX	400	38.48	4.71%
O-EN	459	*42.06*	12.80%
T-DE	405	**36.67**	0.00%
T-FR	472	40.96	10.47%
T-IT	489	41.39	11.39%
T-NL	440	39.54	7.26%
POS Abstraction			
Orig. Lang.	**OOVs**	**PP**	**PP diff.**
MIX	0	8.02	1.22%
O-EN	0	*8.19*	3.31%
T-DE	0	**7.92**	0.00%
T-FR	0	8.10	2.16%
T-IT	0	8.12	2.50%
T-NL	0	8.03	1.42%

Table 7: Fitness of O- vs. T-based LMs to the reference set (DE-EN), different abstraction levels

4.2 Original vs. Translated LMs for MT

The last hypothesis we test is whether a better fitting language model yields a better machine translation system. In other words, we expect the T-based LMs to outperform the O-based LMs when used as part of an MT system. We construct German-to-English, French-to-English, Italian-to-English and Dutch-to-

Hebrew to English translations			
Orig. Lang.	**Unigrams**	**OOV**	**PP**
O-EN	74,305	2,955	*282.75*
T-HE	61,729	3,253	**226.02**

Table 8: Fitness of O- vs. T-based LMs to the reference set (HE-EN)

No Punctuation			
Orig. Lang.	**OOVs**	**PP**	**PP diff.**
O-EN	2,601	442.95	19.2%
T-HE	2,922	358.11	0.0%
NE Abstraction			
Orig. Lang.	**OOVs**	**PP**	**PP diff.**
O-EN	1,794	350.3	17.3%
T-HE	2,038	289.71	0.0%
Noun Abstraction			
Orig. Lang.	**OOVs**	**PP**	**PP diff.**
O-EN	679	93.31	12.4%
T-HE	802	81.72	0.0%
POS Abstraction			
Orig. Lang.	**OOVs**	**PP**	**PP diff.**
O-EN	0	11.47	6.2%
T-HE	0	10.76	0.0%

Table 9: Fitness of O- vs. T-based LMs to the reference set (HE-EN), different abstraction levels

English MT systems using the Moses phrase-based SMT toolkit (Koehn et al., 2007). The systems are trained on the parallel corpora described in Section 3.3. We use the reference sets (Section 3.4) as follows: 1,000 sentences are randomly extracted for minimum error-rate tuning (Och, 2003), and another set of 1,000 sentences is randomly used for evaluation. Each system is built and tuned with six different LMs: MIX, O-based and four T-based (Section 3.2). We use BLEU (Papineni et al., 2002) to evaluate translation quality. The results are listed in Table 10.

These results are consistent: the translated-from-source systems outperform all other systems; mixture models come second; and systems that use original English LMs always perform worst. We test the statistical significance of differences between various MT systems using the bootstrap resampling method (Koehn, 2004). In all experiments, the best system (translated-from-source LM) is significantly better than all

DE to EN	
LM	**BLEU**
MIX	21.95
O-EN	*21.35*
T-DE	**22.42**
T-FR	21.47
T-IT	21.79
T-NL	21.59

IT to EN	
LM	**BLEU**
MIX	26.79
O-EN	*25.69*
T-DE	25.86
T-FR	26.56
T-IT	**27.28**
T-NL	25.77

FR to EN	
LM	**BLEU**
MIX	25.43
O-EN	*24.85*
T-DE	25.03
T-FR	**25.91**
T-IT	25.44
T-NL	25.17

NL to EN	
LM	**BLEU**
MIX	25.17
O-EN	*24.46*
T-DE	25.12
T-FR	24.79
T-IT	24.93
T-NL	**25.73**

Table 10: Machine translation with various LMs

other systems ($p < 0.05$); (even more) significantly better than the O-EN system ($p < 0.01$); and the mixture systems are significantly better than the O-EN systems ($p < 0.01$).

We also construct a Hebrew-to-English MT system using Moses' factored translation model (Koehn and Hoang, 2007). Every token in the training corpus is represented as two factors: surface form and lemma. Moreover, the Hebrew input is fully segmented. The system is built and tuned with O- and T-based LMs. Table 11 depicts the performance of the systems. The T-based LM yields a statistically better BLEU score than the O-based system.

LM	**BLEU**	**p-value**
O-based LM	*11.98*	0.012
T-based LM	**12.57**	

Table 11: Hebrew-to-English MT results

The LMs used in the above experiments are small. We now want to assess whether the benefits of using translated LMs carry over to scenarios where large original corpora exist. We build yet another set of French-to-English MT systems. We use the Hansard SMT translation model and Hansard LMs to train nine MT systems, three with varying sizes of translated texts and six with varying sizes of original texts.

We tune and evaluate on the Hansard reference set. In another set of experiments we use the Europarl French-to-English scenario (using Europarl corpora for the translation model and for tuning and evaluation), but we use the nine Hansard LMs to see whether our findings are consistent also when LMs are trained on out-of-domain (but similar genre) material.

Table 12 shows that the original English LMs should be enlarged by a factor of *ten* to achieve translation quality similar to that of translation-based LMs. In other words, much smaller translated LMs perform better than much larger original ones, and this is true for various LM sizes.

In-Domain		**Out-of-Domain**	
Original French		**Original French**	
Size	**BLEU**	**Size**	**BLEU**
1M	34.05	1M	18.87
5M	35.12	5M	23.90
10M	35.65	10M	24.36
Original English		**Original English**	
Size	**BLEU**	**Size**	**BLEU**
1M	32.57	1M	18.68
5M	33.37	5M	23.02
10M	33.92	10M	23.45
25M	34.71	25M	23.82
50M	34.85	50M	23.95
100M	35.36	100M	24.16

Table 12: The effect of LM size on MT performance

5 Discussion

We use language models computed from different types of corpora to investigate whether their fitness to a reference set of translated-to-English sentences can differentiate between them (and, hence, between the corpora on which they are based). Our main findings are that LMs compiled from manually translated corpora are much better predictors of translated texts than LMs compiled from original-language corpora of the same size. The results are robust, and are sustainable even when the corpora and the reference sentences are abstracted in ways that retain their syntactic structure but ignore specific word meanings. Furthermore, we show that translated LMs are better predictors of translated sentences even when the LMs are compiled from texts translated from languages *other* than the source language. However, LMs based on texts translated from the source language still outperform LMs translated from other languages.

We also show that MT systems based on translated-from-source-language LMs outperform MT systems based on originals LMs or LMs translated from other languages. Again, these results are robust and the improvements are statistically significant. This effect seems to be amplified as translation quality improves. Furthermore, our results show that original LMs require ten times more data to exhibit the same fitness to the reference set and the same translation quality as translated LMs.

More generally, this study confirms that insights drawn from the field of theoretical translation studies, namely the dual claim according to which (1) translations as such differ from originals, and (2) translations from different source languages differ from each other, can be verified experimentally and contribute to the performance of machine translation.

Future research is needed in order to understand *why* this is the case. One plausible hypothesis is that recurrent multiword expressions in the source language are frequently solved by human translations and each of these expressions converges to a set of high-quality translation equivalents which are represented in the LM. Another hypothesis is that since translation-based LMs represent a simplified mode of language use, the error potential is smaller. We therefore expect translation-based LMs to use more unmarked forms.

This work also bears on language typology: we conjecture that LMs compiled from texts translated not from the original language, but from a closely related one, can be better than texts translated from a more distant language. Some of our results support this hypothesis, but more research is needed in order to establish it.

Acknowledgements

This research was supported by the Israel Science Foundation (grant No. 137/06). We are grateful to Alon Lavie for his consistent help.

References

Omar S. Al-Shabab. *Interpretation and the language of translation: creativity and conventions in translation*. Janus, Edinburgh, 1996.

Lalit R. Bahl, Frederick Jelinek, and Robert L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):179–190, 1983.

Mona Baker. Corpus linguistics and translation studies: Implications and applications. In Gill Francis Mona Baker and Elena Tognini-Bonelli, editors, *Text and technology: in honour of John Sinclair*, pages 233–252. John Benjamins, Amsterdam, 1993.

Mona Baker. Corpora in translation studies: An overview and some suggestions for future research. *Target*, 7(2):223–243, September 1995.

Mona Baker. Corpus-based translation studies: The challenges that lie ahead. In Gill Francis Mona Baker and Elena Tognini-Bonelli, editors, *Terminology, LSP and Translation. Studies in language engineering in honour of Juan C. Sager*, pages 175–186. John Benjamins, Amsterdam, 1996.

Marco Baroni and Silvia Bernardini. A new approach to the study of Translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274, September 2006. URL `http://llc.oxfordjournals.org/cgi/content/short/21/3/259?rss=1`.

Thorsten Brants and Peng Xu. Distributed language models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Tutorial Abstracts*, pages 3–4, Boulder, Colorado, May 2009. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/N/N09/N09-4002`.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370, Morristown, NJ, USA, 2005. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1219840.1219885.

William Frawley. Prolegomenon to a theory of translation. In William Frawley, editor, *Translation. Literary, Linguistic and Philosophical Perspectives*, pages 159–175. University of Delaware Press, Newark, 1984.

Martin Gellerstam. Translationese in Swedish novels translated from English. In Lars Wollin and Hans Lindquist, editors, *Translation Studies in Scandinavia*, pages 88–95. CWK Gleerup, Lund, 1986.

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. Identification of translationese: A machine learning approach. In Alexander F. Gelbukh, editor, *Proceedings of CICLing-2010: 11th International Conference on Computational Linguistics and Intelligent Text Processing*, volume 6008 of *Lecture Notes in Computer Science*, pages 503–511. Springer, 2010. ISBN 978-3-642-12115-9. URL `http://dx.doi.org/10.1007/978-3-642-12116-6`.

Frederick Jelinek, Robert L. Mercer, Lalit R. Bahl, and J. K. Baker. Perplexity—a measure of the difficulty of speech recognition tasks. *Journal of the Acoustical Society of America*, 62:S63, November 1977. Supplement 1.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall, second edition, February 2008. ISBN 013122798X. URL `http://www.worldcat.org/isbn/013122798X`.

Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Philipp Koehn. Europarl: A parallel corpus for

statistical machine translation. MT Summit, 2005.

Philipp Koehn and Hieu Hoang. Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 868–876, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D/D07/D07-1091`.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54. Association for Computational Linguistics, 2003.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P07-2045`.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. Automatic detection of translated text and its impact on machine translation. In *Proceedings of MT-Summit XII*, 2009.

Franz Josef Och. Minimum error rate training in statistical machine translation. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167, Morristown, NJ, USA, 2003. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1075096.1075117.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA, 2002. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1073083.1073135.

Anthony Pym and Grzegorz Chrupała. The quantitative analysis of translation flows in the age of an international language. In Albert Branchadell and Lovell M. West, editors, *Less Translated Languages*, pages 27–38. John Benjamins, Amsterdam, 2005.

Diana Santos. On grammatical translationese. In *In Koskenniemi, Kimmo (comp.), Short papers presented at the Tenth Scandinavian Conference on Computational Linguistics (Helsinki*, pages 29–30, 1995.

Andreas Stolcke. SRILM—an extensible language modeling toolkit. In *Procedings of International Conference on Spoken Language Processing*, pages 901–904, 2002. URL `citeseer.ist.psu.edu/stolcke02srilm.html`.

Gideon Toury. *In Search of a Theory of Translation*. The Porter Institute for Poetics and Semiotics, Tel Aviv University, Tel Aviv, 1980.

Gideon Toury. *Descriptive Translation Studies and beyond*. John Benjamins, Amsterdam / Philadelphia, 1995.

Kristina Toutanova and Christopher D. Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora*, pages 63–70, Morristown, NJ, USA, 2000. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1117794.1117802.

Yulia Tsvetkov and Shuly Wintner. Automatic acquisition of parallel corpora from websites with dynamic content. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*,

pages 3389–3392. European Language Resources Association (ELRA), May 2010. ISBN 2-9517408-6-7.

Hans van Halteren. Source language markers in EUROPARL translations. In *COLING '08: Proceedings of the 22nd International Conference on Computational Linguistics*, pages 937–944, Morristown, NJ, USA, 2008. Association for Computational Linguistics. ISBN 978-1-905593-44-6.