Human Language Technologies – The Baltic Perspective
A. Tavast et al. (Eds.)

doi:10.3233/978-1-61499-133-5-50

In-domain Data FTW

Mark FISHEL [1]

Institute of Computer Science, University of Tartu, Estonia
Institute of Computational Linguistics, University of Zurich, Switzerland

Abstract. We describe experiments on Estonian-English statistical machine translation with a strong emphasis on domain adaptation. We show that disregarding text domains can harm a translation system and that even a small in-domain corpus can lead to significant translation quality improvements.

Keywords. statistical machine translation, domain adaptation, parallel corpora

1. Motivation

Languages outside the mainstream computational linguistics research have to do with whatever resources can be laid hands on. In case of text corpora this often means that their domain (i.e. the type of text, e.g. news articles, fiction, legalese, etc.) is disregarded, as the domain choice is limited.

Here we present the development of an online translation system for Estonian-English.[2] Its engine being a statistical machine translation system, parallel corpora for the initial version of the system was selected based on availability, without any analysis of typical user input.

We first introduce a small parallel corpus, created by collecting user input into the `masintolge.ut.ee` translation system and translating it from Estonian into English; it thus constitutes an in-domain corpus in respect to the translation system. Using this corpus we demonstrate the effects of disregarding text domains; we then explore two domain adaptation techniques for our translation system – tuning on in-domain data and weighed combination of general-domain corpora.

2. Related work

Early work on parallel resources involving Estonian was carried out in 2003-2005 and resulted in a medium-sized Estonian-English parallel corpus of legislation texts[3]. Later several multilingual corpora were created, that also included Estonian among other languages: e.g. JRC-Acquis [1], OpenSubtitles and KDE [2]. Recently Estonian has been

[1] Corresponding Author: Mark Fishel, Institute of Computational Linguistics, University of Zurich, Binzmühlestrasse 14, 8050 Zürich, Switzerland; E-mail: fishel@cl.uzh.ch.

[2] `http://masintolge.ut.ee`

[3] `http://www.cl.ut.ee/korpused/paralleel/`

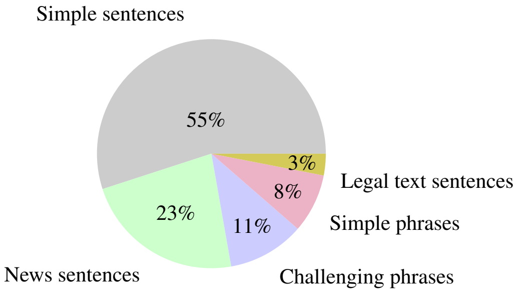

Figure 1. Distribution of text domains within the TempEst corpus (in terms of number of tokens).

added into the Europarl corpus [3], and a substantially larger version of OpenSubtitles has been released.

Previous work on statistical translation from and into Estonian includes [4,5,6]. Public online translation systems that support translation from or into Estonian at the moment of writing this paper are Google Translate[4] and Bing Translator[5].

3. TempEst: a Small Corpus of User Input Translations

Our first task was to assess the domain distribution in the `masintolge.ut.ee` translation system user input. In order to do so and also to estimate the translation quality of the system on in-domain data we created the TempEst corpus.

First we collected the phrases and sentences that were sent as input to the translation system by its users during a period of 6 months. Undesired content (such as input in languages other than Estonian, single words and repeated phrases or sentences) was filtered out, leaving a total of 2650 Estonian phrases and sentences.

Next, we manually analyzed the domain distribution of the Estonian input on a random 20% of the whole set; results are presented in Figure 1. Most of the corpus consists of simpler sentences and phrases like “how are you doing” and “my name is John”; a smaller part consists of intentional challenges (tongue twisters, proverbs, fiction quotes, swearing) and news article and legal text sentences.

As a next step the list of Estonian phrases and sentences was manually translated into English. Translation was performed by a single human translator, who contributed several alternative translations when appropriate – e.g. in case of idiomatic expressions or ambiguous meanings.

After tokenization the Estonian input had 15 737 tokens and the English translations – 22 975 tokens, averaged over the alternative translations. This means that the input is on average slightly shorter than 6 words.

Given the small number of tokens and translation units, the TempEst corpus is too small for being used as a training corpus for statistical machine translation; on the other hand, it can be used for tuning SMT systems as a development corpus and for evaluating translation quality as a test corpus.

[4] `http://translate.google.com`

[5] `http://www.microsofttranslator.com`

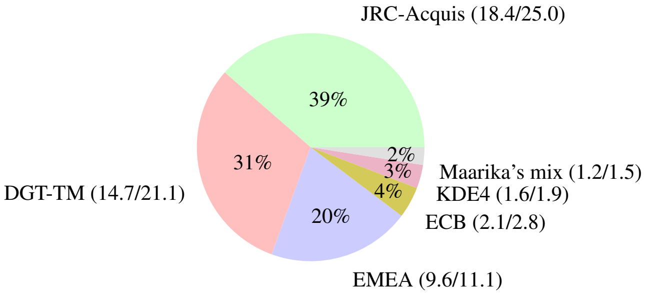

Figure 2. Corpora sizes and domain distribution of the smaller corpus set, without in-domain-like data; sizes and percentage are in terms of Estonian/English tokens ($\times 10^6$) – a total of 47.6 Estonian / 63.3 English.

The newly created TempEst corpus is openly accessible online [6]; in the next section we present several translation experiments that use it.

4. Experiments

All experiments within this paper were performed with phrase-based statistical machine translation models, trained according to the baseline setup of the shared task of WMT'12 [7]. To account for the instability in the tuning stage, five independent MERT runs were performed and the one yielding the best BLEU score on the development corpus was used to translate the test set for evaluation. Language models were trained on the target side of the parallel corpora. Translation quality is estimated using the MultEval package[7], which implements the BLEU, TER and METEOR scores and significance tests for them.

From here on the domain mix of the TempEst corpus will be referred to as in-domain and the collection of training corpora constitutes the general-domain material.

4.1. Disregarding Text Domains

Initial development of `masintolge.ut.ee` was based on all available parallel Estonian-English corpora, which at the time included legal texts (JRC-Acquis [1], DGT-TM [8]), website and documentation texts (from the banking domain in ECB and the medical domain in EMEA [2]), software localization data (KDE4 [2]) and a set of small texts based mostly on Estonian national websites, documents and tourist information[8] was also used. Corpora sizes and the distribution of domains are shown on Figure 2. All of these domains differ from the target domain of the translation system and TempEst.

In the beginning of the system's development a small experiment was performed to select a word alignment method with the best effect on translation quality. The described general-domain corpus mix was split into training, development and testing corpora for training the system, tuning the model weights and estimating translation quality; in other words, general-domain translation quality was evaluated. Results are shown in Table 1.

[6] `http://statmt.ut.ee`
[7] `http://github.com/jhclark/multeval`
[8] collected and provided by Dr. Maarika Traat

Table 1. Translation quality estimates of a general-domain-tuned system, applied to a general-domain test corpus (before/after tuning). Length is given as average percentage of the reference translation length.

	BLEU	**METEOR**	**TER**	**Length (%)**
GIZA++$_{diag}$	56.1 / 57.7	43.3 / 44.0	36.0 / 36.0	97.8 / 100.3
GIZA++$_{inter}$	55.7 / 62.9	42.1 / 45.9	34.5 / 33.6	86.0 / 101.2
Berkeley	60.2 / 63.4	45.1 / 46.5	31.1 / 30.9	93.7 / 99.7

Table 2. Translation quality estimates of a general-domain-tuned system, applied to an in-domain test corpus (before/after tuning). Length is given as average percentage of the reference translation length.

	BLEU	**METEOR**	**TER**	**Length (%)**
GIZA++$_{diag}$	14.1 / 13.7	24.1 / 24.3	67.3 / 71.0	88.3 / 94.0
GIZA++$_{inter}$	11.8 / 10.2	22.0 / 22.4	67.6 / 85.0	77.9 / 107.9
Berkeley	14.1 / 13.8	24.2 / 24.6	65.2 / 70.2	82.2 / 93.3

After the creation of the TempEst corpus, the same comparison of word alignment methods was repeated for in-domain translation: the same translation systems, tuned on a general-domain development set were applied to the in-domain TempEst. These results can be found in Table 2; both un-tuned and tuned system results are presented.

Comparing the two tables, not only would one draw different conclusions about the word alignment methods, but even the effect of the tuning step is different: while MERT significantly increases the scores for general-domain translation, in-domain scores either drop (BLEU, TER) or practically do not change (METEOR) as a result of tuning.

This effect can be explained by the nature of the tuning step: its aim is to increase translation quality estimates for a particular development set – the translation system is therefore tuned on the kind of text, represented in that set. Its performance on texts with different lexical choices, length ratios, word and phrase order patterns, etc. can thus suffer. The effect of tuning on length ratios is explicit in the result tables: while the length of general-domain translations is rather stable in respect to average reference translation length (99.7%–101.2%), in-domain translation length is more unstable (93.3%–107.9%).

Based on these results it is clear that the next essential step is to tune the system using in-domain data.

4.2. *In-domain Tuning*

The TempEst corpus was our only option for an in-domain development set. However, it is too small to split into held-out development and test sets; besides, translation quality estimations would not be comparable to previous results if computed on another corpus.

To guarantee comparable results, we instead performed tuning and evaluation via 2-fold cross-validation: the corpus was split into two halves, then two independent tuning sessions were performed using both halves and finally each half-corpus was translated for evaluation using the system tuned on the other half. Finally the two half-corpus translations were concatenated and further treated as a whole.

Results of cross-validated tuning are given in Table 3, including scores of un-tuned and tuned systems; the positive effect of tuning on in-domain data is clearly seen, as the

Table 3. Translation quality estimates of a system, tuned and evaluated via 2-fold cross-validation over an in-domain corpus (before/after tuning). Length is given as average percentage of the reference translation length.

	BLEU	**METEOR**	**TER**	**Length (%)**
GIZA++$_{diag}$	14.1 / 16.0	24.1 / 25.7	67.3 / 68.0	88.3 / 96.8
GIZA++$_{inter}$	11.8 / 14.6	22.0 / 24.3	67.6 / 71.5	77.9 / 98.4
Berkeley	14.1 / 16.4	24.2 / 25.9	65.2 / 69.1	82.2 / 99.0

quality estimates improve as a result of tuning. Length ratios between the hypothesis and reference translations are also much more stable (96.8%–99.0%).

Interestingly enough, the TER scores still worsen as a result of MERT (even though much less than in case of tuning on general-domain data). However, the TER score is sensitive to sentence length: since it is an edit distance measure, shorter sentences need less edit operations to turn into the reference translation in case they are distant enough. Looking at the length ratios of the translations before and after tuning the systems (Tables 1, 2, 3) it becomes apparent that the TER scores of the un-tuned systems are unrealistically positive because of shorter translations. As for shorter translations from un-tuned systems, these are most likely caused by a high initial value of the word penalty. Thus, MERT reduces the relative weight of the word penalty, translation lengths increase as a result and consequently the TER scores deteriorate.

4.3. Weighted Corpus Combination

All previously described experiments treated the training corpus mix as a homogeneous collection. However, domain difference is not a binary phenomenon and various domains can be more or less similar to each other. This suggests that treating corpora from different domains separately (e.g. giving clear preference to some of them in case of conflicting translations) could reduce the effects of out-of-the-box out-of-domain data usage and lead to translation quality improvement.

In order to do so we selected the TMCombine package [9]. TMCombine takes separate phrase tables, trained on each corpus individually and produces a joint phrase table from a linear combination of the individual tables. Using perplexity of the joint phrase table on a development set as a quality measure it finds a set of weights for the linear combination that would optimize perplexity.

Here we initially applied TMCombine to the same set of training corpora as in the previous experiments – a total of 14, counting the additional small corpora.

At some point after all the presented experiments were performed, Estonian was added to Europarl [3] and a substantially larger version of OpenSubtitles was created [2]. Both contain texts from domains that are much more similar to TempEst than any previously existing parallel corpora. Added to this, a multilingual corpus based on the official journal of the EU was created;[9] the material there is more similar to the JRC-Acquis corpus, but is perhaps slightly more similar to the news domain. The new corpora sizes and domain distribution are given on Figure 3.

Tuning and evaluation with the smaller and bigger corpus set was done again with 2-fold cross-validation over the TempEst corpus. Results are presented in Table 4. In

[9] `http://apertium.eu/data`

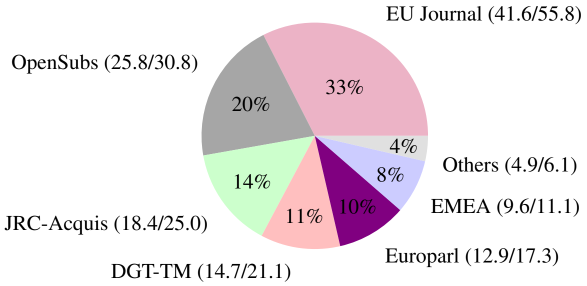

Figure 3. Corpora sizes and domain distribution of the larger corpus set, with in-domain-like data; sizes and percentage are in terms of Estonian/English tokens ($\times 10^6$) – a total of 127.8 Estonian / 167.2 English.

Table 4. Comparison of systems based on non-optimized corpus concatenation ("Uniform") and optimized linear combination by TMCombine ("Optimized").

Without in-domain-like data:

	BLEU	**METEOR**	**TER**
Uniform	16.0	25.7	68.0
Optimized	17.2	26.5	66.7

With in-domain-like data:

	BLEU	**METEOR**	**TER**
Uniform	27.5	33.6	54.7
Optimized	27.4	33.6	55.7

case of the smaller corpus set weighted combination of corpora results in significant improvement of all scores. The highest weights in the linear combination were given to KDE4 (56%), DGT-TM (14%) and JRC-Acquis (12%).

Adding the new corpora provides a vast improvement to translation quality estimates. On the other hand, judging by the results a large in-domain-like corpus (OpenSubtitles) overweights the other corpora on its own, and TMCombine cannot improve over that baseline. Looking at the optimized corpora weights, TMCombine gives 71% of the weight mass to OpenSubtitles; the remaining mass is mostly divided between Europarl (18%), EU Journal (5%) and KDE4 (2%). Also, our intuition of OpenSubtitles (and Europarl to a smaller degree) being closer by domain to TempEst seems to be correct.

Thus TMCombine apparently does not lead to improvements in translation quality when a large in-domain training corpus is available. On the other hand it performs well when there is little or no in-domain training data, which is in general a typical scenario.

5. Final Translation System Evaluation

For the final `masintolge.ut.ee` translation service we selected the engine, trained on the larger corpus collection without optimization; in this section we evaluate this final system. We first compare it to two publicly available translation services, Google Translate and Bing Translator.

Just like our system, the engine of Google Translate uses the statistical machine translation paradigm. It is trained on an in-house corpus collection, which is collected from the web and is supposedly excessively large.[10]

[10] `http://translate.google.com/about`

Table 5. Comparison of the performance of `masintolge.ut.ee` to two publicly available translation services on the TempEst corpus.

	BLEU	**METEOR**	**TER**
`masintolge.ut.ee`	27.5	33.6	54.7
Google Translate	29.3	34.4	51.1
Bing Translator	24.9	31.2	57.7

src:	teleka pult
ref:	tv remote control
m-t:	the tv remote
g-t:	tv remote control
b-t:	*teleka* remote control

src:	kas sa šokolaadi tahad ?
ref:	do you want a chocolate ?
m-t:	do you want the chocolate ?
g-t:	do you want chocolate ?
b-t:	do you want a chocolate ?

src:	mul on paha
ref:	i feel sick
m-t:	i have bad
g-t:	i have a bad
b-t:	i have a bad

src:	mine metsa kuradi tõlge
ref:	go to hell , bloody translation
m-t:	go to the forest fucking translation
g-t:	translation of the devil in the forest
b-t:	go to the translation of forest fucking

Figure 4. Examples of Estonian sentences (**src**) and their translations in the TempEst corpus (**ref**), by `masintolge.ut.ee` (**m-t**), Google Translate (**g-t**) and Bing Translator (**b-t**). Translations with Google Translate and Bing Translator were done on July 3, 2012.

Bing Translator was created by Microsoft purchasing the Yahoo! Babel Fish translation service, which is based on Systran's[11] rule-based translation engine.

To compare `masintolge.ut.ee` to the two translation services the latter were applied to the TempEst corpus; translation quality estimates are presented in Table 5. In this evaluation the scores of Google Translate are higher than of `masintolge.ut.ee`; both have significantly higher scores than Bing Translator. Superiority of Google Translate can be explained at least by a probably larger corpus, since their system has access to the same corpora as we do, plus their own large collection.

Figure 4 presents some translation examples by the three systems. In some cases Bing Translator produces a more grammatical output; in several other cases Google Translator and `masintolge.ut.ee` cope better than Bing Translator with non-conventional input, such as slang.

In order to see, how well `masintolge.ut.ee` and Google Translate perform on different domains within TempEst we evaluate translation quality on separate domains in Table 6. To avoid unreliable estimates on too small corpora legal sentences were grouped with the news domain.

According to this comparison our system has the same quality estimates like Google Translate on simple sentences and intentionally difficult phrases; simple phrases, news article and legal text sentences are translated with higher quality by Google. The results also show that translating simpler sentences and phrases is indeed easier than news, legalese or tongue twisters, proverbs, etc.

[11] `http://www.systran.co.uk/`

Table 6. Translation quality estimation, split into sub-domains of the TempEst corpus for two translation systems: `masintolge.ut.ee` / Google Translate.

	BLEU	**METEOR**	**TER**
Simple sentences	31.3 / 31.7	36.9 / 36.5	48.4 / 48.0
Simple phrases	27.4 / 31.5	34.0 / 34.7	54.6 / 54.3
News, Legalese	13.9 / 21.1	26.8 / 28.9	72.5 / 62.1
Proverbs, Swearing, etc.	17.7 / 18.0	26.8 / 27.5	61.0 / 63.1

6. Conclusions

We have described the development of an online Estonian-English translation service, `masintolge.ut.ee` and introduced a small parallel corpus of user inputs into this service. Using the corpus we have shown the danger of ignoring text domains when using parallel corpora.

Our experiments show that even a small in-domain corpus can be invaluable to developing a translation system. The least anyone can do is perform the tuning step on an in-domain corpus; in case of scarce data using a method of automatic weight optimization for weighted corpus combination also proves to be efficient.

The `masintolge.ut.ee` translation system shows competitive results in case of some text domains, even though in the general evaluation Google Translate has slightly higher scores. In the future we are considering adding the opposite translation direction (English-Estonian) and other language pairs to the system.

References

[1] Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dniel Varga. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of LREC'2006*, pages 2142–2147, Genoa, Italy, 2006.

[2] Jörg Tiedemann. News from OPUS - a collection of multilingual parallel corpora with tools and interfaces. In *Proceedings of RANLP'2009*, volume V, pages 237–248, Borovets, Bulgaria, 2009.

[3] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X*, pages 79–86, Phuket, Thailand, 2005.

[4] Mark Fishel, Heiki-Jaan Kaalep, and Kadri Muischnek. Estonian-english statistical machine translations: the first results. In *Proceedings of NODALIDA'2007*, pages 278–283, Tartu, Estonia, 2007.

[5] Mark Fishel and Harri Kirik. Linguistically motivated unsupervised segmentation for machine translation. In *Proceedings of LREC'2010*, pages 1741–1745, Valletta, Malta, 2010.

[6] Maxim Khalilov, Lauma Pretkalniņa, Natalja Kuvaldina, and Veronika Pereseina. SMT of latvian, lithuanian and estonian languages: a comparative study. In *Proceedings of Baltic HLT'2010*, pages 117–124, Riga, Latvia, 2010.

[7] Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of NAACL WMT'2012*, pages 10–51, Montréal, Canada, 2012.

[8] Ralf Steinberger, Andreas Eisele, Szymon Klocek, Spyridon Pilos, and Patrick Schlter. Dgt-tm: A freely available translation memory in 22 languages. In *Proceedings of LREC'2012*, pages 454–459, Istanbul, Turkey, 2012.

[9] Rico Sennrich. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of EACL'2012*, pages 539–549, Avignon, France, 2012.