Robust parfda Statistical Machine Translation Results

Ergun Biçici
ergun.bicici@boun.edu.tr
Department of Computer Engineering, Boğaziçi University
orcid.org/0000-0002-2293-2031
bicici.github.com

Abstract

We build parallel feature decay algorithms (parfda) Moses statistical machine translation (SMT) models for language pairs in the translation task. parfda obtains results close to the top constrained phrase-based SMT with an average of 2.252 BLEU points difference on WMT 2017 datasets using significantly less computation for building SMT systems than that would be spent using all available corpora. We obtain BLEU upper bounds based on target coverage to identify which systems used additional data. We use PRO for tuning to decrease fluctuations in the results and post-process translation outputs to decrease translation errors due to the casing of words. F_1 scores on the key phrases of the English to Turkish testsuite that we prepared reveal that parfda achieves $2nd$ best results. Truecasing translations before scoring obtained the best results overall.

1 Introduction

Statistical machine translation is widely prone to errors in text including encoding, tokenization, morphological variations and the mass they take, the size of the training and language model datasets used, and model errors. parfda is an instance selection tool based on feature decay algorithms (Biçici and Yuret, 2015) we use to select training and language model instances to build Moses phrase-based SMT systems to translate the test sets in the news translation task at WMT18 (WMT, 2018). As we work towards tools that can be used for multiple languages at the same time, we aim to obtain robust results for comparison and record the statistics of the data and the resources used. Our contributions are:

- a test suite for machine translation that is out of the domain of news task to take the chance of taking a closer look at the current status of SMT technology used by the task participants when translating 10 sentences taken from literary context in Turkish, which shows that parfda phrase-based SMT can obtain $2nd$ best results on this test set,
- parfda results for language pairs in the translation task and data statistics,
- comparison of processing alternatives for translation outputs to obtain better results,
- upperbounds on the translation performance using lowercased coverage to identify which models used data in addition to the parallel corpus,
- a set of rules that fix tokenization errors in Turkish using Moses' (Koehn et al., 2007) tokenization scripts.

We obtain parfda Moses phrase-based SMT (Koehn et al., 2007) results for the language pairs in both directions in the WMT18 news translation task, which include English-Czech (en-cs), English-Estonian (en-et), English-German (en-de), English-Finnish (en-fi), English-Russian (en-ru), and English-Turkish (en-tr). Building a language independent system that can perform well in translation tasks is a challenging task and SMT systems participating at WMT18 have been largely built dependent on the translation direction.

2 parfda

Parallel feature decay algorithms (parfda) (Bicici, 2016) parallelize feature decay algorithms (FDA), a class of instance selection algorithms that use feature decay, for fast deployment of accurate SMT systems. We use parfda to select parallel training data and language model (LM)

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 345–354

https://doi.org/10.18653/v1/W18-64032

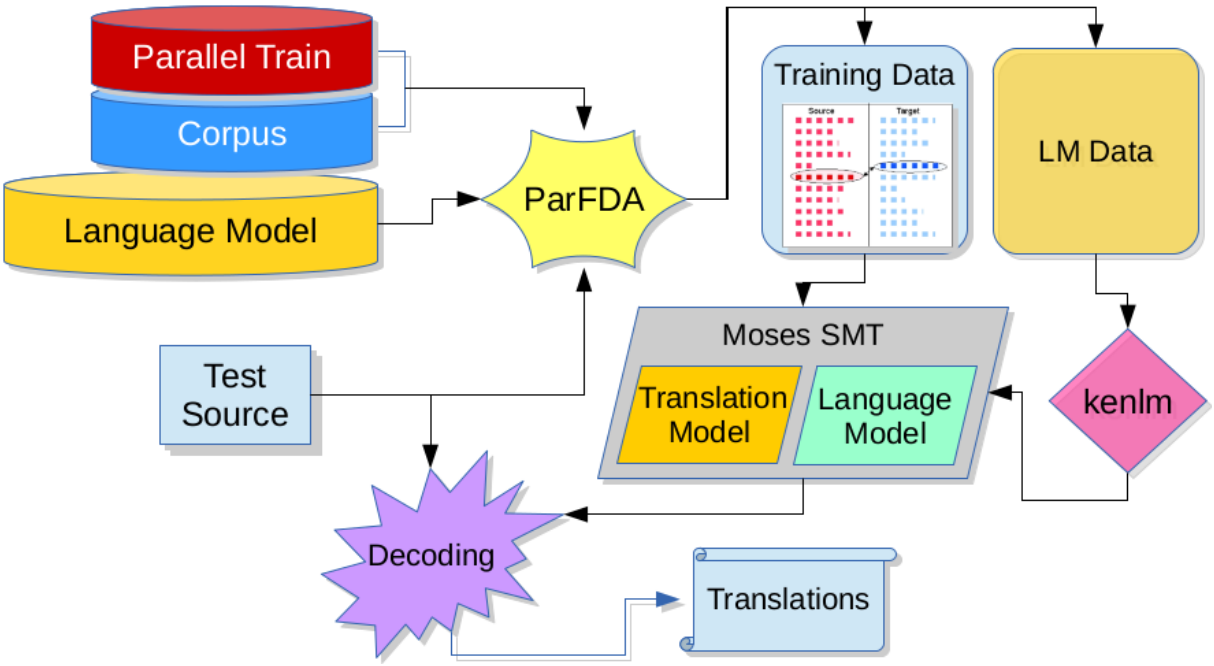

Figure 1: `parfda` Moses SMT workflow.

data for building SMT systems. `parfda` runs separate FDA5 (Biçici and Yuret, 2015) models on randomized subsets of the available data and combines the selections afterwards. Figure 1 depicts `parfda` Moses SMT workflow. The approach also obtained improvements using NMT (Poncelas et al., 2018).

We obtain transductive learning results since we use source sentences of the test set to select data. However, decaying only on the source test set features does not necessarily increase diversity on the target side thus we also decay on the target features that we already select. With the new `parfda` model, we select about 1.7 million instances for training data and about 15 million sentences for each LM data not including the selected training set, which is added later. Table 1 shows size differences with the constrained dataset (C). We use 3-grams to select training data and 2-grams for LM data. TCOV lists the target coverage in terms of the 2-grams of the test set. We also use CzEng17 (Bojar et al., 2016) for en-cs and SETIMES2 (Tiedemann, 2009) for en-tr.

We set the maximum sentence length to 126 and train 6-gram LM using `kenlm` (Heafield et al., 2013). For increasing the robustness of the optimization results, we use PRO (Section 2.1) and we use varying n-best list size. For word alignment, we use mgiza (Gao and Vogel, 2008) where GIZA++ (Och and Ney, 2003) parameters set max-fertility to 10, the number of iterations to 7,3,5,5,7 for IBM models 1,2,3,4, and the HMM model, and learn 50 word classes in three iterations with the mkcls tool during training. The development set contains up to 4000 sentences randomly sampled from previous years' development sets (2011-2017) and remaining come from the development set for WMT18. Table 2 lists the coverage of the test set.

2.1 Robust Optimization Results with PRO

Pairwise ranking optimization (PRO) (Hopkins and May, 2011) is found to obtain scores that monotonically increase, with results that are at least as good as MERT (Och, 2003), and with a standard deviation that is three times lower than MERT. We use PRO for tuning to obtain robust results due to fluctuating scores with MERT. PRO tuning performance graph is compared with MERT performance plot in Figure 2. We used monotonically increasing n-best list size at the start to increase robustness by using multiples of 50 until the 8th iteration, 350 every 10th, and 150 in the remaining. We only need 4 iterations to find parameters whose tuning score reach 1% close to the best tuning parameter set score (Figure 3).

2.2 Testsuite for en-tr and tr-en

We prepared an SMT test suite that is out of the domain of news translation task to take a closer

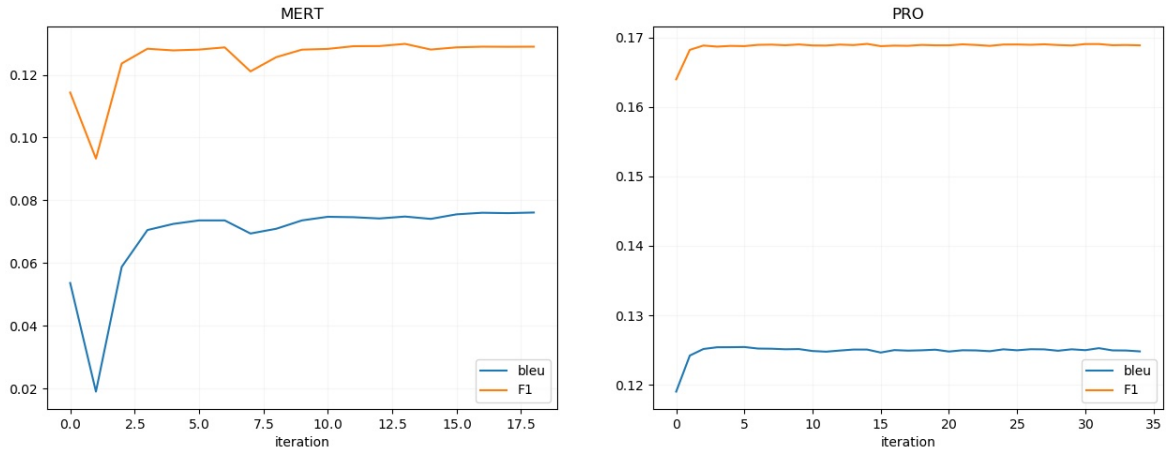

Figure 2: Comparison of MERT and PRO tuning on en-tr using results from 2017 and 2018 respectively.

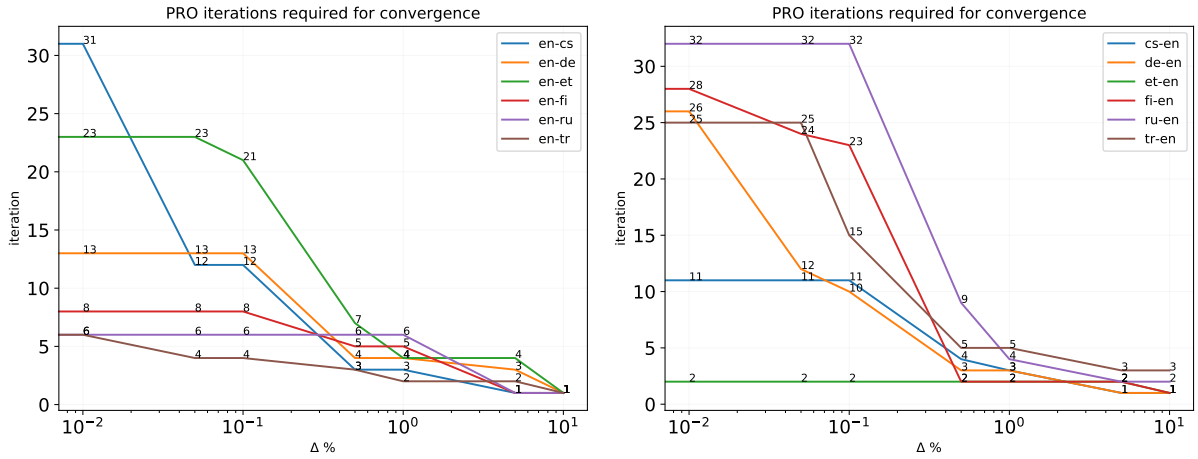

Figure 3: The number of iterations PRO would need to reach $\Delta\%$ close to the best tuning score.

look at the current status of SMT technology used by the task participants to translate 10 sentences taken from literary context in Turkish. The sentences and their translations are provided in Appendix A.

Table 3 details the testsuite results on en-tr and tr-en where the best translations of `parfda` are selected based on their BLEU (Papineni et al., 2002) and F_1 (Biçici, 2011) scores:

en-tr	`lctc` 1 align
en-tr ts	`lctc` 1 align
tr-en	`tc` 2 align
tr-en ts	`tc` 1 align

where `tc` and `lctc` are defined in Section 2.3.

We count tokens of translation as non-translation when they are found in the test source, are not a number or punctuation, and are considered by the SMT model's phrase table or the lexical translation table as a token whose translation differs from the source token. We have access to the lexical tables of `parfda` SMT models and among the tr-en `lctc` entries (Table 4), 2.7% contain a translation the same as the source. According to the testsuite results using translations from task participants, only RWTH and `parfda` contained non-translations and RWTH had only a token non-translated. The scores for up to n-grams in Table 12 show that alibaba.5744 achieves the best results in en-tr and online-B achieves the best results in tr-en in all scores. When we look at some of the OOV tokens in en-tr, we observe that lowercasing and then truecasing might help.

We identified 5 key phrases for both en-tr and tr-en that we would like to see translated correctly (Table 5). Some are trimmed to make them closer to their root form so that suffixes can be added without decreasing identification rates. Appendix A presents F_1 scores based on the identification of them in the translations. We see that even though `parfda` achieves the lowest scores in BLEU, on the key phrases, it provides the 2rd

$S \rightarrow T$	Data	Training Data				LM Data	
		#word S (M)	#word T (M)	#sent (K)	TCOV	#word (M)	TCOV
en-cs	C	618.2	548.9	43274	0.749	1357.0	0.857
en-cs	parfda	83.6	73.9	1777	0.663	416.2	0.809
cs-en	C	548.9	618.2	43274	0.855	11100.4	0.948
cs-en	parfda	78.1	88.1	1773	0.8	502.2	0.896
en-de	C	580.2	548.0	24914	0.791	3078.9	0.892
en-de	parfda	96.6	91.4	1776	0.74	467.3	0.839
de-en	C	548.0	580.2	24914	0.859	11100.4	0.941
de-en	parfda	90.4	94.9	1776	0.82	509.1	0.893
en-et	C	20.4	27.0	1073	0.422	1197.1	0.772
en-et	parfda	27.0	20.4	1072	0.422	375.8	0.654
et-en	C	27.0	20.4	1073	0.673	11100.4	0.943
et-en	parfda	20.4	27.0	1072	0.673	416.4	0.883
en-fi	C	52.3	72.3	2846	0.45	1536.4	0.743
en-fi	parfda	53.4	38.6	1598	0.438	468.0	0.676
fi-en	C	72.3	52.3	2846	0.725	11100.4	0.943
fi-en	parfda	37.2	50.8	1550	0.715	473.2	0.888
en-ru	C	172.6	202.3	8766	0.739	9643.5	0.916
en-ru	parfda	69.3	53.0	1712	0.684	561.3	0.83
ru-en	C	202.3	172.6	8766	0.844	11100.4	0.95
ru-en	parfda	61.7	71.6	1764	0.816	489.6	0.903
en-tr	C	4.6	5.1	208	0.352	4026.5	0.824
en-tr	parfda	5.1	4.6	207	0.352	474.5	0.72
tr-en	C	5.1	4.6	208	0.569	11100.4	0.936
tr-en	parfda	4.6	5.1	207	0.569	442.1	0.877

Table 1: Statistics for the training and LM corpora in the constrained (C) setting compared with the `parfda` selected data. #words is in millions (M) and #sents in thousands (K). TCOV is target 2-gram coverage.

	SCOV					TCOV				
	1	2	3	4	5	1	2	3	4	5
en-cs	0.9635	0.8612	0.6212	0.3305	0.1334	0.9718	0.7587	0.4042	0.165	0.0553
en-de	0.9657	0.8646	0.627	0.3261	0.1253	0.9362	0.7939	0.5254	0.2534	0.0904
en-et	0.8912	0.6821	0.3717	0.1345	0.0376	0.8095	0.4308	0.1664	0.0528	0.0136
en-fi	0.9135	0.7339	0.4477	0.1889	0.0619	0.834	0.4595	0.1895	0.0612	0.0165
en-ru	0.9682	0.8683	0.6444	0.3581	0.1567	0.9703	0.78	0.4572	0.2076	0.0812
en-tr	0.8286	0.5817	0.2807	0.0905	0.0226	0.7944	0.3613	0.12	0.0282	0.006

Table 2: Test set SCOV and TCOV for n-grams.

best in en-tr among 9 models and 4th best among 6 in tr-en. Key phrase identification is important since when scores are averaged, important phrases that are missing only decrease the score by $\frac{1}{|p|N_{|p|}}$ for BLEU calculation for a phrase of length $|p|$ over $N_{|p|}$ phrases with length $|p|$.

2.3 Comparing Text Processing Settings for SMT

Experiment management system (EMS) (Koehn, 2010) of Moses prepares translations as follows:

truecase input
→ translate input
→ clean output (XML tags)
→ detruecase output

Truecasing updates the casing of words according to the most common form observed in the whole training corpus. EMS does not truecase the translations of an SMT model when training data are already truecased. However, each casing of words are a different entry in the phrase table and the casing we are interested in might be missing in the translations. Therefore, truecasing (`tc`) before detruecasing makes sense.

The casing of the text affects the number of tokens in the data sets. A casing of a token might appear in the phrase table but not its lowercased (`lc`) version. In EMS, truecasing is applied on the input. We experiment with truecasing lowercased text (`lctc`) to decrease the number of out-of-vocabulary words in the translations and to reduce the number of unique n-grams, dataset sizes, and the binary LM size by about 2%.

We process tokenized Turkish text using a set of rules since Moses' (Koehn et al., 2007) tokenization scripts can encounter tokenization errors in Turkish. A simpler approach was also tried for fixing tokenization of Turkish by removing space for unbalanced single quotes (Ding et al., 2016). Additionally, we retain the casing of the

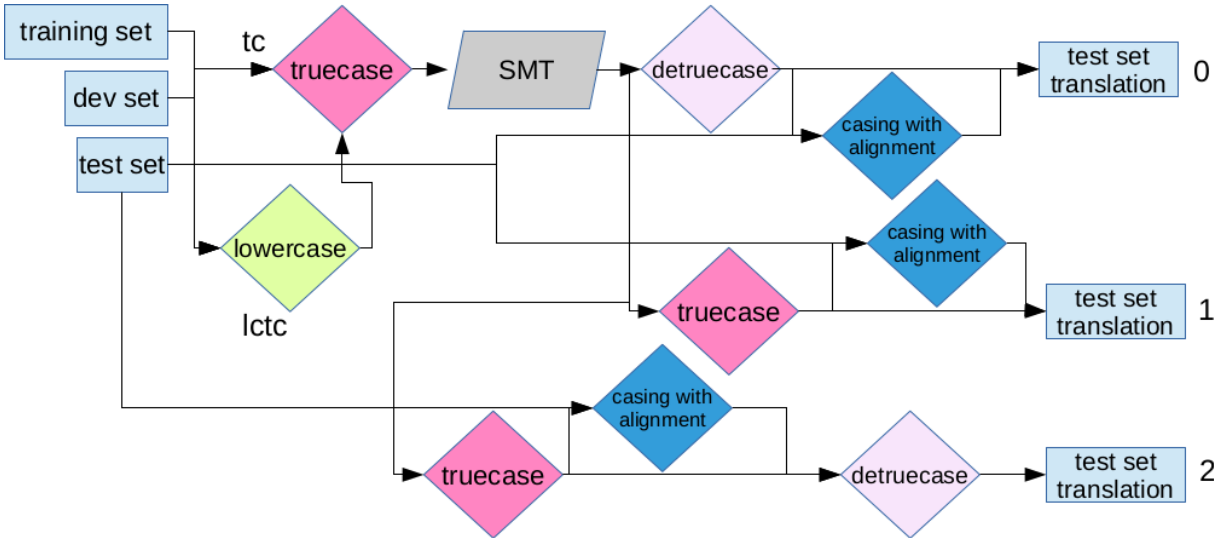

Figure 4: SMT text processing comparison of truecasing after lowercasing (lctc) and truecasing (tc).

testset	setting	% sent. OK	non-trans.	non-trans. %	not in PT/lex	oov	% oov	BLEU	F_1
en-tr	`tc` 1 align	60.9	1941	18.81	22	1324	12.8	0.0746	0.1302
	`lctc` 1 align	61.1	1933	18.74	0	1290	12.5	0.0883	0.1406
en-tr ts	`lctc` 1 align	60.0	5	3.4	0	4	2.7	0.051	0.105
tr-en	`tc` 2 align	38.9	3549	20.7	23	2643	15.4	0.1192	0.1708
	`lctc` 2 align	44.8	3028	17.6	12	2132	12.4	0.1055	0.1567
tr-en ts	`tc` 1 align	20.0	32	17.7	0	24	13.3	0.0	0.1011

Table 3: Best performing en-tr and tr-en translation results detailed with their types of errors.

	setting	N (lexical)	% ($w_S == w_T$)
en-tr	`tc`	928K	2.91
	`lctc`	890K	2.68
tr-en	`tc`	891K	3.02
	`lctc`	860K	2.74

Table 4: Lexical translation table comparison.

	phrase	count
en-tr	Türk Dil Kurumu	6
	Türkçe Sözlü	4
	Yazım Kılavuzu Çalışma Grubu	3
	Yazım Kılavuzu	7
	yazım kural	4
tr-en	Turkish Language Institution	6
	Turkish Language	6
	Turkish Dictionary	4
	Working Group	3
	Writing Manual	4

Table 5: Key phrases we look for in the translations.

test source sentences using the word alignment information (Ding et al., 2016). Using alignment information is more complicated since not all alignments are 1-to-1. We also experiment with finding the casing of the input words in the development and test sets according to the form found in the translation tables to replace them before decoding. Figure 4 compares `tc` and `lctc` approaches to text processing for SMT. Both can use the alignment information for casing words.

Table 6 compares the results using translations that contain the alignment information and the unknown words where `tc` 0 is the baseline. The additional Moses decoder parameter is `--print-alignment-info`. We obtain the highest en-tr score using the alignments for casing but scores decrease for en-de and de-en. For which translation directions it helps can be seen in the `lctc` 0 row. The difference between the base and the lowercased results are the gain we can achieve if we fix casing accordingly. Using `tc` translation as a start, the gain on average is about 1.1 BLEU points (0.011 BLEU). The best setting overall is `tc` 2. The largest room for improvement with `lctc lc` BLEU results are for cs-en and tr-en.

BLEU		cs-en	de-en	et-en	fi-en	ru-en	tr-en	en-cs	en-de	en-et	en-fi	en-ru	en-tr	sum
tc 0	base	**0.2296**	**0.3332**	**0.1766**	**0.1536**	0.247	**0.1286**	**0.1567**	**0.2669**	**0.1182**	**0.1016**	0.1934	0.0823	2.1877
	align	0.2134	0.1851	0.1684	0.1427	0.233	0.1269	0.1455	0.1637	0.1138	0.095	0.1821	0.083	1.8526
tc 1		0.2239	0.318	0.1718	0.1493	0.2464	0.1269	0.1506	0.2543	0.1117	0.0961	0.1918	0.0824	2.1232
	align	0.2127	0.1845	0.1682	0.1416	0.2329	0.1261	0.1455	0.1625	0.1137	0.0948	0.182	0.083	1.8475
tc 2		0.2295	0.3331	0.1765	0.1535	**0.2526**	0.1284	0.1566	0.2669	0.1182	0.1013	**0.1989**	0.0823	**2.1978**
	align	0.2136	0.1945	0.1681	0.1458	0.2347	0.1273	0.1469	0.1645	0.1151	0.0965	0.1827	0.0816	1.8713
	lc	0.2394	0.3447	0.184	0.1614	**0.2632**	0.1373	0.1622	0.2727	**0.1216**	0.1053	**0.2048**	0.089	2.2856
lctc 0		0.1869	0.2736	0.1523	0.129	0.1901	0.1066	0.1339	0.1251	0.0975	0.0875	0.1671	0.0616	1.7112
	align	0.2113	0.2443	0.1661	0.1266	0.211	0.1043	0.1448	0.1603	0.1114	0.0941	0.1644	0.0882	1.8268
lctc 1		0.1817	0.2602	0.1484	0.1252	0.1901	0.1064	0.1294	0.1149	0.0934	0.0828	0.1656	0.0639	1.662
	align	0.2105	0.2437	0.1658	0.1247	0.211	0.1032	0.1447	0.1585	0.1113	0.094	0.1642	**0.0887**	1.8203
lctc 2		0.1896	0.275	0.1552	0.1303	0.1974	0.1083	0.1369	0.127	0.1005	0.0887	0.1712	0.0632	1.7433
	align	0.2121	0.2583	0.1663	0.1304	0.2125	0.1055	0.1468	0.1611	0.1125	0.0954	0.1648	0.0871	1.8528
	lc	**0.2445**	**0.3452**	**0.1871**	**0.1634**	0.2506	**0.1402**	**0.1651**	**0.2781**	0.1212	**0.1056**	0.1803	**0.0978**	2.2791

Table 6: parfda tokenized and cased results with different text processing settings. Baseline is tc 0 (in *italic*). **bold** lists the best for a translation direction.

BLEU	cs-en	de-en	et-en	fi-en	ru-en	tr-en	en-cs	en-de	en-et	en-fi	en-ru	en-tr
parfda 2018	0.2322	0.3343	0.1741	0.1547	0.2485	0.1267	0.1529	0.2674	0.1203	0.0968	0.1970	0.0821
parfda 2018 F_1	0.2551	0.344	0.2123	0.1936	0.2685	0.1768	0.1921	0.2891	0.1613	0.1494	0.2214	0.1314
TopC NMT 2018 lc	0.348	0.499	0.315	0.258	0.358	0.291	0.266	0.489	0.258	0.192	0.348	0.207
TopC NMT 2018	0.339	0.484	0.307	0.249	0.349	0.28	0.26	0.483	0.252	0.182	0.348	0.20
- parfda	0.1068	0.1497	0.1329	0.0943	0.1005	0.1533	0.1071	0.2156	0.1317	0.0852	0.1510	0.1179
avg diff lc	0.1288											

Table 7: parfda results compared with the top results in WMT18 and their difference.[1]

parfda results at WMT18 are in Table 7 using BLEU over tokenized text. We compare with the top constrained submissions at WMT18 in Table 7 and at WMT17 in Table 8. [2] Performance compared with the top constrained (TopC) phrase-based SMT improved to 2.252 in 2017 from 3 BLEU points difference on average compared with WMT16 results, which is likely due to the new parfda model and phrase-based SMT being less common in 2017. parfda Moses SMT system can obtain 0.6 BLEU points close to the top result in Finnish to English translation in 2017. All top models use NMT in 2018 and most use backtranslations, which means that their TCOV is upper bounded by LM TCOV.

3 Translation Upper Bounds with TCOV

We obtain upper bounds on the translation performance based on the target coverage (TCOV) of n-grams of the test set found in the selected parfda training data (Bicici, 2016) but using lowercased text this time. For a given sentence T', the number of OOV tokens are identified:

$$OOV_r = \texttt{round}((1 - \text{TCOV}) * |T'|) \quad (1)$$

where $|T'|$ is the number of tokens in the sentence. We obtain each bound using 500 such instances and repeat for 10 times. TCOV BLEU bound is optimistic since it does not consider reorderings in the translation or differences in sentence length. Each plot in Table 9 locates TCOV BLEU bound obtained from each n-gram and from n-gram TCOVS combined up to and including n and ■ locates the parfda result and ★ locates the top constrained result. In en-de and en-tr, the top model achieves a higher score than the TCOV BLEU bound, which indicates that data additional to the constrained training data was used. In both, backtranslations were used.

4 Conclusion

We use parfda for selecting instances for building SMT systems using less computation overall and results at WMT18 provides new data about using the current phrase-based SMT technology towards rapid SMT system development. Our data processing experiments show that lowercasing and then truecasing data can improve SMT models and translation results provided that we can find the casing correctly and truecasing translations before scoring can improve the results. Our

[2]Due to different tokenization rules used by mteval-v14.pl in matrix.statmt.org, parfda BLEU scores are higher than the scores in Table 6.

[2]We use the results from matrix.statmt.org.

BLEU	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	en-cs	en-de	en-fi	en-lv	en-ru	en-tr
parfda 2017	0.2276	0.2613	0.1987	0.1549	0.2812	0.1172	0.1381	0.1851	0.1282	0.1303	0.218	0.097
TopC NMT 2017	0.309	0.351		0.19	0.308	0.179	0.228	0.283	0.207	0.183	0.298	0.165
- parfda	0.0814	0.0897		0.0351	0.0268	0.0618	0.0899	0.0979	0.0788	0.0527	0.0800	0.0680
avg diff	0.0693											
TopC phrase 2017	0.265		0.205	0.168	0.315	0.126	0.191	0.216	0.145	0.142	0.253	0.098
- parfda	0.0374		0.0063	0.0131	0.0338	0.0088	0.0529	0.0309	0.0168	0.0127	0.035	0.001
avg diff	0.02252											

Table 8: parfda results compared with the top results in WMT17 and their difference.

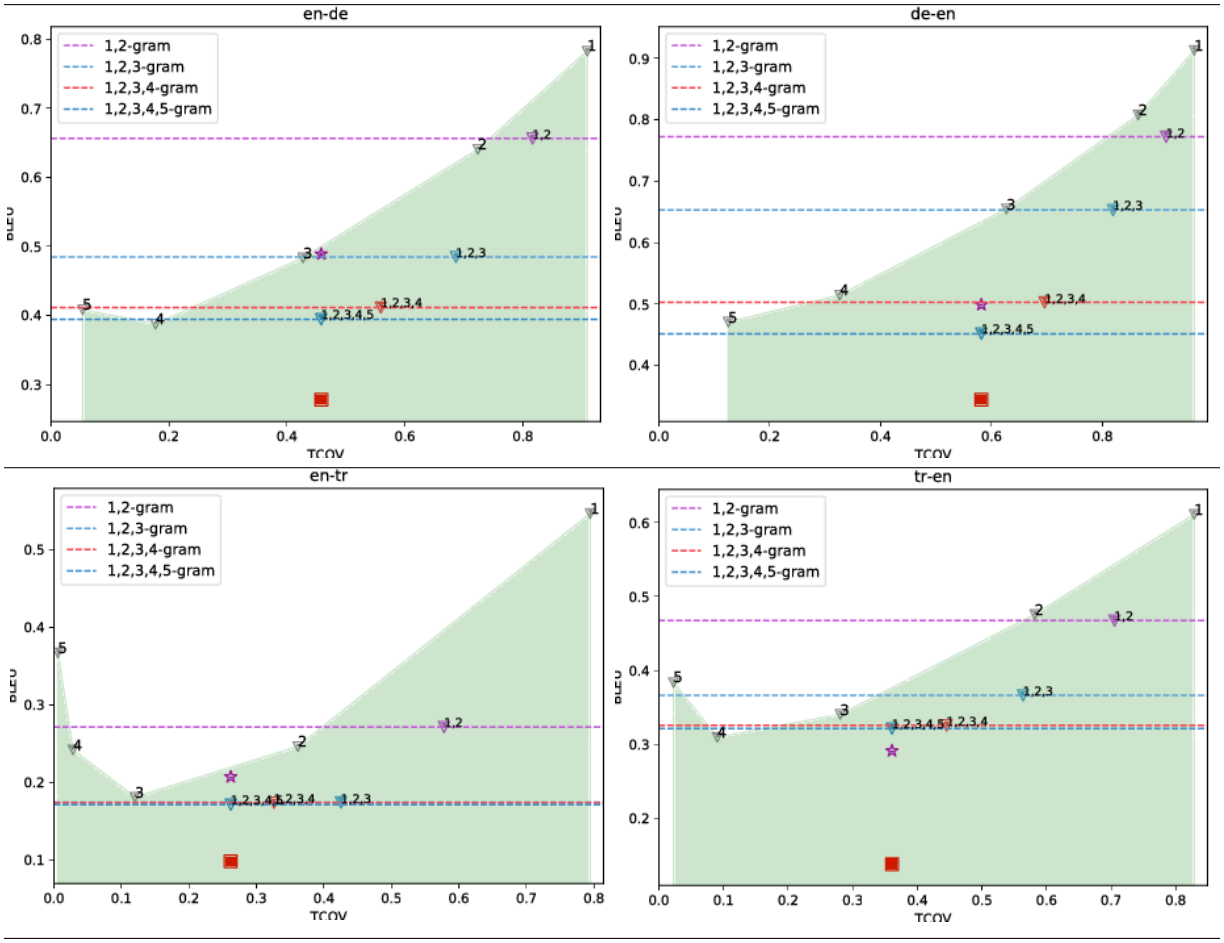

Table 9: parfda results (■) and OOV_r TCOV BLEU upper bounds for de and tr.

method of tuning with PRO provides robust results and the BLEU bounds we obtain show which systems used additional training data. We are often interested to conserve the semantic content in the translations and parfda Moses phrase-based SMT achieves $2nd$ best results on the tr-en test-suite in our evaluations with key phrases.

Acknowledgments

The research reported here received financial support in part from the Scientific and Technological Research Council of Turkey (TÜBİTAK) without contribution to the content nor responsibility thereof. We also thank the reviewers' comments.

References

2018. *Proc. of the Third Conference on Machine Translation*. Association for Computational Linguistics, Brussels, Belgium.

Ergun Biçici. 2011. *The Regression Model of Machine Translation*. Ph.D. thesis, Koç University. Supervisor: Deniz Yuret.

Ergun Biçici and Deniz Yuret. 2015. Optimizing instance selection for statistical machine translation with feature decay algorithms. *IEEE/ACM Transactions On Audio, Speech, and Language Processing (TASLP)*, 23:339–350.

Ergun Bicici. 2016. ParFDA for instance selection for statistical machine translation. In *Proc. of the*

First Conference on Statistical Machine Translation (WMT16), pages 252–258, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Ondřej Dušek, Tom Kocmi, Jindřich Libovický, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš. 2016. *CzEng 1.6: Enlarged Czech-English Parallel Corpus with Processing Tools Dockered*. Springer International Publishing, Cham.

Shuoyang Ding, Kevin Duh, Huda Khayrallah, Philipp Koehn, and Matt Post. 2016. The jhu machine translation systems for wmt 2016. In *Proceedings of the First Conference on Machine Translation*, pages 272–280, Berlin, Germany. Association for Computational Linguistics.

Qin Gao and Stephan Vogel. 2008. *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, chapter Parallel Implementations of Word Alignment Tool. Association for Computational Linguistics.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proc. of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria.

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1352–1362. Association for Computational Linguistics.

Philipp Koehn. 2010. An experimental management system. *The Prague Bulletin of Mathematical Linguistics*, 94:87–96.

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. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proc. of the Demo and Poster Sessions*, pages 177–180. Association for Computational Linguistics.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. *Association for Computational Linguistics*, 1:160–167.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA, USA.

Alberto Poncelas, Gideon Maillette de Buy Wenniger, and Andy Way. 2018. Feature decay algorithms for neural machine translation. In *Proc. of the 21th Conference of the European Association for Machine Translation (EAMT)*, Spain.

Jorg Tiedemann. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

A en-tr and tr-en Testsuite Sentences

	English
1	The Turkish Language Institution's Turkish Dictionary and Writing Manual Working Group has reached the twenty-seventh edition of the Written Manual.
2	To say upfront; to make the writing rules lasting and rooted to reach unity in writing, no changes have been made in the established rules in this edition as well.
3	Turkish Language Institution's principles of taking the manual published in 1996 as the basis for traditional writing structures and the principle of taking the middle path to end discussions in writing are also followed in this edition.
4	This edition of the Writing Manual is prepared in parallel with the last edition of the Turkish Dictionary and new issues that emerge with the use of language are attached to rules and issues that were not addressed in previous editions are converted into rules.
5	Turkish Dictionary and Writing Manual Working Group reviews hundreds of examples about every rule and considers usage statistics with the principle of viewing writing as habit and custom.
6	After using Latin alphabet for more than eighty years we can say that Turkish writing has been traditionalized against some Latin sourced problems.
7	Without a doubt, Turkish Language Institution's work for 80 years has played an important role in this.
8	The task of preparing, writing, and distributing writing manual is given to Turkish Language Institution according to the ç subitem of the 10th item in 664 numbered decree law based on the 134th item in the constitution.
9	Since its establishment, Turkish Language Institution has been trying to fulfill its duty on determining writing rules and publication of writing manuals.
10	Turkish Language Institution Turkish Dictionary and Writing Manual Working Group has diligently worked on each writing rule and writing to end discussions in writing and to spread writing rules and styles that everybody will accept and use.

	Turkish
1	Türk Dil Kurumu Güncel Türkçe Sözlük ve Yazım Kılavuzu Çalışma Grubunun yaptığı çalışmalarla Yazım Kılavuzu yirmi yedinci baskısına ulaşmış bulunuyor.
2	Hemen belirtelim; yazım kurallarını kalıcı hâle getirmek, kökleştirmek, böylece yazımda birliği sağlamak amacıyla bu baskıda da yerleşmiş kurallarla ilgili olarak herhangi bir değişikliğe gidilmemiştir.
3	Türk Dil Kurumunun 1996 yılında yayımladığı kılavuzda yazımı gelenekleşmiş biçimleri esas alma, yazımdaki tartışmalara son verme amacıyla yazımda orta yolun tutturulması ilkesi bu baskıda da gözetilmiştir.
4	Yazım Kılavuzu'nun bu baskısı Türkçe Sözlük'ün son baskısıyla eş güdüm içerisinde hazırlanmış, dilde yaşanan gelişmeler sonucunda ortaya çıkan yazımla ilgili yeni sorunlar bir kurala bağlanmış, önceki baskılarda değinilmeyen konular yazım kuralı hâline getirilmiştir.
5	Güncel Türkçe Sözlük ve Yazım Kılavuzu Çalışma Grubu, her kuralla ilgili yüzlerce örneği gözden geçirirken yazımın bir alışkanlık ve gelenek olduğu ilkesiyle kullanım sıklıklarını göz önünde bulundurmuştur.
6	Seksen yılı aşkın bir süredir kullanmakta olduğumuz Latin kaynaklı Türk yazısıyla kimi sorunlara karşın artık yazımın gelenekleştiğini söyleyebiliriz.
7	Bunda hiç kuşkusuz, Türk Dil Kurumunun seksen yıla ulaşan çalışmaları önemli bir rol oynamaktadır.
8	Ülkemizde yazım kılavuzu hazırlamak, yazmak ve yaymlamak görevi, Anayasa'nın 134. maddesine dayalı olarak çıkarılan 664 sayılı Kanun Hükmünde Kararname'nin 10. maddesinin ç fıkrasıyla Türk Dil Kurumuna verilmiştir.
9	Türk Dil Kurumu, kuruluşundan bu yana yazım kurallarının belirlenmesinde ve yazım kılavuzlarının yayımlanmasında kendisine düşen görevi yerine getirmeye çalışmaktadır.
10	Türk Dil Kurumu Güncel Türkçe Sözlük ve Yazım Kılavuzu Çalışma Grubu, yazımda yaşanan tartışmaları sona erdirmek ve herkesin benimseyeceği, kullanacağı yazım kurallarını ve yazım biçimlerini yaygınlaştırmak ilkesiyle her kuralın, her yazılışın üzerinde titizlikle durmuştur.

Table 10: Testsuite English and Turkish for en-tr / tr-en.

en-tr				
alibaba.5732	Türk Dil Kurumu	1.0	6	6
	Türkçe Sözlü	0.86	3	4
	Yazım Kılavuzu Çalışma Grubu	1.0	3	3
	Yazım Kılavuzu	0.6	3	7
	yazım kural	1.0	4	4
	F_1	0.88	19	24
alibaba.5744	Türk Dil Kurumu	1.0	6	6
	Türkçe Sözlü	0.86	3	4
	Yazım Kılavuzu Çalışma Grubu	1.0	3	3
	Yazım Kılavuzu	0.6	3	7
	yazım kural	1.0	4	4
	F_1	0.88	19	24
parfda	Türk Dil Kurumu	1.0	6	6
	Türkçe Sözlü	1.0	4	4
	F_1	0.59	10	24
uedin.5644	Türk Dil Kurumu	1.0	6	6
	yazım kural	0.86	3	4
	F_1	0.54	9	24
online-B	Türk Dil Kurumu	0.91	5	6
	Türkçe Sözlü	0.86	3	4
	yazım kural	0.4	1	4
	F_1	0.54	9	24
NICT.5695	Türk Dil Kurumu	1.0	6	6
	yazım kural	0.67	2	4
	F_1	0.5	8	24
RWTH.5632	Türk Dil Kurumu	1.0	6	6
	F_1	0.4	6	24
online-A	Türk Dil Kurumu	0.8	4	6
	F_1	0.29	4	24
online-G	Türkçe Sözlü	0.86	3	4
	F_1	0.22	3	24
tr-en				
online-B	Turkish Language Institution	0.5	2	6
	Turkish Language	0.86	8	6
	Turkish Dictionary	0.67	2	4
	Working Group	1.0	3	3
	Writing Manual	0.4	1	4
	F_1	0.82	16	23
NICT.5708	Turkish Language Institution	0.67	3	6
	Turkish Language	1.0	6	6
	Working Group	1.0	3	3
	Writing Manual	0.67	2	4
	F_1	0.76	14	23
uedin.5709	Turkish Language Institution	0.67	3	6
	Turkish Language	1.0	6	6
	Working Group	1.0	3	3
	F_1	0.69	12	23
parfda	Turkish Language Institution	0.29	1	6
	Turkish Language	0.8	4	6
	Working Group	1.0	3	3
	F_1	0.52	8	23
online-G	Turkish Dictionary	0.86	3	4
	F_1	0.23	3	23
online-A	Turkish Language	0.29	1	6
	F_1	0.08	1	23

Table 11: Testsuite F_1 scores with key phrases.

		BLEU lc				BLEU				F_1 lc				F_1			
	model	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
en-tr	alibaba.5732	0.4591	0.4482	0.3177	0.3112	0.2287	0.2255	0.1701	0.1684	0.3597	0.3537	0.2827	0.2789	0.2323	0.2294	0.1981	0.1958
	alibaba.5744	0.4778	0.4669	0.3297	0.3231	0.2394	0.2362	0.1785	0.1768	0.3717	0.3658	0.2936	0.2898	0.2416	0.2388	0.2055	0.2033
	online-A	0.3977	0.3612	0.2486	0.2027	0.1469	0.109	0.0714	0.0571	0.292	0.2532	0.2146	0.1829	0.1649	0.141	0.135	0.1156
	online-B	0.4412	0.423	0.3005	0.2831	0.2045	0.185	0.1296	0.1118	0.3419	0.3196	0.2631	0.2419	0.2083	0.1896	0.1711	0.1557
	online-G	0.4105	0.3778	0.2493	0.2172	0.1544	0.1308	0.0934	0.0824	0.3099	0.2788	0.2313	0.2064	0.1817	0.1629	0.15	0.1346
	parfda	0.3367	0.3258	0.1818	0.1718	0.1055	0.0934	0.0558	0.051	0.2417	0.2292	0.1775	0.1657	0.1375	0.1286	0.1123	0.105
	uedin	0.448	0.4334	0.2971	0.2808	0.191	0.1708	0.1231	0.1007	0.3428	0.3255	0.2572	0.2415	0.2033	0.189	0.1678	0.1551
	NICT	0.4376	0.4231	0.2851	0.2686	0.1947	0.1786	0.132	0.117	0.3404	0.3233	0.262	0.2472	0.2101	0.1972	0.1741	0.1629
	RWTH	0.3752	0.3643	0.2065	0.2001	0.119	0.1125	0.0676	0.0648	0.2507	0.2417	0.1803	0.1725	0.1395	0.1336	0.1142	0.1094
tr-en	parfda	0.413	0.4006	0.2122	0.1963	0.0963	0.0861	0.0	0.0	0.2586	0.2404	0.1782	0.1656	0.1343	0.1249	0.1086	0.1011
	online-A	0.5452	0.4877	0.3759	0.3056	0.2602	0.1988	0.1778	0.1369	0.4391	0.3573	0.3367	0.2661	0.2698	0.213	0.2253	0.1782
	online-B	0.5765	0.5656	0.4297	0.418	0.3244	0.3152	0.2459	0.239	0.4825	0.4732	0.3863	0.3786	0.3197	0.3135	0.2725	0.2679
	online-G	0.5414	0.4972	0.3843	0.3362	0.2803	0.2364	0.2038	0.168	0.4497	0.3929	0.353	0.3024	0.288	0.2441	0.243	0.2045
	uedin	0.5333	0.5227	0.3606	0.3385	0.2507	0.2222	0.1738	0.1426	0.424	0.4046	0.3253	0.3038	0.2603	0.2386	0.216	0.1961
	NICT	0.5339	0.526	0.3544	0.3478	0.2294	0.2244	0.1422	0.1399	0.4239	0.4199	0.3185	0.316	0.2502	0.2488	0.2064	0.2055

Table 12: Testsuite BLEU and F_1 results.