ParFDA for Fast Deployment of Accurate Statistical Machine Translation Systems, Benchmarks, and Statistics

Ergun Biçici
ADAPT Research Center
School of Computing
Dublin City University, Ireland
ergun.bicici@computing.dcu.ie

Qun Liu
ADAPT Research Center
School of Computing
Dublin City University, Ireland
qliu@computing.dcu.ie

Andy Way
ADAPT Research Center
School of Computing
Dublin City University, Ireland
away@computing.dcu.ie

Abstract

We build parallel FDA5 (ParFDA) Moses statistical machine translation (SMT) systems for all language pairs in the workshop on statistical machine translation (Bojar et al., 2015) (WMT15) translation task and obtain results close to the top with an average of 3.176 BLEU points difference using significantly less resources for building SMT systems. ParFDA is a parallel implementation feature decay algorithms (FDA) developed for fast deployment of accurate SMT systems (Biçici, 2013; Biçici et al., 2014; Biçici and Yuret, 2015). ParFDA Moses SMT system we built is able to obtain the top TER performance in French to English translation. We make the data for building ParFDA Moses SMT systems for WMT15 available: `https://github.com/bicici/ParFDAWMT15`.

1 Parallel FDA5 (ParFDA)

Statistical machine translation performance is influenced by the data: if you already have the translations for the source being translated in your training set or even portions of it, then the translation task becomes easier. If some token does not appear in your language model (LM), then it becomes harder for the SMT engine to find its correct position in the translation. The importance of ParFDA increases with the proliferation of training material available for building SMT systems. Table 1 presents the statistics of the available training and LM corpora for the constrained (C) systems in WMT15 (Bojar et al., 2015) as well as the statistics of the ParFDA selected training and LM data.

ParFDA (Biçici, 2013; Biçici et al., 2014) runs separate FDA5 (Biçici and Yuret, 2015) models on randomized subsets of the training data and combines the selections afterwards. FDA5 is available at `http://github.com/bicici/FDA`. We run ParFDA SMT experiments using Moses (Koehn et al., 2007) in all language pairs in WMT15 (Bojar et al., 2015) and obtain SMT performance close to the top constrained Moses systems. ParFDA allows rapid prototyping of SMT systems for a given target domain or task.

We use ParFDA for selecting parallel training data and LM data for building SMT systems. We select the LM training data with ParFDA based on the following observation (Biçici, 2013):

> No word not appearing in the training set can appear in the translation.

Thus we are only interested in correctly ordering the words appearing in the training corpus and collecting the sentences that contain them for building the LM. At the same time, a compact and more relevant LM corpus is also useful for modeling longer range dependencies with higher order n-gram models. We use 3-grams for selecting training data and 2-grams for LM corpus selection.

2 Results

We run ParFDA SMT experiments for all language pairs in both directions in the WMT15 translation task (Bojar et al., 2015), which include English-Czech (en-cs), English-German (en-de), English-Finnish (en-fi), English-French (en-fr), and English-Russian (en-ru). We truecase all of the corpora, use 150-best lists during tuning, set the LM order to a value in $[8, 10]$ for all language pairs, and train the LM using SRILM (Stolcke, 2002) with `-unk` option. We set the maximum sentence length filter to 126 and for GIZA++ (Och and Ney, 2003), max-fertility is set to 10, with the number of iterations set to 7,3,5,5,7 for IBM models 1,2,3,4, and the HMM model and 70 word classes are learned over 3 iterations with the mkcls tool during training. The development set con-

$S \rightarrow T$	Data	Training Data					LM Data	
		#word S (M)	#word T (M)	#sent (K)	SCOV	TCOV	#word (M)	TCOV
en-cs	C	253.8	224.1	16083	0.832	0.716	841.2	0.862
en-cs	ParFDA	49.0	42.1	1206	0.828	0.648	447.2	0.834
cs-en	C	224.1	253.8	16083	0.716	0.832	5178.5	0.96
cs-en	ParFDA	42.0	46.3	1206	0.71	0.786	1034.2	0.934
en-de	C	116.3	109.8	4525	0.814	0.72	2380.6	0.899
en-de	ParFDA	37.6	33.1	904	0.814	0.681	513.1	0.854
de-en	C	109.8	116.3	4525	0.72	0.814	5111.2	0.951
de-en	ParFDA	33.3	33.1	904	0.72	0.775	969.1	0.923
en-fi	C	52.8	37.9	2072	0.684	0.419	52.7	0.559
en-fi	ParFDA	37.2	26.4	1035	0.684	0.41	79.1	0.559
fi-en	C	37.9	52.8	2072	0.419	0.684	5054.2	0.951
fi-en	ParFDA	25.1	34.5	1035	0.419	0.669	985.9	0.921
en-fr	C	1096.9	1288.5	40353	0.887	0.905	2989.4	0.956
en-fr	ParFDA	58.8	63.2	1261	0.882	0.857	797.1	0.937
fr-en	C	1288.5	1096.9	40353	0.905	0.887	5961.6	0.962
fr-en	ParFDA	72.4	60.1	1261	0.901	0.836	865.3	0.933
en-ru	C	51.3	48.0	2563	0.814	0.683	848.7	0.881
en-ru	ParFDA	37.2	33.1	1281	0.814	0.672	434.8	0.857
ru-en	C	48.0	51.3	2563	0.683	0.814	5047.8	0.958
ru-en	ParFDA	33.8	36.0	1281	0.683	0.803	996.3	0.933

Table 1: Data statistics for the available training and LM corpora in the constrained (C) setting compared with the ParFDA selected training and LM data. #words is in millions (M) and #sents in thousands (K).

tains up to 5000 sentences randomly sampled from previous years' development sets (2010-2014) and remaining come from the development set for WMT15.

2.1 Statistics

The statistics for the ParFDA selected training data and the available training data for the constrained translation task are given in Table 1. For en and fr, we have access to the LDC Gigaword corpora (Parker et al., 2011; Graff et al., 2011), from which we extract only the story type news. The size of the LM corpora includes both the LDC and the monolingual LM corpora provided by WMT15. Table 1 shows the significant size differences between the constrained dataset (C) and the ParFDA selected data and also present the source and target coverage (SCOV and TCOV) in terms of the 2-grams of the test set. The quality of the training corpus can be measured by TCOV, which is found to correlate well with the BLEU performance achievable (Biçici, 2011).

The space and time required for building the ParFDA Moses SMT systems are quantified in Table 2 where size is in MB and time in minutes. PT stands for the phrase table. We used Moses version 3.0, from `www.statmt.org/moses`. Building a ParFDA Moses SMT system can take about half a day.

2.2 Translation Results

ParFDA Moses SMT results for each translation direction together with the LM order used and the top constrained submissions to WMT15 are given in Table 3 [1], where BLEUc is cased BLEU. ParFDA significantly reduces the time required for training, development, and deployment of an SMT system for a given translation task. The average difference to the top constrained submission in WMT15 is 3.176 BLEU points whereas the difference was 3.49 BLEU points in WMT14 (Biçici et al., 2014). Performance improvement over last year's results is likely due to using higher order n-grams for data selection. ParFDA Moses SMT system is able to obtain the top TER performance in fr-en.

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

$S \to T$	Time (Min)							Space (MB)		
	ParFDA			Moses			Overall	Moses		
	Train	LM	Total	Train	Tune	Total		PT	LM	ALL
en-cs	10	73	83	999	1085	2154	2237	3914	4826	41930
cs-en	11	524	535	965	413	1445	1980	3789	6586	39661
en-de	9	146	155	852	359	1279	1434	3333	4867	36638
de-en	6	232	238	797	421	1285	1523	3065	6233	34316
en-fi	7	0	7	591	569	1212	1219	2605	18746	44334
fi-en	5	308	313	543	164	744	1057	2278	6115	22933
en-fr	22	233	255	2313	331	2730	2985	5628	7359	76970
fr-en	26	330	356	2810	851	3749	4105	6173	6731	86442
en-ru	11	463	474	704	643	1429	1903	4081	4719	43479
ru-en	42	341	383	704	361	1140	1523	4039	6463	40948

Table 2: The space and time required for building the ParFDA Moses SMT systems. The sizes are in MB and time in minutes. PT stands for the phrase table. ALL does not contain the size of the LM.

BLEUc	$S \to en$					$en \to T$				
	cs-en	de-en	fi-en	fr-en	ru-en	en-cs	en-de	en-fi	en-fr	en-ru
ParFDA	0.204	0.2441	0.1541	0.3263	0.2598	0.148	0.1761	0.1135	0.3195	0.22
TopC	0.262	0.293	0.179	0.331	0.279	0.184	0.249	0.127	0.336	0.243
diff	0.058	0.0489	0.0249	0.0047	0.0192	0.036	0.0729	0.0135	0.0165	0.023
LM order	8	8	8	8	8	8	8	10	8	8

Table 3: BLEUc for ParFDA results, for the top constrained result in WMT15 (TopWMTC, from `matrix.statmt.org`), their difference, and the ParFDA LM order used are presented. Average difference is 3.176 BLEU points

2.3 LM Data Quality

A LM selected for a given translation task allows us to train higher order language models, model longer range dependencies better, and achieve lower perplexity as shown in Table 4. We compare the perplexity of the ParFDA selected LM with a LM trained on the ParFDA selected training data and a LM trained using all of the available training corpora. We build LM using SRILM with interpolated Kneser-Ney discounting (`-kndiscount -interpolate`). We also use `-unk` option to build open-vocabulary LM. We are able to achieve significant reductions in the number of OOV tokens and the perplexity, reaching up to 78% reduction in the number of OOV tokens and up to 63% reduction in the perplexity. ParFDA can achieve larger reductions in perplexity than the 27% that can be achieved using a morphological analyzer and disambiguator for Turkish (Yuret and Biçici, 2009) and can decrease the OOV rate at a similar rate. Table 4 also presents the average log probability of tokens and the log probability of token `<unk>`. The increase in the ratio between them in the last column shows that OOV in ParFDA LM are not just less but also less likely at the same time.

3 Conclusion

We use ParFDA for solving computational scalability problems caused by the abundance of training data for SMT models and LMs and still achieve SMT performance that is on par with the top performing SMT systems. ParFDA raises the bar of expectations from SMT with highly accurate translations and lower the bar to entry for SMT into new domains and tasks by allowing fast deployment of SMT systems. ParFDA enables a shift from general purpose SMT systems towards task adaptive SMT solutions. We make the data for building ParFDA Moses SMT systems for WMT15 available: `https://github.com/bicici/ParFDAWMT15`.

Acknowledgments

This work is supported in part by SFI as part of the ADAPT research center

$S \to T$	order	OOV Rate C train	OOV Rate FDA5 train	OOV Rate FDA5 LM	OOV Rate %red	perplexity C train	perplexity FDA5 train	perplexity FDA5 LM	perplexity %red	avg log probability C train	avg log probability FDA5 train	avg log probability FDA5 LM	<unk> log probability C train	<unk> log probability FDA5 train	<unk> log probability FDA5 LM	$\frac{\texttt{<unk>}}{\text{avg}}$ %inc
en-cs	3	.038	.055	.014	.64	763	694	444	.42	-2.91	-2.89	-2.66	-4.94	-5.58	-5.69	.26
	4					716	668	403	.44	-2.89	-2.87	-2.62				.27
	5					703	662	396	.44	-2.88	-2.87	-2.61				.27
	8					699	660	394	.44	-2.88	-2.86	-2.61				.27
cs-en	3	.035	.046	.014	.62	281	255	196	.3	-2.46	-2.42	-2.3	-4.84	-5.33	-5.83	.29
	4					260	243	157	.39	-2.43	-2.4	-2.2				.33
	5					251	237	150	.4	-2.41	-2.39	-2.18				.33
	8					247	236	148	.4	-2.41	-2.39	-2.18				.33
en-de	3	.092	.107	.034	.63	425	383	303	.29	-2.68	-2.64	-2.5	-5.69	-5.92	-5.52	.04
	4					414	377	268	.35	-2.67	-2.64	-2.45				.06
	5					412	376	262	.37	-2.67	-2.64	-2.44				.06
	8					412	376	261	.37	-2.67	-2.64	-2.43				.06
de-en	3	.05	.06	.025	.5	289	265	205	.29	-2.48	-2.45	-2.32	-5.69	-5.85	-5.81	.09
	4					277	258	164	.41	-2.46	-2.44	-2.22				.13
	5					275	257	156	.43	-2.46	-2.43	-2.2				.14
	8					275	257	154	.44	-2.46	-2.43	-2.2				.14
en-fi	3	.203	.213	.128	.37	1413	1290	1347	.05	-3.44	-3.42	-3.31	-4.17	-5.45	-4.2	.05
	4					1403	1285	1323	.06	-3.44	-3.41	-3.3				.05
	5					1401	1284	1320	.06	-3.44	-3.41	-3.3				.05
	8					1400	1284	1319	.06	-3.44	-3.41	-3.3				.05
fi-en	3	.087	.107	.019	**.78**	505	465	228	.55	-2.75	-2.72	-2.37	-4.34	-5.86	-5.91	.58
	4					485	449	188	.61	-2.73	-2.71	-2.28				.63
	5					482	447	179	**.63**	-2.73	-2.71	-2.26				.64
	8					481	446	177	**.63**	-2.73	-2.71	-2.26				.65
en-fr	3	.019	.031	.01	.49	196	146	155	.21	-2.3	-2.18	-2.19	-5.28	-5.56	-5.36	.07
	4					173	137	125	.27	-2.25	-2.15	-2.1				.08
	5					167	136	119	.29	-2.23	-2.15	-2.08				.09
	8					165	136	117	.29	-2.23	-2.15	-2.07				.09
fr-en	3	.022	.031	.01	.52	290	217	220	.24	-2.47	-2.35	-2.35	-5.28	-5.44	-5.31	.06
	4					266	208	187	.3	-2.44	-2.33	-2.28				.08
	5					260	207	181	.3	-2.43	-2.33	-2.26				.08
	8					258	207	180	.3	-2.42	-2.33	-2.26				.08
en-ru	3	.049	.054	.014	.71	547	515	313	.43	-2.77	-2.75	-2.51	-3.57	-4.87	-5.45	.69
	4					537	507	273	.49	-2.77	-2.75	-2.44				.73
	5					536	507	264	.51	-2.77	-2.74	-2.43				**.74**
	8					535	507	259	.52	-2.77	-2.74	-2.42				**.74**
ru-en	3	.041	.046	.017	.58	225	214	188	.16	-2.37	-2.35	-2.28	-3.65	-4.9	-5.79	.65
	4					216	207	148	.31	-2.35	-2.33	-2.18				.71
	5					215	206	140	.35	-2.35	-2.33	-2.15				.73
	8					215	206	138	.36	-2.34	-2.33	-2.15				.73

Table 4: Perplexity comparison of the LM built from the training corpus (train), ParFDA selected training data (FDA5 train), and the ParFDA selected LM data (FDA5 LM). %red is proportion of reduction.

(www.adaptcentre.ie, 07/CE/I1142) at Dublin City University and in part by SFI for the project "Monolingual and Bilingual Text Quality Judgments with Translation Performance Prediction" (computing.dcu.ie/~ebicici/Projects/TIDA_RTM.html, 13/TIDA/I2740). We also thank the SFI/HEA Irish Centre for High-End Computing (ICHEC, www.ichec.ie) for the provision of computational facilities and support.

References

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 Biçici, Qun Liu, and Andy Way. 2014. Parallel FDA5 for fast deployment of accurate statistical machine translation systems. In *Proc. of the Ninth Workshop on Statistical Machine Translation*, pages 59–65, Baltimore, USA, June.

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

Ergun Biçici. 2013. Feature decay algorithms for fast deployment of accurate statistical machine translation systems. In *Proc. of the Eighth Workshop on Statistical Machine Translation*, Sofia, Bulgaria, August.

Ondrej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Pavel Pecina, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proc. of the Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal, September.

David Graff, ngelo Mendona, and Denise DiPersio. 2011. French Gigaword third edition, Linguistic Data Consortium.

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 *ACL*, pages 177–180, Prague, Czech Republic, June.

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

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English Gigaword fifth edition, Linguistic Data Consortium.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proc. Intl. Conf. on Spoken Language Processing*, pages 901–904.

Deniz Yuret and Ergun Biçici. 2009. Modeling morphologically rich languages using split words and unstructured dependencies. In *Proc. of the ACL-IJCNLP 2009 Conference Short Papers*, pages 345–348, Suntec, Singapore, August.