# Parallel FDA5 for Fast Deployment of Accurate Statistical Machine Translation Systems

**Ergun Biçici**
Centre for Next Generation Localisation
School of Computing
Dublin City University
ergun.bicici@computing.dcu.ie

**Qun Liu**
Centre for Next Generation Localisation
School of Computing
Dublin City University
qliu@computing.dcu.ie

**Andy Way**
Centre for Next Generation Localisation
School of Computing
Dublin City University
away@computing.dcu.ie

## Abstract

We use parallel FDA5, an efficiently parameterized and optimized parallel implementation of feature decay algorithms for fast deployment of accurate statistical machine translation systems, taking only about half a day for each translation direction. We build Parallel FDA5 Moses SMT systems for all language pairs in the WMT14 translation task and obtain SMT performance close to the top Moses systems with an average of 3.49 BLEU points difference using significantly less resources for training and development.

## 1 Introduction

Parallel FDA5 is developed for fast deployment of accurate statistical machine translation systems using an efficiently parameterized and optimized parallel implementation of feature decay algorithms (Biçici and Yuret, 2014). Parallel FDA5 takes about half a day for each translation direction. We achieve SMT performance that is on par with the top constrained Moses SMT systems.

Statistical machine translation (SMT) is a data intensive problem. If you have the translations for the source sentences you are translating in your training set or even portions of it, then the translation task becomes easier. If some tokens are not found in the training data then you cannot translate them and if some translated word do not appear in your language model (LM) corpus, then it becomes harder for the SMT engine to find its correct position in the translation. The importance of parallel FDA5 increases with the proliferation of training material available for building SMT systems. Table 2 presents the statistics of the available training and LM corpora for the constrained (C) systems as well as the statistics of the Parallel FDA5 selected training and LM corpora.

Parallel FDA5 runs separate FDA5 models on randomized subsets of the training data and combines the selections afterwards. We run parallel FDA5 SMT experiments using Moses (Koehn et al., 2007) in all language pairs in WMT14 (Bojar et al., 2014) and obtain SMT performance close to the top constrained Moses systems training using all of the training material. Parallel FDA5 allows rapid prototyping of SMT systems for a given target domain or task and can be very useful for MT in target domains with limited resources or in disaster and crisis situations (Lewis et al., 2011).

## 2 Parallel FDA5 for Instance Selection

### 2.1 FDA5

FDA is developed mainly for building high performance SMT systems using fewer yet relevant data that is selected for increasing the coverage of the test set features while maximizing their diversity (Biçici and Yuret, 2011; Biçici, 2011). Parallel FDA parallelize instance selection and significantly reduces the time to deploy accurate MT systems in the presence of large training data from weeks to half a day and still achieve state-of-the-art SMT performance (Biçici, 2013). FDA5 is developed for efficient parameterization, optimization, and implementation of FDA (Biçici and Yuret, 2014). FDA5 can be used in both transductive learning scenarios where test set is used to select the training data or in active learning scenarios where training set itself is used to obtain a sorting of the training data and select.

We run transductive learning experiments in this work such that the instance selection is performed for the given test set. According to SMT experiments performed on the 2 million sentence English-German section of the Europarl corpus (Biçici and Yuret, 2014), FDA5 can increase the performance by 0.41 BLEU points compared to using all of the available training data and by

**Algorithm 1:** Parallel FDA5

```
Input: Parallel training sentences $\mathcal{U}$, test set
       features $\mathcal{F}$, and desired number of
       training instances $N$.
Output: Subset of the parallel sentences to be
        used as the training data $\mathcal{L} \subseteq \mathcal{U}$.
$\mathcal{U} \leftarrow \texttt{shuffle}(\mathcal{U})$
$\boldsymbol{\mathcal{U}}, M \leftarrow \texttt{split}(\mathcal{U}, N)$
$\mathbf{L} \leftarrow \{\}$
foreach $\mathcal{U}_i \in \boldsymbol{\mathcal{U}}$ do
    $\langle \mathcal{L}_i, \mathbf{s}_i \rangle \leftarrow \texttt{FDA5}(\mathcal{U}_i, \mathcal{F}, M)$
    $\mathbf{L} \leftarrow \mathbf{L} \cup \langle \mathcal{L}_i, \mathbf{s}_i \rangle$
$\mathcal{L} \leftarrow \texttt{merge}(\mathbf{L})$
```

3.22 BLEU points compared to random selection. FDA5 is also used for selecting the training set in the WMT14 medical translation task (Calixto et al., 2014) and the tuning set in the WMT14 German-English translation task (Li et al., 2014).

FDA5 has 5 parameters that effect the instance scores based on the three formulas used:

- Initialization:

$$\texttt{init}(f) = \log(|\mathcal{U}|/C_{\mathcal{U}}(f))^i \, |f|^l \qquad (1)$$

- Decay:

$$\texttt{decay}(f) = \texttt{init}(f)(1+C_{\mathcal{L}}(f))^{-c} d^{C_{\mathcal{L}}(f)} \qquad (2)$$

- Sentence score:

$$\texttt{sentScore}(S) = \frac{1}{|S|^s} \sum_{f \in F(S)} \mathit{fvalue}(f) \qquad (3)$$

$C_{\mathcal{L}}(f)$ returns the count of feature $f$ in $\mathcal{L}$. `d` is the feature score polynomial decay factor, `c` is the feature score exponential decay factor, `s` is the sentence score length exponent, `i` is the initial feature score idf exponent, and `l` is the initial feature score $n$-gram length exponent. FDA5 is available at `http://github.com/bicici/FDA` and the FDA5 optimizer is available at `http://github.com/bicici/FDAOptimization`.

### 2.2 Parallel FDA5

Parallel FDA5 (ParFDA5) is presented in Algorithm 1, which first shuffles the training sentences, $\mathcal{U}$ and runs individual FDA5 models on the multiple splits from which equal number of sentences, $M$, are selected. We use ParFDA5 for selecting parallel training data and LM data for building SMT systems. `merge` combines $k$ sorted arrays, $\mathcal{L}_i$, into one sorted array in $O(Mk \log k)$ using their scores, $\mathbf{s}_i$, where $Mk$ is the total number of elements in all of the input arrays. [1] ParFDA5 makes FDA5 more scalable to domains with large training corpora and allows rapid deployment of SMT systems. By selecting from random splits of the original corpus, we work with different $n$-gram feature distributions in each split and prevent feature values from becoming negligible, which can enhance the diversity.

### 2.3 Language Model Data Selection

We select the LM training data with ParFDA5 based on the following observation (Biçici, 2013):

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

It is impossible for an SMT system to translate a word unseen in the training corpus nor can it translate it with a word not found in the target side of the training set [2]. 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 1-gram features for LM corpus selection since we don't know which phrases will be generated by the translation model. After the LM corpus selection, the target side of the parallel training data is added to the LM corpus.

## 3 Results

We run ParFDA5 SMT experiments for all language pairs in both directions in the WMT14 translation task (Bojar et al., 2014), which include English-Czech (en-cs), English-German (en-de), English-French (en-fr), English-Hindi (en-hi), and English-Russian (en-ru). We true-case all of the corpora, use 150-best lists during tuning, set the LM order to a value between 7 and 10 for all language pairs, and train the LM using SRILM (Stolcke, 2002). We set the maximum sentence length filter to 126 and for GIZA++ (Och and Ney, 2003),

[1] (Cormen et al., 2009), question 6.5-9. Merging k sorted lists into one sorted list using a min-heap for k-way merging.

[2] Unless the translation is a verbatim copy of the source.

| $S \rightarrow T$ | | Training Data | | | | | LM Data | |
|---|---|---|---|---|---|---|---|---|
| | Data | #word S (M) | #word T (M) | #sent (K) | SCOV | TCOV | #word (M) | TCOV |
| en-cs | C | 253.5 | 223.4 | 16068 | 0.8282 | 0.7046 | 717.0 | 0.8539 |
| en-cs | ParFDA5 | 22.0 | 19.6 | 1205 | 0.8161 | 0.6062 | 325.8 | 0.8238 |
| cs-en | C | 223.4 | 253.5 | 16068 | 0.7046 | 0.8282 | 5541.9 | 0.9552 |
| cs-en | ParFDA5 | 19.3 | 22.0 | 1205 | 0.7046 | 0.7581 | 351.0 | 0.9132 |
| en-de | C | 116.0 | 109.5 | 4511 | 0.812 | 0.7101 | 1573.8 | 0.8921 |
| en-de | ParFDA5 | 16.7 | 16.8 | 845 | 0.8033 | 0.6316 | 206.9 | 0.8184 |
| de-en | C | 109.5 | 116.0 | 4511 | 0.7101 | 0.812 | 5446.8 | 0.9525 |
| de-en | ParFDA5 | 17.8 | 19.6 | 845 | 0.7087 | 0.753 | 339.5 | 0.9082 |
| en-fr | C | 1096.1 | 1287.8 | 40344 | 0.8885 | 0.9163 | 2534.5 | 0.9611 |
| en-fr | ParFDA5 | 22.6 | 26.6 | 1008 | 0.8735 | 0.8412 | 737.4 | 0.9491 |
| fr-en | C | 1287.8 | 1096.1 | 40344 | 0.9163 | 0.8885 | 6255.8 | 0.9675 |
| fr-en | ParFDA5 | 20.9 | 19.3 | 1008 | 0.8963 | 0.7845 | 463.4 | 0.9282 |
| en-hi | C | 3.4 | 5.0 | 306 | 0.5467 | 0.5986 | 36.3 | 0.7972 |
| en-hi | ParFDA5 | 3.3 | 4.9 | 254 | 0.5467 | 0.5976 | 41.2 | 0.8115 |
| hi-en | C | 5.0 | 3.4 | 306 | 0.5986 | 0.5467 | 5350.4 | 0.9473 |
| hi-en | ParFDA5 | 5.0 | 3.3 | 284 | 0.5985 | 0.5466 | 966.8 | 0.9209 |
| en-ru | C | 49.6 | 46.1 | 2531 | 0.7992 | 0.6823 | 590.8 | 0.8679 |
| en-ru | ParFDA5 | 19.6 | 18.6 | 1107 | 0.7991 | 0.6388 | 282.1 | 0.8447 |
| ru-en | C | 46.1 | 49.6 | 2531 | 0.6823 | 0.7992 | 5380.6 | 0.9567 |
| ru-en | ParFDA5 | 16.6 | 19.4 | 1107 | 0.6821 | 0.7586 | 225.1 | 0.9009 |

Table 2: The data statistics for the available training and LM corpora for the constrained (C) submissions compared with the ParFDA5 selected training and LM corpora statistics. #words is in millions (M) and #sents is in thousands (K).

| | $S \rightarrow T$ | d | c | s | i | l |
|---|---|---|---|---|---|---|
| Training, $n = 2$ | en-de | 1.0 | 0.5817 | 1.4176 | 5.0001 | -3.154 |
| | de-en | 1.0 | 1.0924 | 1.3604 | 5.0001 | -4.341 |
| | en-cs | 1.0 | 0.0676 | 0.8299 | 5.0001 | -0.8788 |
| | cs-en | 1.0 | 1.5063 | 0.7777 | 3.223 | -2.3824 |
| | en-ru | 1.0 | 0.6519 | 1.6877 | 5.0001 | -1.1888 |
| | ru-en | 1.0 | 1.607 | 3.0001 | 0.0 | -1.8247 |
| | en-hi | 1.0 | 3.0001 | 3.0001 | 1.5701 | -1.5699 |
| | hi-en | 1.0 | 0.0 | 1.1001 | 5.0001 | -0.8264 |
| | en-fr | 1.0 | 0.8143 | 0.801 | 3.5996 | -1.3394 |
| | fr-en | 1.0 | 0.19 | 1.0106 | 5.0001 | 1.238 |
| LM, $n = 1$ | en-de | 1.0 | 0.1924 | 1.0487 | 5.0001 | 4.9404 |
| | de-en | 1.0 | 1.7877 | 3.0001 | 3.1213 | -0.4147 |
| | en-cs | 1.0 | 0.4988 | 1.1586 | 5.0001 | -5.0001 |
| | cs-en | 0.9255 | 0.2787 | 0.7439 | 3.7264 | -2.0564 |
| | en-ru | 1.0 | 1.4419 | 2.239 | 1.5543 | -0.5097 |
| | ru-en | 1.0 | 2.4844 | 3.0001 | 4.6669 | 3.7978 |
| | en-hi | 1.0 | 0.0 | 0.0 | 5.0001 | -4.944 |
| | hi-en | 1.0 | 0.3053 | 3.0001 | 5.0001 | 4.1216 |
| | en-fr | 1.0 | 3.0001 | 2.0452 | 3.0229 | 3.4364 |
| | fr-en | 1.0 | 0.7467 | 0.7641 | 5.0001 | 5.0001 |

Table 1: Optimized ParFDA5 parameters for selecting the training set using 2-grams or the LM corpus using 1-grams.

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 contains 5000 sentences, 2000 of which are randomly sampled from previous years' development sets (2008-2012) and 3000 come from the development set for WMT14.

### 3.1 Optimized ParFDA5 Parameters

Table 1 presents the optimized ParFDA5 parameters obtained using the development set. Translation direction specific differences are visible. A negative value for l shows that FDA5 prefers shorter features, which we observe mainly when the target language is English. We also observe higher exponential decay rates when the target language is mainly English. For optimizing the parameters for selecting LM corpus instances, we still use a parallel corpus and instead of optimizing for TCOV, we optimize for SCOV such that we select instances that are relevant for the target training corpus but still maximize the coverage of source features and be able to represent the source sentences within a translation task. The selected LM corpus is prepared for a translation task.

### 3.2 Data Selection

We select the same number of sentences with Parallel FDA (Biçici, 2013), which is roughly $15\%$ of the training corpus for en-de, $35\%$ for ru-en, $6\%$ for cs-en, and $2\%$ for en-fr. After the training set selection, we select the LM data using the target side of the training set as the target domain to select LM instances for. 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. We select 15 million sentences for each LM not including the se-

|  | Time (Min) |  |  |  |  |  |  | Space (MB) |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
| $S \to T$ | ParFDA5 |  |  | Moses |  |  | Overall | Moses |  |  |
|  | Train | LM | Total | Train | Tune | Total |  | PT | LM | ALL |
| en-cs | 5 | 28 | 34 | 375 | 702 | 1162 | 1196 | 1871 | 5865 | 19746 |
| cs-en | 7 | 65 | 72 | 358 | 448 | 867 | 939 | 1808 | 4906 | 18650 |
| en-de | 8 | 29 | 38 | 302 | 1059 | 1459 | 1497 | 1676 | 2923 | 18313 |
| de-en | 8 | 85 | 93 | 358 | 474 | 924 | 1017 | 1854 | 5219 | 19247 |
| en-fr | 23 | 60 | 84 | 488 | 781 | 1372 | 1456 | 2309 | 9577 | 24362 |
| fr-en | 21 | 99 | 120 | 315 | 490 | 897 | 1017 | 1845 | 4888 | 17466 |
| en-hi | 2 | 9 | 11 | 91 | 366 | 511 | 522 | 269 | 817 | 4292 |
| hi-en | 1 | 36 | 37 | 91 | 330 | 467 | 504 | 285 | 9697 | 3845 |
| en-ru | 11 | 25 | 35 | 358 | 369 | 837 | 872 | 2174 | 4770 | 21283 |
| ru-en | 10 | 62 | 71 | 309 | 510 | 895 | 966 | 1939 | 2735 | 19537 |

Table 3: The space and time required for building the ParFDA5 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 | fr-en | hi-en | ru-en | en-cs | en-de | en-fr | en-hi | en-ru |
| WMT14C | 0.288 | 0.28 | 0.35 | 0.139 | 0.318 | 0.21 | 0.201 | 0.358 | 0.111 | 0.287 |
| ParFDA5 | 0.256 | 0.239 | 0.319 | 0.105 | 0.282 | 0.172 | 0.168 | 0.325 | 0.07 | 0.257 |
| diff | 0.032 | 0.041 | 0.031 | 0.034 | 0.036 | 0.038 | 0.033 | 0.033 | 0.041 | 0.03 |
| LM order | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 7 | 10 | 9 |

Table 4: BLEUc for the top constrained result in WMT14 (WMT14C) and for ParFDA5 results, their difference to WMT14C, and the LM order used are presented. Average difference is 3.49 BLEU points.

lected training set, which is added later. The statistics of the ParFDA5 selected training data and the available training data for the constrained translation task is given in Table 2. The size of the LM corpora includes both the LDC and the monolingual LM corpora provided by WMT14. Table 2 shows the significant size differences between the constrained dataset (C) and the ParFDA5 selected data. Table 2 also present the source and target coverage (SCOV and TCOV) in terms of the 2-grams of the test set observed in the training data or the LM data. The quality of the training corpus can be measured by TCOV, which is found to correlate well with the BLEU performance achievable (Biçici and Yuret, 2011; Biçici, 2011).

### 3.3 Computing Statistics

We quantify the time and space requirements for running ParFDA5 SMT systems for each translation direction. The space and time required for building the ParFDA5 Moses SMT systems are given in Table 3 where the sizes are in MB and the time in minutes. PT stands for the phrase table. We used Moses version 2.1.1, from `www.statmt.org/moses`. Building a ParFDA5 Moses SMT system takes about half a day.

### 3.4 Translation Results

The results of our two ParFDA5 SMT experiments for each language pair and their tokenized BLEU performance, BLEUc, together with the LM order used and the top constrained submissions to the WMT14 are given in Table 4 [3], which use phrase-based Moses for comparison [4]. We observed significant gains (+0.23 BLEU points) using higher order LMs last year (Biçici, 2013) and therefore we use LMs of order 7 to 10. The test set contains 10,000 sentences and only 3000 of which are used for evaluation, which can make the transductive learning application of ParFDA5 harder. In the transductive learning setting, ParFDA5 is selecting target test task specific SMT resources and therefore, having irrelevant instances in the test set may decrease the performance by causing FDA5 to select more domain specific data and less task specific. ParFDA5 significantly reduces the time required for training, development, and deployment of an SMT system for a given translation

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

[4] Phrase-based Moses systems usually rank in the top 3.

| | | | | | | | ppl | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | OOV | | | | log OOV = $-19$ | | | | log OOV = $-11$ | | | |
| Translation | T | order | train | FDA5 | FDA5 LM | % red. | train | FDA5 | FDA5 LM | % red. | train | FDA5 | FDA5 LM | % red. |
| en-cs | en | 3 | 866 | 1205 | 525 | 0.39 | 1764 | 1731 | 938 | 0.47 | 1370 | 1218 | 805 | 0.41 |
| | | 4 | | | | | 1788 | 1746 | 877 | 0.51 | 1389 | 1229 | 753 | 0.46 |
| | | 5 | | | | | 1799 | 1752 | 868 | 0.52 | 1398 | 1233 | 745 | 0.47 |
| | | 6 | | | | | 1802 | 1753 | 867 | 0.52 | 1400 | 1234 | 744 | 0.47 |
| cs-en | cs | 3 | 557 | 706 | 276 | 0.5 | 480 | 419 | 333 | 0.31 | 408 | 342 | 307 | 0.25 |
| | | 4 | | | | | 487 | 422 | 292 | 0.4 | 415 | 344 | 269 | 0.35 |
| | | 5 | | | | | 495 | 424 | 285 | 0.42 | 421 | 346 | 263 | 0.38 |
| | | 6 | | | | | 497 | 425 | 284 | 0.43 | 423 | 346 | 262 | 0.38 |
| en-de | en | 3 | 1666 | 2116 | 744 | 0.55 | 1323 | 1605 | 747 | 0.44 | 831 | 890 | 607 | 0.27 |
| | | 4 | | | | | 1307 | 1596 | 689 | 0.47 | 821 | 885 | 560 | 0.32 |
| | | 5 | | | | | 1307 | 1596 | 680 | 0.48 | 822 | 885 | 553 | 0.33 |
| | | 6 | | | | | 1308 | 1596 | 679 | 0.48 | 822 | 885 | 552 | 0.33 |
| de-en | de | 3 | 691 | 849 | 417 | 0.4 | 482 | 498 | 394 | 0.18 | 386 | 379 | 345 | 0.11 |
| | | 4 | | | | | 470 | 490 | 344 | 0.27 | 376 | 373 | 301 | 0.2 |
| | | 5 | | | | | 470 | 490 | 336 | 0.29 | 377 | 373 | 293 | 0.22 |
| | | 6 | | | | | 471 | 490 | 334 | 0.29 | 377 | 373 | 292 | 0.23 |
| en-fr | en | 3 | 270 | 411 | 153 | 0.43 | 185 | 167 | 173 | 0.07 | 173 | 151 | 166 | 0.04 |
| | | 4 | | | | | 170 | 160 | 135 | 0.21 | 159 | 144 | 130 | 0.19 |
| | | 5 | | | | | 171 | 160 | 126 | 0.27 | 160 | 145 | 121 | 0.24 |
| fr-en | fr | 3 | 306 | 604 | 179 | 0.42 | 349 | 325 | 275 | 0.21 | 320 | 275 | 261 | 0.19 |
| | | 4 | | | | | 338 | 321 | 235 | 0.3 | 310 | 271 | 224 | 0.28 |
| | | 5 | | | | | 342 | 322 | 228 | 0.33 | 314 | 272 | 217 | 0.31 |
| en-hi | en | 3 | 2035 | 2123 | 950 | 0.53 | 242 | 246 | 114 | 0.53 | 168 | 168 | 96 | 0.43 |
| | | 4 | | | | | 237 | 241 | 87 | 0.63 | 164 | 165 | 73 | 0.55 |
| | | 5 | | | | | 238 | 242 | 78 | 0.67 | 165 | 165 | 66 | 0.6 |
| | | 6 | | | | | 239 | 242 | 75 | 0.68 | 165 | 165 | 64 | 0.62 |
| hi-en | hi | 3 | 1842 | 1860 | 623 | **0.66** | 1894 | 1898 | 482 | 0.75 | 915 | 911 | 377 | 0.59 |
| | | 4 | | | | | 1910 | 1914 | 398 | 0.79 | 923 | 919 | 312 | 0.66 |
| | | 5 | | | | | 1915 | 1919 | 378 | **0.8** | 925 | 921 | 296 | 0.68 |
| | | 6 | | | | | 1915 | 1919 | 378 | **0.8** | 926 | 921 | 296 | 0.68 |
| en-ru | en | 3 | 959 | 1176 | 585 | 0.39 | 1067 | 1171 | 668 | 0.37 | 814 | 840 | 566 | 0.3 |
| | | 4 | | | | | 1053 | 1159 | 603 | 0.43 | 803 | 831 | 511 | 0.36 |
| | | 5 | | | | | 1052 | 1159 | 591 | 0.44 | 802 | 831 | 501 | 0.38 |
| | | 6 | | | | | 1052 | 1159 | 588 | 0.44 | 802 | 831 | 498 | 0.38 |
| ru-en | ru | 3 | 558 | 689 | 340 | 0.39 | 385 | 398 | 363 | 0.06 | 334 | 334 | 333 | 0.0 |
| | | 4 | | | | | 377 | 391 | 325 | 0.14 | 327 | 328 | 298 | 0.09 |
| | | 5 | | | | | 378 | 392 | 318 | 0.16 | 328 | 329 | 292 | 0.11 |
| | | 6 | | | | | 378 | 392 | 318 | 0.16 | 328 | 329 | 291 | 0.11 |

Table 5: Perplexity comparison of the LM built from the training corpus (train), ParFDA5 selected training corpus (FDA5), and the ParFDA5 selected LM corpus (FDA5 LM). % red. column lists the percentage of reduction.

task. The average difference to the top constrained submission in WMT14 is 3.49 BLEU points. For en-ru and en-cs, true-casing the LM using a true-caser trained on all of the available training data decreased the performance by 0.5 and 0.9 BLEU points respectively and for cs-en and fr-en, increased the performance by 0.2 and 0.5 BLEU points. We use the true-cased LM results using a true-caser trained on all of the available training data for all language pairs where for hi-en, the true-caser is trained on the ParFDA5 selected training data.

### 3.5 LM Data Quality

A LM training data selected for a given translation task allows us to train higher order language models, model longer range dependencies better, and at the same time, achieve lower perplexity as given in Table 5. We compare the perplexity of the ParFDA5 selected LM with a LM trained on the ParFDA5 selected training data and a LM trained using all of the available training corpora. To be able to compare the perplexities, we take the OOV tokens into consideration during calculations (Biçici, 2013). We present results for the cases when we handle OOV words with a cost of $-19$ or $-11$ each in Table 5. We are able to achieve significant reductions in the number of OOV tokens and the perplexity, reaching up to 66% reduction in the number of OOV tokens and up to 80% reduction in the perplexity.

| BLEUc | $S \rightarrow en$ | | | | $en \rightarrow T$ | | | |
|---|---|---|---|---|---|---|---|---|
| | cs-en | de-en | fr-en | ru-en | en-cs | en-de | en-fr | en-ru |
| ParFDA5 | 0.256 | 0.239 | 0.319 | 0.282 | 0.172 | 0.168 | 0.325 | 0.257 |
| ParFDA | 0.243 | 0.241 | 0.254 | 0.223 | 0.171 | 0.179 | 0.238 | 0.173 |
| diff | 0.013 | -0.002 | 0.065 | 0.059 | 0.001 | -0.011 | 0.087 | 0.084 |

Table 7: Parallel FDA5 WMT14 results compared with parallel FDA WMT13 results. Training set sizes are given in millions (M) of words on the target side. Average difference is 3.7 BLEU points.

| BLEUc | $S \rightarrow en$ | | $en \rightarrow T$ | |
|---|---|---|---|---|
| | cs-en | fr-en | en-cs | en-fr |
| ParFDA5 | 0.256 | 0.319 | 0.172 | 0.325 |
| ParFDA5 $15\%$ | 0.248 | 0.321 | 0.178 | 0.333 |
| diff | -0.008 | 0.002 | 0.006 | 0.008 |

Table 6: ParFDA5 results, ParFDA5 results using $15\%$ of the training set, and their difference.

### 3.6 Using $15\%$ of the Available Training Set

In the FDA5 results (Biçici and Yuret, 2014), we found that selecting $15\%$ of the best training set size maximizes the performance for the English-German out-of-domain translation task and achieves 0.41 BLEU points improvement over a baseline system using all of the available training data. We run additional experiments selecting $15\%$ of the training data for fr-en and cs-en language pairs to see the effect of increased training sets selected with ParFDA5. The results are given in Table 6 where most of the results improve. The slight performance decrease for cs-en may be due to using a true-caser trained on only the selected training data. We observe larger gains in the $en \rightarrow T$ translations.

### 3.7 ParFDA5 versus Parallel FDA

We compare this year's results with the results we obtained last year (Biçici, 2013) in Table 7. The task setting is different in WMT14 since the test set contains 10,000 sentences but only 3000 of these are used as the actual test set, which can make the transductive learning application of ParFDA5 harder. We select the same number of instances for the training sets but 5 million more instances for the LM corpus this year. The average difference to the top constrained submission in WMT13 was 2.88 BLEU points (Biçici, 2013) and this has increased to 3.49 BLEU points in WMT14. On average, the performance improved 3.7 BLEU points when compared with ParFDA results last year. For the fr-en, en-fr, and en-ru translation directions, we observe increases in the performance. This may be due to better modeling of the target domain by better parameterization and optimization that FDA5 is providing. We observe some decrease in the performance in en-de and de-en results. Since the training material remained the same for WMT13 and WMT14 and the modeling power of FDA5 increased, building a domain specific rather than a task specific ParFDA5 model may be the reason for the decrease.

## 4 Conclusion

We use parallel FDA5 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. Parallel FDA5 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 in about half a day. Parallel FDA5 enables a shift from general purpose SMT systems towards task adaptive SMT solutions.

## Acknowledgments

This work is supported in part by SFI (07/CE/I1142) as part of the CNGL Centre for Global Intelligent Content (www.cngl.org) at Dublin City University and in part by the European Commission through the QTLaunchPad FP7 project (No: 296347). We also thank the SFI/HEA Irish Centre for High-End Computing (ICHEC) for the provision of computational facilities and support.

## References

Ergun Biçici and Deniz Yuret. 2011. Instance selection for machine translation using feature decay algorithms. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 272–283, Edinburgh, Scotland, July. Association for Computational Linguistics.

Ergun Biçici and Deniz Yuret. 2014. Optimizing instance selection for statistical machine translation with feature decay algorithms. *IEEE/ACM Transactions On Audio, Speech, and Language Processing (TASLP)*.

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 *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Sofia, Bulgaria, August. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Matouš Macháček, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, and Lucia Specia. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proc. of the Ninth Workshop on Statistical Machine Translation*, Balrimore, USA, June. Association for Computational Linguistics.

Iacer Calixto, Ali Hosseinzadeh Vahid, Xiaojun Zhang, Jian Zhang, Xiaofeng Wu, Andy Way, and Qun Liu. 2014. Experiments in medical translation shared task at wmt 2014. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA, June. Association for Computational Linguistics.

Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. 2009. *Introduction to Algorithms (3. ed.)*. MIT Press.

David Graff, Ângelo Mendonça, 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.

William Lewis, Robert Munro, and Stephan Vogel. 2011. Crisis mt: Developing a cookbook for mt in crisis situations. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 501–511, Edinburgh, Scotland, July. Association for Computational Linguistics.

Liangyou Li, Xiaofeng Wu, Santiago Cortes Vaillo, Jun Xie, Jia Xu, Andy Way, and Qun Liu. 2014. The dcu-ictcas-tsinghua mt system at wmt 2014 on german-english translation task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, USA, June. Association for Computational Linguistics.

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.