Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020), pages 285–293
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

On the Exploration of English to Urdu Machine Translation

Sadaf Abdul Rauf[1,2], **Syeda Abida**[1], **Noor-e-Hira**[1], **Syeda Zahra**[1]
Dania Parvez[1], **Javeria Bashir**[1] **and Qurat-ul-ain Majid**[1]
[1] Fatima Jinnah Women University, Pakistan
[2] LIMSI-CNRS, France
{firstName.lastName}@gmail.com

Abstract

Machine Translation is the inevitable technology to reduce communication barriers in today's world. It has made substantial progress in recent years and is being widely used in commercial as well as non-profit sectors. Such is only the case for European and other high resource languages. For English-Urdu language pair, the technology is in its infancy stage due to scarcity of resources. Present research is an important milestone in English-Urdu machine translation, as we present results for four major domains including Biomedical, Religious, Technological and General using Statistical and Neural Machine Translation. We performed series of experiments in attempts to optimize the performance of each system and also to study the impact of data sources on the systems. Finally, we established a comparison of the data sources and the effect of language model size on statistical machine translation performance.

Keywords: English , Urdu, Statistical Machine Translation (SMT), Neural Machine Translation (NMT)

1. Introduction

Machine translation (MT) for low resource languages has been a challenging task (Irvine, 2013; Zoph et al., 2016). The dimensionality of difficulty increases when it comes to translating between a morphologically rich and morphologically poor language (Habash and Sadat, 2006). In this study, we will be presenting one such pair, English to Urdu translation, with English being a morphologically simple language while Urdu is a language with rich inflectional and derivational morphology. In case of Urdu-English translation topological distance between both languages is the biggest hurdle to get best results (Jawaid et al., 2016; Khan et al., 2017).

Findings of WMT 2011 evaluation (Callison-Burch et al., 2011) reported Urdu-English translation to be a relatively difficult problem. With some works on rule based systems (RBMT) (Tafseer and Alvi, 2002; Karamat, 2006; Naila Ata, 2007) and a small cascade of works on phrase based SMT systems (Jawaid and Zeman, 2011; Ali et al., 2013; Jawaid et al., 2014a), hierarchical MT systems (Khan et al., 2013; Jawaid et al., 2014a) and NMT using transfer learning from a high resource language (Zoph et al., 2016), it is still an arena requiring much work. Present study is a consolidated study in this regard.

In this study we present results of some of the unexplored areas with reference to this language pair. Previous works have built general domain translation systems, we present a domain analysis on Technological, Religious and General domain translations (Section 5). This study is also an attempt to initiate the field of MT for Bio-medical domain despite zero resources available for the language pair. Effect of smaller and larger language models on translations are also explored.

We have explored and used all the freely available English-Urdu corpora and also developed various small corpora by using human translations, synthetic corpora by machine translation and Hindi to Urdu transliteration. Starting with a brief review of previous works we describe the resources used in Section 3 followed by detailed results in Section 4 The paper concludes with a brief discussion on results.

2. Related Works

Perhaps, Tafseer and Alvi (2002) presents one of the earliest attempts on English to Urdu translation based on transforming the parse tree of the English sentence to Urdu using transformation rules. Issues relating to translation for verbs in context of English to Urdu RBMT using lexical functional grammar are discussed by (Karamat, 2006). A minimal English to Urdu RBMT system is presented in (Naila Ata, 2007) (Jawaid and Zeman, 2011) used phrase based models to solve the long distance word reordering problem between the two languages. They used Emille (Baker et al., 2002), Treebank (Marcus et al., 1993), Quran and Bible corpora and report improvement in BLEU scores by the proposed reordering scheme. Our general domain systems are built using these above mentioned corpora.

(Jawaid and Zeman, 2011) used phrase based models to solve the long distance word reordering problem between the two languages. They used Emille (Baker et al., 2002), Treebank (Marcus et al., 1993), Quran and bible corpora and report improvement in BLEU scores by the proposed reordering scheme. We also use these corpora in our general domain systems. Building up on previous work (Jawaid et al., 2014a) present a comparison of phrase based versus hierarchical systems. They have added AFRL corpus (not free) to the earlier system and reported the hierarchical systems to outperform phrase based systems. (Ali et al., 2010; Ali et al., 2013) built SMT using parallel ahadith corpus from Sahih bukhari and Sahih Muslim. (Khan et al., 2013) also presented a hierarchical SMT system.

Several other studies have also contributed, for instance (Shahnawaz and Mishra, 2013) and (Khan Jadoon et al., 2017) present neural systems trained on small corpora.

Category	Corpora	Size	Tokens		Sentences			
		(Mbs)	(Millions)					
			UR	EN	Train	Dev	Test	Total
General	Emille	1.5	0.12	0.09	5583	176	118	5877
	Treebank	2.3	0.18	0.13	5408	170	115	5693
	Indic	8.8	0.63	0.49	33244	1000	1000	35244
	NLT	3.1	0.22	0.19	10662	336	226	11224
	OPUS	4.7	0.38	0.33	46805	1501	1002	49308
	TDIL	0.42	0.03	0.02	1141	37	25	1203
	Flickr_H	0.42	0.03	0.03	2578	82	55	2715
	Flickr_G	0.41	0.04	0.03	2578	82	55	2715
	Transliterations	0.99	0.08	0.07	3441	516	172	4129
	Total	22.64	1.71	1.38	111440	3990	2768	118108
Bio-Medical	Emille	0.92	0.07	0.05	2970	78	77	3125
	Scielo	9.1	0.65	0.60	21680	650	492	22822
	Jang Health News	1.9	0.14	0.12	5450	360	264	6074
	EMEA	14.3	1.03	0.82	51775	1363	1363	54501
	Total	26.22	1.89	1.59	81875	2451	2196	86522
Religious	Quran	2.9	0.24	0.03	6000	214	200	6414
	Bible	2.5	0.20	0.21	7400	300	257	7957
	QBJ	55.5	1.13	1.02	47198	1250	1062	49510
	Tanzil	1000	23.1	19.0	710904	22449	14967	748320
	Total	1060.9	24.67	20.26	771502	24213	16486	812201
Techno-logy	Gnome	0.85	0.06	0.05	13186	417	278	13881
	Ubuntu	0.16	0.02	0.01	2873	90	62	3025
	Total	1.01	0.08	0.06	16059	507	340	16906
Mono-lingual	Jawaid	717.4	95.4	-	-	-	-	5464575
	NLT	5.4	0.63	-	-	-	-	62063
	Jhang	3.3	0.39	-	-	-	-	32984
	All Urdu corpus	199.4	26.2	-	-	-	-	934631
	Total	925.5	122.7	-	-	-	-	6494253

Table 1: Corpus Details: Training, development, test and monolingual data used for each domain.

3. Data Collection

Data collection and its cleaning is an important but a challenging part for NLP, including machine translation. Our Data collection scheme included 1) an extensive search of all the freely available parallel corpora. 2) Synthetic parallel corpus creation using a good translation system and 3) transliteration from a highly similar language, Hindi.
We have categorised the corpora in four categories, General, Biomedical, Religious and Technology, each explained in subsections 3.1, 3.2, 3.3, and 3.4 respectively. Corpus details are summarized in table 1.

3.1. General

This section lists the corpora and their details for general category.

1. The Emille[1] corpus (Baker et al., 2002) is a collection of annotated, parallel and monolingual data in written and spoken form. It consists of multi domain corpora (social, legal, educational, health, etc.) in fourteen South Asian languages and is distributed by ELRA (European Language Resource Association). This first crowd sourced corpus enabled initial work on Indian languages. We used English-Urdu part of this dataset consisting of 9000 sentences. Health documents from Emille corpus were separated and used as the BioMedical corpus.

2. CLE[2] released Urdu translations of Wall Street Journal part of The Penn *Treebank* corpus (Marcus et al., 1993). The Urdu corpus was available online and we were able to get English sentences from LDC Treebank.

3. Indic[3] is a freely available multi-domain parallel corpus created by using crowd-sourcing (Post et al., 2012).

4. TDIL[4] is an Indian Language Technology Proliferation and Deployment Center. We were able to get a sample of this corpus in domains of tourism, art, culture and architecture etc.

5. Opus[5] project (Tiedemann, 2012) provides freely

[1] The Emille/CIIL Corpus:ID:ELRA-W0037

[2] http://www.cle.org.pk/software/ling_resources

[3] http://joshua-decoder.org/indian-parallel-corpora/

[4] http://tdil-dc.in/index.php—?lang=en

[5] http://opus.nlpl.eu/

available annotated corpora to the research community. We used their English-Urdu corpus comprising of Tanzil, Tatoeba, OpenSubtitles {2016, 2018}, Ubuntu, GNOME and Global Voices. Tanzil was a religious corpus, whereas Ubuntu and Gnome were technology related corpora. We further sub categorized these according to the domains as shown in table 1.

6. Flickr corpora are the human and automatic translations of the flickr 8[6] Image to text Corpus. The human translations are done from English captions to Urdu by human translators and Google translate was used for automatic translations.

7. National Language Translations (NLT) are the translation documents obtained from a translation agency. We collected translations of various articles, books, survey reports etc. The data collected was in raw form, it was cleaned and sentence aligned.

8. UMC002 Hindi-Urdu transliterations. Hindi and Urdu are almost similar languages having different writing scripts. To overcome data scarceness we experimented with transliterations from Hindi to Urdu. A similar scheme has been used by (Durrani et al., 2014) but in the opposite direction, .i.e they transliterated from Urdu to Hindi.

3.2. Bio-Medical

Since no prior work exists in the Biomedical domain for English-Urdu, consequently there were no separate parallel corpora available. However, *Emille* corpus had a small part comprising of 0.055M English and 0.075 Urdu words respectively in health domain. We used these as Biomedical corpus.

Furthermore, we *developed* Biomedical parallel corpora by using ideas from unsupervised learning techniques successfully used for other language pairs, where translations are used as additional bi-texts to cover up for data scarcity (Lambert et al., 2011) and domain adaptation (Abdul Rauf et al., 2016; Hira et al., 2019). We collected Biomedical parallel corpora from various sources and translated them. We are working on using domain adapted translation and language models for the biomedical domain, however, the translations used in this work are done using google translate. We used the following corpora:

1. Scielo[7] corpus contains documents retrieved from the scielo database comprising of titles and abstracts of published articles in bio-medical domain. Our Scielo corpus comprises of 0.022M sentences. Overall it contains 0.60M English and 0.65M Urdu words.

2. Jang[8] group of news is a Pakistan based media corporation. Their newspapers are published in both Urdu and English independently,but they are not the translations of each other. We cleaned and extracted 6k English sentences from the *health news* section and translated to Urdu to be used as parallel corpus. We got a corpus of 0.11M words in English and 0.14M words in Urdu.

3. EMEA[9] is a parallel corpus extracted out of documents published by European Medical Agency. The corpus is freely available in a number of language pairs but is not available in Urdu. We downloaded English part of corpus available in plain text and selected data related to medicines, disease, treatment and instructions. We automatically translated the extracted dataset and produced Urdu parallel translations. At the end of translation process we got a parallel dataset comprising of 1.03M words in Urdu and 0.82M words in English.

3.3. Religious

This section lists the corpora and their details for religious category.

1. UMC005 (Jawaid and Zeman, 2011) provides 6414 sentence pairs from Bible and 7957 sentence pairs form Quran corpus.

2. QBJ corpus, which is another collection of Quran+Bible+Joshua was also available online with their own test and dev sets. The data consists of 1.02M English words and 1.13M Urdu words.

3. Tanzil is a collection of online Quranic Translations by different scholars and is a sub part of OPUS corpus. The corpus contains 878 bi-texts with total of 0.75M sentence fragments having 19.0M English tokens and 23.1M Urdu tokens.

3.4. Technology

This consists of English-Urdu Parallel corpus from localization files of Ubuntu and Gnome. Ubuntu contains 3.03k sentences and 0.1M, 0.2M English and Urdu tokens respectively, Gnome has 0.05M English and 0.06M Urdu tokens.

3.5. Monolingual Urdu Corpus

Monolingual corpus is an essential resource for building language models for SMT. We used the corpus developed by (Jawaid et al., 2014b). This corpus consists of 95.4 million Urdu words, representing 5.4 million sentences of various domains including science, news, religion and education.

We also collected Urdu monolingual documents from Jang (0.03M sentences) and other sources comprising of (0.06M sentences) as shown at the end of table 1. Urdu side of all parallel corpora was also used to build the large language model used in the indicated experiments in results.

3.6. Data Preprocessing

Data cleaning and preprocessing is highly important for the performance of MT systems. The corpora provided by Emillie, NLT and Penn Tree-bank were partially parallel

[6] https://forms.illinois.edu/sec/1713398

[7] http://www.statmt.org/wmt16/biomedical-translation-task.html

[8] https://jang.com.pk/

[9] http://opus.nlpl.eu/EMEA.php

so we sentence aligned them using LF sentence aligner. [10] Due to the topological distance between the two languages we were not able to get fully aligned parallel corpus using LF aligner, thus manual alignment was done to ensure correctness.

4. Experimental Framework

To demonstrate the performance of MT systems on the corpora collected and generated in this work, we performed a number of experiments for SMT and a few experiments for NMT. This section provides the description of the experimental frameworks and settings used for building SMT and NMT systems.

4.1. Statistical Machine Translation:

The goal of SMT is to produce a target sentence e from a source sentence f. Among all possible target language sentences the one with the highest probability is chosen:

$$\begin{aligned} e^* &= \arg\max_e \Pr(e|f) && (1) \\ &= \arg\max_e \Pr(f|e)\Pr(e) && (2) \end{aligned}$$

where $\Pr(f|e)$ is the translation model and $\Pr(e)$ is the target language model (LM). This approach is usually referred to as the *noisy source-channel* approach in SMT (Brown et al., 1993). Bilingual corpora are needed to train the translation model and monolingual texts to train the target language model.

Common practice is to use phrases as translation units (Koehn et al., 2003; Och and Ney, 2003a) instead of the original word-based approach. A phrase is defined as a group of source words $\tilde{f}$ that should be translated together into a group of target words $\tilde{e}$. The translation model in phrase-based systems includes the phrase translation probabilities in both directions, i.e. $P(\tilde{e}|\tilde{f})$ and $P(\tilde{f}|\tilde{e})$. The use of a maximum entropy approach simplifies the introduction of several additional models explaining the translation process :

$$\begin{aligned} e^* &= \arg\max Pr(e|f) \\ &= \arg\max_e \{exp(\sum_i \lambda_i h_i(e,f))\} && (3) \end{aligned}$$

The feature functions h_i are the system models and the λ_i weights are typically optimized to maximize a scoring function on a development set. In our system fourteen features functions were used, namely phrase and lexical translation probabilities in both directions, seven features for the lexicalized distortion model, a word and a phrase penalty, and a target language model.

To built standard phrase-based SMT systems we used Moses toolkit (Koehn et al., 2007), with the default settings for all the parameters. A 5-gram KenLM (Heafield, 2011) language model was used. For individual systems the language models were trained on the target side of the corpus. For experiments on size of the language model, all the available monolingual and target side corpus was used (122.5M Urdu words).

Word-alignment was done using Giza++ (Och and Ney, 2003b) with grow-diag-final-and symmetrization method. Maximum sentence length was chosen to be 100. A distortion limit of 6 with 100-best list was used. Msd-bidirectional-fe feature was used for lexical reordering with the phrase limit of 5. Systems were tuned on the development data using the MERT (Och, 2003). BLEU (Papineni et al., 2002) scores were computed on dev and test sets of the corpora, as well as on standard test sets. BLEU scores were calculated using *multi-bleu.perl*. Scoring is case sensitive and includes punctuation.

4.2. Neural Machine Translation:

We used OpenNMT[11] (Klein et al., 2017) for building Neural MT systems. Two layered encoder-decoder architecture with global attention (Luong et al., 2015) was used. We used RNN size of 500 and LSTM for cell structure for both encoder and decoder, applying dropout of 0.3 for each input cell. Translations were evaluated on BLEU scores to enable comparison with the corresponding SMT systems.

4.3. Development and Test sets

Most of the corpora available online had their own development (dev) and test sets, so we evaluated the systems according to these *dev* and *test* sets. To be able to compare the systems in each domain, we created *Standard test set (STS)* for each domain comprising of 1k sentences. We randomly selected sentences from test sets of each data source of the particular domain. This was done on the basis of data set size and combined these specific sized chunks so that each data-set is represented on the basis of its size in the standard test set. We also used the test set of *CLE*[9] which was used to evaluate the general domain systems and the standard *Scielo* test set for Bio-Medical domain.

5. Results and Discussion

One of the endeavours of our study is to present domain specific translation results. As is common in machine learning approaches, the domain of the system being built depends on the data used to train the system. MT performance quickly degrades when the testing domain is different from the training domain. The reason for this degradation is that the statistical models closely approximate the empirical distributions of the training data (Abdul Rauf et al., 2016). MT system trained on parallel data from the news domain may not give appropriate translations when used to translate articles from the medical domain.

This study intends to build MT systems for four different domains namely Bio-medical, Religious, Technological and General domain. We evaluated our systems on the development and test data along with the standard test set for each domain, and the CLE test set as explained in section 4.3 We will be giving weight-age to Standard test scores as they are representative of system performance on the whole domain rather than the test set created from the data itself (see section 4.3)

[10] https://sourceforge.net/projects/aligner/

[11] http://opennmt.net/OpenNMT/

[9] http://www.cle.org.pk/software/ling_ resources/testingcorpusmt.htm

Corpus	Tokens (UR) (millions)	BLEU									
		SMT					largeLM				
		Dev	Test				Dev	Test			
			Self	STS	CLE	Scielo		Self	STS	CLE	Scielo
General											
Emille	0.12	35.38	5.67	2.58	1.97	NA	40.09	8.86	5.34	2.55	NA
Treebank	0.18	18.14	20.90	3.73	5.10	NA	20.66	24.62	6.27	6.13	NA
Indic	0.63	11.67	12.23	8.28	4.41	NA	21.86	22.77	15.39	5.56	NA
NLT	0.22	15.09	8.52	4.80	4.04	NA	20.76	10.96	8.26	4.23	NA
OPUS	0.38	12.27	14.08	4.79	3.06	NA	15.99	18.16	10.15	6.22	NA
TDIL	0.03	5.85	3.01	1.70	1.21	NA	7.93	4.66	2.58	1.51	NA
FlickrHumanTrans	0.03	3.02	2.39	2.04	0.1	NA	3.94	2.78	2.27	0.00	NA
FlickrGooglTrans	0.03	35.71	27.58	0.56	0.1	NA	36.23	28.18	1.98	0.90	NA
Transliteration	0.08	54.30	47.34	2.08	0.95	NA	54.53	48.90	2.21	1.32	NA
Bio-Medical											
Scielo	0.65	39.20	34.33	25.95	NA	27.97	46.59	41.13	37.10	NA	-
Jang	0.14	33.46	49.78	17.78	NA	17.99	40.62	61.76	30.25	NA	-
EMEA	1.03	40.59	48.66	44.45	NA	19.25	54.56	54.43	50.15	NA	-
Emille	0.07	20.88	3.41	12.90	NA	10.81	29.15	3.23	24.25	NA	-
Religious											
Quran	0.24	16.03	12.44	12.54	NA	NA	23.33	20.84	19.63	NA	NA
Bible	0.20	17.69	11.16	11.17	NA	NA	31.07	23.55	23.45	NA	NA
QBJ	1.13	10.37	10.05	9.98	NA	NA	20.29	21.96	22.08	NA	NA
Tanzil	23.1	19.93	17.46	17.08	NA	NA	-	-	-	NA	NA
Technology											
Gnome	0.06	78.58	79.42	79.42	NA	NA	83.25	83.15	12.81	NA	NA
Ubuntu	0.02	10.05	5.36	5.36	NA	NA	13.43	12.60	14.61	NA	NA

Table 2: BLEU Scores of all Standalone corpora on SMT Systems for English to Urdu translation.

5.1. Standalone SMT Systems

To build the best domain specific SMT system, we first explored the performance of each corpora for standalone SMT systems. Table 2 lists the BLEU scores for each system. As already mentioned we are interested in the scores obtained on standard test set, it is observed that $Indic$ showed the best performance among the systems built on general domain corpus. Whereas; $Treebank$, $Transliteration$ and $FlickrGoogleTranslate$, despite outperforming on self test have shown a decline in performance for standard test set. The standard test set includes part of the test sentences from each corpora basis of data set size. Indic has the most tokens, resultantly the standard test set includes sentences from Indic the most. This explains the best performance on STS.

For the systems built on Bio-medical corpora, $EMEA$ showed the best performance on standard set. Interestingly, in this domain we see reasonable BLEU scores on all test sets, including STS. Similar phenomenon of better scores for EMEA on STS is observed, which corresponds to more sentences from $EMEA$ test set in STS.SMT system trained on $Jang$ shows an abrupt decline in the performance for standard test set while achieving the best BLEU point (49.78) among all the other SMT systems built for this domain, when chosen on test set scores. This is particularly interesting as $Jang$ has 0.14M tokens while $EMEA$ has 1.03M Urdu tokens.

$Tanzil$ and $Genome$ showed the best performance for Religious and technology domains respectively. While overfitting is observed in these two domains. The performance of the systems, built for these two domains, have shown a uniform trend for both self and standard test sets.

5.1.1. Effect of size of Language Model

Along with, the exploration of best SMT system for each category we also investigated the effect of the size of language model on each standalone SMT system. To explore this dimension, a large language model was also build by concatenating the Urdu text of all the bi-texts and the monolingual corpus mentioned in section 3.5. The scores for large LM are shown in the third column in table 2. It is observed that the BLEU scores of all the standalone systems approximately doubled with large LM. Figure 1 shows these results graphically for each domain. These results highlight the effect of bigger language model on SMT quality, obviously a bigger language model helps improve translation quality by improving the grammar of the output sentences.

5.2. Concatenated SMT Systems

After building standalone systems for each corpus, we selected the corpora which resulted in best BLEU scores, for building systems by concatenating different combinations of corpora. We selected systems on the basis of best score among the standalone systems from each domain (baseline system) and concatenated them with system having second highest BLEU score. Table 3 reports these results.

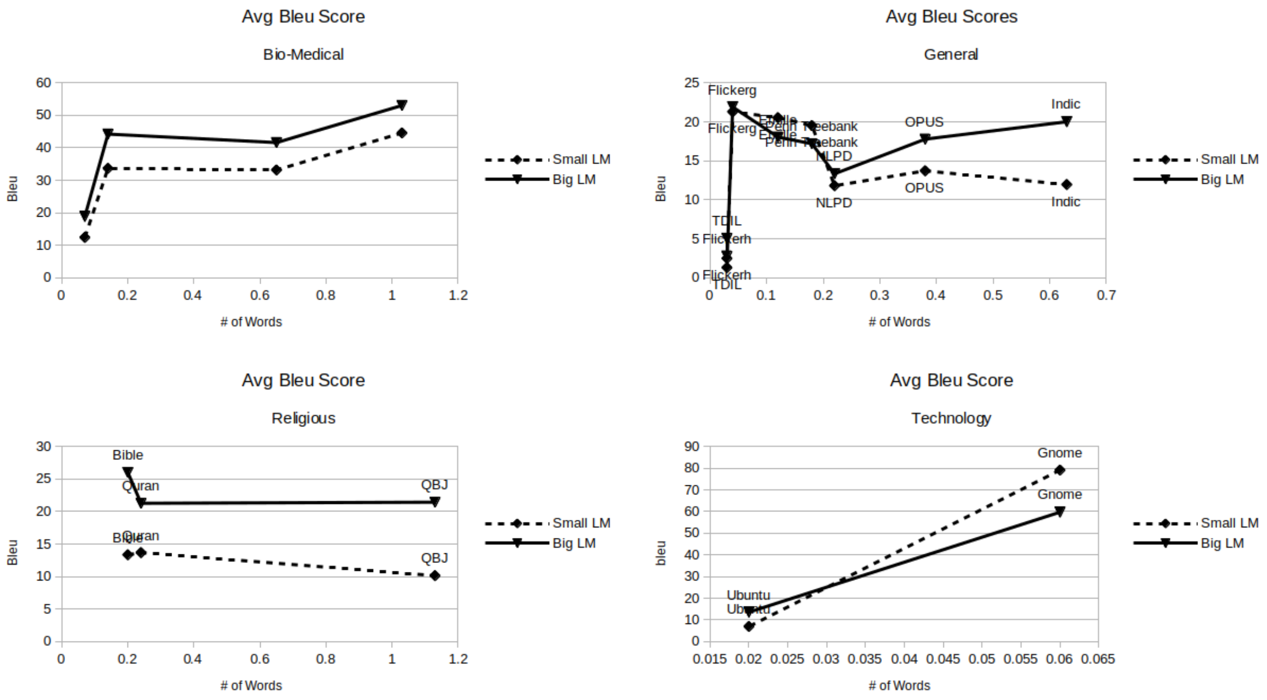

Figure 1: Comparison of systems built on small and large language model (x-axis represents words in millions)

SMT Results on baseline +					
Category	**Baseline**	**Tokens**	**BLEU**		
			Dev	Test	STS
General	Emille	0.12	35.38	5.67	2.58
	+Treebank	0.3	23.62	13.32	4.24
	+Treebank+NLT	0.52	21.08	11.93	6.96
	+Treebank+NLT+Indic	1.15	15.77	12.06	10.08
	+Treebank+NLT+Indic+TDIL+OPUS	1.56	15.21	12.53	-
Bio-Medical	EMEA	1.03	40.59	48.66	44.45
	+Scielo	1.86	40.09	43.83	50.34
	+Scielo+jang	1.68	39.91	44.23	49.76
	+Scileo+jang+Emille	1.89	39.54	44.13	50.71
Religious	Tanzil	23.1	19.93	17.46	17.08
	+Bible	23.3	-	-	-
	+Bible+Quran	23.54	-	-	-
	+Bible+Quran+QBJ	24.67	19.52	17.18	18.36
Techno-logy	Gnome	0.06	78.58	79.42	79.42
	+Ubuntu	0.08	69.38	58.47	58.56

Table 3: Results of SMT on baselines and addition of bitexts.

5.2.1. Bio-Medical Domain

Bio-medical domain is an interesting domain as the corpora are not of same type. $Emille$ are the health domain sentences taken from the $Emille$ corpus, $Jang$ sentences are taken from a semi-parallel comparable corpus and then sentence aligned and human corrected. Whereas, $EMEA$ and $Scielo$ are synthetic forward translated corpora.

$EMEA$ was chosen as baseline, for bio-medical domain, having the highest score 44.45 amongst other three standalone systems. Then, we built a system on $EMEA$ concatenated with the second best system $Scielo$, having score of 25.95 (table 2). The BLEU score of the resultant system $EMEA$+$Scielo$ is 50.34 (table 3). We can see an improvement in the score after concatenation of these two data-sets. Note that this system is built with only forward translated synthetic corpus, and we get an appreciable BLEU score.

This system $EMEA + Scielo$, is further concatenated with $jang$ corpus (standalone score 17.78) and the resultant score of the $EMEA + scielo + jang$ system is 49.76, which is a bit lower than the previous system's score. Contrary to the standard test set scores, addition of bitexts did not improve scores for dev and test, rather resulted in a de-

Corpus	Size(M)	Dev	Test	STS	CLE
FlickrHumanTrans	0.42	3.02	2.39	2.04	0.11
FlickrGooglTrans	0.41	35.71	27.58	0.56	2.35
Transliteration	0.99	54.30	47.34	2.08	0.95
FlickrHumanTrans + Transliteration	1.4	39.23	40.91	2.24	1.44
FlickrGooglTrans + Transliteration	1.4	48.61	45.87	2.66	2.13
Emille	1.5	35.38	5.67	2.58	1.97
Emille + Transliteration	2.5	38.97	30.02	2.36	2.35
Emille + Treebank	3.8	23.62	13.32	4.24	-
Emille + Treebank + FlickrGooglTrans	4.21	24.69	14.28	4.28	6.53

Table 4: Bleu scores using human translations vs machine translations as training data

cline of BLEU score. $EMEA$ and $Scielo$ are translated from standard biomedical corpora as described in data preprocessing section 3.6 The sentences of these corpora specially of EMEA consists of concise short sentences of similar nature (we found certain redundancies in these corpora). That is the reason their concatenation gave a big increase as it mounted to adding more data. On the other hand, we created $Jang$ corpus by automatic translation of news and tips in health section of a national English news paper. This could be a reason that when concatenated with $EMEA + Scielo$ the combined score reduced to 49.76 from 50.34.

Finally, concatenating with $Emille$, having BLEU score 12.90 for standalone model, the score for the resultant system is 50.71 which is highest among all other systems. $Emille$ is again a standard biomedical corpus comprising of health documents from the EMILLE corpus (section 3.6), and its concatenation improved the overall BLEU score. An increase of 6.26 points upon the addition of just 0.86M words of $Scielo+Jang+Emille$ corpora to 1.03M words of EMEA (baseline), has been observed which is a significant gain. These are encouraging results for the development of standard corpora for the Bio-medical domain.

5.2.2. Religious, General and Technology Domain

For the religious domain we have two corpora namely $Tanzil$ and the other is concatenation of $Quran$, $Bible$ and $Joshua(QBJ)$. Firstly we built two standalone systems for both corpora as shown in Table:2, Tanzil having BLEU score of 17.46 on test set and 19.93 on dev set. BLEU score of QBJ is 10.05 on test set and 10.37 on dev set. We did not create standard test set to evaluate these two corpora as there is a huge difference between the size of corpora, if we generate standard corpus out of these by evenly distributing them; the standard test set will mostly consist of bitexts from $Tanzil$. In this case $Tanzil$ will perform well for that specific standard test set but QBJ would not be able to perform well. After standalone evaluation we concatenated both data sets to see the impact of corpora on each other. We got 18.36 BLEU score that is better then the standalone systems. Again the performance of system increases with the increase of the size of corpora.

For the general domain, we considered $Emille$ as a baseline on the basis of higher score 35.38 on dev set and 5.67 on test set so its average BLEU score is higher then the rest of standalone corpus (table 2). We concatenated it with the $Treebank$ having score 18.14 on dev set and 20.90 on test set,and got 23.63 score on dev set and 13.32 on test set. We further concatenated this system with NLT whose standalone BLEU score are 15.09 on dev and 8.52 on test set, and got scores of 21.08 on dev and 11.93 on test set. Finally we concatenate our last data set $Indic$ having score 11.67 on dev and 12.23 on test set. Following the same trend as seen in the biomedical domain, we see a steady improvement in the standard test scores by the addition of bitexts.

Interestingly, technology domain gave the best results. $Gnome$ being the baseline of the domain achieved 78.58 BLEU score on dev data and 79.42 on both test sets. Whereas, $Ubuntu$ had a standalone BLEU score of 10.05 and 5.36 on dev and test of both test sets (table 2). $Gnome$ corpus had a maximum sentence length of 40 to 50 whereas all other data sets had sentence size of 100 words. A combination of two yields a great improvement with respect to $Ubuntu$ but a decrease for $Gnome$.

5.3. Impact of Various Corpora

We performed series of experiments using transliterations, human and machine translated data to compare the performance of such systems. These results are reported in Table 4. On the standard test set transliterations and human translations were better than google translations having scores on 2.08, 2.04 and 0.56 respectively. When evaluated on dev and test sets of individual corpora, surprisingly $FlickrHumanTrans$ performed worst of all with minimum BLEU scores of 3.02 on dev and 2.39 on test. These are the captions from flickr 8k dataset and often the English side is not grammatically correct. More interestingly, the same corpus when translated using google gave 35.71 on dev, 27.58 on test. $Transliteration$ of Hindi UMC002 corpus to Urdu gave the best scores of 54.30 and 47.34 on dev and test respectively.

$FlickrHumanTrans$ is further combined with the $Transliteration$ data set which is machine transliterated data, to build another system in order to observe the effect of machine transliterated data on the human translations. The performance of the resultant system is far better than the baseline system $FlickrHumanTrans$ yielding 2.24 on standard test set, 39.23 on dev and 40.91 on test. Further, we concatenated transliterations with the baseline

NMT Results on baseline + additional bitexts for Bio-Medical					
Baseline	Tokens	BLEU			
		Dev	Test	STS	Scielo
EMEA	1.03	26.44	40.27	39.81	5.24
+Scielo	1.68	26.96	35.22	45.90	14.37
+Scielo+jang	1.82	27.22	35.01	47.46	16.09
+Scileo+jang+Emille	1.89	27.55	34.62	46.28	16.72

Table 5: Results on NMT on addition of bitexts for Bio-medical domains.

$FlickrGooglTrans$ and a good improvement in scores is observed i.e. 2.66 BLEU on standard test data, 48.61 on dev data, 45.87 on test data.

Now, we address the question of effect on performance by addition of these corpora to the already available resources. $Emille$ is the already available human translated corpus, when combined with our transliterated data set, an improvement of almost 4.00 and 24.35 BLEU points on dev and test is observed, however on standard test a decline of 0.22 points is observed. Similar trend is observed when machine translated data is added to $Emille + Treebank$ yielding improvements on all datasets.

5.4. NMT Systems

We are presenting NMT system performance only for Bio-Medical domain. Table 5 shows the results of our experiments for NMT. We maintained the same baseline and corpus concatenation combination as used in SMT experiments. The results of Bio-Medical NMT are lower than the corresponding SMT systems (Table 3). This is expected as NMT systems don't perform well with small amounts of corpus. A unanimous observation is that addition of bitexts improves the systems across all dev and test sets, a slight deviation to this trend is observed when $Emille$ is added to $EMEA + Scielo + jang$ (last row in Table 5).

6. Conclusion

We presented domain based results on SMT and NMT systems for translation from English to Urdu. This is the first work being reported on several domains for the English-Urdu language pair. We collected corpora for four main domains namely Bio-medical, Religious, Technology and General. We experimented with various methods to reduce data scarcity which include, the use of automatic translations and transliterations. We also collected and compiled human translations from translation agencies as well as produced human translations of Flickr 8k dataset. We performed series of experiments in attempts to optimize the performance of each system and also to study the impact of data sources on the systems. Finally, we established a comparison of the data sources and the effect of Language Model size on statistical machine translation performance.

7. Bibliographical References

Abdul Rauf, S., Schwenk, H., Lambert, P., and Nawaz, M. (2016). Empirical use of information retrieval to build synthetic data for smt domain adaptation. *IEEE Transactions on Audio, Speech and Language Processing*.

Ali, A., Siddiq, S., and Malik, M. K. (2010). Development of Parallel Corpus and English to Urdu Statistical Machine Translation. *International Journal of Engineering*, (05):3–6.

Ali, A., Hussain, A., and Kamran Malik, M. (2013). Model for English-Urdu statistical machine translation. *World Applied Sciences Journal*, 24(10):1362–1367.

Baker, P., Hardie, A., McEnery, T., Cunningham, H., and Gaizauskas, R. J. (2002). Emille, a 67-million word corpus of indic languages: Data collection, mark-up and harmonisation. In *LREC*.

Brown, P., Della Pietra, S., Della Pietra, V. J., and Mercer, R. (1993). The mathematics of statistical machine translation. *Computational Linguistics*, 19(2):263–311.

Callison-Burch, C., Koehn, P., Monz, C., and Zaidan, O. F. (2011). Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64. Association for Computational Linguistics.

Durrani, N., Haddow, B., Koehn, P., and Heafield, K. (2014). Edinburgh's phrase-based machine translation systems for wmt-14. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 97–104.

Habash, N. and Sadat, F. (2006). Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 49–52. Association for Computational Linguistics.

Heafield, K. (2011). Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics.

Hira, N.-e., Abdul Rauf, S., Kiani, K., Zafar, A., and Nawaz, R. (2019). Exploring transfer learning and domain data selection for the biomedical translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 158–165, Florence, Italy, August. Association for Computational Linguistics.

Irvine, A. (2013). Statistical Machine Translation in Low Resource Settings. *Proceedings of the 2013 NAACL HLT Student Research Workshop*, (June):54–61.

Jawaid, B. and Zeman, D. (2011). Word-Order Issues in English-to-Urdu Statistical Machine Translation. *Prague Bulletin of Mathematical Linguistics*, 95(-1):87–106.

Jawaid, B., Kamran, A., and Bojar, O. (2014a). English to Urdu Statistical Machine Translation: Establishing a Baseline. *Proceedings of the Fifth Workshop on South and Southeast Asian Natural Language Processing*, pages 37–42.

Jawaid, B., Kamran, A., and Bojar, O. (2014b). Urdu monolingual corpus. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics ('UFAL), Faculty of Mathematics and Physics, Charles University.

Jawaid, B., Kamran, A., and Bojar, O. (2016). Enriching Source for English-to-Urdu Machine Translation. *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*, pages 54–63.

Karamat, N. (2006). *Verb transfer for English to Urdu Machine Translation*. Ph.D. thesis, National University of Computer & Emerging Sciences.

Khan, N., Anwar, M. W., Bajwa, U. I., and Durrani, N. (2013). English to urdu hierarchical phrase-based statistical machine translation. In *Proceedings of the 4th Workshop on South and Southeast Asian Natural Language Processing*, pages 72–76.

Khan, N. J., Anwar, W., and Durrani, N. (2017). Machine Translation Approaches and Survey for Indian Languages. *arXiv preprint arXiv:1701.04290*, 18(1):47–78.

Khan Jadoon, N., Anwar, W., Bajwa, U. I., and Ahmad, F. (2017). Statistical machine translation of Indian languages: a survey. *Neural Computing and Applications*.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). Opennmt: Open-source toolkit for neural machine translation. *CoRR*, abs/1701.02810.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrased-based machine translation. pages 127–133.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Meeting of the Association for Computational Linguistics*, pages 177–180.

Lambert, P., Schwenk, H., Servan, C., and Abdul-Rauf, S. (2011). Investigations on translation model adaptation using monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 284–293, Edinburgh, Scotland, July. Association for Computational Linguistics.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Naila Ata, Bushra Jawaid, A. K. (2007). Rule based english to urdu machine translation. *In Proceedings of Conference on Language and Technology (CLT'07). 2007.*

Och, F. J. and Ney, H. (2003a). A systematic comparison of various statistical alignement models. *Computational Linguistics*, 29(1):19–51.

Och, F. J. and Ney, H. (2003b). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 160–167. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Post, M., Callison-Burch, C., and Osborne, M. (2012). Constructing parallel corpora for six indian languages via crowdsourcing. *Wmt-2012*, pages 401–409.

Shahnawaz and Mishra. (2013). Statistical Machine Translation System for English to Urdu. *Int. J. Adv. Intell. Paradigms*, 5(3):182–203.

Tafseer, A. and Alvi, S. (2002). English to urdu translation system. *manuscript, University of Karachi*.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In Nicoletta Calzolari (Conference Chair) and Khalid Choukri and Thierry Declerck and Mehmet Ugur Dogan and Bente Maegaard and Joseph Mariani and Jan Odijk and Stelios Piperidis, editor, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer Learning for Low-Resource Neural Machine Translation.