English to Urdu Statistical Machine Translation: Establishing a Baseline

Bushra Jawaid, Amir Kamran and Ondřej Bojar
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské nám. 25, Praha 1, CZ-118 00, Czech Republic
jawaid,kamran,bojar@ufal.mff.cuni.cz

Abstract

The aim of this paper is to categorize and present the existence of resources for English-to-Urdu machine translation (MT) and to establish an empirical baseline for this task. By doing so, we hope to set up a common ground for MT research with Urdu to allow for a congruent progress in this field. We build baseline phrase-based MT (PBMT) and hierarchical MT systems and report the results on 3 official independent test sets. On all test sets, hierarchial MT significantly outperformed PBMT. The highest single-reference BLEU score is achieved by the hierarchical system and reaches 21.58% but this figure depends on the randomly selected test set. Our manual evaluation of 175 sentences suggests that in 45% of sentences, the hierarchical MT is ranked better than the PBMT output compared to 21% of sentences where PBMT wins, the rest being equal.

1 Introduction

Statistical Machine Translation (SMT) has always been a challenging task for language pairs with significant word ordering differences and rich inflectional morphology. The language pair such as English and Urdu, despite of descending from the same family of Indo-European languages, differs heavily in syntactic strucure and morphological characteristics. English is relatively fixed word order language and follows subject-verb-object (SVO) structure whereas Urdu uses restricted free word order language and most commonly follows the SOV pattern. Urdu word order is restricted for only few parts of speeches such as adjectives always precede nouns and postpositions follow nouns. Unlike English, Urdu is a pro-drop language. The morphology of Urdu is similar to other Indo-European languages, e.g. by having inflectional morphological system.

To the best of our knowledge, the research on English-to-Urdu machine translation has been very much fragmented, preventing the authors to build upon the works of others. Our underlying motivation for this paper is to establish a common ground and provide a concise summary of available data resources and set up reproducible baseline results of several available test sets. With this basis, future Urdu MT research should be able to stepwise improve the state of the art, in contrast with the scattered experiments done so far (Khan et al., 2013; Ali et al., 2013; Ali and Malik, 2010).

In Section 2, the experimental setup and data processing tools are described. Existing corpora are introduced in Section 3, automatic results are reported in Section 4 and manual evaluation is discussed in Section 5.

2 Experimental Setup

This section briefly introduces the selection of SMT models that are used to build the baseline English-Urdu SMT system and also explains the processing of parallel data before passing it to the MT system.

2.1 Two Models of SMT

The state-of-the-art MT toolkit Moses[1] (Koehn et al., 2007), offers two mainstream models of SMT: phrase-based (PBMT) and syntax-based (SBMT) that includes the hierarchical model.

The PBMT model operates only on mapping of source phrases (short sequences of words) to target phrases. For dealing with word order differences, two rather weak models are available: lexicalized and distance-based. The lexicalized reordering models (Tillmann, 2004) are considered more advanced as they condition reordering on the actual phrases, whereas the latter model makes the reordering cost (paid when picking source phrases out of sequence) dependent only on the length of the jump. The distance-based model is suited well for local reordering but it is fairly weak in capturing any long distance reorderings.

The syntax-based model (SBMT) builds upon Synchronous Context-Free Grammar (SCFG) that synchronously generates source and target sentences. The grammar rules can either consist of linguistically motivated non-terminals such as NP, VP etc. or the generic non-terminal "X" in which case the model is called "hierarchical phrase-based" (Chiang, 2005; Chiang, 2007). In either case, the model is capable of capturing long-distance reordering much better than the lexicalized reordering of PBMT.

2.2 Data Processing and MT Training

For the training of our en-ur translation systems, the standard training pipeline of Moses is used along with the GIZA++ (Och and Ney, 2000) alignment toolkit and a 5-gram SRILM language model (Stolcke, 2002). The source texts were processed using the Treex platform (Popel and Žabokrtský, 2010)[2], which included tokenization and lemmatization.

The target side of the corpus is tokenized using a simple tokenization script[3] by Dan Zeman and it is lemmatized using the Urdu Shallow Parser[4] developed by Language Technologies Research Center of IIIT Hyderabad.

The alignments are learnt from the lemmatized version of the corpus. In all other cases, word forms (i.e. no morphological decomposition) in their true case (i.e. names capitalized but sentence starts lowercased) are used. The lexicalized reordering model uses the feature set called "msd-bidirectional-fe".

3 Dataset

Parallel and monolingual data resources are very scarce for low-resource language pairs such as English-Urdu. This section highlights the existing parallel and monolingual data resources that can be utilized for training SMT models. The number of official test sets are also exhibited.

3.1 Parallel Corpus

Our parallel corpus consists of around 79K sentences collected from five different sources. The collection comes from several domains such as News, Religion, Technology, Language and Culture etc. 95% of the data is used for training, whereas the rest is evenly split into dev and test sets.

- **Emille:** EMILLE (Enabling Minority Language Engineering) (Baker et al., 2002) is a collection of monolingual (written and spoken), parallel and annotated corpora of fourteen South Asian languages which is distributed by the European Language Resources Association (ELRA). The Urdu-English part are documents produced by the British Departments of Health, Social Services, Education and Skills, and Transport, Local Government and the Regions of British government translated into Urdu.

  In this work, the manually sentence aligned version of English-Urdu Emille corpus Jawaid and Zeman (2011) is used.

[1] `http://statmt.org/moses/`

[2] `http://ufal.mff.cuni.cz/treex/`

[3] The tokenization script can be downloaded from: http://hdl.handle.net/11858/00-097C-0000-0023-65A9-5

[4] `http://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow_parser.php`

- **IPC:** The Indic Parallel Corpus (Post et al., 2012)[5] is a collection of Wikipedia documents of six Indian sub-continent languages translated into English through crowdsourcing in the Amazon Mechanical Turk (MTurk) platform.

  The English-Urdu part generally contains four (in some cases three) English translations for each Urdu sentence. In a separate MTurk task, the Turkers voted which of the English translations is the best one. The official training, dev and devtest sets is first merged and afterwards the voting list is used to retrieve only the winning English sentence ignoring sentences with no votes altogether. The official testset is left unaltered to report our final results on this data.

- **Quran:** The publicly available parallel English and Urdu translation of Quranic data[6] is used, which is collected by Jawaid and Zeman (2011) in their work. The data consists of 6K aligned parallel sentences.

- **Penn Treebank:** Penn Treebank (Marcus et al., 1993) is an annotated corpus of around 4.5 million words originating from Wall Street Journal (WSJ), Brown corpus, Switchboard and ATIS. The entire treebank in English is released by the Linguistic Data Consortium (LDC). A subset of the WSJ section whose Urdu translations are provided by Center for Language Engineering (CLE)[7] is used. Out of 2,499 WSJ stories in the Treebank, only 317 are available in Urdu.

- **Afrl:** Afrl, the largest of the parallel resources we were able to get, is not publicly available. The corpus originally consists of 87K sentences coming from mix of several domains mainly news articles. The sentence alignments are manually checked of almost two thirds of the corpus, around 4K misaligned and 30K duplicate sentences are discarded.

The statistics shown in Table 1 are reported after removing duplicated sentences from each source. Almost all parallel corpora contained at least tens or hundreds of duplicate sentences. Afrl on the other hand contained larger chunks of Emille and also smaller subset of Penn Treebank. Around 3K sentences from Afrl that were seen in Emille are discarded but the Penn Treebank subset of Afrl is left intact because it provides different Urdu translations.

Each parallel corpus is randomly split into train, dev and test sets according to its relative size.

Corpus	Sentences	Tokens		% of Data	Train	Dev	Test
		EN	UR				
AFRL	50,313	960,683	1,022,563	63.6%	47,769	1,272	1,272
EMILLE	8,629	152,273	199,320	10.9%	8,193	218	218
IPC	7,478	118,644	132,968	9.46%	7,098	190	190
QURAN	6,364	251,387	269,947	8.05%	6,040	162	162
PENN	6,204	158,727	179,457	7.86%	5,888	158	158
TOTAL	78,988	-	-	100%	74,988	2,000	2,000

Table 1: Statistics of English-Urdu parallel corpora.

3.2 Monolingual Corpus

Jawaid et al. (2014) release[8] a plain and annotated Urdu monolingual corpus of around 95.4 million tokens distributed in around 5.4 million sentences. The monolingual corpus is a mix

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

[6] http://ufal.mff.cuni.cz/legacy/umc/005-en-ur/

[7] http://www.cle.org.pk/software/ling_resources/UrduNepaliEnglishParallelCorpus.htm

[8] http://hdl.handle.net/11858/00-097C-0000-0023-65A9-5

of domains such as News, Religion, Blogs, Literature, Science, Education etc. Only plain text monolingual data is used to build our language model.

3.3 Official Testsets

In addition to the testset that is created from the parallel corpora resources, results are reported on three official testsets.

NIST 2008 Open Machine Translation (OpenMT) Evaluation[9] has distributed test data from 2 domains: Newswire and Web. The Web data is collected from user forums, discussion groups and blogs, whereas Newswire data is a mix of newswire stories and data from web. The test data contain 4 English translations for each Urdu sentence, the first English translation is picked in all cases. Because the majority of test sets are created in order to faciliatate Urdu-to-English MT, most of them contain multiple English references against each Urdu source.

Another testset is released with the IPC. Only those sentences are used whose ids are present in the voting list. The domain of the IPC test set is discussed in Section 3.1.

CLE[10] has published small test set from News domain specifically for MT evaluation. The test data contains 3 Urdu references against each source. All reference translations are used for the evaluation.

Table 2 shows the number of sentences in each test set that are used for the final evaluation. We also report the coverage of each test set (calculated on vocabulary size) i.e. how many source words in a test set were seen in the training data. The notions used in Table 2 to introduce coverage are explained in Section 4.

		NIST 2008		IPC	CLE
		NewsWire	Web Test		
Sentences		400	600	544	400
Coverage	ALL	84%	91%	90%	87%
	Except-Afrl	80%	87%	88%	84%

Table 2: Statistics of official English-Urdu test sets.

4 Results

The BLEU metric (Papineni et al., 2002) has been used to evaluate the performance of the systems. Models are trained on two different datasets: all parallel corpora (referred as "ALL") and parallel data excluding Afrl corpus (referred as "Except-Afrl"). The latter model is trained due to the fact that Afrl corpus is publicly not available. The community working on English-Urdu machine translation can thus have one common baseline that could be used to evaluate their improved systems in the future. Including Afrl allows us to see the gains in performance thanks to the additional data.

Table 3 shows the baseline results of phrase-based and hierarchical systems when trained on both datasets. The results are reported on two test sets: the test set of 2,000 sentences (called Large in Table 3) as shown in Table 1 and its subset of 728 sentences which excludes 1,272 test sentences from Afrl (called Small in Table 3).

PBMT performs better when integrated with lexicalized reordering model but Hierarchical MT outperforms both PBMT setups on both smaller and larger test sets. The absolute BLEU scores drop by up to 6 points when Afrl is removed from the training data, however they return back to ∼20 when Afrl is also removed from the test set. This highlights the importance of data overall and the match in domain in particular, as supported by the differences in vocabulary coverage (see the column "Coverage" in Table 3).

Table 4 shows the results of the best performing setups (i.e. phrase-based with lexicalized reordering model and hierarchical model) trained on both training datasets and evaluated on the

[9] http://catalog.ldc.upenn.edu/LDC2010T21

[10] http://www.cle.org.pk/software/ling_resources/testingcorpusmt.htm

Parallel Corpora	Test Set	Phrase-based	Phrase-based-LexReo	Hierarchical	Coverage
ALL	Large	18.30±0.74	19.19±0.72	21.35±0.84	92%
Except-Afrl	Large	12.85±0.74	13.78±0.73	15.11±0.82	78%
Except-Afrl	Small	18.41±1.25	19.67±1.27	21.21±1.55	91%

Table 3: Results of Phrase-based, Phrase-based with Lexical Reordering and Hierarchical MT systems.

official test sets. The BLEU score for CLE test set is reported using all 3 reference translations at once as well as the average of single-reference BLEUs, taking each reference translation separately. IPC and NIST2008 results are evaluated on a single reference.

The hierarchical MT performs significantly better than the phrase-based MT on all test sets. The lowest scores were achieved on the NIST2008 test set but it is difficult to pinpoint any specific reason (other than some domain difference) because the coverage is comparable to other test sets (see Table 2). Across all the test sets, Afrl corpus brings about 2 points BLEU absolute.

		CLE		IPC	NIST2008
		3 refs	1 ref (avg.)	1 ref	1 ref
ALL	Phrase-based-LexReo	18.19±1.19	11.12±1.02	15.82±1.36	15.13±0.95
	Hierarchical	19.29±1.31	11.81±1.09	18.70±1.64	16.69±1.06
Except-Afrl	Phrase-based-LexReo	16.53±1.13	9.92±0.96	13.82±1.20	11.65±0.87
	Hierarchical	18.48±1.28	11.30±1.03	16.91±1.54	13.01±0.84

Table 4: Results of Phrase-based and Hierarchical systems on official test sets.

5 Manual Evaluation

To manually analyze the output of best performing models sample of 175 sentences is randomly selected from the large test set translated using both PBMT with lexical reordering and hierarchical models trained on "ALL" data sets. QuickJudge[11] is used to rank the outputs. The annotator is shown the source, reference and output from both machine translation systems, the identity of the MT systems is not known. There are four permitted outcomes of the ranking: both systems marked as equally good; both systems are equally bad or the output of one of the systems is better than the other one. Here is the summary of annotation by a single annotator:

- Out of 175 sentences, 41 sentences received equally bad translations from both systems.
- 17 items are marked as equally good.
- In 79 cases, the hierarchical MT is ranked better than the phrase-based MT.
- In the remaining 38 cases, the phrase-based MT is ranked better than the hierarchical MT.

The results from the manual ranking show that the hierarchical systems wins twice more often than PBMT. The two systems tie in about one third of input sentences, of which about 70 % are cases where the translations are bad.

6 Conclusion

In this work, a collection of sizeable English-Urdu corpora is summarized for statistical machine translation. These resources are used to build baseline phrase-based and hierarchical MT systems for translation into Urdu and the results are reported on 3 independent official test sets. This

[11] `http://ufal.mff.cuni.cz/project/euromatrix/quickjudge/`

can hopefully serve as a baseline for a wider community of researchers. The output of both translation models is manually analyzed and it confirms that the hierarchical model is preferred over phrase-based MT for English-to-Urdu translation.

Acknowledgments

This work has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013). This work was also supported by the grant FP7-ICT-2011-7-288487 (MosesCore) of the European Union.

References

Aasim Ali and Muhmmad Kamran Malik. 2010. Development of parallel corpus and english to urdu statistical machine translation. *Int. J. of Engineering & Technology IJET-IJENS*, 10:31–33.

Aasim Ali, Arshad Hussain, and Muhammad Kamran Malik. 2013. Model for english-urdu statistical machine translation. *World Applied Sciences*, 24:1362–1367.

Paul Baker, Andrew Hardie, Tony McEnery, Hamish Cunningham, and Robert J. Gaizauskas. 2002. Emille, a 67-million word corpus of indic languages: Data collection, mark-up and harmonisation. In *LREC*. European Language Resources Association.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proc. of ACL*, pages 263–270.

David Chiang. 2007. Hierarchical phrase-based translation. *Comput. Linguist.*, 33(2):201–228, June.

Bushra Jawaid and Daniel Zeman. 2011. Word-order issues in english-to-urdu statistical machine translation. Number 95, pages 87–106, Praha, Czechia.

Bushra Jawaid, Amir Kamran, and Ondřej Bojar. 2014. A Tagged Corpus and a Tagger for Urdu (to appear). Reykjavík, Iceland. European Language Resources Association. In print.

Nadeem Khan, Waqas Anwar, Usama Ijaz Bajwa, and Nadir Durrani. 2013. English to urdu hierarchical phrase-based statistical machine translation. In *The 4th Workshop on South and Southeast Asian NLP (WSSANLP), IJCNLP*, pages 72–76, Nagoya, Japan.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proc. of ACL Companion Volume, Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.*, 19(2):313–330, June.

Franz Josef Och and Hermann Ney. 2000. A Comparison of Alignment Models for Statistical Machine Translation. In *Proc. of COLING*, pages 1086–1090. ACL.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proc. of ACL*, pages 311–318.

Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: Modular NLP Framework. In *Lecture Notes in Artificial Intelligence, Proceedings of the 7th International Conference on Advances in Natural Language Processing (IceTAL 2010)*, volume 6233 of *Lecture Notes in Computer Science*, pages 293–304. Springer.

Matt Post, Chris Callison-Burch, and Miles Osborne. 2012. Constructing parallel corpora for six indian languages via crowdsourcing. In *Proc. of WMT, ACL*, pages 401–409, Montréal, Canada.

Andreas Stolcke. 2002. Srilm - an extensible language modeling toolkit. In *In Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP) 2002*, pages 901–904.

Christoph Tillmann. 2004. A unigram orientation model for statistical machine translation. In *Proc. of HLT-NAACL Short Papers*, pages 101–104.