Learning Trilingual Dictionaries for Urdu – Roman Urdu – English

Moiz Rauf and **Sebastian Padó**
Institut für Maschinelle Sprachverarbeitung
University of Stuttgart, Germany
{moiz.rauf,pado}@ims.uni-stuttgart.de

Abstract

In this paper, we present an effort to generate a joint Urdu, Roman Urdu and English trilingual lexicon using automated methods. We make a case for using statistical machine translation approaches and parallel corpora for dictionary creation. To this purpose, we use word alignment tools on the corpus and evaluate translations using human evaluators. Despite different writing script and considerable noise in the corpus our results show promise with over 85% accuracy of Roman Urdu–Urdu and 45% English–Urdu pairs.

1 Introduction

Bilingual lexicons serve an integral role in cross lingual information retrieval and bringing NLP to low resourced languages. The process of dictionary generation has greatly benefited from improvements in statistical translation methods. However, for low resourced languages the large parallel and monolingual corpora necessary to learn these dictionaries are hard to come by and remain a critical hurdle (Lam et al., 2015). In this paper, we have developed such a resource for Urdu, English and Roman Urdu (Urdu written in Latin script) language pairs. Urdu is an Indo-Aryan language with an extended Persio-Arabic script. It is the national language of Pakistan (Rasul, 2013), while English has been established as the medium used in educational and official settings in the country (Rafi, 2013; Muhammad Asghar and Mahmood, 2013). Roman Urdu despite not being an official script, plays an important role in communication and is widely popular on social media platforms (Bilal et al., 2017). Unlike the standard script, the romanized version lacks uniform orthography and contains discrepancies in particular for vowel sounds (Tafseer, 2009).
Despite recent interest in understanding the behavior and characteristics of Roman Urdu (Baseer et al., 2016; Irvine et al., 2012), there is a lack of infrastructure for exploring its relationship to English and Urdu. Furthermore, to the best of our knowledge no parallel lexicons exist for these languages. To address this limitation, the primary focus of this study was to employ automatic approaches for building a trilingual dictionary to capture lexical relationships between the three languages. The dictionary is available at https://github.com/MoizRauf/Urdu--Roman-Urdu--English--Dictionary.

2 Dictionary Extraction

Brown et al. (1990); Gale and Church (1991); Melamed (1998); Otero (2007) showed that by using statistical alignment methods, accurate translation pairs can be captured from parallel datasets without the need for additional bilingual lexicons. To offset the lack of parallel resources for language pairs, (Tanaka and Umemura, 1994; Varga et al., 2009; Tsunakawa et al., 2013) showed the potential of using an intermediate *pivot* language dictionary for looking up words and creating parallel lexicons. Our experiments to build this joint lexicon employed both parallel and pivot based approaches. Table 1s provide details on individual datasets for the two language pair (En–Ur and Ur–Rom).

Ur–Rom. To capture the colloquial nature of Roman Urdu and its relation to Urdu, we used the annotated SMS text parallel corpus developed by Irvine et al. (2012). The dataset contains 4,195 SMS messages that

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/.

Corpus	Language Pair	Source	SentPairs	Aligned Word Pairs
Jawaid and Zeman (2011)	En–Ur	Web	7957	10223
Post et al. (2012)	En–Ur	Wikipedia	33798	27025
Irvine et al. (2012)	Ur–Rom	SMS	4195	8321
Trilingual Lexicon	Rom–Ur–En	NA	NA	5916

Table 1: Statistics of parallel corpora and resulting lexicon.

Language Pair	Correct (5)	Somewhat Correct (4)	Undecided (3)	Somewhat Incorrect (2)	Wrong (1)
En–Ur	42%	7%	4%	4%	43%
Ur–Rom	84%	4%	4%	3%	5%

Table 2: Overall distribution of sample translations on 5-point quality scale.

were normalized and converted to Urdu by MTurk annotators.
To generate transliteration pairs between the original and romanized language, we used a *sub-string transducer* (Sherif and Kondrak, 2007). Almost all alignment pairs were one-to-one and monotonic. This statistical transliteration system enabled us to capture loan/foreign words and named entities present in the text. We also reduced spelling inconsistencies by grouping all similar roman variations through a *lemmatizer* (Sharf and Rahman, 2017). We then only considered the lemma with the highest frequency as base representation and replaced all occurrences of the words in that group by the canonical representation. Such a normalization phase was shown to reduce noise and improve performance of various classification tasks by (Sohail et al., 2018; Singh et al., 2018). Finally, each unique pair was automatically assigned a confidence score based on transliteration rules listed in Tafseer (2009).

En–Ur. In order to obtain word pairs for En–Ur we used various pre-existing parallel corpora (Jawaid and Zeman, 2011; Post et al., 2012). Both datasets vary in domain, genre and size. We used the GIZA++ (Och and Ney, 2003) alignment system to extract 10223 and 27025 En–Ur translation pairs from both datasets respectively. Additionally, we manually curated 500 seed pairs to bootstrap an iterative self-learning system *VecMap* (Artetxe et al., 2017), that exploits structural similarities of embedding spaces. The method mapped Urdu and English embeddings (Grave et al., 2018) and induced translations using Cross-domain Similarity Local Scaling (CSLS) (Conneau et al., 2018).

Trilingual (Rom–Ur–En). To create a unified lexicon, we used *Urdu* as the pivot language. We extracted English translations for every Ur–Rom pair from *VecMap* and aligned En–Ur sentences. We finally considered the candidate with higher cosine similarity. The resulting triplets also included borrowed words and named entities.

3 Evaluation & Results

To evaluate the validity of our dictionary, we employed a manual evaluation approach similar to that of (Sjöbergh, 2005; Charitakis, 2007). Three native Urdu speakers fluent in English evaluated 1000 randomly selected parallel pairs. The evaluation was carried out using a five-point Likert scale for correctness.

Analysis of Correctness Ratings We measured agreement between our annotators by computing average pairwise Spearman's correlation. We obtained agreements of 0.66 (En–Ur) and 0.75 (Ur–Rom), which are statistically significant and comparable to those observed by Zhao et al. (2003); Lew and Szarowska (2017). We report the union of annotator's judgment for En–Ur pairs and Ur–Rom pairs, respectively (cf. Table 2). Our translations follow a bi-modal distribution, with almost all judgments either Correct (5) or Wrong (1). The results show that our evaluators judged 88% of our Ur–Rom transliteration pairs as correct or nearly so, while En–Ur pairs were considered correct or nearly so for 49% of the cases.

4 Error Analysis

Based of our empirical evaluation, we can conclude that our method was able to identify a substantial number of Ur–Rom pairs, while our En–Ur translations were considerably more incorrect. To better

Word Class	Count	Example
Noun	78	ملکہ (queen) – majesty, مدار(pivot) – earth's
Verb	28	کرنا(To do) – choosing, نکلتے(to leave) – absorb
Loan terms	32	موٹر (motor) – ramp, کنگ(king) – burger
Named Entities	23	میمن (memon) – mujeeb, چاغی(Chaghi) – Paktika
Incomplete Urdu Lemma	23	ستان (-stan) – vetitum, مو(mo) – choreography
Adj & Adv	16	فرشتوں (angels) – earthly, ٹھوکر(stumble) – stuporous
Total	200	

Table 3: Distribution of error types for wrong En–Ur translations

understand the behavior of our translation candidates, we manually performed an error analysis.

An initial inspection revealed that the informal nature of Roman Urdu terms was the cause for most low scoring Ur–Rom transliterations. Case such as (saadgimeri → saadgi, meri) presented ambiguous compound terms, while (chaghi1islamabd → chaghi, islamabd) was an example of segmentation issue which were partially captured by our system.

To better understand the true and false positives in our En–Ur translations, we sampled 200 instances marked *incorrect* by our annotators. We grouped the terms based on *part-of-speech (POS) tags* of the English terms (cf. Table 3). We observed that *Noun* was the predominant incorrect class, followed by instances where the Urdu lemmas were incomplete. While, in 16% of translations Urdu terms were loaned/borrowed English words (e.g. motor → ramp, lady → liberty, watt → kw etc). Furthermore, our approach has similar behavior as that of Artetxe et al. (2017) for named entities, the model selected related location based translations (e.g. Chaghi → Paktika, Qasur → Peshawar) for source terms.
Majority of all proposed translations have had some semantic relatedness with the source term (e.g. labor → capitalist, king → burger). We further assessed the prevalent semantic relationships of our sampled error translations. Similar to Peirsman and Padó (2008), we classified our translation pairs into semantic categories (cf. Table 4). In the majority of cases (65%), there is still some relationship in the translation pair, such as antonymy (e.g. un-aware → knowledgeable, change → unchanged), co-hyponymy (e.g. rice → vegetables, red → yellow), or least unspecific, often syntagmatic, relatedness (e.g. pain → through, reward → reaping). However, a substantial portion of pairs (35%) remained for which we could not find a reasonable association.

5 Conclusion

In this work, we present a joint Urdu, Roman Urdu and English lexicon building approach. Our results and error analysis have revealed that our method encapsulates significant information to capture bilingual semantic relationships and that it is a concrete bootstrapping lexicon which can be built upon and has utility in various linguistic tasks. In the future, we would like to employ alternative strategies that could complement our En–Ur translations and add additional information to enrich our dataset. Additionally, we would like to explore this lexicon in more practical settings.

Relation	Count	Example	Meta-Relation
antonym	6	تبدیل(change) – unchanged	taxonomic similarity
near-synonym	9	ارضی (terrestrial) – geomorphic	
co-hyponym	10	پاؤں (foot) – shoulder	
related	103	مخالفت (opposition) – vehemently	relatedness
unrelated	72	اکیلی (alone) – shown	error
Total	200		

Table 4: Distribution of semantic classes for wrong En–Ur translations

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 451–462, Vancouver, BC.

F. Baseer, A. Habib, and J. Ashraf. 2016. Romanized Urdu corpus development (RUCD) model: Edit-distance based most frequent unique unigram extraction approach using real-time interactive dataset. In *2016 Sixth International Conference on Innovative Computing Technology (INTECH)*, pages 513–518.

Anas Bilal, Aimal Rextin, Ahmad Kakakhel, and Mehwish Nasim. 2017. Roman-txt: Forms and functions of Roman Urdu texting. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '17, pages 15:1–15:9, New York, NY, USA. ACM.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Comput. Linguist.*, 16(2):79–85.

Konstantinos Charitakis. 2007. Using parallel corpora to create a Greek-English dictionary with Uplug. In *Proceedings of the 16th Nordic Conference of Computational Linguistics*, pages 212–215, Tartu, Estonia.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of ICLR 2018*, Vancouver, BC.

William A. Gale and Kenneth W. Church. 1991. Identifying word correspondence in parallel texts. In *Proceedings of the Workshop on Speech and Natural Language*, pages 152–157, Pacific Grove, BA.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation*, Miyazaki, Japan.

Ann Irvine, Jonathan Weese, and Chris Callison-Burch. 2012. Processing informal, romanized Pakistani text messages. In *Proceedings of the NAACL Workshop on Language in Social Media*, Montreal, Canada.

Bushra Jawaid and Daniel Zeman. 2011. Word-order issues in English-to-Urdu statistical machine translation. *The Prague Bulletin of Mathematical Linguistics*, 95(1):87–106.

Khang Nhut Lam, Feras Al Tarouti, and Jugal Kalita. 2015. Automatically creating a large number of new bilingual dictionaries. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2174–2180, Austin, Texas.

Robert Lew and Agnieszka Szarowska. 2017. Evaluating online bilingual dictionaries: The case of popular free English-Polish dictionaries. *ReCALL*, 29(2):138–159.

I. Dan Melamed. 1998. A word-to-word model of translational equivalence. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 490–497, Madrid, Spain.

Zobina Muhammad Asghar and Muhammad Asim Mahmood. 2013. Urdu in anglicized world: A corpus based study. *International Journal of English and Literature*, 4:134–140.

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

Pablo Gamallo Otero. 2007. Learning bilingual lexicons from comparable English and Spanish corpora. In *Proceedings of MT Summit XI*, pages 191–198.

Yves Peirsman and Sebastian Padó. 2008. Semantic relations in bilingual lexicons. *ACM Trans. Speech Lang. Process.*, 8(2):3:1–3:21.

Matt Post, Chris Callison-Burch, and Miles Osborne. 2012. Constructing parallel corpora for six Indian languages via crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 401–409, Montréal, Canada.

Muhammad Shaban Rafi. 2013. Urdu and English contact in an e-discourse: Changes and implications. *Gomal university journal of research*, 2:78–86.

Sarwet Rasul. 2013. Borrowing and code-mixing in Pakistani children's magazines: practices and fuctions. *Pakistaaniat: a journal of Pakistan studies*, pages 46–72.

Zareen Sharf and Saif Ur Rahman. 2017. Lexical normalization of Roman Urdu text. *International Journal of Computer Science and Network Security*, 17(12):213–221.

Tarek Sherif and Grzegorz Kondrak. 2007. Substring-based transliteration. In *Proceedings of ACL*, pages 944–951, Prague, Czech Republic.

Rajat Singh, Nurendra Choudhary, and Manish Shrivastava. 2018. Automatic normalization of word variations in code-mixed social media text. *CoRR*, abs/1804.00804.

Jonas Sjöbergh. 2005. Creating a free digital Japanese-Swedish lexicon. In *Proceedings of PACLING*, pages 296–300, Tokyo, Japan.

Omayya Sohail, Inam Elahi, Ahsan Ijaz, Asim Karim, and Faisal Kamiran. 2018. Text classification in an under-resourced language via lexical normalization and feature pooling. In *22nd Pacific Asia Conference on Information Systems, PACIS 2018*, page 96.

Ahmed Tafseer. 2009. Roman to Urdu transliteration using wordlist. In *Proceedings of the Conference on Language and Technology*, pages 305–309, Lahore, Pakistan.

Kumiko Tanaka and Kyoji Umemura. 1994. Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of the 15th Conference on Computational Linguistics*, pages 297–303, Kyoto, Japan.

Takashi Tsunakawa, Yosuke Yamamoto, and Hiroyuki Kaji. 2013. Improving calculation of contextual similarity for constructing a bilingual dictionary via a third language. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1057–1061, Nagoya, Japan.

István Varga, Shoichi Yokoyama, and Chikara Hashimoto. 2009. Dictionary generation for less-frequent language pairs using WordNet. *Literary and Linguistic Computing*, 24(4):449–466.

Bing Zhao, Klaus Zechner, Stephan Vogel, and Alex Waibel. 2003. Efficient optimization for bilingual sentence alignment based on linear regression. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 81–87.