Journal of Autonomous Intelligence (2020) Volume 3 Issue 2.
doi:10.32629/jai.v3i2.279

Original Article

Research on Chinese-Urdu Machine Translation Based on Deep Learning

Zeeshan[1]*, Zeshan Ali[1], Jawad[1], Muhammad Zakira[1]

School of Information Science and Engineering, Xinjiang University Urumqi, Xinjiang, China

ABSTRACT

Urdu is Pakistan 's national language. However, Chinese expertise is very negligible in Pakistan and the Asian nations. Yet fewer research has been undertaken in the area of computer translation on Chinese to Urdu. In order to solve the above problems, we designed of an electronic dictionary for Chinese-Urdu, and studied the sentence-level machine translation technology which is based on deep learning. The Design of an electronic dictionary Chinese-Urdu machine translation system we collected and constructed an electronic dictionary containing 24000 entries from Chinese to Urdu. For Sentence we used English as an intermediate language, and based on the existing parallel corpus of Chinese to English and English to Urdu, we constructed a bilingual parallel corpus containing 66000 sentences from Chinese to Urdu. The Corpus has trained by using two NMT Models (LSTM,Transformer Model) and the above two translation model were compared to the desired translation, with the help of bilingual valuation understudy (BLEU) score. On NMT, The LSTM Model is gain of 0.067 to 0.41 in BLEU score while on Transformer model, there is gain of 0.077 to 0.52 in BLEU which is better than from LSTM Model score. Furthermore, we compared the proposed model with Google and Microsoft translation.

Keywords: *Chinese; Urdu; Neural Machine Translation; Deep Learning; Bilingual Electronic Dictionary*

ARTICLE INFO
Received: Mar 21, 2021
Accepted: Sep 15, 2021
Available online: Sep 18, 2021

***CORRESPONDING AUTHOR**
Zeeshan, School of Information Science and Engineering, Xinjiang University Urumqi, Xinjiang, China;
zeshanali531@gmail.com;

CITATION
Zeeshan, Zeshan Ali, Jawad, Muhammad Zakira. Research on Chinese-Urdu machine translation based on deep learning. Journal of Autonomous Intelligence 2020; 3(2): 34-44. doi: 10.32629/jai.v3i2.279

1. Background and Motivation

Urdu is a derivational word from Turkish and it's recommend "crowd" (Lashkar-لشکر). Urdu, an Indo-European language of the Indo Aryan hover of family members, is spoken in India and Pakistan. Urdu is the national language of Pakistan which is spoken by in excess of 100 million of locals in Pakistan in various areas and for various purposes. Next to Pakistan Urdu language is spoken in India, Afghanistan and Middle East. Among all the dialects in the worldwide its miles most eagerly simply like Hindi language. Urdu and Hindi each have begun from the tongue of Delhi place and other than the moment data these dialects rate their morphology. Since Hindi has embraced numerous words from Sanskrit, Urdu has acquired an immense number of jargon objects from Persian and Arabic. Urdu is similarly getting wide assortment of jargon from Turkish, Portuguese and English. There are quite various expressions that have watched a spot in Urdu Language, routinely through the Persian Language, have in any case nuanced meanings and uses. One of the significant components of Urdu language structure constitution is its expression request SOV (trouble, thing, and action word). This request flaunts a couple of adaptability as the difficult pronouns are regularly dropped. Things in Urdu Language syntax have two types of gender orientation (singular/plural) and 3 cases (vocative, direct and oblique). All things in Urdu, when utilized inside a sentence, may be bent for assortment

and case. Addition indicates sexual orientation on action words and descriptors (for example Paagal → paagalpan,"insanity", ghabraaanaa→ghabraahat, "strain") and not bizarre postfixes might be utilized to control things from different words. These structures are manly and female things .One uncommon point is that the acquired Arabic and Persian plurals type of the thing are never arched in Urdu, .Verb understanding is appeared with trouble or with direct article at a specific model. Action words in Urdu language have two nonfinite structures root and infinitive. The infinitives contain a verbal stem and a postfix. The stem may furthermore itself contain verbal root and addition. (for example Aanaa "to move",jaanaa "to move") .For this situation/aa-/and/jaa-/are the premise and/naa-/is addition. "Syntactic factor of view, the infinitive styles of all action words are stamped manly. In Urdu Subjunctive is the limited verbal structure which passes on the week guesses at the Urdu a piece of speaker. Another two verbal structures that might be limited or nonfinite are the perfective and imperfective participles. The imperfective molecule closes with-taa, -tee, -tii, -tiin. In instance of perfective molecule closes with e.g -aa, -ee, -ii, -iin. But the circumstance is novel each time any verbal stem that prompts a vowel, we should transfer a/y/sooner than manly particular completion. The fate action word structures in Urdu can't be chosen from the action word stems as a substitute it is chosen from subjunctive structures. The endings for the future structures are (for example-gaa, -hmm, -gii). The strained and segment are frequently demonstrated through the utilization of unpredictable assistant action words. Basic utilization of semi assistant components moreover offers different semantic implications.

1.1 The statement of the problems

Although a whole lot of studies and observe has been accomplished about the history and idea of linguistics for Urdu language. However in Pakistan in other Asian Countries there is a huge difference between an official language and what people actually speak and understand. Although Chinese is not the first or Second official language, many (especially working class) people have a very limited Chinese level and barely speak a few sentences with terrible pronunciation .In fact in Pakistan and the above mentioned countries insight regarding Chinese language is quite nominal. All in all, there are two problems in Chinese-Urdu machine learning.

Problem 1: There have been less work in Chinese to Urdu machine translation field. As now nowadays machine translation had made a great progress in translating different languages from sources to target but unfortunately Urdu get less attention as compared to other international languages there is less work done on Chines to Urdu and need more attention to improve the communication between two different languages speakers.

Problem 2: Local language translation software for other foreign Languages is limited. The number of domestic translation software for native and some national languages is limited in the present with the help of deep learning and NMT has gained incredible performance. Such an objective cannot guarantee is the sufficiency of the generated translation in MNT model.

2. Related Work

Truth be told in Pakistan and the previously mentioned nations understanding with respect to the Chinese language is very ostensible. Till date, the majority of the analyst works and endeavors made in the territory of slant examination manage English and Urdu text[1-7]. This is because of the way that mining and assessment of assumptions from text need a high commitment of lexical assets of that language. All however, in contrast to English, Urdu is a Low asset language, and thus, the production of semantic vocabulary for Urdu text is a significant and testing task[1]. The cutting-edge investigation, the presentation report of pattern frameworks is interpreting Indian dialects based content (Bengali, Hindi, Malayalam, Punjabi, Tamil, Telugu, Gujarati, and Urdu) into English content with a normal of 10% Rightness for all language pairs[8].In 2013 Kalchbrenner[9] proposed repetitive, constant interpretation models for machine interpretation. This model uses a convolutional neural system (CNN) to encode a given piece of information text into a whole vector and afterward utilizes a recurrentpetitive neural system (RNN) as a decoder to change over the vector into yield language. In 2014[10], long short term memory (LSTM) was brought into NMT. To take care of the issue

of producing fixed-length vectors for encoders, they bring consideration instrument into NMT[11]. The consideration component permits the neural system to pay more thought to the important pieces of the info, and dispose of random parts. From that point forward, the exhibition of the neural machine interpretation technique has been fundamentally improved. In this Sutskever a multilayer LSTM is utilized to encode the input sentence into a fixed-size Heading and afterward unravel it into yield by another LSTM. The utilization of LSTM Productively settled the issue of inclination evaporating, which concurs with the model to catch information over broadened space in a sentence.

Muhammad Bilal[12] utilize the three-characterization models are utilized for text arrangement utilizing Waikato Condition for Information Examination (WEKA). Opinions written in Roman Urdu and English for the blog. These suppositions are archives which is utilized for preparing dataset, marked models, and messaging information. Because of testing these three distinct models and the outcomes for each situation are examined. The outcomes show that Credulous Bayesian beat Choice Tree and KNN as far as more exactness, accuracy, review, and Measure. Mehreen Alam[13] address this troublesome and convert Roman-Urdu to Urdu literal interpretation into an arrangement to succession learning trouble. The Urdu corpus was made and pass it to neural machine interpretation that speculation sentences up to length 10 while accomplishing great BLEU score. Neelam Mukhtar depicts Urdu language is helpless dialects, for example, Urdu is generally disregarded by the examination network. In the wake of gathering information from numerous web journals of around 14 distinct kinds, the information is being noted with the assistance of human annotators. Three notable AI calculation Backing Vector Machine, Choice tree and k-Closest Neighbor (k-NN) which is utilized for test, correlation. Its show that k-NN execution is better than helping Vector Machine and Choice tree regarding exactness, accuracy, review, and f-measure[14].

Muhammad Usman[15] additionally portray five notable arrangement strategies on Urdu language corpus. The corpus contains 21769 news records of seven classes (Business, Amusement, Culture, Wellbeing, Sports, and Peculiar). In the wake of preprocessing 93400 highlights are taking out from the information to apply AI calculations up to 94% accuracy. Yang and Dahl their work, first word prepared with an immense mono-lingual corpus, at that point the word installing is adjusted with bilingually in a setting depended DNN Gee system. Word catching lexical interpretation data and demonstrating setting data for improve the word arrangement execution. Tragically, the better word arrangement results produced yet can't give huge execution a start to finish SMT assessment task[16]. Auli[17] improved the neural system language model, so as to utilize both the source and target-side data. In their work, not just the objective word installing is utilized as the contribution of the system, yet additionally the current objective word. Liu[18] propose an improved neural system for SMT translating. Mikolov[19] is right off the bat used to create the source and target word embedding, which take a shot at one covered up layer neural system to get an interpretation certainty score. Our work will be demonstrated an electronic instrument for simplicity to the business class and locals of Pakistan and other neighbor nations to comprehend the Chinese language. It gives a major bit of leeway to those individuals who are working together China in Pakistan or other piece of Asia on the grounds that through this correspondence is simple. Because of this electronic device the Chinese Urdu Dialects Local can comprehend there Correspondence and just as in Culture.

3. Chinese to Urdu Dictionary—C2U

Data resources are very rare for low-resource language pairs such as Urdu and there is 2 Dataset for the Urdu:

Monolingual Corpus: the Urdu monolingual corpus of around 95.4 million tokens distributed in around in different Website the monolingual corpus is a mix of domains such as News, Blogs, Literature, Science, Education[1] etc. Only words monolingual data is used to build our C2U Model[20].

IPC: The Indic Parallel Corpus is a collection of Wikipedia documents of six Indian sub-continent languages translated into English through crowd sourcing in the Amazon Mechanical Turk (MTurk) plat-

form[21].

However the above two Dataset is just only for Monolingual Language therefor it's not enough to satisfy user desire, so we have created our two datasets- **ETCLW**, **UTCLW**.The ETCLW Dataset Data is taken from learning Chinese language website[2] but UTCLW dataset is developed manually because there is no freely data available. For UTCLW we used a middle language English like (Urdu English Chinese)[3] and with the help of English we manually translate English word to Chinese. In each dataset consists of 24000 words and its description which is collected from five different sources, give the reference[22]. This sector highlights the present dataset resources that can be applied for developing C2U Dictionary. The number of official test sets are also exhibited. The words and its description collected from several domains such as News, daily life, Technology, Language and Culture etc.

ETCLW: ETCLW (English to Chinese Language words) is a collection of words of fourteen South Asian languages which is distributed by the European Language Resources Association. The English part is documents produced by the British Departments of Health, Social Services, Education and Skills, and Transport, Local Government and the Regions of British government which we used is translated into Chinese.

Table 1. C2U CTCLW Corpus Dataset

Words	Nouns	Verbs	Particles	Punctuation
24000	12985	5700	4878	1892

UTCLW: UTCLW (Urdu to Chinese language Words) is a collection of words which is handwritten, because it is not openly available for UTCLW dataset 10k word selected from monolingual corpus dataset, 4k words from IPC and the 10k words collected from daily use Urdupot[4] internet from collected from different data source.(**Table 2**)

废弃的车辆/چھوڑ دیا گاڑی , 非生物因素/غیر متوقع عنصر, 通路/رسائی سڑک, 进入大海/سمندر تک رسائی , 事故/حادثہ,累加器/جمع کرنے والا , 酸化/امیڈریشن, 酸度/املتا, 酸度/تیزاب کی ڈگری , 酸雨/تیزابی بارش , 声学滤波器/صوتی فلٹر , 教条主义/ کٹر مذہبی, 教条/عقیدہ/مذہب/قول/اصول عقائد, 沮丧/اچاٹ/سر نگون/بجھا ہوا/جھکا ہوا , 倒台/زوال/ /اتار/ڈھال ڈھلوان/نشیب/下坡,نقشی/نقشین/سجا/سجی/点缀, بربادی/تباہی/تنزل/زوال , 挪用/پچانا/کرنا برد خرد/کھانا/ہضم کرنا/ڈکار جانا/چرانا ,联邦/وفاقی

3.1 Chinese to Urdu parallel corpus

Using English as an intermediate language, and based on the existing parallel corpus of Chinese to English and English to Urdu. Our dataset consists of 66000 Chinese-Urdu parallel corpus which is come from the combination of all the below datasets which are define below.

HC Corpora: The HC Corpus[5] was a great resource that contains natural language text from various newspapers, social media posts and blog pages in multiple languages. This is a cleaned version of the raw data from newspaper subset of the HC corpus. The corpus contains 16,806,041 sentences/paragraphs in 67 languages. we search a similar article in Chinese as well as in Urdu and taken 26000 parallel carpus for our project.

WiLI-Dataset: WiLI-2018[6], the Wikipedia language identification benchmark dataset, contains 235000 paragraphs of 235 languages. Each language in this dataset contains 1000 rows/paragraphs. After same data selection and preprocessing we select same Chinese Urdu 45 paragraph with the help of middle language English.

The 45 paragraphs have 25000 parallel sentences which give a good contribution in our work.

UNHD: Urdu Nastaliq Handwritten Dataset. In addition, dataset contains 15000 sentences written manually and Urdu Part is derived from **UNHD**[7](Urdu Nastaliq Handwritten Dataset) in which are showing below **Table 3**.

Table 3. Urdu Nastaliq Handwritten Dataset

该榜单上已经包括圣彼得堡艺术界的杰出人物，他们的成就超越了通常在欧洲公认的城市范围，绕过了俄罗斯的名声。 罗斯巴特的新球员-大胆的艺术家基里尔米勒。 整个城市都知道基里尔•米勒（基里尔·米勒），一个大胡子的穿着红色衣服，可以在俄罗斯博物馆，夏季花园或时髦的聚会和表演中看到。 无论展览在哪里，基里尔•米勒（基里尔·米勒）的作品总是吸引很多人。 基里尔·米勒（基里尔·米勒）是纯粹的圣彼得堡社会和哲学讲故事者之一，也是新神话的创造者。 基里尔·米勒（基里尔·米勒）是 80 年代末 90 年代初圣彼得堡前卫的杰出人物。 此外，他是一个城市人，让人在街上微笑，振奋人心。 最近，他接管街头风琴并成为圣彼得堡的音乐人，因为他已经准备好担负起他在波西米亚时代的所有存在，哲学和形象的复杂角色。	اس فہرست میں پہلے ہی سینٹ پیٹرزبرگ کے فن منظر کی نمایاں شخصیات شامل ہیں ، جن کی کامیابیاں شہر کے دائرے سے باہر تک پہنچ جاتی ہیں ، جو اکثر یوروپ میں پہچانا جاتا ہے ، روس میں شہرت کو نظرانداز کرتے ہوئے۔ روزبالٹ میں نیا کھلاڑی - بولڈ آرٹسٹ کریل ملر۔ پورا شہر کریل ملر کو جانتا ہے ، داڑھی والا ایک شخص سرخ رنگ کا ملبوس ہے ، جسے روسی میوزیم ، یا سمر گارڈن ، یا فیشن پارٹیوں اور شوز میں دیکھ سکتا ہے۔ کریل ملر کا کام لوگوں کے ہجوم میں ہمیشہ آتا ہے ، خواہ اس کی نمائش کہیں بھی نہ ہو۔ کریل ملر خالصتا St. سینٹ پیٹرزبرگ کے سماجی اور فلسفیانہ کہانی سنانے والوں اور نئے افسانوں کے تخلیق کاروں میں سے ایک ہے۔ کریل ملر سینٹ پیٹرزبرگ کے 80 کی دہائی کے اوائل میں 90 کے اوائل میں سینٹ پیٹرزبرگ کے بدقسمت آدمی تھے۔ مزید یہ کہ وہ شہر کا آدمی ہے ، جو لوگوں کو سڑک پر مسکراہٹ دیتا ہے اور ہر ایک کی روح بلند کرتا ہے۔ حال ہی میں اس نے اسٹریٹ آرگنائز لیا اور سینٹ پیٹرزبرگ کا میوزک مین بن گیا ، کیوں کہ وہ اپنے تمام بوہیمین وجود ، فلسفہ اور شبیہہ کے ساتھ اس پیچیدہ کردار کے لئے تیار تھا۔ - کیرل ، کیوں آپ اس شہر کے چاروں طرف سرخ رنگ میں چہل قدمی کرتے ہیں ، مثال کے طور پر پیلے رنگ یا فیروزی نہیں؟

CTUS: We created a third corpus that is in fact a merge of the above described corpora. The idea is to test whether more sentences from different domains increase or decrease the efficiency of a translation model. We named this new corpus CTUS Chinese to Urdu Sentences.

4. Proposed work for C2U Dictionary

In our work of Chinese to Urdu Word base Dictionary machine translation, we used a neural network and Neural Machine Translation approach. Neural Machine Translation approach is the classical approach of machine translation. In Neural Machine Translation Approach, system is fed with parallel Corpus through that parallel Corpus machine generate output model. Many systems have been developed using Neural Machine Translation, in which main systems are as Systran, Eurotra and Japanese MT System. Neural networks are a possible solution to the machine translation problem. Neural networks have the ability of learning by examples. Our Urdu Word base Dictionary machine translation system uses Neural Network with Neural Machine Translation approach. Neural networks are very efficient in pattern matching. Our C2U Model uses Neural Machine Translation approach.

4.1 C2U system architecture and description

The block diagram of our Chinese to Urdu Word base Dictionary Machine Translation System is shown in figure (to see **Figure 1**). There are Parallel Words and its description dataset, training data and output

model. The C2U model for training we used 24000 words dataset in which 70% for training, 20% for validation and 10% testing. we used open source code of openNMT8 with help of Python to train our model.

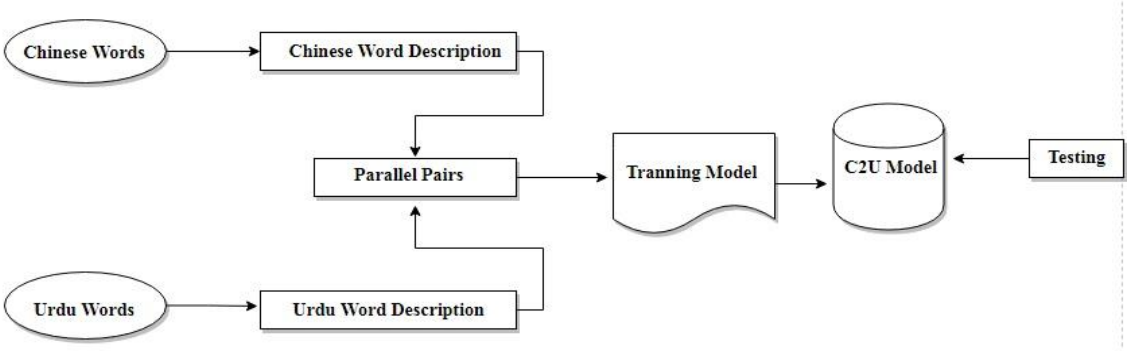

Figure 1. C2U Model

4.1 C2U system architecture and description

Table 4. C2U Dictionary Example

ترک کردہ صنعتی سائٹ 废弃的工业现场/Fèiqì de gōngyè xiànchǎng 无法被任何目的使用,被污染物污染的场所 Wúfǎ bèi rènhé mùdì shǐyòng, bèi wūrǎn wù wūrǎn de chǎngsuǒ
غیر متوقع عنصر 非生物因素/非生物因素 物理，化学和其他非生命环境因素 Wùlǐ, huàxué hé qítā fēi shēngmìng huánjìng yīnsù
کتاب 书,shu 一组纸张绑在一起铰接在一个边缘，包含印刷或书面材料，图片等。 Yī zǔ zhǐzhāng bǎng zài yīqǐ jiǎojiē zài yīgè biānyuán, bāohán yìnshuā huò shūmiàn cáiliào, túpiàn děng.
بک مارک 簿记/Bùjì 记录商业帐户和交易的艺术或科学。 Jìlù shāngyè zhànghù hé jiāoyì de yìshù huò kēxué.

5. Training Model for Chinese to Urdu Machine Translation with LSTM

The amount of parallel corpus and it's nice plays significant role in excellent of translation output. For low useful resource languages like Urdu, it's miles extraordinarily difficult to discover enough parallel corpus for schooling, validation and trying out of translation engine. For our experiment, we used the above datasets corpus. The corpus consists of 66k sentence pairs dividing into three categories training, validation and testing. For data training we are taking default OpenNMT encoder and decoder, LSTM layers, and RNN. We start our research basing on open-source codebase[23].This codebase is written in Python, using pytorch, an open-source software library[9]. For NMT, we use a LSTM Model and Transformer Model with outstanding network connections as well as good mechanism to train a translation model. The stats of corpus are given in below Tables: Validation-Data, Training-Data, and Test-Data of Chinese to Urdu Sentence Translation.

Training-Data

We Filtered parallel Corpus data with the help of open source code[10] for giving better result in translation. we take 66k Parallel corpus dataset for training which is describe in below.

Table 5. Training-Data

	Chinese	**Urdu**
Sentences	66000	
Sentence length	35	42
Min, Max words in sentence	450	520
Total words	125305	185305
Unique words	95005	98300

Validation-Data

Validation files are required and used to evaluate the convergence of the training. it usually contains no more than 5000 sentences. We used 20% of training data for validation which is showing in below.

Table 6. Validation-Data

	Chinese	**Urdu**
Sentences	13200	
Sentence length	35	42
Min, Max words in sentence	350	370
Total words	43777	48530
Unique words	2960	3043

Test-Data

We select 10% of data for testing for our Chinese to Urdu Translation model which data is showing below.

Table 7. Test-Data of Chinese to Urdu Sentence Translation

	Chinese	**Urdu**
Sentences	6600	
Sentence length	30	32
Min, Max words in sentence	150	180
Total words	21797	27305
Unique words	1454	1682

5.1 Training transformer model for Chinese-Urdu machine translation

At this steps, we trained another model of neural networks. The efficiency of translation module is also tested to know about the impression across other Model. The core idea behind transformer model is self-attention. LSTMs have some issues parallelization, long and short range dependencies and distance between positions is linear. To solve these problems, attention mechanism is introduced in neural networks. Each word has hidden state which is passed along the way while translating the sentence instead of decoding whole sentence in a single hidden unit/layer. To solve problem of parallelization, transformers used convolutional neural networks (CNN) along with attention mechanism. The mathematical form attention is given as:

$$Attention(Q,K,V) = softmax\left(QK^T\middle|\sqrt{d_k}\right)V \quad (1)$$

Transformer uses attention mechanism in three ways[24]:

In "encoder-decoder attention" layers, the queries come from the previous decoder layer, and the memory keys and values come from the output of the encoder. This allows every position in the decoder to attend over all positions in the input sequence. This mimics the typical encoder-decoder attention mechanisms in sequence-to-sequence models.

The encoder contains self-attention layers. In a self-attention layer all of the keys, values and queries come from the same place, in this case, the output of the previous layer in the encoder. Each position in the encoder can attend to all positions in the previous layer of the encoder. Similarly, self-attention layers in the decoder allow each position in the decoder to attend to all positions in the decoder up to and including that position. We need to prevent leftward information flow in the decoder to preserve the auto-regressive property. We implement this inside of scaled dot-product attention by masking out (setting to $-\infty$) all values in the input of the softmax which correspond to illegal connections.

We implemented above models of NMT using OpenNMT-py[11] toolkit. For LSTM, we used default settings while for transformer model followings were parameter settings. Selection of these parameters helped researchers to mimic behavior of Google translator as reported in WNMT-18[25]. To run these experiments, GPU Tesla k40 is used with 20 gb ram and graphic card of 12 gb. For LSTM model, it took around 20-24 hours to train model and for transformer model it took around 70 hours due to large number of hidden layers and rnn size.

Table 8. NMT parameter selection

layers = 4	batch_size = 4096	learning_rate = 2	max_grad_norm = 0
rnn_size = 512	batch_type = tokens	label_smoothing = 0.1	param_init = 0
word_vec_size= 512	dropout = 0.1	encoder_type=transformer	param_init_glorot

6. C2U Result Discussion and Accuracy Graph

The results of the evaluation are shown in **Figure 1** and **Figure 3**, the input text for which multi words as well as and the description of that words matches could be found in Model. We have done different test on different dataset size with help of Accuracy Formula[12]. We done 3 test on different data size of words first we test on 5k words dataset which give us 0.23 blue score Accuracy result then we done our second test over 15k words dataset which give us 0.56 blue score Accuracy then we done our third test over 24k words dataset which give us 0.68 blue score accuracy because some of words having in dataset same translation or occasionally have not present in dataset. This value is the maximum coverage that C2U Dictionary could achieve, given perfect alignment for all matches. Sometimes it does not give us the perfect result just because the open dataset is created by us so the alignment rarely fails. The C2U Dictionary gives us perfect result in accordance to the given words in different languages. See **Figure 2**

$$P = \frac{M}{Wt} \ldots\ldots\ldots\ldots\ldots\ldots 1 \quad (2)$$

Where P is a Precision, is number of words from the candidate that are found in the reference, and Wt is the total number of words in the candidate. So M is divide with Total size of words and then give us Precision which we called Accuracy.

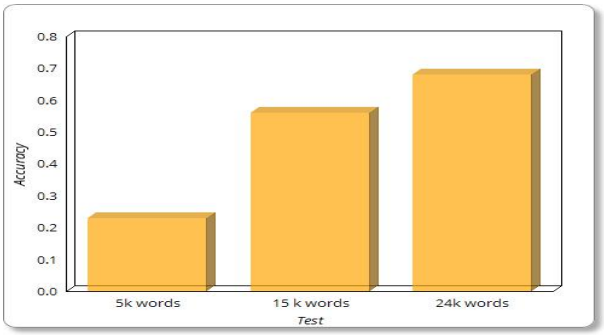

Figure 2. C2U Accuracy Graph

6.1 Results of Chinese Urdu sentence to sentence translation

Our transliteration technique improved baseline score up to 0.067 to 0.52 in terms of BLEU score. We have applied our technique in Neural Machine Translation two models to show the effectiveness of the technique.

Table 8. NMT Results

Evaluation Measures	LSTM Model	Transformer Model
BLEU Score	0.41	0.52
Precision	0.73	0.83
Recall	0.53	0.57
F1	0.61	0.69

6.2 Comparison of Chinese Urdu translation model to other translation model

We also Compare our two NMT Translation Model with Google translation System and also with Microsoft Translation System. For Comparison we select our high Blue Score Sentence and its clear showing that our system is giving good translation. Which is describe in **Table 9**.

Table 9. Comparison of NMT Model

	Sentences	Bleu Score
Source	必须将希腊人的这一问题视为这种追求的一种形式。 اس مسئلے کو یونانیوں افراد سے اس طرح کے تعاقب کی شکل میں رابطہ کرنا بہت ضروری ہے۔	
LSTM Model	یہ یونانی افراد کے ساتھ اس طرح کے حصول کے لئے ایک فارم کے طور پر اس مسئلے پر غور کرنا ضروری ہے.	0.41
Transformer Model	یہ یونانیوں افراد کے ساتھ اس طرح کے حصول کے لئے ایک فارم کے طور پر اس مسئلے پر رابطہ کرنا بہت ضروری ہے.	0.52
Google Translation	یونانیوں کے اس مسئلے کو اس تعاقب کی ایک شکل سمجھنا چاہئے۔	0.49
Microsoft translation	یونانی مسئلے کو اس حصول کی ایک شکل کے طور پر دیکھا جانا چاہئے.	0.40

6.3 Graphically representation Chinese Urdu translation model

In NMT we done our work with two models LSTM and Transformer model. We have trained Chinese to Urdu language dataset which is 66k parallel corpus. The validation part of a source and target which have been taken as 20% of training corpus, then for testing the model we have identified 10k sentences randomly from the corpus then we make 7 different test for our models . We collected different results against each Model for each test we give a name of CTU and also compare the translation with manually as well as in translation model. We also calculated the BLEU score for the each CTU.The details of the each CTU test below in **Table 10** and **Figure 3**.

Table 10. Bleu Score of LSTM and Transformer Model

Test	LSTM Model	Transformer Model
	BLEU SCORE	
CTU 1	0.0678	0.0778
CTU 2	0.0947	0.1347
CTU 3	0.1255	0.1655
CTU 4	0.2089	0.2074
CTU 5	0.2824	0.3024
CTU 6	0.3629	0.4529
CTU 7	0.4187	0.5287

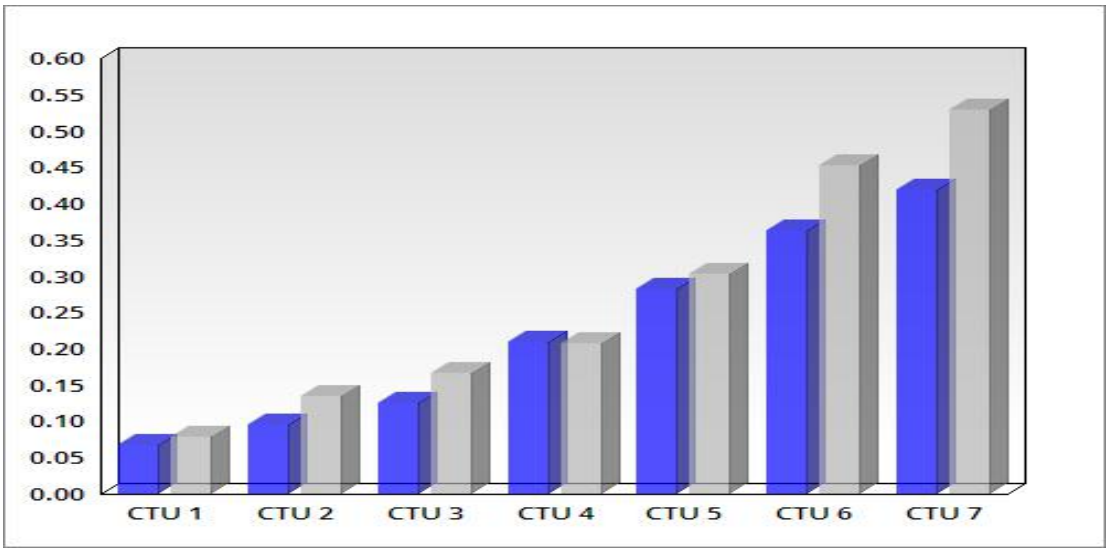

Figure 3. Bar representation of BLEU score of different CTU.

7. Conclusion

Our work will be proved an electronic tool for ease to the business class and natives of Pakistan and other neighbor countries to understand the Chinese language. Basically we have created C2U bilingual translation system which translates Urdu, Chinese words and explanation. The set up reproducible basic word results of several available test of datasets. With this basis, C2U research should be able to stepwise improve the state of the art, in distinction with the spread experiments. So we have taken training data from several sources which is be made up of different variety of sentences. The training part of method is conducted in the shape of data-test, and it has proved practical as the BLEU score has been increased with the number of data-test; the accuracy of the system is obtained after seven data-test, which is suitably matched to other machine translation systems. The BLEU score of Chinese to the Urdu translation system is going to be improved by applying some more techniques, which are used to generate the best model of translation.

References

[1] https://www.pinterest.com/lodhran/urdu-articles/

[2] https://www.chineseclass101.com/chinese-word-lists/

[3] https://www.urdupod101.com/urdu-word-lists/?coreX=100

[4] https://www.urdupod101.com/urdu-word-lists/?coreX=100

[5] https://www.kaggle.com/alvations/old-newspapers

[6] https://www.kaggle.com/zarajamshaid/language-identi

fication-datasst

[7] https://sites.google.com/site/researchonurdulanguage1/

databases/offline-urdu-database/offline-handwritten-urdu-dataset

[8] https://github.com/OpenNMT/OpenNMT-py

[9] https://github.com/OpenNMT/OpenNMT-py

[10] https://github.com/sharmaroshan/Text-Classification

[11] https://github.com/OpenNMT/OpenNMT-py

[12] https://en.wikipedia.org/wiki/BLEU

1. Zubair AM, Aurangzeb K, Shakeel A, *et al*. A unified framework for creating domain dependent polarity lexicons from user generated reviews [J]. PLOS ONE 2015; 10(10): e0140204.
2. Badaro, Gilbert, Baly, *et al*. A large scale arabic sentiment lexicon for arabic opinion mining [C]. Acm Emnlp Workshop on Arabic Natural Language Processing. ACM 2014.
3. Akshat Bakliwal, Piyush Arora, Vasudeva Varma. Hindi subjective lexicon: A lexical resource for Hindi polarity classification [J]. 2012.
4. Dashtipour K, Poria S, Hussain A, *et al*. Multilingual sentiment analysis: State of the Art and independent comparison of techniques [J]. Cognitive Computation 2016; 8(4): 757-771.
5. Dehkharghani R, Saygin Y, Yanikoglu B, *et al*. Sentiturknet: A Turkish polarity lexicon for sentiment analysis [J]. Language Resources and Evaluation 2016; 50(3): 667-685.
6. Mukhtar N, Khan MA. Effective use of evaluation measures for the validation of best classifier in Urdu sentiment analysis [J]. Cognitive Computation 2017; 9(2): 1-11.
7. Torii Y, Das D, Bandyopadhyay S, *et al*. Developing Japanese wordnet affect for analyzing emotions [C]. Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis. Association for Computational Linguistics 2011.
8. Antony PJ. Machine translation approaches and survey for Indian languages [J]. Chinese Journal of Computational Linguistics 2013; 18(1): 47-78.
9. Nalisnick E, Ravi S. Learning the dimensionality of word embeddings [J]. Computer Science 2015.
10. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks [J]. Advances in Neural Information Processing Systems 2014.
11. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [J]. Computer Science 2014.
12. Bilal M, Israr H, Shahid M, *et al*. Sentiment classification of Roman-Urdu opinions using Navie Baysian, Decision tree and KNN classification techniques [J]. Journal of King Saud University Computer & Information Sciences 2015; 28(3): 330-344.
13. Alam M, Hussain SU. Sequence to sequence networks for Roman-Urdu to Urdu transliteration [J]. 2017.
14. Mukhtar N, Khan MA. Urdu sentiment analysis using supervised machine learning approach [J]. International Journal of Pattern Recognition and Artificial Intelligence 2017.
15. Hussain SA, Zaman S, Ayub M. A self organizing map based Urdu Nasakh character recognition [C]. International Conference on Emerging Technologies, IEEE 2009.
16. Auli, Michael, Michel Galley, *et al*. Joint language and translation modeling with recurrent neural networks. 2013.
17. Yang N, Liu S, Li M, *et al.* Word alignment modeling with context dependent deep neural network [C]. Meeting of the Association for Computational Linguistics 2013.
18. Mikolov T, Martin Karafiát, Burget L, *et al*. Recurrent neural network based language model [C]. INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010. DBLP 2010.
19. PAPINENI, Blue S. A method for automatic evaluation of machine translation [C]. Meeting of the Association for Computational Linguistics. Association for Computational Linguistics 2002.
20. Baker, Paul, Hardie, *et al*. EMILLE: A 67-million word corpus of Indic languages: Data collection, mark-up and harmonization [J]. Blood 2002; 87(11): 4723-30.
21. Cohn T, Callison-Burch C, Lapata M. Constructing Corpora for the development and evaluation of paraphrase systems [J]. Computational Linguistics 2008; 34(4): 597-614.
22. Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units [J]. Computer Science 2015.
23. Luong MT, Pham H, Manning CD. Effective approaches to attention-based neural machine translation [J]. Computer Science 2015.
24. Vaswani, Ashish, Noam Shazeer, *et al*. Attention is all you need. In Advances in Neural Information Processing Systems 2017; pp: 5998-6008.
25. Senellart J, Zhang D, Wang B, *et al*. Opennmt system description for WNMT 2018: 800 words/sec on a single-core CPU. In Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, Melbourne 2018.