COMPOUND SENTENCE SEGMENTATION AND SENTENCE BOUNDARY DETECTION IN URDU

ASAD IQBAL, ASAD HABIB, JAWAD ASHRAF

Institute of Information Technology, Kohat University of Science and Technology, Pakistan

ABSTRACT:

The raw Urdu corpus comprises of irregular and large sentences which need to be properly segmented in order to make them useful in Natural Language Engineering (NLE). This makes the Compound Sentences Segmentation (CSS) timely and vital research topic. The existing online text processing tools are developed mostly for computationally developed languages such as English, Japanese and Spanish etc., where sentence segmentation is mostly done on the basis of delimiters.

Our proposed approach uses special characters as sentence delimiters and computationally extracted sentence-end-letters and sentence-end-words as identifiers for segmentation of large and compound sentences. The raw and un-annotated input text is passed through preprocessing and word segmentation. Urdu word segmentation itself is a complex task including knotty problems such as space insertion and space deletion etc. Main and subordinate clauses are identified and marked for subsequent processing. The resultant text is further processed in order to identify, extract and then segment large as well as compound sentences into regular Urdu sentences.

Urdu computational research is in its infancy. Our work is pioneering in Urdu CSS and results achieved by our proposed approach are promising. For experimentation, we used a general genre raw Urdu corpus containing 2616 sentences and 291503 words. We achieved 34% improvement in reduction of average sentence length from 111 w/s to 38 w/s (words per sentence). This increased the number of sentences by almost three times to 7536 shorter and computationally easy to manage sentences. Resultant text reliability and coherence are verified by Urdu language experts.

Keywords: Urdu sentence segmentation, sentence tokenization, word tokenization, compound sentence segmentation, Urdu conjunction extraction, Urdu sentence delimiter identification.

1. INTRODUCTION:

Urdu Compound Sentence Segmentation using words and conjunctions as delimiters is a complex task. Most of the available raw corpora contain large sentences which are combination of sentences with conjunctions or without conjunctions. Such sentences are called compound sentences. Such sentences make it challenging for automated and computational processes such as text summarization, parsing and named entity recognition etc. [7][19].

There are some online tools available that segment sentences on the basis of sentence termination marks. For example;

1.1. Automatic Sentence Segmentation
1.2. Morph Adorner Sentence Splitter Example

Automatic sentence segmentation alters simple text into separate sentence per line format by simply adding return code after sentence termination mark but most tools cannot handle abbreviation's like *Dr., Mr., p.m., Prof., a.m.* This online tool covers most of those abbreviations. It also provides editing facility after resulting text to cover up the remaining abbreviations [1]. Morph Adorner Sentence Splitter uses punctuation marks to split sentences but punctuation marks not always define sentence termination mark for example ellipses, abbreviations, acronym, or decimal system and in some of the poems not even have any termination mark in it. Morph Adorner Sentence Splitter works best for plain English text and unlike Automatic Sentence Segmentation it covers more abbreviations [2].

Unlike Urdu English is not even suffering from this problem too much as most of the sentences are already separated from each other, Urdu on the other hand has even sentences in a form of paragraph which in self contains many sentences and our proposed idea is to identify those sentences and convert them into more than one sentences on the basis of these words [3,4].

2. LITERATURE REVIEW:

Sentence segmentation is a relatively new topic of research in computationally developing languages. We could not find any automated sentence boundary segmentation tool in Urdu language. Aroonmanakun W., analyzed sentence and word segmentation in Thai language [5]. They considered sentence discourse where combination of phrases and some clues are used for each discourse segmentation process. Baseer et al. presents a sophisticated Romanized Urdu Corpus utilizes tokens with the uppermost frequency of occurrence in the data set, which was collected from participants who uses Romanized Urdu as a mean of communication [6]. Xu et al. used IBM word alignment for sentence segmentation for translation tasks of English and Chinese in the news domain [7]. The focus of this technique is to use lexicon information to identify sentence segmentation point to split compound sentences into multiple sentences. This paper proposed a technique to split large sentences into multiple ones because the longer the sentences the more problems are faced by their proposed system, subsequently resulting into higher computation cost and compromised quality in word alignment. Habib et al. presented a novel approach that records properly segmented words in order to avoid the classical space insertion and space deletion problems in native script Urdu text [13][15]. Their proposed solutions can be of direct value in Urdu sentence segmentation.

Xue, N. and Yang, Y., focuses on how Chinese uses comma, exclamation marks and questions marks for sentence boundary indication. Proposed model is being tested and trained on data provided by Chinese tree bank and the accuracy achieved by this model is up to 90% [8,9]. Rehman, Z. and Anwar, W., uses rule based algorithm and Unigram statistical model to deal with Urdu sentence boundary disambiguate. Initial result before testing and training were 90% precision, 92.45% F1-measure and 86% recall, but after same testing data and training, results improved to 99.36% precision, 97.89% F1-measure and 96.45% recall [10]. Kiss and Strunk presented language-independent approach. In this paper assumptions are made that once abbreviations are identified most of the ambiguities will be eliminated while detecting sentence boundary. In order to detect high accuracy abbreviations, proposed system define three rules which required independence of context and type of candidate. The system was tested on different text types and eleven different languages [11]. A number of related research points out to interesting aspects of text segmentation and optimized input systems. Habib et al. proposed an optimized system and input methods with respect to various modern devices for Urdu composing [18]. Adnan et al. assessed the realization of smartphones learning objects in computing adaptive learning paths of undergraduate university students. Jurish and Würzner introduced a "WASTE" method for segmentation of text into tokens and sentences. Hidden Markov Model is used as a segment boundary detection. Model parameters were defining from pre-segmented corpora. Such corpora are available as an aligned multi-lingual corpora and treebanks [12].

Hearst, M.A., A presents technique called TextTiling in which text is segmented into multi-paragraph units that is subtopics or passages. Identification of sub-topics is done using lexicon co-occurrence and distribution patterns. This segmentation can further be used for text summarization and information retrieval [16]. In Evang, K., et al. technique the accuracy achieved by rule based model Tokenization is considered as no problem, but issue regarding rule-based is its language specific rules and maintenance. This paper used unsupervised feature learning combining it with supervised sequence labeling on character level to accomplish segmentation and high accuracy word goal. Evaluation of proposed system is done on three different languages with the error rate of 0.76% (Italian), 0.27% (English), and 0.35% (Dutch) [16]. Xu, L.F., et al. proposes an idea of segmenting long sentences into short one using conjunctions in them. Long sentences took a lot of machine translation resources for processing. Punctuations were used previously for segmentation but dealing with long sentences that was not enough. This paper presents rule-based approach on conjunctions to segment Chinese long sentences. 901 conjunctions were found in 10 patent papers during experimentation. Using rule-based approach, 89% accuracy was achieved [17]. Gibson E., et al proposes Comprehension of sentences technique. Comprehension of sentences means applying constraints using available computational resources to integrate variety of information sources. Four types of comprehension techniques are explained in this paper. (1) phrase-level contingent frequency constraints, (2) locality-based computational resource constraints, (3) contextual constraints, and (4) lexical constraints [19].

3. URDU SENTENCE SEGMENTATION OPERATIONS:

Sentence Segmentation is a process of identifying sentence boundaries. Previous researcher's focuses on using only punctuations as a boundary detection. Our proposed research not only uses punctuations, it also uses words as a delimiters and conjunctions as boundary markers to segment sentences appropriately.

3.1. RAW CORPUS:

Raw corpus has been collected by several means which includes websites, books, Urdu magazines, newspapers etc., we categories them into more general form that is online and offline category. Online category involves websites, online books and newspapers. While offline involves Magazines, newspapers, books etc. The collected raw corpus contains large and compound sentences including combination of conjunction and non-conjunction sentences. These sentences require two different methodologies for segmentation. The former is compound sentence segmentation and the latter is sentence tokenization which are described in the following.

3.2. WORDS/CHARACTER SEGMENTATION:

Conjunctions and delimiter words and characters are identified using words and character segmentation techniques. Individual words and characters in text are split on the basis of complex processing including space, joiner and non-joiner properties keeping in view the complex problems of Urdu specific space insertion and space deletion [3][10][19]. The segmented words and characters are then analyzed for conjunctions and delimiting words and characters.

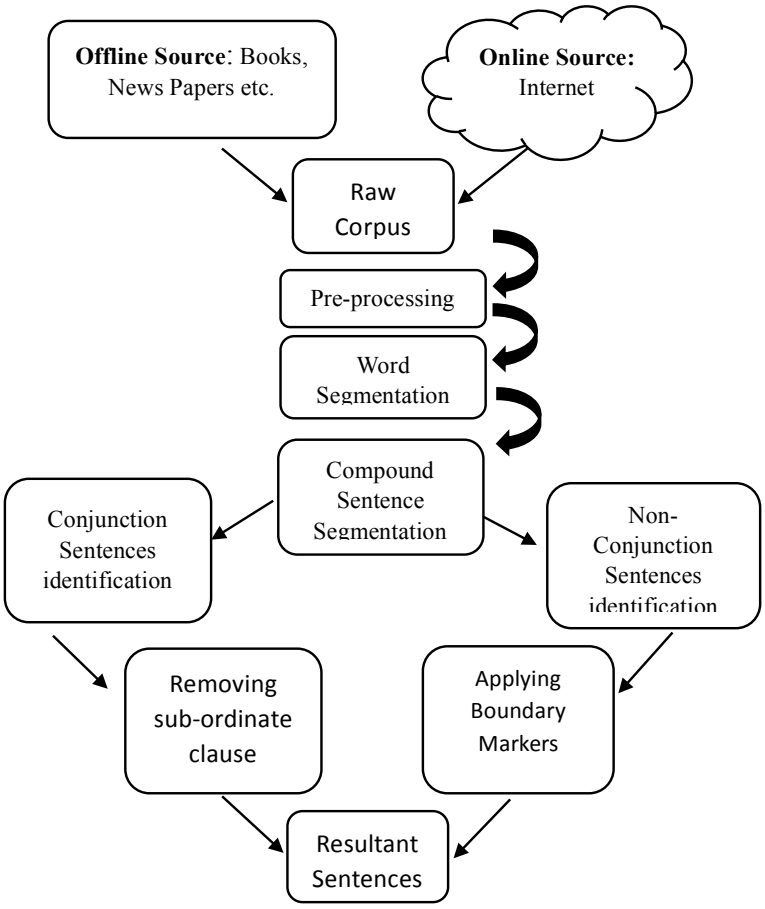

Figure.1. **Compound Sentence Segmentation and Sentence Tokenization Architecture**

3.3. COMPOUND SENTENCE IDENTIFICATION:

The issues with longer sentences is high consumption of resources such as processing time. We look for conjunction containing sentences and we analyze that the subordinate clause of sentences are mostly the explanation of main clause, which makes sentence extra-large. The options in dealing with compound sentences are two, first is to eliminate those sentences completely but by doing so, we may be risking useful information. Second option is to trim those sentences by eliminating sub-ordinate clause i.e. the explanatory part of the sentence.

We use a pattern matching approach for identification of conjunctions containing sentences. We came up with the list of conjunction words which we generated from pre tagged corpus, using this list of words conjunction containing sentences are easily detectable. List of conjunction words we are given below.

کیونکہ ، لیکن ، یعنی ، گویا ، یعنی ، چنانچہ ، بلکہ ، مگر

Naturally occurring raw corpus text examples are mentioned in the following.

3.4. With Conjunction part:

مگر سابق کونسلر عبدالنعیم نے کہا کہ تمام علاقے کی بجلی کاٹنا اس موقع پر حالات کو قابو کرنے کے لئے پولیس کی نفری بھی پہنچ گئی تھی عوام کے ساتھ سراسر زیادتی ہے جن لوگوں پر واجبات ہیں صرف انکی بجلی کاٹنی چاہئے۔

3.5. Without Conjunction part:

اس موقع پر حالات کو قابو کرنے کے لئے پولیس کی نفری بھی پہنچ گئی تھی ۔

We exclude "اور" from conjunction list because it is not only used as a joiner in a sentence but as a joiner between two nouns too for example;

جماعت اسلامی اور الخدمت فاؤنڈیشن

The remaining sentences in corpus contain combination of multiple sentences. These sentences do not contain conjunctions but they need to be segmented. For this purpose, we pass resultant text to the next process called Sentence Tokenization.

3.6. SENTENCE TOKENIZATION:

Sentence tokenization segments large sentences from single into multiple ones. We target words as delimiters to identify sentence boundaries. For example;

ھنگو(بیورورپورٹ)ڈپٹی ڈسٹرکٹ ایجوکیشن آفیسر عقل بادشاہ نے کہاکہ ہمارے بھتیجے کی وفات پر ھنگو کے عوام اور تمام مکاتب فکر کی جانب سے تعزیتی پیغامات پر اہل علاقہ کے نہایت مشکورہیں ان خیالات کا اظہار ڈپٹی ایجوکیشن آفیسر عقل بادشاہ نے اپنی رہائش گاہ پر صحافیوں سے بات چیت کرتے ہوئے کیا انہوں نے کہا کہ واقعہ دلخراش ہے مگر اللہ تعالیٰکی مہربانی سے ان کے خاندان اور سوگواران نے بزرگوں ،اہل علاقہ کے دعاؤں سے پورے خاندان کو صبر جمیل عطا ہوئی کہاکہ فیاض مرحوم کے سوگوران ان تمام حضرات جنہوں نے نمازہ جنازہ میں شرکت کی یا بذریعہ ٹیلی فون تعزیت کی ان سب کے دلی مشکور ہیں۔

Above example is a single large sentence which is a combination of multiple sentences with 110 words in it. In available corpus Urdu is full of these kind of sentences. These delimiting words are identified manually from single chunk of file because analyzing large file manually is time consuming and it may take weeks to generate list of delimiting words and still that list will not be enough. Delimiting boundary list is not limited to this corpus only to achieve better results in segmenting large sentences this corpus also needs constant updating with constant manual updating of delimiting words also. We generate that list using a chunk of file is to carry on our experimentation and we accomplish significant results. For example, the above single sentence is analyzed for delimiting words and on the basis of this list above sentence is segmented down into 4 sentences with average length of 27 words per each sentence.

ھنگو(بیورورپورٹ)ڈپٹی ڈسٹرکٹ ایجوکیشن آفیسر عقل بادشاہ نے کہاکہ ہمارے بھتیجے کی وفات پر ھنگو کے عوام اور تمام مکاتب فکر کی جانب سے تعزیتی پیغامات پر اہل علاقہ کے نہایت مشکورہیں۔

ان خیالات کا اظہار ڈپٹی ایجوکیشن آفیسر عقل بادشاہ نے اپنی رہائش گاہ پر صحافیوں سے بات چیت کرتے ہوئے کیا۔ انہوں نے کہا کہ واقعہ دلخراش ہے مگر اللہ تعالیٰکی مہربانی سے ان کے خاندان اور سوگواران نے بزرگوں ،اہل علاقہ کے دعاؤں سے پورے خاندان کو صبر جمیل عطا ہوئی کہاکہ فیاض مرحوم کے سوگوران ان تمام حضرات جنہوں نے نمازہ جنازہ میں شرکت کی۔ یا بذریعہ ٹیلی فون تعزیت کی ان سب کے دلی مشکور ہیں۔

3.6.1. Sentence Boundary Maker:

Complexity of delimiting words of sentence boundary markers rises when generated list of delimiting words turned into a part of another word and that word appears in the middle or start of sentence not at the end. To deal with these kind of sentences, we need to take our approach to another level as we cannot use unigram. For example, the following list of delimiter words are unigram. We cannot use them as boundary markers alone because these words can be a part of other words and that results in ambiguity. We require n-gram approach. For example;

کریں ، آسکے ، کئے ، جائے ، دی

1) ہمارے مسائل ترجیحی بنیادوں پر حل **کریں** بصورت دیگر ہم پارلیمینٹ ہاؤس اور گور نر ہاؤس کے سامنے احتجاجی دھرنے پر مجبور ہونگے ۔

2) نہوں نے کوئی ایسا کام نہیں کیا جس کو یادگار کہا **جاسکے** بلکہ **اسکے** اپنے صوبائی اسمبلی کے ممبر باغی ہوچکے ہیں

3) تعلیمی اداروں سے نقل کا ناسُور ختم **کئے** بغیر ترقی ممکن نہیں

4) یہاں پر انفرادی کاموں کی **بجائے** اجتماعی کاموں کو ترجیح دی جاتی ہے

5) یہاں پر **انفرادی** کاموں کی بجائے اجتماعی کاموں کو ترجیح دی جاتی ہے

In example 1, 2 and 3 the delimiting words appear in the middle of sentences. In example 2 we also realize that there are two delimiters used one as a separate word and other as a part of other word for example جاسکے, which is a combination of word اسکے and ج. Same apply to example 4 and 5 i.e. بجائے and انفرادی. These delimiting words are also a combination of words i.e. بجائے is a combination of جائے and ب and انفرادی is a combination of دی and انفرا. But that was not all there another issue Urdu delimiting words were having, some of these words were not even used in the sense of termination mark for example اسکے in the above example 2 was not used in sense of a sentence boundary marker but it was used in a sense of "his" in that sentence, but the ratio of these words are low and they can also be handled by n-gram approach so this does not create any problem. Below Table 1 represents sentences boundary markers of unigram, bigram and trigram. We only include few of the boundary markers with their frequency from a single experiment.

S. No.	Frequency (sorted)	Sentence Boundary Marker List
1	259	ہیں
2	67	تھا
3	57	گے
4	36	تھے
5	20	دیا گیا
6	13	تھی
7	13	ہوگا
8	11	ہوگی
9	5	رہاہے
10	1	کرتاہے
11	1	حوالے کئے
12	1	حاصل کیے
13	1	کرائی ہے
14	1	تلاش شروع کردی
15	1	دی جائیں

Table 1: **Sentence boundary markers list with corresponding frequencies**

رابطہ کریں ، تلاش شروع کردی ، میسر آسکے ، حوالے کئے ، کرایا جائے

Some sentences contain two delimiting words and our model will add termination marks after both of these boundary markers, but that will not affect the sentence, because if sentence ends with first delimiting word it will still retain its meanings. For example;

3.6.1.1. **With two delimiting words**:

جو دس سال اقتدار میں رہے آج ان سب کو گرینڈ اپوزیشن میں بیٹھا دیا گیا ہے۔

3.6.1.2. **Without second delimiting word**:

جو دس سال اقتدار میں رہے آج ان سب کو گرینڈ اپوزیشن میں بیٹھا دیا گیا ۔

It can be observed that from semantic point of view, the sentence preserves its meanings.

4. EVALUATION:

The lack of appropriate hardware resources impeded delays in compiling results of our experiments. We used a personal computer running Microsoft Windows 8 for processing our general genre raw Urdu corpus containing 2616 sentences and 291503 words. Computationally exhaustive iterative processing was not possible on such a workstation. It took more than two and a half days to process our raw corpus file and still it was not completed when the system suddenly shut down due to hardware failure. The only possible solution for this problem was the customary divide and conquer approach.

We divided the experimentation file into 14 smaller chunks and experimented chunk-wise to identify compound and large sentences. Delimiting words were extracted with their respective frequencies in the text. Compound sentences are identified by the list of conjunctions present in them. All statistical results were accumulated and manually verified by Urdu language experts. Consolidated results of the respective 14 chunks are shown in the following table 2.

S.No.	Chunk (File) Size		بلکہ	چنانچہ	یعنی	گویا	لیکن	کیونکہ	مگر	اور
	No. of Sentences	No. of Words								
1	184	20548	16	0	0	0	8	12	22	485
2	175	21794	10	0	2	0	11	8	20	517
3	274	28034	14	0	1	0	13	10	41	697
4	285	26372	15	0	1	0	8	4	29	661
5	270	27165	13	0	1	0	11	9	41	677
6	264	22956	15	0	1	0	9	3	22	582
7	211	25773	6	0	2	0	10	7	27	629
8	163	25621	17	0	1	1	17	7	36	626
9	184	22228	7	0	1	1	6	5	45	568
10	124	15326	7	0	2	0	8	2	15	335
11	104	12364	4	0	0	0	2	2	19	284
12	99	10172	6	0	1	0	2	4	17	277
13	142	15075	4	0	1	0	14	9	18	400
14	137	18075	3	0	0	0	11	6	28	457
∑	2616	291503	137	0	14	2	130	88	380	7195

Table 2**: Conjunctions and delimiting words extracted from the raw Urdu corpus**

In table 2, we also generate frequency of **اور** as a conjunction and it appears 485 times in a text. This was done due to realization of the fact that **اور** is not only used as a conjunction but also as a part of other words. For example;

اورزئی ، دلاورشاہ ، پشاور ، نچھاور ، مشاورت ، اورکزئی

These are the words that contain conjunction word "اور". These words are the combination of two words. For example, in اورکزئی and اورزئی, the word اورis at the beginning and it is a combination of کزئی + اور and زئی + اور respectively. Similarly, in مشاورت and دلاورشاہ, the word اور is in the middle of two words that is between مش + ت and دل + شاہ respectively. In the same manner, in words such as نچھاور and پشاور, the word اور appears in the end of these words. Our proposed model identifies them as a conjunction which increases the computational cost and algorithmic complexity. 135 times out of 485, the word "اور" is used as a conjunction that is 27% of its total occurrences. The remaining 73% i.e. 353 times, this word appeared as a part of other words or sometimes in a non-conjunction way. So we exclude it from list of conjunctions. Identifying them as a conjunction and as a part of other word is done manually.

The word "اور" and other words posing similar problem is an interesting discovery in this research. "چنانچہ" is also a conjunction word but in our existing corpus it does not appear even once. Our corpus is growing continuously. So therefore we hope this and other similar words will be encountered in the subsequent experiments. Thus we did not exclude it from the list of conjunction word.

Complexity of Urdu language increases when dealing with boundary markers. The Urdu boundary marking words may not always appear at the end of sentences but it can also be a part of other words which may be anywhere in sentences. Also, they may appear in the middle of sentences. Reaching a computational solution becomes more difficult when sometimes these are not used in the sense of boundary markers. For example, ہیں appears about 259 times in a single chunk of file. We carried out the same experiment on all 14 chunks. However, for simplicity we consider the example

of only first chunk here. It was not sure whether ہیں always appears as a boundary marker or a part of other words. In our experimentation we realize, that 154 time ہیں appeared as a boundary marker that is 59% and 105 times i.e. 41% it appeared as part of another word for example نہیں, تنخواہیں, انہیں, تمہیں. The experimentation continued on other boundary markers such as تھا. Frequency of تھا in a text is 67 and it appears in text about 26 times that is 38% as a boundary marker and 41 times that is 62% as part of other words.

Most of the boundary markers are part of other words and to handle this we use bi-gram approach i.e. we combine it with the second closest word but using only bi-gram approach did not solve our problem because combination of these two words may also appear in middle of sentences, to tackle this issue we combine third closest word i.e. we use tri-gram approach but still that did not solve our problem and the process continues. The solution to this problem was to use n-gram approach. Similarly, تھے appears in a text for 36 times and 100% it is used as a boundary marker. تھی appears 13 times in which 8 time that is 61% it appears as a boundary marker and 5 times that is 39% as part of other word or boundary marker i.e. تھیں and the same goes for other boundary markers. The above mention gazetteer list of delimiting words contains only 15 boundary markers and we have about 103 list of boundary markers created manually and the list is still in a growing process the more the list grows the more effective will be the results. Considering the corpus, we have, 103 is not a huge list of boundary markers but these are enough for our experimentation. The more the list grows the more processing is required because our workstations cannot handle that kind of processing very effectively and that makes our experimentation process slower, so we restricted list of boundary markers to 103.

Initially there were 184 sentences with 20548 words in experiment 1. After compound and tokenization process we have 722 sentences. We got results for about 569 sentences that is 78% of sentences can be categorized as accurate or inaccurate sentences out of which 461 were accurately marked boundary markers that is 63% and 108 were marked inaccurate that is only 14%. This inaccuracy was due to existence of those boundary markers as a part of other words. Remaining 22% sentence were the one or two word sentences that appears because of two delimiting words in a single sentence. For example;

اُن کے لیے نہ تو قریبی علاقوں میں کوئی پارکنگ موجود ہے اور نہ ہی وہاں پر عارضی انتظام کیا گیا ہے۔

ہے is second boundary marker and just a single word marked as separate sentence. There were 153 such kind of sentences which means that 22% sentences were useless and discarded accordingly.

5. CONCLUSION:

Our work regarding Compound Sentence Segmentation and Tokenization of large Sentences was pioneering work in Urdu. The results generated by our proposed system are promising. The results generated for a single chunk of file were generated manually. Therefore, we did not include other chunks of files. We realize that with having powerful servers and with increasing delimiting words gazetteer list, we can improve our results further. Beside generating statistical results, in future we will also analyze our model and its results by language expert, by comparing our automatically tokenized sentences with human manually tokenized sentences to analyze its coherence and readability.

REFERENCES:

[1] Yasumasa, S. Kansai. 2016. University of Graduate School of Foreign Language Education and Research. Automatic Sentence Segmentation, accessed (Feb 19). DOI: http://www.someya-net.com/00-class09/sentenceDiv.html.

[2] Brian, L., Zillig, P., Ramsay, S., Mueller, M., and Smutniak, F., 2016. Academic Technologies and Research Services. Morph Adorner Sentence Splitter, accessed (Feb 19). DOI: http://morphadorner.northwestern.edu/sentencesplitter/example/.

[3] Malik, A.A. and Habib, A., 2013. Urdu to English Machine Translation using Bilingual Evaluation Understudy. *International Journal of Computer Applications*, *82*(7).

[4] Palmer, D.D., 2000. Tokenisation and sentence segmentation. In *Handbook of Natural Language Processing*. CRC Press.

[5] Aroonmanakun, W., 2007, December. Thoughts on word and sentence segmentation in Thai. In *Proceedings of the Seventh Symposium on Natural language Processing, Pattaya, Thailand, December 13–15* (pp. 85-90).

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

[7] Xu, J., Zens, R. and Ney, H., 2005, May. Sentence segmentation using IBM word alignment model 1. In Proceedings of EAMT (pp. 280-287).

[8] A. Gul, A. Habib, J. Ashraf, "Identification and extraction of Compose-Time Anomalies in Million Words Raw Urdu Corpus and Their Proposed Solutions", *proceedings of the 3rd International Multidisciplinary Research Conference (IMRC)*, 2016.

[9] Xue, N. and Yang, Y., 2011. Chinese sentence segmentation as comma classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2* (pp. 631-635). Association for Computational Linguistics.

[10] Rehman, Z. and Anwar, W., 2012. A hybrid approach for urdu sentence boundary disambiguation. *International Arab Journal of Information Technology*, *9*(3), pp.250-255.

[11] Kiss, T. and Strunk, J., 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, *32*(4), pp.485-525.

[12] Jurish, B. and Würzner, K.M., 2013. Word and Sentence Tokenization with Hidden Markov Models. *JLCL*, *28*(2), pp.61-83.

[13] Habib, A. Iwatate, M., Asahara, M. Matsumoto, Y., 2012. Keypad for large letter-set languages and small touch-screen devices (case study: Urdu). *International Journal of Computer Science 9(3), ISSN: 1694-0814.*

[14] Hearst, M.A., 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, *23*(1), pp.33-64.

[15] Habib, A. Iwatate, M., Asahara, M. Matsumoto, Y. W.K., 2013. Optimized and Hygienic Touch Screen Keyboard for Large Letter Set Languages. *Proceedings of 7th ACM International Conference on Ubiquitous Information Management and Communication (ICUIMC) Kota Kinabalu, Malaysia.*

[16] Evang, K., Basile, V., Chrupała, G. and Bos, J., 2013, October. Elephant: Sequence labeling for word and sentence segmentation. In *EMNLP 2013*.

[17] Xu, L.F., Zhu, Y., Yang, L.J. and Jin, Y.H., 2014. Research on Sentence Segmentation with Conjunctions in Patent Machine Translation. In *Applied Mechanics and Materials* (Vol. 513, pp. 4605-4609). Trans Tech Publications.

[18] Habib, A. Iwatate, M., Asahara, M. Matsumoto, Y., 2011. Different input systems for different devices: Optimized touch-screen keypad designs for Urdu scripts. *Proceedings of Workshop on Text Input Methods WTIM2011, IJCNLP, Chiang Mai, Thailand.*

[19] Gibson, E. and Pearlmutter, N.J., 1998. Constraints on sentence comprehension. *Trends in cognitive sciences*, *2*(7), pp.262-268.

[20] Adnan, M. Habib, A. Mukhtar, H. Ali, G., 2017. Assessing the Realization of Smartphones Learning Objects in Students' Adaptive Learning Paths. *International Journal of Engineering Research, 6(4), ISBN:978-81-932091-0-3.*