Urdu Summary Corpus

Muhammad Humayoun[1], **Rao Muhammad Adeel Nawab**[2], **Muhammad Uzair**[2], **Saba Aslam**[2], **Omer Farzand**[2]
[1]IRIT (Institut de Recherche en Informatique de Toulouse), Université Paul Sabatier, Toulouse, France
[2]Department of Computer Science, COMSATS Institute of Information Technology, Lahore, Pakistan
muhammad.humayoun@irit.fr, adeelnawab@ciitlahore.edu.pk, uzairnaroo@gmail.com,
saba_12@hotmail.fr, umerfarzand@gmail.com

Abstract

Language resources, such as corpora, are important for various natural language processing tasks. Urdu has millions of speakers around the world but it is under-resourced in terms of standard evaluation resources. This paper reports the construction of a benchmark corpus for Urdu summaries (abstracts) to facilitate the development and evaluation of single document summarization systems for Urdu language. In Urdu, space does not always mark word boundary. Therefore, we created two versions of the same corpus. In the first version, words are separated by space. In contrast, proper word boundaries are manually tagged in the second version. We further apply normalization, part-of-speech tagging, morphological analysis, lemmatization, and stemming for the articles and their summaries in both versions. In order to apply these annotations, we re-implemented some NLP tools for Urdu. We provide Urdu Summary Corpus, all these annotations and the needed software tools (as open-source) for researchers to run experiments and to evaluate their work including but not limited to single-document summarization task.

Keywords: benchmark corpus, abstracts, automatic text summarization, single-document summarization

1. Introduction

Urdu is an Indo-Aryan language, widely spoken in South Asia. It is also spoken all over the world due to the large South Asian Diaspora. Urdu has more than 100 million speakers[1]. It is written in a modified Perso-Arabic script from right to left. It requires specific rendering to be viewed properly. Normally, it is written in Nastalique, a highly complex writing system that is cursive and context-sensitive.

Urdu has a complex morphology that inherits grammatical forms and vocabulary from Arabic, Persian, and native languages of South Asia (Humayoun et al., 2007). There is no capitalization in Urdu. This makes identifying proper nouns, titles, acronyms, and abbreviations a difficult task. Diacritics (vowels) are hardly present in the text and words are guessed with the help of the context of surrounding words. In terms of syntax, it has a free word order (Subject Object Verb). Despite spoken by millions of people, Urdu is an under-resourced language. A sentence illustrating Urdu is given below:

اُردو پاکستان کی قومی زبان ہے -

Urdu is the national language of Pakistan.

The availability of benchmark corpora plays an important role in the development of tools and techniques for various NLP tasks. For automatic text summarization, the shared tasks offered by Document Understanding Conference (DUC)[2] and Text Analysis Conference (TAC)[3] provide different sets of good quality benchmark corpora mainly for English. These corpora consist of single and multi-document summaries written by humans. These summaries are abstractive as well as extractive[4]. These sets of benchmark corpora have been used for the development and evaluation of summarization systems (mainly for English) and results are published regularly.

Unfortunately, unavailability of standard evaluation resources (in general) is one of the main hinders for doing research in the Urdu language. As a first step, this work develops a benchmark Urdu summary corpus. It contains 50 articles and their corresponding abstractive summaries covering various domains. These are human-written single-document abstracts. There is only one summary for an article.

Urdu is one of the languages that suffers from word segmentation problem, i.e., space does not always identify the word boundary (Durrani and Hussain, 2010). Therefore, we produced two versions of the corpus. In the first version, words are separated by space; whereas in the second version, proper word boundaries are manually tagged. This may allow researchers to measure the effect of word segmentation on summarization task as well as on other language processing tasks for Urdu.

Automatic summarization can be further improved by preprocessing work, such as normalization of text, part-of-speech tagging, morphological analysis, lemmatization and stemming (Leite et al., 2007; Torres-Moreno, 2012; Torres-Moreno, 2014; Nuzumlal and Özgür, 2014; Eryiğit et al., 2008). Therefore, we further apply the needed software tools and provide the annotated output corpus freely. We also provide the source code of these software tools in one package. Note that such pre-processing tools are scarce for Urdu. In addition, these tools sometimes require fine-tuning or re-implementing because of the occasional up-

[1] According to Ethnologue: www.ethnologue.com/language/urdLast visited: 04-03-2016

[2] DUC: www-nlpir.nist.gov/projects/duc/ Last visited: 04-03-2016

[3] TAC: http://www.nist.gov/tac/ Last visited: 04-03-2016

[4] DUC-2002 is the last version of DUC that included the evaluation of single-document informative summaries. In later years, only headline-length single-document summaries were analysed (Steinberger and Ježek, 2012).

dates (if any at all). Therefore, we think that providing such utilities are beneficial for researchers working on Urdu (see Section 3. for details). More precisely, Urdu Summary Corpus consists of:

- Fifty Urdu articles (and their summaries) that are normalized
- Fifty abstractive single-document summaries (one for each article)
- Fifty part-of-speech tagged articles
- Fifty morphologically analyzed articles
- Fifty lemmatized articles
- Fifty stemmed articles

Whereas, the accompanying software package consists of:

- A normalizing utility
- A POS tagger
- A table-lookup based morphological analyzer and lemmatizer
- A stemmer which is self-implemented from Assas-Band stemmer (Akram et al., 2009)

Urdu Summary Corpus and the source-code of the tools are made freely available here[5]. Details of articles and summary sizes, article sources, peer evaluation scores for human-written abstractive summaries are given in Appendix (§ 6.).

2. Summary Corpus Creation

Fifty articles for Urdu Summary Corpus (USC) were collected from various online sources mainly news portals and blogs. The sources were chosen based on the following merits:

1. They contain real text as written by the native speakers.
2. Authors of different backgrounds have written the text.
3. Getting permission from authors to re-distribute the texts was easy as compared to print media.

We tried to make the document collection balanced with the inclusion of diverse categories (see Table 1).

In general, an abstractive summary should completely convey the meaning of an original text. According to (Newfields, 2001), "*A summary is to express the main ideas of the original text using reported speech. Summary contains the essential points, but it does not contain writer's own point of view about the original text.*"

For the summary writing, a group of volunteers was selected. They were native speakers of Urdu, either (1) academicians teaching Urdu in colleges, or, (2) university students that have an interest in Urdu literature.

Category	Articles
News	6
Current Affairs	6
Health	6
Sports	10
Science & Technology	10
Tourism	3
Religion	4
Miscellaneous	5
Total	50

Table 1: Categories of the articles used in Urdu Summary Corpus

For summary writing, we did not pose any size restriction for human-written summaries[6]. We asked the writers to produce good summaries; doesn't matter if a summary is large, medium or small in size. However, the summary must not exceed half the size of an article.

In addition, we asked the writers to follow three basic steps:

1. After reading a text, identify the key phrases.
2. Paraphrase these key phrases at a sentence-level if needed.
3. Add sequential markers in between, if needed, to create a proper flow.

These steps are influenced by the six editing operations in human abstracting, which are: sentence reduction, sentence combination, syntactic transformation, lexical paraphrasing, generalization & specification, and reordering (Steinberger and Ježek, 2012).

2.1. Quality of Human Written Summaries

Each summary was evaluated by five peer contributors on a scale of 1 to 5. The scale values were represented as, 1: very bad summary, 2: poor summary, 3: adequate summary, 4: good summary, 5: excellent summary. As suggested by (Steinberger and Ježek, 2012), we asked peer contributors to consider the following aspects when assigning score:

1. Is the summary grammatical?
2. Is the summary non-redundant?
3. Is the summary free from references such as anaphora?
4. Is the summary coherent and properly structured?

The average scores given by peer contributors range from 3.8 to 4.8. Statistics of articles and Human-Written abstracts are shown in Table 2. For more details, see Appendix (§ 6.).

3. Preprocessing Challenges

3.1. Normalization of Urdu Text for Summary Corpus

As already mentioned, Urdu script is an extension of Persian and Arabic script. Because of this, some characters got

[5] `https://github.com/humsha/USCorpus/`

[6] As it turns out, not restricting the size of the summaries makes it difficult to compare with DUC summary datasets.

	Tokens in Article	Tokens in HW Summary	Compression Rate
Smallest article	159	79	49.69%
Largest article	2,518	761	30.20%
Tokens in all articles	29,889	11,683	39.08%

Table 2: Statistics of Articles and Human-Written (HW) Summaries. Tokenization is performed on space.

Space segmented words	6,074
Properly Segmented words	5,478
Difference	596 (9.8%)

Table 3: Token difference of two versions of Urdu Summary Corpus

assigned multiple Unicode for these similar scripts, resulting in orthographic variations. For example, unlike Arabic, characters like ک (kaaf) can never be written as ك in Urdu. In addition, some characters jointly form a compound character and all such characters usually got assigned their unique Unicode. For example, آ (alif-madd) can be written with: (1) two characters (unicode: 0627+0653) and (2) with one character (Unicode: 0622) (Gulzar, 2007). The Unicode Normalization standard (Davis and Whistler, 2014) defines four normalization forms; Normalization Form C (NFC) is one of them. We used it for the normalization of Urdu text for summary corpus as suggested by (Gulzar, 2007). Diacritic marks present in the collection were also removed in this step.

3.2. Word Segmentation for Urdu Summary Corpus

Urdu alphabet could be separated into two groups: joiners and non-joiners. Joiners join together with neighboring characters if space is not inserted between them. In contrast, non-joiners are not affected by the neighboring characters and maintain their shape. There are two types of word segmentation issues: (1) space insertion and (2) space deletion. Urdu faces both of them.

Space insertion problem is caused when an Urdu word contains multiple morphemes. If a morpheme ends with a joiner, it joins together with the first character of the next morpheme. Therefore, writer has to insert a space to retain proper shape of the morphemes; see (a) in Figure 1. In contrast, space omission problem is caused due to non-joiners. If a word ends with a non-joiner the shape of morphemes remain intact whether space is inserted between them or not. Therefore, writers generally do not insert spaces. See (b) in Figure 1.

How severe is word segmentation problem in Urdu? This question is answered by a small study of Urdu words (Durrani and Hussain, 2010). On a corpus of 5000 words, it is found that 52.84% words needed corrections to properly segment the text[7]. With regards to Urdu Summary Corpus, we did not perform such experiment. However, we found the token difference of 9.8% between both versions, as shown in Table 3.

Lastly, Urdu speakers tend to disagree whether compound words are one word or more (Durrani and Hussain, 2010). Therefore, we do not mark compound words as single words. To be precise, compound words with the following patterns are marked as two words. (1) X-e-Y, called ezafe (Butt et al., 2008) (e.g. صورتِحَال, Eng: situation), (2) X-Y (e.g. بات چیت, Eng: talk), and (3) X-X, (e.g. جگہ جگہ, Eng: on multiple places). Compound words with X-o-Y pattern are marked as three words, e.g. خیروعافیت (خیر, و and عافیت; Eng: being fine).

3.3. POS tagging

A stand-alone tagger for Urdu is reported in (Jawaid et al., 2014), with a reported accuracy of 88.74%, though the tool is not available publicly. This tagger extends the work of (Jawaid and Bojar, 2012) which uses an existing tagger, morphological analyzer, and shallow parser together. A voting scheme is further employed to disambiguate between these different tools. A large tagged corpus of Urdu is released publicly by (Jawaid et al., 2014). We used a large fragment of this tagged corpus to train a model on *Stanford POS tagger*[8] to automatically tag Urdu summary corpus[9].

3.4. Morphological Analysis and Lemmatization

A lemmatizer converts inflected surface forms of a word to its lemma or root form. Whereas, a morphological analyzer gives more details such as word class, number, gender, etc, in addition to lemma. We used Urdu Morphological Analyzer (Humayoun et al., 2007) for both tasks. In this open-source tool, words are built through lexical functions. The lexical functions are coded manually with respect to the inflection tables of nouns, verbs, adjectives, etc. These lexical functions are then connected with the lexicon which is built semi-automatically, and further, manually checked for mistakes. Therefore, there seem to be less possibility of a word to be incorrectly analyzed (and lemmatized) with this tool if it is present in the lexicon. The lexicon contains $5,000$ words, capable of handling $140,000$ word forms. The statistics on Urdu Summary Corpus are shown in Table 4.

Urdu Morphological Analyzer is built in Haskell (Marlow, 2010) (using Functional Morphology Toolkit (Forsberg and Ranta, 2004)), but it is not updated from a long time. Therefore, it is not possible to compile it due to the use of obsolete libraries. Fortunately, the analyzer provides full-form lexicon in text format. We built a table-lookup based analyzer and lemmatizer from it.

3.5. Stemming

A stemmer produces a truncated form for all inflected surface forms of a word. Assas-band (Akram et al., 2009) is

[7] 18.72% words suffering space omission problem and 34.12% words suffering space insertion problem.

[8] http://nlp.stanford.edu/software/tagger.shtml

[9] We also used this tagged corpus to build a simple table-lookup based POS tagger, in which, bigram and unigram counts (both for text and tags) are used to build the model.

شادیشدہ
No space in between (incorrect)
شادی شدہ
Space in between (correct)
Translation: Married

(a)

میرے شاگرد چائے پیتے ہیں
Translation: My students take tea
میرےشاگردچائےپیتےہیں
Translation: Mystudentstaketea
Both versions are considered correct

(b)

Figure 1: Examples of Space Insertion Problem (a) and Space Omission Problem (b)

	Surface forms	Words analyzed	Coverage
SS Corpus	42,075	27,118	64.5%
PS Corpus	41,947	27,350	65.2%

Table 4: Statistics of Morphological Analysis (and lemmatization) for Urdu Summary Corpus. (Space segmented: SS, Properly Segmented: PS)

a rule based Urdu stemmer. It is available publicly in two forms: A web-based tool and a desktop tool. However, both versions restrict the input to be only one word (manually entered) at a time, making it impossible to use it for a longer text. Therefore, we re-implemented the stemmer and tested it on selected inputs in comparison with the publicly available web-based tool, to ensure the correctness. As a final step, stemming is applied on the articles.

4. Conclusion

Urdu Summary Corpus is a pioneering effort, including 50 articles, corresponding annotated text, and corresponding abstractive summaries to foster the work in Urdu NLP. In order to achieve this, we also re-implemented many NLP related software tools. We provide two versions of the corpus: (1) properly segmented and (2) space segmented. This may allow researchers to measure the effect of proper word segmentation on the language processing of Urdu. The benchmark corpus is small yet pioneering effort in the context of Urdu and it is distributed freely. In future, we plan to increase the corpus size by adding more articles. Currently, there is only one abstractive summary per article. In future, we also plan to increase the number of summaries per article.

5. Bibliographical References

Akram, Q.-u.-A., Naseer, A., and Hussain, S., (2009). *Proceedings of the 7th Workshop on Asian Language Resources (ALR7)*, chapter Assas-band, an Affix-Exception-List Based Urdu Stemmer, pages 40–47. Association for Computational Linguistics.

Butt, M., Sulger, S., Butt, M., and (editors, T. H. K. (2008). Urdu ezafe and the morphology-syntax interface. In *In Proceedings of LFG08*. CSLI Publications.

Davis, M. and Whistler, K. (2014). Unicode normalization forms, unicode standard annex #15. Technical report, The Unicode Consortium, June.

Durrani, N. and Hussain, S. (2010). Urdu word segmentation. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA*, pages 528–536.

Eryiğit, G., Nivre, J., and Oflazer, K. (2008). Dependency parsing of Turkish. *Computational Linguistics*, 34(3):357–389.

Forsberg, M. and Ranta, A. (2004). Functional morphology. In *Proceedings of the Ninth ACM SIGPLAN International Conference of Functional Programming, ICFP'04*, pages 213–223. ACM.

Gulzar, A. (2007). Urdu normalization utility v1.0. Technical report, Center for Language Engineering, Al-kwarzimi Institute of Computer Science (KICS), University of Engineering, Lahore, Pakistan. www.cle.org.pk/software/langproc/urdunormalization.htm.

Humayoun, M., Hammarström, H., and Ranta, A. (2007). Urdu morphology, orthography and lexicon extraction. *CAASL-2: The Second Workshop on Computational Approaches to Arabic Script-based Languages, LSA Linguistic Institute. Stanford University, California, USA.*, pages 21–22. http://www.lama.univ-savoie.fr/ humayoun/UrduMorph/.

Jawaid, B. and Bojar, O. (2012). Tagger voting for urdu. In *Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing*, pages 135–144, Mumbai, India, December. COLING 2012.

Jawaid, B., Kamran, A., and Bojar, O. (2014). A tagged corpus and a tagger for urdu. In Nicoletta Calzolari, et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Leite, D. S., Rino, L. H. M., Pardo, T. A. S., Gracas, M., and Nunes, V. (2007). Extractive automatic summarization: Does more linguistic knowledge make a difference? In C. Biemann, et al., editors, *Proceedings of the HLT/NAACL Workshop on TextGraphs-2: Graph-Based Algorithms for Natural Language Processing*, pages 17–24, Rochester, NY, USA. ACL.

Marlow, S. (2010). Haskell 2010: language and libraries. Technical report, http://www.haskell.org.

Newfields, T. (2001). Teaching summarizing skills: Some practical hints. *ELJ Journal.*, 2(2):1 – 7. www.tnewfields.info/Articles/sum.htm.

Nuzumlal, M. Y. and Özgür, A. (2014). Analyzing stemming approaches for turkish multi-document summarization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 702–706, Doha, Qatar.

Steinberger, J. and Ježek, K. (2012). Evaluation measures for text summarization. *Computing and Informatics*, 28(2):251–275.

Torres-Moreno, J. (2012). Artex is another text summarizer. *CoRR*, abs/1210.3312.

Torres-Moreno, J. (2014). Three statistical summarizers at CLEF-INEX 2013 tweet contextualization track. In

Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014., pages 565–573.

6. Appendix

	Article Name	R1	R2	R3	R4	R5	Avg		Tokens in Article	Tokens in Abstracts	Compression
1	Current Affairs 1	4	5	4	4	5	4.4		236	110	46.6
2	Current Affairs 2	4	5	4	5	3	4.2		841	179	21.3
3	Current Affairs 3	4	5	4	4	4	4.2		361	137	38
4	Current Affairs 4	5	4	4	4	4	4.2		256	129	50.4
5	Current Affairs 5	4	4	4	3	5	4		336	205	61
6	Current Affairs 6	5	4	4	5	5	4.6		312	101	32.4
7	Health 1	4	4	4	4	4	4		477	261	54.7
8	Health 2	4	5	4	5	4	4.4		547	206	37.7
9	Health 3	4	5	4	5	3	4.2		474	201	42.4
10	Health 4	3	4	4	5	5	4.2		307	138	45
11	Health 5	4	5	4	4	3	4		1319	475	36
12	Health 6	4	5	4	4	3	4		288	131	45.5
13	History 1	3	4	4	5	4	4		1047	428	40.9
14	News 1	4	4	4	4	3	3.8		537	300	55.9
15	News 2	4	4	4	4	4	4		914	389	42.6
16	News 3	4	5	4	4	3	4		441	186	42.2
17	News 4	4	4	4	5	4	4.2		629	238	37.8
18	News 5	4	4	3	4	4	3.8		496	235	47.4
19	News 6	4	2	4	5	5	4		178	80	44.9
20	Pakistan	4	5	4	4	5	4.4		760	247	32.5
21	Politics 1	3	5	4	4	5	4.2		550	171	31.1
22	Religion 2	5	5	4	3	5	4.4		367	157	42.8
23	Religion 4	5	5	4	4	4	4.4		743	313	42.1
24	Religion 5	5	5	4	4	3	4.2		517	195	37.7
25	Religion 6	5	5	3	4	5	4.4		796	270	33.9
26	Science&IT 1	4	4	4	4	3	3.8		338	217	64.2
27	Science&IT 2	5	5	4	4	5	4.6		988	313	31.7
28	Sports 1	4	5	4	5	4	4.4		494	164	33.2
29	Sports 2	4	4	4	4	4	4		394	185	47
30	Sports 3	5	4	4	4	4	4.2		726	342	47.1
31	Sports 4	4	4	4	4	3	3.8		1026	383	37.3
32	Sports 5	3	4	3	4	5	3.8		419	151	36
33	Sports 6	4	4	4	4	3	3.8		454	177	39
34	Sports 7	4	4	4	4	4	4		301	138	45.8
35	Sports 8	4	5	4	4	4	4.2		686	254	37
36	Sports 9	4	4	4	4	3	3.8		975	322	33
37	Sports 10	5	5	4	4	4	4.4		456	227	49.8
38	Technology 1	5	4	4	4	5	4.4		655	256	39.1
39	Technology 2	5	5	4	4	4	4.4		353	166	47
40	Technology 3	5	5	4	3	5	4.4		1337	353	26.4
41	Technology 5	3	3	4	5	5	4		346	146	42.2
42	Technology 6	5	5	4	5	5	4.8		318	120	37.7
43	Technology 7	4	5	4	4	4	4.2		744	326	43.8
44	Technology 8	5	4	4	4	4	4.2		216	118	54.6
45	Technology 9	4	5	4	5	5	4.6		189	103	54.5
46	Technology 10	5	4	5	4	3	4.2		234	129	55.1
47	Technology 11	4	4	4	4	5	4.2		159	79	49.7
48	Tourism 1	5	4	4	3	3	3.8		525	282	53.7
49	Tourism 2	5	5	4	4	4	4.4		991	351	35.4
50	Tourism 3	3	4	4	4	4	3.8		2518	761	30.2