Urdu Summary Corpus

Single-document Abstracts for Text Summarization

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

Introduction

- Language resources, such as benchmark corpora are important for various natural language processing tasks.
- Urdu is widely spoken but it is under-resourced in terms of standard evaluation resources.
- This work reports the construction of a benchmark corpus for Urdu summaries (abstracts) to facilitate the development and evaluation of single document summarization systems.
- A pioneering effort.

Contribution

Urdu Summary Corpus

- 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

Two versions of the corpus

- Space does not always mark word boundary in Urdu.
- Therefore, we created two versions of the same corpus.
  - Version 1: words are separated by space.
  - Version 2: proper word boundaries are manually tagged.

Software Package

- A normalizing utility
- A POS tagger
- A table-lookup based morphological analyzer and lemmatizer built from Humayoun's lemmatizer [2]
- A stemmer which is self-implemented from Assas-Band stemmer [1]

The source code these software tools are provided in one package. Such pre-processing tools are scarce for Urdu. In addition, these tools sometimes require fine-tuning or re-implementing because of the occasional updates (if any at all). Therefore, we think that providing such utilities are beneficial for researchers working on Urdu.

Urdu Summary Corpus and the source-code of the tools are made freely available at

https://github.com/humsha/USCorpus/

Urdu Language

1. Urdu is an Indo-Aryan language, widely spoken in South Asia and all over the world (due to the large South Asian Diaspora).
2. It has more than 100 million speakers.
3. It is written in a modified Perso-Arabic script from right to left.
4. 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.
5. Urdu has a complex morphology that inherits grammatical forms and vocabulary from Arabic, Persian, and native languages of South Asia.
6. There is no capitalization.
7. Diacritics (vowels) are hardly present in the text and words are guessed with the help of the context of surrounding words.
8. It has a free word order (Subject Object Verb).

An Example Sentence

اُردو پاکستان کی قومی زبان ہے ۔
Urdu is the national language of Pakistan.

Summary Corpus Creation

Articles 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:

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

- A group of volunteers was selected for the summary writing.
- They were native speakers of Urdu:
  - Academicians teaching Urdu in colleges
  - University students that have an interest in Urdu literature
- No size restriction was imposed for the human-written summaries but the summary must not exceed half the size of an article. (As it turns out, not restricting the size of the summaries makes it difficult to compare with DUC summary datasets.)

Three basic steps for summary creation: Influenced by the six editing operations in human abstracting [3]:

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.

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 [3], 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.

	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%

Statistics of articles and Human-Written abstracts

Preprocessing Challenges

Normalization of Urdu Summary Corpus Urdu script is an extension of Persian and Arabic script, and because of this, some characters got assigned multiple Unicode for these similar scripts, resulting in orthographic variations. Therefore, normalization was performed. Diacritic marks present in the collection were also removed in this step.

Word Segmentation for Urdu Summary Corpus Two types of word segmentation issues: (1) space insertion (2) space deletion. Urdu faces both of them. It means that Urdu words cannot always be separated by space.

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

(a) Space Insertion Problem

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

(b) Space Omission Problem

Token difference of two versions of Urdu Summary Corpus:

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

Morphological Analysis and Lemmatization We used Urdu Morphological Analyzer [2] for both. In this open-source tool, the lexicon contains $5,000$ words, capable of handling $140,000$ word forms. This tool has not been 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 and applied on the Urdu Summary Corpus, and made the tool and tagged corpus publicly available.

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

Coverage on Urdu Summary Corpus, SS=Space Segmented, PS=Properly Segmented

Stemming Assas-band [1] is 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. This self-implemented version is provided along-with the Urdu Summary Corpus. As a final step, stemming is applied on the articles.

Conclusion

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. We also plan to increase the number of summaries per article.

References

[1] Qurat-ul-Ain Akram, Asma Naseer, and Sarmad Hussain. *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, 2009.

[2] Muhammad Humayoun, Harald Hammarström, and Aarne Ranta. 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, 2007. http://www.lama.univ-savoie.fr/ humayoun/UrduMorph/.

[3] Josef Steinberger and Karel Ježek. Evaluation measures for text summarization. *Computing and Informatics*, 28(2):251–275, 2012.