# Urdu Text Summarizer using Sentence Weight Algorithm for Word Processors

Aqil Burney, Badar Sami, Nadeem Mahmood, Zain Abbas
Department of Computer Science,
University of Karachi, Karachi, Pakistan

Kashif Rizwan
Department of Computer Science & IT,
Federal Urdu University, Karachi, Pakistan

## ABSTRACT

For the past few years, a lot of work has been done in Urdu text processing. However, many areas are still open for research in Urdu text processing. Microsoft® has done a lot of work in text processing in its product MS Word®. It also supports Urdu Language but the major drawback is that many of these features do not support Urdu text. One such feature is “Auto Summarize Tool”. In this paper, we present an Add-in “Auto Summarizer for Urdu language” for MS Word. The Add-in has been specially designed to summarize of news, informative articles such as scientific writings, economical items, and sports commentary.

## Keywords

Text summarizer, Urdu language processing, sentence weight algorithm, Urdu summary, Urdu Stop words.

## 1. INTRODUCTION

A variety of word processors are available in the market to carry out desktop publishing tasks. Due to user friendliness, features and simplicity, MS Word® [1] has emerged as the leading tool used for desktop publishing and document processing. It is due to the extensive usage of MS Word that we have chosen it to carry out this research.

MS Word was first released in 1983 under the name Multi-Tool Word [2] for Xenix systems. Consequent versions were later written for many other platforms including IBM PCs running DOS in 1983, for Apple Macintosh in 1984, for AT&T Unix PC in 1985, for Atari ST in 1986, and for MS Windows® in 1989. MS Word is a vital part of the Microsoft Office® suite [3] along with several other applications. MS Word has a built in “Auto-summarize” feature. This feature highlights phrases that it considers valuable in any document. The length of the summary generated by this feature can be specified by the user as a percentage of the amount of text present in the original document.

Urdu is spoken by around 100 million people around the world, predominantly in Pakistan and India [7] [8] [15]. It is the state language of Pakistan. It belongs to the Indo-Aryan branch of the family of Indo-European languages. Urdu is closely related to Hindi and number of Hindi-Urdu speakers in world is over 500 million. It has borrowed a great deal of vocabulary from Persian, Arabic and Sanskrit languages. Urdu language contains 38 alphabets, 25 consonants and 12 vowels.

A summary, synopsis, or recap is a shorter version of the original. It highlights the major points from a longer text, speech, or event. The purpose of summarization is to facilitate the audience or reader in comprehending the essence of the document in a short period of time. Summaries are supposed to be written in a balanced and objective way, mirroring the genre’s aim by portraying original from the author’s point of view. Nonfiction summaries generally do not offer analysis or assessment. Summarizers generate these condensed versions by arranging the text distillations in a syntactic manner. They exclude unessential examples, descriptions and digressions. The opening sentence introduce the topic, whereas the final sentence sum up the theme, taking into account and building upon the knowledge gained from the body of the text.

In recent years, a summarizing industry has sprung up. Leading firms are working on building summarization software mainly concentrating on business literature. Although they adhere to the nonfiction guidelines mentioned above, but they also provide numerical ratings and evaluations of the titles covered. Shorter, more concise nonfiction summaries are often referred as abstracts. Abstracts may also be generated through the summarization software.

Auto summarizing applications are generally developed in the form of plug-in that can be readily associated with word processors. Programmers typically implement plug-in using libraries installed in an area prescribed by the host application. Generally all word processing applications support plug-in functionality. There are many advantages of plug-in; some of them are listed below [1]:

- Enable third party designers to develop feature for desktop publishing applications through extension.
- Facilitate addition of new features.
- Reduce the size of the parent application.
- Separate source code from application because of incompatible software licenses.

According to Microsoft Word 97 team [4], Auto-summarize feature works by counting words and ranking sentences in any document. It does so by identifying the most frequent words in the document (excluding “a”, “the” and such words). It then assigns a “score” to each word in a way that the most common word receives the highest score and so on. It then “averages” sentences by adding scores of the constituent words and dividing sum by the number of words in that sentence.

In this paper, we present an Add-in that does auto-summarization work for document in Urdu language. The

The rest of the paper is organized as follows: Section 2 reviews some related work carried out in summarization. Section 3 describes the Sentence Weight algorithm. Section 4 presents a detailed list of stop words in Urdu language. Section 5 summarizes the results produced by our Add-in.

Section 6 concludes the discussion which is followed by acknowledgment and references.

# 2. RELATED WORK

Different techniques have been proposed for auto-summarization of text [7] [11] [13] [14]. Each technique uses a different approach for text distillation. Some of these techniques are discussed below:

## 2.1 Luhn's method

Luhn's method [7] [8] [9] focus technical literature for summarization. It takes into consideration the frequency and relative position of the significant words. However, it doesn't account for the semantics of those words [8]. It is based on the assumption that frequency of occurrence of a word is a useful measurement of its significance in an article. The method was originally proposed to work with limited capability machines so semantic information pertaining to the words wasn't taken into account. It is a simple and straightforward algorithm that is economical to implement but has high time complexity. Luhn's method is useful in situations where insignificant words have low frequency or high (e.g. "the", "and"). In such situations it becomes easy to remove such words. Similarly, minimum and maximum frequency threshold can be set and finally comparison with common word list may be done.

## 2.2 Weighting methods:

Edmundson [12] developed four methods for computing weights for technical literature. These methods are

- Cue Method
- Key Method
- Title Method
- Location Method

Four weights are computed using these algorithms. The weight of each sentence is calculated through linear combination of these weights. The sentences with highest weights are included in the summary or abstract.

## 2.3 Naive-Bayes Method

Kupiec [11] described a method that was derived from [10]. The method was able to learn from data. The classification function used by the method categorized each sentence as worthy of extraction or not, using a Naive-Bayes classifier. The features provided by the method were compliant with the weighting methods with the additional capability of handling sentence length and uppercase letters.

# 3. SENTENCE WEIGHT ALGORITHM

Sentence weight algorithm is a statistical method, in which each sentence in the text is given a specific weight or rank to decide its inclusion in the summary. This algorithm is used by MS Word in "Auto Summarize" feature to generate summary of English text. The listing of the algorithm is provided is section 3.3.

## 3.1 Stop words

A stop word [13] is generally a token in any language that does not have any linguistic meaning. For example, in English "of", "is", "an", and "the" etc. are stop words. Stop words are small, simple words that make a sentence grammatically correct. They give a correct form or "structure" to any sentence. If the stop words are removed from a sentence, the meaning of the sentence is still understandable.

## 3.2 Content words

Content words in any text are the words that have meaning. Rather than indicating a syntactic function, these words have a state able lexical meaning. These include words such as nouns, verbs, or adjectives. Content words are the keywords of any sentence. Without the content words any sentence will lose its connotation and sense. An example of content and stop words is provided in Fig. 1.

## 3.3 Algorithm

The algorithm of calculating weight of sentences is listed below:

i. Calculate the total words.

ii. Find all the stop words.

iii. Calculate the content words by subtract stop words from the total words using Eq. (1).

iv. Calculate weight percentage using Eq. (2).

v. Sort the sentences with respect to percentages in descending order.

vi. Pick the required number of sentences after sorting

vii. Again sort the sentences selected in order of occurrence in original document to get the summarized document.

The Content words in step (iii) are of the algorithm are computed using the formula

$$\textit{Content Words} = \textit{Total word} - \textit{Stop words} \qquad (1)$$

where Content words, Total Words and Stop words are the respective count of the words for a sentence.

The sentence weight in the step (iv) of the algorithm is calculated using the following formula

$$\textit{Sentence Weight} = \frac{\textit{Content words}}{\textit{Total words}} * 100 \qquad (2)$$

Where, Sentence Weight is the weight of a single sentence, Content words is the number of content words in that sentence and Total Words is number of words in the sentence.

# 4. URDU STOP WORDS

As discussed previously, stop words are functional words of a language and meaningless in context of text classification. They are eliminated from the lexicon in order to reduce its size by using a list of most frequent words known as Stop Word list. A lot of research work has been carried out in English language to find stop words and more than 400 stop words have been identified.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Content Words | | | | | | | | | |
| Will | you | SELL | my | CAR | because | I've | GONE | to | FRANCE. |
| Stop Words | | | | | | | | | |

**Fig. 1: Content words and Stop words in an English sentence**

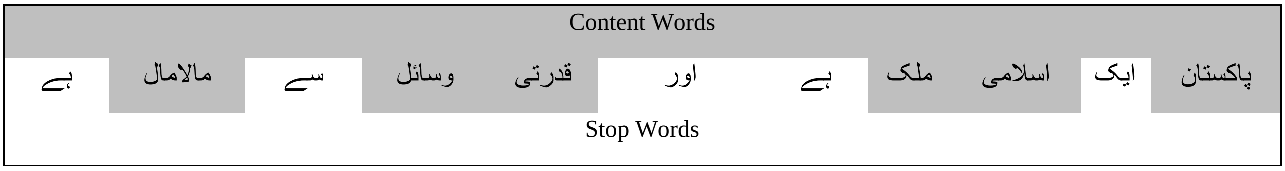

**Fig. 2: Content words and Stop words in an Urdu sentence**

Like English, Urdu language has stop words as well, for example "کا", "ہے", "ایک", "گا" etc. However, no considerable work has been carried out to find the stop words in Urdu language. In order to overcome this problem, we have collected English stop words and translate them into Urdu. An example of content and stop words in Urdu in provided in Fig. 2 and detailed list of the translated stop words is presented in Table 1.

**Table 1: Stop words list**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| ابھی | جو | گی | اپنے | گئے | بہت | طرف | ہماری | پائے |
| اپنا | دوسری | گیا | کب | گا | بھی | سے | ہر | پر |
| اس | دی | گے | لگیں | ہے | بعد | سکتے | وہ | تھی |
| ان | دیا | لئے | والے | یہ | بجائے | سکتی | نے | تھا |
| اندر | ذریعے | لگی | ہمارا | ہونے | باہر | سکتا | نہیں | تو |
| اور | رہا | لگے | ہوسکتا | ہوں | کا | ہمارے | تمام | کیا |
| ایسے | رہی | مگر | ہوسکتی | ہیں | کریں | ہو | تک | کی |
| ایک | رہے | میں | ہوسکتے | کیسے | ہونا | تب | کہ | ہوا |
| آئے | ساتھ | نا | تھے | کیوں | ہوتا | نہ | جب | کے |
| پھر | بغیر | جارہے | رکھ | کیسا | کوئی | ذریعے | بارے | جا |
| اسطرح | بلکہ | جبکہ | رکھتا | کیطرف | براں | جارہا | ذریعہ | کسی |
| اسکا | بند | جس | رکھتاہوں | کیلئے | بائیں | تمہیں | دوسرے | کررہی |
| اسکی | بیچ | جوکہ | رکھتی | کیونکہ | دونوں | کررہے | جارہی | برآں |
| اسکے | پچھلا | جیسا | رکھتے | کےبعد | تمہی | دوران | کررہا | یہاں |
| ◌ٓآس | انہیں | بن | پسند | تھوڑا | چکے | حکمیہ | دوسروں | سکا |
| اب | اونچا | بنا | پل | تھوڑی | چلا | خاموش | دیتا | سکنا |
| اجازت | اونچائی | بنارہا | پوچھا | تھوڑے | چلو | ختم | دیتی | سکی |
| اچھا | اونچی | بنارہی | پوچھتا | تین | چلیں | در | دیتے | سکے |
| اچھی | اونچے | بنارہے | پوچھتی | جانا | چلے | درجات | دیر | سلسلہ |
| اچھے | اٹھانا | بنانا | پوچھتے | جانتا | چھوٹا | درجہ | دیکھنا | سوچ |

**Table 2:** **Stop words list (continued)**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| اختتام | اہم | بند | پوچھنا | جانتی | چھوٹوں | درجے | دیکھو | سوچا |
| ادھر | آئی | بندکرنا | پوچھو | جانتے | چھوٹی | درحقیقت | دیکھی | سوچتا |
| ارد | آئے | بندکرو | پوچھوں | جاننا | چھوٹے | درست | دیکھیں | سوچتی |
| اردگرد | آج | بندی | پوچھیں | جسطرح | چھہ | دس | دینا | سوچتے |
| ارکان | آخر | بڑا | پورا | جگہ | چیزیں | دفعہ | دے | سوچنا |
| استعمال | آخرکار | بڑوں | پہلا | جگہوں | حاصل | دکھائیں | راستوں | سوچو |
| استعمالات | آدمی | بڑی | پہلی | جگہیں | حاضر | دکھاتا | راستہ | سوچی |
| اشیا | آنا | بڑے | پہلےسی | جلدی | حال | دکھاتی | راستے | سوچیں |
| اطراف | آٹھ | بھر | پہلےسے | جناب | حال | دکھاتے | رکن | سیدھا |
| افراد | آیا | بھرا | پہلےسےہی | جوان | حالات | دکھانا | رکھا | سیدھی |
| اکثر | با | بھراہوا | پیش | جونہی | حالیہ | دکھاو | رکھی | سیدھے |
| اکٹھا | باترتیب | بھرپور | تازہ | جیساکہ | حصوں | دکھایا | رکھے | سیکنڈ |
| اکٹھی | باری | بہتر | تر | چار | حصہ | دلچسپ | زیادہ | شاید |
| اکٹھے | بالا | بہتری | ترتیب | چاہا | حصے | دلچسپی | سات | شخص |
| اکیلا | بالترتیب | بہترین | ترین | چاہنا | حقائق | دلچسپیاں | سادہ | شد |
| اکیلی | برس | پاس | تعداد | چاہے | حقیتیں | مناسب | سارا | شروع |
| اکیلے | بغیر | پانا | تقریباً | چکا | حقیقت | دو | سارے | شروعات |
| اگرچہ | بلند | پانچ | تم | چکی | حکم | دور | سال | شے |
| الگ | بلندوبالا | پرانا | تنہا | چکیں | حکماً | دوسرا | سالوں | صاف |
| صحیح | قبیلہ | کونسے | لازمی | مسئلے | نیا | طریق | کرتی | کہتے |
| صفر | قسم | کھولا | لگتا | مسائل | وار | طریقوں | کرتے | کہنا |
| صورت | کئی | کھولنا | لگتی | مستعمل | وار | طریقہ | کرتےہو | کہنا |
| صورتحال | کئے | کھولو | لگتے | مشتمل | ٹھیک | طریقے | کرنا | کہو |
| صورتوں | کافی | کھولی | لگنا | مطلق | ڈھونڈا | طور | کرو | کہوں |
| صورتیں | کام | کھولیں | لگی | معلوم | ڈھونڈلیا | طورپر | کریں | کہی |
| ضرور | کبھی | کھولے | لگے | مکمل | ڈھونڈنا | ظاہر | کرے | کہیں |
| ضرورت | کرا | کہا | لمبا | ملا | ڈھونڈو | عدد | کل | کہیں |
| ضرورتاً | کرتا | کہتا | لمبی | ممکن | ڈھونڈی | عظیم | کم | کہے |
| ضروری | کرتاہوں | کہتی | لمبے | ممکنات | ڈھونڈیں | علاقوں | کمتر | کیے |

**Table 3:** **Stop words list (continued)**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| لمحات | ممکنہ | ہم | لے | ناپسند | ہورہے | علاقہ | کمرا | کےذریعے |
| لمحہ | مڑا | ہوئی | متعلق | ناگزیر | ہوگئی | علاقے | کمروں | گئی |
| لو | مڑنا | ہوئے | محترم | نسبت | ہوگئے | علاوہ | کمرہ | گرد |
| لوگ | مڑے | ہوتی | محترمہ | نقطہ | ہوگیا | عموماً | کمرے | گروپ |
| لوگوں | مہربان | ہوتے | محسوس | نکالنا | ہونی | عمومی | کمسن | گروہ |
| لڑکپن | میرا | ہوچکا | مختلف | نکتہ | ہی | فرد | کون | گروہوں |
| لی | میری | ہوچکی | مزید | نو | یقیناً | فی | کونسا | گنتی |
| لیا | میرے | ہوچکے | مسئلہ | نوجوان | یقینی | قبل | کونسی | لازماً |
| لینا | نئی | ہورہا | لیں | نئے | ہورہی | باعث | سب | |

**Table 4:** **List of documents used for summarization**

| S. No. | Title | Type | Reference |
|---|---|---|---|
| 1 | جنوبی کوریا کی زبردست جنگی مشقیں | News | [16] |
| 2 | ہمالیہ سے بلند پاک چین دوستی | News | [16] |
| 3 | تھر کا کوئلہ، پانچ سو برس تک بجلی | News | [16] |
| 4 | حج انتظامات میں کرپشن | News | [16] |
| 5 | برف کے دور میں بھی زندگی پنپ رہی تھی | Article | [16] |
| 6 | ترکی اسرائیل تعلقات بھتری کے امکانات | News | [16] |
| 7 | پچھلے تین سال میں کابل ترقی کی راہ پر | Article | [16] |
| 8 | پاکستان: شراب کی قیمتوں میں اضافہ | News | [16] |
| 9 | جیکسن کے کیرئر کے نشیب و فراز | News/ article | [16] |
| 10 | سقراط نے زہر کا جام پینا کیوںپسند کیا؟ | Fiction article | [17] |
| 11 | دنیا کا پہلا فلسفی | Fiction article | [17] |
| 12 | آداب و اطوار از علی عباس جلالپوری | Article | [17] |
| 13 | کیا اسلام اور سائنس میں تضاد ہے | Research article | [17] |
| 14 | خواب کی حقیقت ؟؟؟ | Article | [17] |
| 15 | کیا ماضی کی سائنس زیادہ ترقی یافتہ تھی؟ از ڈاکٹر مطلوب حسین۔ | Article | [17] |
| 16 | زندگی کی پیدا ئش: خدا کی تخلیق یا ارتقا؟ | Research art | [17] |
| 17 | پرہیز علاج سے بہتر ہے | Article | [17] |
| 18 | لیموں قدرت کا انمول تحفہ | Article | [17] |
| 19 | آم :گرمیوں کا سب سے مقبول پھل | Article | [17] |
| 20 | دنیا دوبارہ جڑی بوٹیوں کے فطری اور بے ضرر علاج کی طرف متوجہ ہور ہی ہے | Article | [17] |

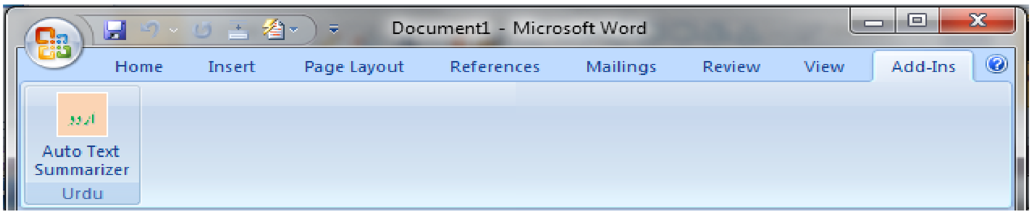

Fig. 3: Main menu of the Add-in (Auto Summarizer Control)

## 5. RESULTS

Once the Auto-Summarizer was complete, we integrated it as Add-in to the MS Word main menu as shown in fig. 3. To verify the accuracy of the add-in twenty different types of documents were selected. The documents include News, articles, informative Articles, and fiction Articles that were collected as random from [15] [16]. List of the documents collected is provided in Table 4.

It was observed that the summary generated by the summarizer is well formed and easy to understand. More importantly, the summary generated especially very close to the original in case of News articles. The result was cross checked by human verification. Five human verifiers were used for this purpose. The process clearly indicated that the sentences selected in the summary by the summarizer, also selected by at least one of the human verifiers generally. To be precise, if one verifier was selected the result was about 80% accurate. If two verifiers out of five were selected, the accuracy was over 50%.

The decreasing value of accuracy in case of increasing the number of verifiers is due to the fact that all the verifiers have their own perception regarding each document and all of the generate a different summary. The human verification facts pertaining to one document are provided below:

- The original document contained 718 words, 3415 characters in 47 lines.
- The summary generated by Add-in was about 25% of original document. It contained 139 words, 698 characters in 10 lines.
- Summary generated by the verifiers contained about 12 lines.
- There were 7 common lines between the human and auto summaries.
- The similarity result stood around 64%.

## 6. CONCLUSION AND FUTURE WORK

The work presented here is an initial step towards building a true Urdu summarizer. In the technique presented here, our main focus remained in the elimination of stop words as we had used the Sentence Weight algorithm.

Many advanced algorithms for summarization of text have been developed and are waiting to be implemented. Some algorithms also eliminate the words that are rarely used in the text. Others propose to enhance the overall accuracy by calculating the frequency of content words that exist in the document title, and rating the sentences containing these words higher. In future, we plan to implement such advanced algorithms to improve the efficiency of the Auto-Summarizer.

## 7. ACKNOWLEDGEMENTS

We appreciate the efforts of Basharat ullah, Saima Asghar and Waseem of Department of Computer Science, University of Karachi, Pakistan in research and development of the Add-in.

## 8. REFERENCES

[1] Edwards, B. 2008. Microsoft Word Turns 25. *PC World*.

[2] Allen, A. R. 2001. Microsoft in the 1980’s. In *A History of the Personal Computer: The People and the Technology* (pp. 12/25–12/26). Allan Publishing.

[3] Office 2010 Availability. Microsoft Office 2010 Engineering. Microsoft. 2010.

[4] Gore, Karenna. 1997. Cognito Auto Sum. Retrieved from http://www.slate.com/id/2419

[5] Urdu language - Britannica Online Encyclopedia. Retrieved from http://www.britannica.com/EBchecked/topic/619612/Urdu-language.

[6] Durrani, N., and Hussain, S. 2010. Urdu Word Segmentation, Human Language Technologies: *The 2010 Annual Conference of the North American Chapter of the ACL*, pp. 528–536.

[7] Luhn, H. P. 1958. The Automatic Creation of Literature Abstracts, *IBM Journal*, pp. 159-165.

[8] Wieling, M., and Groningen, R. 2004. *Automatic Text Summarization: A Solid Base*, Presented paper.

[9] Das, D., and Martins, F.T. 2007. A Survey on Automatic Text Summarization. *Literature Survey for the Language and Statistics II course at CMU*.

[10] Edmundson, H.P. 1968. New methods in automatic extracting. *Journal of the Association for Computing Machinery*, 16(2), pp. 264-285.

[11] Kupiec, J., Pedersen, J., and Chen, F. 1995. A trainable document summarizer. In *Proc. SIGIR '95*, pp. 68-73. NY. USA.

[12] Russell, S. J., and Norvig, P. 1995. *Artificial intelligence: a Modern Approach*, Englewood Cliffs, NJ: Prentice-Hall International Inc.

[13] Hussain, S. 2008. Resources for Urdu Language Processing. *The 6th Workshop on Asian Language Resources.*

[14] Patel, A., Siddiqui, T., and Tiwary, U.S. 2007. A language independent approach to multilingual text summarization. *Conference RIAO2007*, Pittsburgh PA, U.S.A.

[15] ‘A Guide to Urdu’, Retrieved from http://www.bbc.co.uk/languages/other/urdu/guide/

[16] News and research articles retrieved from http://www.bbc.co.uk/urdu

[17] Research articles retrieved from http://www.urduweb.org