Kazakh text summarization using Fuzzy logic

Altanbek Zulkhazhav[⋆], Zhanibek Kozhirbayev[⋆†],
Zhandos Yessenbayev[†], Altynbek Sharipbay[⋆]

[⋆] Faculty of Information Technologies,
L.N.Gumilyov Eurasian National University, Kazakhstan
[†] National Laboratory Astana, Nazarbayev University, Kazakhstan
altinbekpin@gmail.com, {zhanibek.kozhirbayev,zhyessenbayev}@nu.edu.kz,
sharalt@gmail.com

Abstract. In this paper we present an extractive summarization method for the Kazakh language based on fuzzy logic. We aimed to extract and concatenate important sentences from the primary text to obtain its shorter form. With the rapid growth of information on the Internet there is a demand on its efficient and cost-effective summarization. Therefore the creation of automatic summarization methods is considered as a very important task of natural language processing. Our approach is based on the preprocessing of the sentences by applying morphological analysis and pronoun resolution techniques in order to avoid their early rejections. Afterwards, we determine the features of the processed sentences need for exploiting fuzzy logic methods. Additionally, since there is no available data for the given task, we collected and manually annotated our own dataset from the different Internet resources in the Kazakh language for the experimentation. We also applied our method on CNN/Daily Mail dataset. The ROUGE-N indicators were calculated to assess the quality of the proposed method. The ROUGE-L(f-score) score by the proposed method with pronoun resolution for the former dataset is 0.40, whereas for the latter one it is 0.38.

Key words: extractive text summarization, natural language processing, fuzzy logic.

1 Introduction

With the rapid growth of information on the Internet, it becomes extremely difficult for users to get what they really intend. Therefore, the creation of automatic summarization methods is considered as a very important task of natural language processing. This allows the user to quickly understand large amount of information. Automatic text summarization is a task to extract the most important part of the source text in a shorter way. It can be performed in two ways: extraction and abstraction. The extraction method selects sentences or phrases that have high marks of importance, and combines them into a new short text, without changing the selected units. In the method of abstraction, the main content is extracted from the source text and paraphrased using linguistic methods for semantic analysis and text interpretation.

In this work we experiment with extractive summarization method for the Kazakh language based on fuzzy logic. We collected and manually annotated our own dataset from the different Internet resources in the Kazakh language. Additionally, we conduct experiment on CNN/Daily Mail dataset [1], which is an open dataset for use in text summarization experiments in English. Our approach is based on the preprocessing of the sentences by applying morphological analysis and pronoun resolution techniques. Afterwards, we determine features to extract important sentences from the text. The architecture of the extractive text summarization approach based on fuzzy logic is shown in Figure 1.

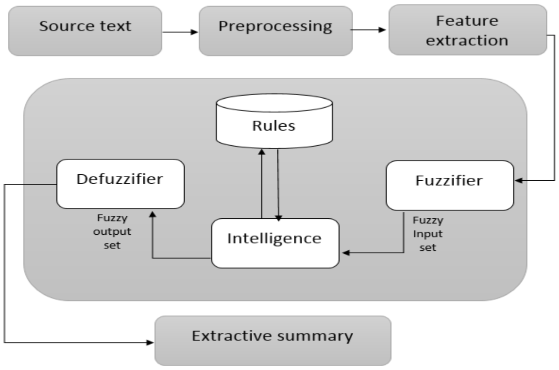

Fig. 1. Fuzzy logic system architecture for extractive text summarization.

This paper is organized as follows: Section II presents a brief review of the related works. Section III presents a methodology of proposed techniques for automatic text summarization. More precisely, it describes the following stages: text preprocessing, feature extraction and calculation, summarization using fuzzy logic. The dataset collection and experimental setup are described in Section IV. Experiment details and obtained results of the extractive text summarization task are also presented in this section. Summary of the performed experiments and areas of further research are given in Section V.

2 Related work

This Section presents a brief review about automatic text summarization techniques. Many different techniques have been proposed for this task that utilizes a variety of different approaches. For a thorough review of works on text summarization the reader is advised to consult a very recent survey by [2].

We limit ourselves to a brief review of extractive text summarization, as our main aim is to score and select text units which have highest scores as summary. Shallow features [3], hidden markov models [4], discourse structure model [5], maximum marginal relevance [6], fuzzy logic [7] and swarm intelligence [8], conceptual graphs [9, 10] approaches are proposed to deal with this task.

A lot of research has been done with respect to the Kazakh language [11–16], but there are a few researches regarding automatic text summarization. [17] implemented and compared different summarization techniques based on TextRank algorithm, namely: General TextRank, BM25, LongestCommonSubstring. They conducted experiments on corpora of news articles parsed from the web written in Russian and Kazakh. [18] performed an experiment to summarize articles from online news websites tengrinews.kz with extractive way.

3 Methodology

3.1 Text prepossessing

In this work we experiment with news articles of the most popular Kazakhstani news websites. To prepare our collected data set for the summarization task, we preprocess the text by deleting unnecessary indents, spaces, punctuation marks and other specific characters as described in [19, 20]. Afterwards, we perform a segmentation similar to [21], which involves a breakdown of the text into sentences and tokenization of each sentence. The next steps in the preprocessing pipeline include such tasks as lemmatization, numerals identification, named entity recognition and finding pronouns. For these we compiled a morphologically dictionary and devised the empirical rules. Finally, we developed a rule-based algorithm for the pronoun resolution which for each found pronoun basically scans several previous sentences and calculates the most probable word or phrase it refers to. This is necessary to improve the quality of summarization, since pronouns such as "*I*", "*he*", "*she*", "*they*" usually are referred to "stop words" and, thus, removed from the text in the early stages. However, they indicate specific persons and carry certain significance. As a result of morphological and syntactic analysis, pronouns were replaced with the names of the persons they indicate. The pseudo-code of the algorithm used to pronoun replacement is illustrated in Figure 2. We remove stop words that are often found in the text, but do not represent a special meaning for determining the importance of the content. Deletion of affixes from a word by stemming concludes the text preprocessing.

3.2 Feature extraction and calculation

After preprocessing the text, it is necessary to extract features and calculate the functions of the sentence, the results of which are vectors of seven elements for each sentence. The elements of each vector take values in the interval [0, 1]. We consider the following features:

- **Title feature (F1):** It is defined as a ratio of the number of matches of the Title words (Tw) in the current sentence (S) to the number of words (w) of the Title (T) [22]:

$$F1(S) = \frac{Number\ of\ Tw\ in\ S}{Number\ of\ words\ in\ T} \quad (1)$$

Algorithm 1: Pronoun replacement

```
Input: text
Result: text with proper nouns in place of pronouns
for sentence in text.sentences:
    for word in sentence.words
        if isPronoun(word)):
            antecedents =findAllAntecedentsFromText(text, word)
            candidate = chooseMostSuitableCandidate(antecedents)
            word = replace (word, candidate)
        end
     end
end
return text
```

Fig. 2. Algorithm for pronoun replacement.

- **Sentence Length (F2):** It is defined as a ratio of the number of words (w) in the current sentence (S) to the number of words in the longest sentence (LS) in the text [22]:

$$F2(S) = \frac{Number\ of\ w\ in\ S}{Number\ of\ w\ in\ LS} \tag{2}$$

  This function is necessary for filtering from the selection of short and incomplete sentences, such as an author of the article, date of the article, etc.
- **Sentence position (F3):** It is defined as a maximum of the next two relations [22]:

$$F3(S) = max(\frac{1}{Position\ of\ S}; \frac{1}{Number\ of\ S\ -\ Position\ of\ S\ +\ 1}) \tag{3}$$

  If the sentence is at the beginning of the text, then the first expression is the maximum, if the sentence is at the end of the text, then the maximum value will be taken by the second expression. This function is important when selecting, as more informative sentences are usually located at the beginning or at the end of the text.
- **Thematic word (F4):** It is defined as a ratio of the number of thematic words (Thw) in the current sentence (S) to the maximum number of thematic words (Thw) calculated on all sentences (S) of the text [22]:

$$F4(S) = \frac{Number\ of\ Thw\ in\ S}{max(Number\ of\ Thw\ in\ all\ S)} \tag{4}$$

  Thematic words are the most frequently used words in the text. They are directly related to the main theme of the text. We chose the five most frequent words in the text as thematic ones.
- **Term Weight (F5):** It is defined as a ratio of the sum of the frequencies of term occurrences (TO) in a sentence (S) to the sum of the frequency of term occurrences in the text [22]:

$$F5(S) = \frac{\sum(Frequency\ of\ TO\ in\ S)}{\sum(Frequency\ of\ TO\ in\ all\ S)} \tag{5}$$

To calculate the weight of a sentence, we find the frequency with which the term appears in the sentence and the frequency with which the same (current) term appears in the text

- **Proper Noun (F6):** It is defined as a ratio of the number of proper nouns (PN) in a sentence (S) to the length (L) of a sentence [22]:

$$F6(S) = \frac{Number\ of\ PN\ in\ S}{L\ of\ S} \tag{6}$$

Proper nouns found in the proposal carry a lot of information about personal facts. Therefore, the sentences with the most proper nouns are an important part of the content.

- **Numerical Data (F7):** It is defined as a ratio of the number of numerical data (ND) in the sentence (S) to the length (L) of the sentence [22]:

$$F7(S) = \frac{Number\ of\ ND\ in\ S}{L\ of\ S} \tag{7}$$

Typically, numerical data has specific important values for summarization. Therefore, numerical data in the text could not be skipped.

The importance of sentences regarding features is presented in Table 1

Table 1. Importance of sentences (rule examples).

Features	**Low**	**Medium**	**High**
topic/title	poor	average	decent or good
thematic word	poor	average	decent or good
term freq.	poor	average or mediocre	decent or good
proper noun	-	-	good
numerical data	-	-	not(poor)
sentence length	-	-	not(poor)
sentence position	-	-	not(poor)

3.3 Fuzzy logic system design

Fuzzy logic system design includes the following concepts: fuzzy set, membership function, fuzzy logic operations, linguistic variables, linguistic terms, fuzzy logical values, fuzzy logic conclusion [23]. A typical fuzzy logic system consists of the following components:

- fuzzifier;
- logical conclusion on the base of fuzzy knowledge;

– defuzzifier.

The fuzzifier determines a correspondence between the clear numerical value of the input variable and the value of the membership function of the corresponding term of the linguistic variable. In our case, the linguistic variables are the seven functions defined by us above. They take meanings from a variety of words such as “poor”, “mediocre”, “average”, “decent”, “good”. These words are called term-sets and take values in the interval [0,1] (Figure 3). In short, fuzzification is the process of transition from a clear to a fuzzy representation [23].

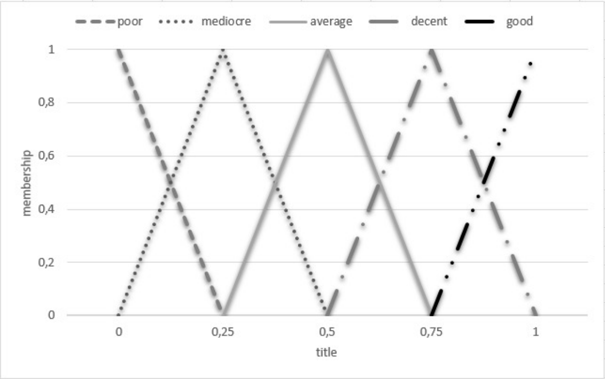

Fig. 3. Linguistic variable “Title feature”.

The fuzzifier depends on the membership function for the corresponding linguistic terms. One of the main problems of using fuzzy logic is a choice of the membership functions of the linguistic variables. The main types of the membership functions are triangular, trapezoidal, piecewise linear, Gaussian, sigmoid, and other functions. The choice of the membership function of a particular variable is a poorly formalized problem, the solution of which is based on intuition and experience [23]. For our task, we have prepared a more appropriate triangular membership function used to specify uncertainties of the type: “approximately equal”, “average value”, “located in the interval”, “similar to the object”, “similar to the object”, etc.

The quality of fuzzy inference depends on the correct construction of “IF-THEN” rules. We obtained rules for a fuzzy knowledge base on the basis of analysis of manually written summaries. Since all the membership functions of linguistic variables are known to us, and the rules we need are defined, we proceed to the aggregation process. Aggregation is a procedure for determining the truth degree of conditions according to each rules of the fuzzy inference system. The values of the membership functions of the linguistic variable terms obtained at the stage of fuzzification are used. If the condition of a fuzzy production rule is a simple fuzzy statement, then the degree of its truth corresponds to the value of the membership function of the corresponding term of the linguistic variable.

If a condition is a compound statement, then the truth degree of the complex statement is determined on the basis of the known truth values of its elementary statements using the fuzzy logic operations introduced earlier [23]. After the logical inference, we obtain fuzzy values by accessing the fuzzy knowledge base. Also we obtain clear values for the output using the defuzzification of the fuzzy values of the linguistic variables (Figure 1).

Defuzzification is the procedure for converting a fuzzy set to a clear number. In the theory of fuzzy sets, the defuzzification is similar to finding the position characteristics (expectation, mode, median) of random variables in probability theory. The simplest way to perform the defuzzification is to select a clear number corresponding to the maximum membership function [23].

For software implementation of text summarization based on fuzzy logic, we used the python language and the skfuzzy package [24]. We constructed the membership function for each function value from five fuzzy sets: poor, mediocre, average, decent, good. Example of the membership function of the header function (Figure 3)

The last step of the fuzzy inference is defuzzification, i.e. output membership function, which we have broken into three: low, medium, high (Figure 4). The pseudo-code of the algorithm used to pronoun replacement is illustrated in Figure 5.

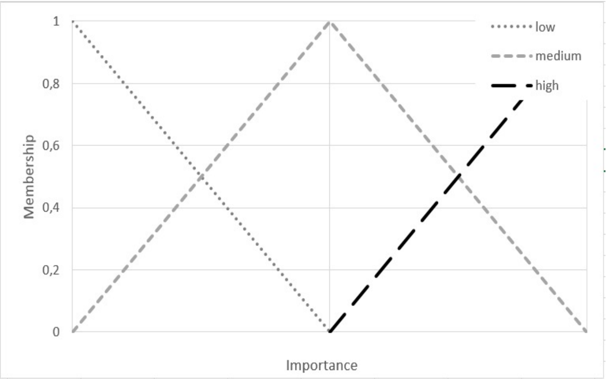

Fig. 4. Linguistic variable “Title feature”.

4 Experimental Setup and Results

4.1 Data set

The data set for the given task was collected from the news articles of the most popular Kazakhstani online news websites, namely kt.kz, bnews.kz, qazaquni.kz

Algorithm 2: Summary extraction algorithm

```
Input: source text
Result: summary of text
sentences = getSentences(text)
for sentence in sentences:
    tokens = doLinguisticAnalysis(sentence)
    features.add(getTopicFeature(tokens))
    features.add(getSentencePosition(tokens))
    features.add(getSentenceLength(tokens))
    features.add(getTFFeature(tokens))
    features.add(getThematicFeature(tokens))
    features.add(getProperNounFeature(tokens))
    features.add(getNumeralsFeature(tokens))
    for feature in features:
fuzzyCalculationInputs.add(doFuzzification(feature))
    end
    importanceValue = doFuzzyCalculationsByRules(fuzzyCalculationInputs)
    importantPropertyOfSentences.add(importanceValue)
end

numberOfSummarySentences = max(sentences.count*0.3, 3)
return doDefuzzifier(first numberOfSummarySentences with maximum value)
```

Fig. 5. Algorithm for summary extraction.

and qazaqtimes.com. The articles cover a wide range of topics and hence represent styles with high variety. Human annotators were asked to write an extractive summary of the article with respect to the style it was written in. Moreover, since we utilized the ROUGE package evaluation metric [25] which uses the reference summary or ideal summary, the extractive summary pair has to be verified by at least two annotators. The professional activity of each annotator also has to be taken into account as well. We also assess the performance of our approach on the selected part of CNN/Daily Mail dataset, which is a popular and free dataset for use in text summarization experiments. This dataset consists of news articles paired with multi-sentence summaries. For the pronoun resolution we used Stanford CoreNLP toolkit [26].

The average number of sentences in articles and the average number of sentences in summaries for both dataset are presented in Table 2.

4.2 Results

In this section, we present our experimental results for the automatic text summarization. We compare results obtained through applying our approach on both the Kazakh news dataset and CNN/Daily Mail dataset.

Table 2. Data set characteristics.

Data set	Number of articles	Average number of sentences in articles	Average number of sentences in summaries
Kazakh news	100	14	4.1
CNN/Daily Mail	100	39	3.75

ROUGE metrics were used for a preliminary assessment of the work quality. More precisely, ROUGE-L considers sentence level structure similarity and determines the longest co-occurring in sequence n-grams in automatic way. ROUGE-1 shows the overlap of unigram between the system and reference summaries, whereas ROUGE-2 indicates for bigrams [25].

Table 3. Rouge scores for the Kazakh news dataset.

Rouge	metrics	Without pronoun resolution	With pronoun resolution
Rouge-1	precision	0.38	0.39
	recall	0.39	0.44
	f-score	0.36	0.38
Rouge-2	precision	0.32	0.34
	recall	0.33	0.37
	f-score	0.31	0.34
Rouge-L	precision	0.38	0.39
	recall	0.40	0.44
	f-score	0.35	0.40

Table 3 lists the results of the experiments on text summarization for the Kazakh news dataset. As it can be seen the proposed pronoun resolution achieves better result rather than without pronoun resolution.The significant increase is indicated for Rouge-L (f-score): from 0.35 to 0.40. Rouge-1 (f-score) and Rouge-2 (f-score) scores are rised to 0.02 and 0.03, respectively. Moreover, we applied our method to CNN/Daily Mail dataset. Rouge-1 (f-score) and Rouge-L (f-score) scores show the same result, which are 0.38, whereas Rouge-2 (f-score) shows slightly worse result. An example of the automatic text summarization for the Kazakh language is shown in Table 5.

Table 4. Rouge scores for the CNN / Daily Mail dataset with pronoun resolution.

Rouge	metrics	CNN/Daily Mail dataset
Rouge-1	precision	0.35
	recall	0.41
	f-score	0.38
Rouge-2	precision	0.33
	recall	0.36
	f-score	0.34
Rouge-L	precision	0.32
	recall	0.4
	f-score	0.38

Table 5. Example of the automatic text summarization.

Manual summarization	Automatic summarization	Sentence weight (importance indicator)	Feature vectors
Биылғы мамыр-шілде айларында 2 миллион тоннадан астам көмір тасымалданды. Бұл өткен жылды сәйкес мерзімімен салыстырғанда 35% жоғары көрсеткіш. Маусымда 850 мы тонна көмір тасымалданса, ол өткен жылды сәйкес мерзімімен салыстырғанда 74% жоғары көрсеткішті құрады.	Жыл басынан бері 144 миллион тонна жүк тиеліп, өткен жылмен салыстырғанда 7% өсті	0.76666 - high	[0.31, 0.62, 0.5, 1.0, 1.0, 0.25, 0.25]
	Биылғы мамыр-шілде айларында 2 миллион тоннадан астам көмір тасымалданды.	0.766127 - high	[0.0, 0.77, 0.25, 0.0, 0.62, 0.1, 0.2]
	Маусымда 850 мы тонна көмір тасымалданса, ол өткен жылды сәйкес мерзімімен салыстырғанда 74% жоғары көрсеткішті құрады	0.766127 - high	[0.0, 0.46, 0.16, 0.33, 0.45, 0.17, 0.17]

5 Conclsuion

The research in the field of computational linguistics for the Kazakh language is expanding rapidly. Therefore, the results of the work will be very popular for a quick public perception of the summary content of a large flow of information. The algorithm of the extractive method of abstracting using fuzzy logic has proved to be effective for the tasks of automatic summarizing of news articles dataset in the Kazakh language. In this work we presented an extractive summarization method on the basis of fuzzy logic. Our approach is advanced by applying morphological analysis and pronoun resolution techniques. The experiments conducted on Kazakh news and CNN/ Daily Mail dataset show perspective results. Nevertheless, the algorithm requires improvements and the inclusion of additional methods in the algorithm, which will allow not only to extract important content, but also to paraphrase in order to more closely correspond to the manual abstract. As a future work, we aim to increase our dataset. Moreover, we want to convert the extracted summaries to abstractive one using neural network techniques.

Acknowledgments

This work was supported by colleagues from the University of Texas at San Antonio: Ramin Sahba, professor Mo Jamshidi. The work of Zhanibek Kozhirbayev and Zhandos Yessenbayev has been conducted under the Research grant $N^{\underline{o}}$AP05134272 funded by the Ministry of Education and Science of the Republic of Kazakhstan.

References

1. Chen, D., Bolton, J., Manning, C.D.: A thorough examination of the cnn/daily mail reading comprehension task. arXiv preprint arXiv:1606.02858 (2016)
2. Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E.D., Gutierrez, J.B., Kochut, K.: Text summarization techniques: a brief survey. arXiv preprint arXiv:1707.02268 (2017)
3. Barzilay, R., Elhadad, M.: Using lexical chains for text summarization. Advances in automatic text summarization (1999) 111–121
4. Conroy, J.M., O'leary, D.P.: Text summarization via hidden markov models. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ACM (2001) 406–407
5. Filippova, K., Mieskes, M., Nastase, V., Ponzetto, S.P., Strube, M.: Cascaded filtering for topic-driven multi-document summarization. In: Proceedings of the Document Understanding Conference. (2007) 26–27
6. Carbonell, J., Goldstein, J.: The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, ACM (1998) 335–336
7. Sahba, R., Ebadi, N., Jamshidi, M., Rad, P.: Automatic text summarization using customizable fuzzy features and attention on the context and vocabulary. In: 2018 World Automation Congress (WAC), IEEE (2018) 1–5
8. Binwahlan, M.S., Salim, N., Suanmali, L.: Integrating of the diversity and swarm based methods for text summarization. In: The 5th postgraduate annual research seminar (PARS). (2009) 17–19
9. Miranda-Jiménez, S., Gelbukh, A., Sidorov, G.: Conceptual graphs as framework for summarizing short texts. International Journal of Conceptual Structures and Smart Applications (IJCSSA) **2** (2014) 55–75

10. Calvo, H., Carrillo-Mendoza, P., Gelbukh, A.: On redundancy in multi-document summarization. Journal of Intelligent & Fuzzy Systems (2018) 1–11
11. Myrzakhmetov, B., Kozhirbayev, Z.: Extended language modeling experiments for kazakh. In: CEUR Workshop Proceedings. Volume 2303. (2018)
12. Kozhirbayev, Z., Karabalayeva, M., Yessenbayev, Z.: Spoken term detection for kazakh language, The 4-th International Conference on Computer Processing of Turkic Languages (2016)
13. Kozhirbayev, Z., Yessenbayev, Z., Karabalayeva, M.: Kazakh and russian languages identification using long short-term memory recurrent neural networks, 11th IEEE International Conference on Application of Information and (2017)
14. Karabalayeva, M., Yessenbayev, Z., Kozhirbayev, Z.: Regarding the impact of kazakh phonetic transcription on the performance of automatic speech recognition systems. (2017)
15. Makazhanov, A., Myrzakhmetov, B., Kozhirbayev, Z.: On various approaches to machine translation from russian to kazakh. (2017)
16. Kozhirbayev, Z., Erol, B.A., Sharipbay, A., Jamshidi, M.: Speaker recognition for robotic control via an iot device. In: 2018 World Automation Congress (WAC), IEEE (2018) 1–5
17. Mussina, A., Aubakirov, S., Ahmed-Zaki, D., Trigo, P.: Automatic document summarization based on statistical information. Journal of Mathematics, Mechanics and Computer Science **96** (2019) 76–87
18. Orynbayeva, A.: Automatic summarization. In: Proceedings of the 15th International Scientific Conference on Information Technologies and Management, Riga, Latvia (2017) 87–88
19. Kozhirbayev, Z., Yessenbayev, Z., Aibek, M.: Document and word-level language identification for noisy user generated text. In: Proceedings of the IEEE 12th International Conference Application of Information and Communication Technologies, Almaty, Kazakhstan (2018)
20. Myrzakhmetov, B., Yessenbayev, Z., Aibek, M.: Initial normalization of user generated content: Case study in a multilingual setting. In: Proceedings of the IEEE 12th International Conference Application of Information and Communication Technologies, Almaty, Kazakhstan (2018)
21. Assylbekov, Z., Myrzakhmetov, B., Makazhanov, A.: Experiments with russian to kazakh sentence alignment, The 4-th International Conference on Computer Processing of Turkic Languages (2016)
22. Suanmali, L., Salim, N., Binwahlan, M.S.: Fuzzy logic based method for improving text summarization. arXiv preprint arXiv:0906.4690 (2009)
23. Zadeh, L.A.: The concept of a linguistic variable and its application to approximate reasoningi. Information sciences **8** (1975) 199–249
24. Warner, J., Sexauer, J.: scikit fuzzy, twmeggs, ams, a. Unnikrishnan, G. Castelo, F. Batista, TG Badger, & H. Mishra (2017, October). Jdwarner/scikit-fuzzy: Scikit-fuzzy 0.3 **1** (2017)
25. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out (2004)
26. Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., McClosky, D.: The stanford corenlp natural language processing toolkit. In: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. (2014) 55–60