Expert Systems With Applications 115 (2019) 264–275

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A text summarization method based on fuzzy rules and applicable to automated assessment

Fábio Bif Goularte[a,*], Silvia Modesto Nassar[a], Renato Fileto[a], Horacio Saggion[b]

[a] Department of Informatics and Statistics, Federal University of Santa Catarina, Florianópolis, Santa Catarina, Brazil
[b] Natural Language Processing Group, Department of Communication and Information Technologies, Pompeu Fabra University, Spain

ARTICLE INFO

Article history:
Received 26 July 2017
Revised 28 June 2018
Accepted 21 July 2018
Available online 27 July 2018

Keywords:
Text assessment
Computer-assisted assessment
Automatic text summarization
Fuzzy logic

ABSTRACT

In the last two decades, the text summarization task has gained much importance because of the large amount of online data, and its potential to extract useful information and knowledge in a way that could be easily handled by humans and used for a myriad of purposes, including expert systems for text assessment. This paper presents an automatic process for text assessment that relies on fuzzy rules on a variety of extracted features to find the most important information in the assessed texts. The automatically produced summaries of these texts are compared with reference summaries created by domain experts. Differently from other proposals in the literature, our method summarizes text by investigating correlated features to reduce dimensionality, and consequently the number of fuzzy rules used for text summarization. Thus, the proposed approach for text summarization with a relatively small number of fuzzy rules can benefit development and use of future expert systems able to automatically assess writing. The proposed summarization method has been trained and tested in experiments using a dataset of Brazilian Portuguese texts provided by students in response to tasks assigned to them in a Virtual Learning Environment (VLE). The proposed approach was compared with other methods including a naive baseline, Score, Model and Sentence, using ROUGE measures. The results show that the proposal provides better f-measure (with 95% CI) than aforementioned methods.

1. Introduction

Methods to automatically summarize, compare and assess text data have become progressively important as increasingly large textual datasets have been collected and made available by a variety of systems (e.g., digital libraries (DLs), Virtual Learning Environments (VLEs), social media) (Wilbik & Keller, 2012). Recently, Natural Language Processing (NLP) and Information Extraction (IE) methods have been proposed to compute the similarity between free-texts at low computational costs (Karanikolas, 2010). This fact has motivated the development of automatic techniques for efficiently assessing short texts from large datasets (Basu, Jacobs, & Vanderwende, 2013; Brew & Leacock, 2013; Burrows, Gurevych, & Stein, 2015; Heilman & Madnani, 2015; Jordan & Mitchell, 2009; Leacock & Chodorow, 2003; Mohler, Bunescu, & Mihalcea, 2011; Selvi & Bnerjee, 2010), including techniques based on automatic text summarization (Abdi, Idris, Alguliev, & Aliguliyev, 2015; He, Hui, & Quan, 2009; Yang et al., 2013; Yang, Wen, Chen, Sutinen et al., 2012). The automatic assessment of natural language text may improve and speed-up processes that require human work, as manual correction of free-text tasks are usually assessed by teachers. However, there is still a lack of proposals for the automatic assessment of large free-text. This happens because large free-text is composed of concepts - unstructured and varied size text segments with complete sense - in an uneven information flow, in the sense that some words may be more important than others. In addition, the sequence of ideas may be non-linear.

The assessment of large texts requires attention for identifying key concepts, what may be tedious and time consuming for human assessors. According to Boran, Akay, and Yager (2016), one solution for this problem is generating explicit and concise summaries that simulate the results of human cognitive mechanisms applied in the reading process. Automatic Text Summarization (ATS) tries to identify, extract and summarize the most important concepts in a reduced version of the original text (Sankarasubramaniam, Ramanathan, & Ghosh, 2014).

In this paper, we propose an automated method for text assessment that starts with text summarization and then compares the automatically generated summaries with reference texts provided by domain experts. The proposed summarization method is a fuzzy rule-based system that identifies and selects the most informative sentences and concepts in each text document. The

* Corresponding author.
E-mail addresses: fabio.bif@posgrad.ufsc.br (F.B. Goularte), silvia.nassar@ufsc.br (S.M. Nassar), r.fileto@ufsc.br (R. Fileto), horacio.saggion@upf.edu (H. Saggion).

https://doi.org/10.1016/j.eswa.2018.07.047
0957-4174/© 2018 Elsevier Ltd. All rights reserved.

proposed method can combine features extracted from the text by using a multivariate linear regression model. It reduces the number of fuzzy rules without performance degradation.

This novel fuzzy-based summarization method identifies key sentences and concepts in the essays by using a relatively small number of rules, compared with other methods from the literature. These traits of the proposed summarization method can be exploited in the development and use of expert systems and intelligent systems for automatic assessment of text. By automatically comparing essays' summaries with gold answers, we facilitate evaluation of large documents in VLEs. The assessment is based on cooccurrence of concepts in the text summaries. The automatically generated text summaries can reduce the cognitive efforts associated with reading, particularly for large volumes of textual information (Baralis, Cagliero, & Farinetti, 2015; Saraswathi, Hemamalini, Janani, & Priyadharshini, 2011). Furthermore, a free-text summary is able to bring indicators for an assessment based on key concepts expected by experts in texts about particular subjects, for example, student's texts (Goularte, Wilges, Nassar, & Cislaghi, 2014). It makes the proposed method particularly suitable for VLEs, among other possibilities.

The efficiency of the proposed method was evaluated in a case study that assesses automatically texts prepared by students in response to tasks proposed in a VLE. This assessment is based on the similarity between summaries extracted from the student texts and the ones from reference textual summaries provided by expert teachers. The proposed process was used to automatically identify the main concepts in the texts to be assessed, and compare the resulting concept-based summaries with those provided as ground true by expert teachers for the respective text writing tasks. This work contributes to the Computer-Assisted Assessment (CAA) area by proposing an automatic assessment method for large free-text based on summary extraction techniques and comparison of the extracted concepts.

The remainder of this paper is organized as follows. Section 2 provides the foundations necessary for understanding our proposal. Section 3 discusses related works. Section 4 describes our method in a top-down way, i.e., first in terms of main modules and then their details. Section 5 reports the experiments for performance evaluation and discuss their results. Finally, Section 6 closes the paper with the conclusions and future works.

2. Foundations

This section provides a short review of Computer-Assisted Assessment (CAA), and its relationship with Automatic Text Summarization (ATS). When applied to natural language text produced by students, CAA may rely on summaries produced via ATS to automate and speed-up the assessment process. This is one of the major motivations of the assessment method proposed in this paper.

2.1. Computer-assisted assessment

The assessment of the student progress in e-learning is not simple because there are many kinds of assessments (e.g., discussion, papers, projects, quizzes and tests, and groupwork) to involve both formative and summative ways (Kearns, 2012; Noorbehbahani & Kardan, 2011).

According to Almahy and Salim (2014) and Carbonaro (2010), the growing volume of material in Virtual Learning Environments (VLE), the increasing adoption of distance education models (e.g. MOOCs massively on-line open courses) , and the need to enroll greater numbers and varieties of students are some of the factors that contribute to specific demands in the automatic evaluation (Basu et al., 2013). Thus, Computer-Assisted Assessment (CAA) has been used to reduce the cognitive overload and time spent by the instructors in some assessment-related activities (Chen, Lee, & Chen, 2005; Shrivastav & Hiltz, 2013).

CAA refers to the use of computers and software to help assess student progress (Brown, Bull, & Race, 2013). It has evolved since the times when students used to fill their answers to tests on paper, which was then processed by card readers, to be analyzed by computerized systems. CAA in distance learning environments has been performed mainly by automated tests, mainly with single-answer questions, multiple-choice questions, and short text answers (Rodrigues & Oliveira, 2014; Rodrigues & Rocha, 2011). However, the student learning progress assessed only through objective tests is not enough, because this type of test cannot measure higher cognitive skills such as analysis, evaluation and synthesis (Palmer & Richardson, 2003; Pascual-Nieto, Santos, Perez-Marin, & Boticario, 2011). Furthermore, objective tests limit the feedback that can be given to each student (Chang, 2005).

CAA has been applied successfully to objective tests that assess key competencies with the possibility to provide timely feedback (Sim, Holifield, & Brown, 2004). However, these tests are not able to demonstrate the understanding of all learning concepts which may be taught in a VLE (Noorbehbahani & Kardan, 2011). This issue rises great interest in CAA with free-text questions to assess student skills such as memory, organization, analysis, and synthesis (combination) of ideas. An assessment method widely used and essential to any teaching-learning process is free-text questions (Rodrigues & Oliveira, 2014), but it has been rarely used in VLEs, because it takes a lot of time, due to the amount of text to assess when there are many students enrolled. In addition, issues associated with the subjective nature of language make it difficult to ensure an evaluation process with fair criteria. The manual correction of free-text questions by teachers and tutors requires human work that cannot be reused, and demand a high-level concentration for long periods of time. On top of this, the assessor's concentration level is subject to fluctuations, what may lead to different degrees of accuracy for similar questions and answers for these very same questions (de Ávila & Soares, 2013; Rodrigues & Araújo, 2012). These issues can become even more evident if questions are corrected by different human assessors, specially when the answers are lengthy.

Several research works have proposed CAA systems based on NLP for free text (Pérez-Marín, Pascual-Nieto, & Rodríguez, 2009). NLP techniques are used for the recognition of language components that develop naturally in humans and human communities (Jurafsky & Martin, 2009).

2.2. Automatic text summarization

A summary is as a text derived from one document (monodocument) or more documents (multi-document), which provides important information about the original text(s). Normally, it is no longer than half of the original text(s) (Radev, Hovy, & McKeown, 2002). The ATS generates a text summary by using computational methods usually based on information retrieval techniques, text mining, and NLP (Aggarwal & Zhai, 2012).

There are two classical techniques to produce a summary: extractive and abstractive (Lloret & Palomar, 2012; Nenkova & McKeown, 2012). The extractive technique uses linguistic heuristics, statistics and empirical methods to select relevant sentences. The abstractive technique, on the other hand, uses formal methods and linguist models to analyze relationships between sentences of the source text. The summary produced by the extractive technique is a copy of some sentences of the original text, while the summary produced by abstractive technique is a reformulated version of the original text. The linguistic heuristics, statistics and empirical methods are computationally efficient, while formal methods and linguistic models, can produce better summaries because they deal with terms semantics.

Text summary assessment usually is performed through one of two ways, intrinsic or extrinsic. The intrinsic assessment compares the free-text with one or more reference texts and determines how much they are similar. On the one hand, the extrinsic evaluation suggests aspects that cannot be assessed automatically by systems (e.g. reading comprehension, cohesion between sentences, grammatical structure, and coherence). Automatized approaches for summary assessment such as n-grams are mostly used because they present a high correlation with human assessment, and are less costly (Lin & Hovy, 2003). Both, intrinsic and extrinsic summary evaluation can be applied in CAA.

3. Related works

This section analyzes the state of the art of the procedures, text features, comparison measures and algorithms that are used in CAA systems and ATS based on fuzzy logics.

Fuzzy logic has a direct relation to NLP (Cambria & White, 2014; Carvalho, Batista, & Coheur, 2012) for tasks such as text summarization (Binwahlan, Salim, & Suanmali, 2010; Kyoomarsi, Khosravi, Eslami, & Davoudi, 2010; Leite & Rino, 2009; Witte & Bergler, 2007), sentiment aalysis (Haque et al., 2014), knowledge representation (Ramos-Soto, Tintarev, De Oliveira, Reiter, & Van Deemter, 2016), word meaning inference (Pope III, 2016), and speech recognition Pope III et al. (2016). This is because Fuzzy logic provides descriptions from data sets using linguistic concepts defined as fuzzy sets and partitions, which deal with the imprecision and ambiguity of human language (Ramos-Soto, Bugarín, & Barro, 2016).

A systematic review of the literature (Kitchenham, 2004) carried out to identify, evaluate, and select relevant research related to the topics of this study. The works selected are discussed in the following.

3.1. CAA systems

The literature about CAA includes relevant works on the learning assessment theme in the Web context, through either individual tools or VLEs. The following works employ varied techniques to implement CAA systems. Alencar, Magalhães, and Diniz (2013) developed a system for correcting objective tests on mobile devices based on image processing. de Ávila and Soares (2013) use textual comparison algorithms combined with text preprocessing for assessing and correcting writing tasks in VLEs. The answer submitted by the student is compared with a standard response by using basic similarity measures. However, their system only copes with short free-text answers. Rodrigues and Rocha (2011) present a tool combined domain ontologies and genetic algorithms to build conceptual maps from free-text questions. These maps are used in the manual assessment process, without well-defined similarity measures. Morais and Arajo (2013) propose an assessment approach based on data mining and decision trees. According to the authors, the developed methodology is able to understand the patterns and answer questions from teachers. However, their approach does not perform the detailed assessment.

3.2. Fuzzy summarization

There are different approaches to ATS and numerous published works, but only some of them use fuzzy logic. The work of Witte and Bergler (2007) propose a clustering algorithm for the analysis of document collections. The algorithm is supported in fuzzy set theory and according to the authors, it has a flexible data structure and easy adjustment for filtering context-sensitive information. The international *Document Understanding Conference* (DUC) datasets from 2003 (DUC, 2003) to 2006 (DUC, 2006) were selected for the evaluation of automatic summaries produced by the algorithm. In the evaluation, they used the ROUGE package and the results were compared to other systems. The authors argue further that the algorithm performance was above average and very close to the best system, from 0.30 to 0.40 in f-measure.

Kyoomarsi, Khosravi, Eslami, Dehkordy, and Tajoddin (2008) propose a fuzzy summarization method to generate summaries of texts applied to TOEFL tests (*Test of English as a Foreign Language*) TOEFL (2017). The authors compare the fuzzy method results with a machine learning method (C4.5 and Nave Bayes algorithms). The evaluation was performed by judgment of five foreign language specialists (English language teachers). The specialists analyzed ten original texts and their automatic summaries of both methods and assigned a score. The score represents the degree to which the main concepts of the original texts were present in automatic summaries. The results of the fuzzy method exceeded the results of machine learning-based method. All specialists consider the summary produced by the fuzzy method to be of better quality, when 77% of concepts were included, and 67% in machine learning method.

Suanmali, Binwahlan, and Salim (2009) used the dataset DUC 2002 to test their method summarization. Nine characteristics served as the entrance to the inference system (FIS) with Gaussian membership functions (MF). Suanmali, Salim, and Binwahlan (2009) used the same test to Suanmali et al. (2009), but with eight characteristics and triangular membership functions. In both cases, they used the ROUGE package for evaluation and observing that the higher precision, recall, and f-measure value were obtained by their fuzzy method.

Leite and Rino (2009) describe an extractive summarization approach in which the fuzzy knowledge-base was generated by a Genetic Algorithm (GA). Called SuPor-2 Fuzzy, the summarizer includes eleven measures, each modeled in three fuzzy sets (low, medium and high) by triangular membership functions. The TeMário corpus was used for training and evaluated with ROUGE. The compression ratio of the summaries produced by the SuPor-2 Fuzzy corresponded to 30%, because this is equivalent to the amount of information in the reference summaries. In the evaluation, the SuPor-2 was compared with other summarizers systems, and the best results were achieved by their proposed method (0.74 and 0.73 respectively for recall ROUGE-1 and ROUGE-2).

Kiani and Akbarzadeh (2006) propose a similar approach to Leite and Rino (2009), with similar results, but evaluation based just on human appreciation, not formal quality measures. The method proposed by Kiani and Akbarzadeh (2006) uses combination of GA and Genetic Programming (GP) to optimize rule sets and membership functions of fuzzy systems. The method achieved 0.75 in f-measure using a corpus with 3 documents.

Binwahlan et al. (2010) presents a hybrid model for text summarization based on three models: (i) the Maximal Marginal Importance (MMI) diversity-based method, (ii) the swarm diversity-based method, and (iii) the fuzzy swarm based method. In the first method, sentences are selected from the binary tree by traversing all levels and applying MMI on the sentences in each level. Then, summary sentences are ordered according with their order in the original document. The second method, which is based on particle swarm optimization is used to adjust weights of the features based on their importance. In the third method, the fuzzy algorithm is used to calculate sentence score. The inputs of the fuzzy system are weights that are found by particle swarm optimization. In each method, sentences are ranked according to their scores, and then the best sentences are extracted from each summarizer. At last, a score is calculated for each summarizer sentence and the best sentences are extracted.

Table 1
Works based on fuzzy summarization.

Id	Author	Placement	Type*	Features in Table 2 (n)
1.	Kiani and Akbarzadeh (2006)	IEEE IC Fuzzy Systems	MU	b, 2c, d, i, g (5)
2.	Witte and Bergler (2007)	Advances in AI	MU	e, i, c (3)
3.	Kyoomarsi et al. (2008)	IEEE/ACIS ICIS	MO	a, b, 2c, d, e, i, m, n, 2o (9)
4.	Suanmali et al. (2009)	IC on HIS	MO	2a, 2b, c, d, e, j, n (7)
5.	Suanmali et al. (2009)	IJCSIS	MO	a, b, c, d, e, i, j, m (8)
6.	Leite and Rino (2009)	Congress of the BSC	MO	2a, 2c, d, e, 2i, 2n, p (7)
7.	Binwahlan et al. (2010)	IPM journal	MO	2a, 3b, 3n (3)
8.	Kyoomarsi et al. (2010)	IJFS	MO	4a, 2c, 2d, p (4)
9.	Suanmali et al. (2011)	IEEE IC on DASC	MO	a, b, c, d, e, i, j, m (8)
10.	Kyoomarsi et al. (2011)	IEEE IC on ICM	MO	a, b, c, d, e, i, o, m (8)
11.	Hannah et al. (2011)	SEMCCO	MO	a, b, d, e, i, j, m (7)
12.	Megala et al. (2014)	IJCSIT	MO	a, 3c, 2e, g, h, i, m (7)
13.	Kumar et al. (2014)	Applied Soft Computing	MU	a, d, 3m (5)
14.	Megala et al. (2015)	IJIRCSE	MO	a, 2c, e, d, g, 3h, i, m (8)
15.	Abbasi-ghalehtaki et al. (2016)	Swarm and Evol. Computation	MO	a, b, c, d, e, i, j, m (8)

* MO = Mono document, MU = Multi document.

Kyoomarsi et al. (2010) developed a text summarization method based on fuzzy logic and WordNet. Two measures were defined by analyzing the synonyms of the words in the sentences and the entire text. The method was patterned with nine input variables and four outputs. The results with ROUGE and the DUC 2003 dataset achieved 0.60 in precision. Already in Suanmali, Salim, and Binwahlan (2011) the approach suggested combines fuzzy logic, GA and Semantic Role Annotation. In relation to ROUGE-1 measure, the proposed method by Suanmali et al. (2011) achieved 0.47 in f-measure.

Kyoomarsi, Rahmani, Eslami, Dehkordy, and Tajoddin (2011) propose a text summarization method based on Cellular Automata (CA). This method also uses GA and fuzzy logic. Three summarizing methods are examined: (i) text summarization based on the GA method, (ii) text summarization based on the fuzzy logic method, and (iii) text summarization based on the CA method. Each method used a set of eight features to calculate sentence score, and then the best sentences in the ranking are extracted from each summarizer. The authors used a sample of 17 English scientific articles to evaluate their methods (16 documents for training and 1 for testing). The fuzzy method achieved 0.46 in precision, while the methods using CA and GA achieved 0.31 and 0.59 in precision, respectively.

Hannah, Geetha, and Mukherjee (2011) describe a text summarization fuzzy method using seven characteristics. The characteristics were categorized into three sets fuzzy (lower, middle and higher) with trapezoidal membership functions. The fuzzy method classifies sentences in not important, medium and major importance. The results with ROUGE and the DUC 2002 dataset showed a performance 0.48 in f-measure.

Megala, Kavitha, and Marimuthu (2014) compare the performance of two text summarization methods, one of them used the fuzzy logic, the other artificial neural network. The authors selected trials legal texts for the experiments and proceeded with a non-automated evaluation. The neural network summarization method achieved 0.42 in f-measure, while the fuzzy method 0.46. In Megala, Kavitha, and Marimuthu (2015) the researchers proposed a summarizer system using extract measures with fuzzy logic to produce the summary and Conditional Random Field (CRF) for classifying the summary segments. The system was tested with the input of 30 Legal Judgements (documents from the service, industry and constitutional law) and the results showed that the system classified accurately all segments in the case, but in the analysis, the performance was 0.26 in f-measure.

Kumar, Salim, Abuobieda, and Albaham (2014) describe a multi-document summarization approach based on news components using fuzzy cross-document relations. This approach has three phases: (i) component sentence extraction, (ii) Cross-document Structure Theory (CST) model which describes the semantic relations between textual units, and (iii) sentence scoring using fuzzy reasoning. Firstly, from a set of input documents D, component sentences are extracted by using the gazetteer list and named entity recognition. Then, the CST relation is identified using a Genetic-Case Base Reasoning (CBR) model with five different features (namely Cosine similarity, Word overlap, Length type, noun phrase similarity, and verb phrase similarity) for each sentence pair taken from a document cluster. The fuzzy reasoning model is used for sentence scoring. The highest ranked sentences are selected until the desired summary length is met. The authors used the dataset obtained from CSTBank for the training and testing. The results with ROUGE-1 and DUC 2002 dataset achieved 0.33 in f-measure.

Abbasi-ghalehtaki, Khotanlou, and Esmaeilpour (2016) used Cellular Learning Automata (CLA), an evolutionary algorithm and Fuzzy logic to propose a new model for ATS. This model uses CLA for reducing the redundancy problem, Particle Swarm Optimization (PSO) to assign weights to the features in terms of their importance, and then fuzzy logic for scoring sentences. The proposed method used a linear combination of features such as thematic words, keywords, proper nouns, sentence length, and sentence position for selecting the important sentences. The authors used de documents from the DUC 2002 dataset for training and testing the model. Text summarization based on fuzzy PSO CLA achieved better performance than other current methods, namely 0.48 in f-measure.

Based on the characterization of text summarization process discussed in Section 3 (e.g., classifications of text summarization, performance evaluation, databases, membership function - MF, number of fuzzy rules, fuzzy inference system - FIS, textual features and metrics used), Tables 1–3 summarizes findings of our literature review.

All, except Witte and Bergler (2007), employ extraction techniques using several features. For example, in Kyoomarsi et al. (2008) nine characteristics were selected to derive eleven measures (Table 1, line 3). In fuzzy summarization, this may seem a problem because the amount measures impacts on the system performance. Therefore, determining which measures is better to achieve the informational content of texts is important (Saggion & Poibeau, 2013), and factors such as text structure and the language used can influence the performance measures. There are other linguistic and statistical characteristics that may be found in the literature, however, the results of the summarization process depend on how they are exploited and combined as well as the purpose of the application.

Table 2
Description of text features.

Id	Features	Comments
a	Frequency of words	It assumes that the main idea of a text can be expressed by the representative frequency of words, using statistical methods or involves morphological analysis (e.g., tf-isf or tf-idf).
b	Title similarity	Sentences with words of the title or heading are indicative document theme (e.g., words that co-occur in the title and sentence over the words in the title).
c	Position	It may involve the positioning of the sentence with regard to paragraph with respect to a section of text or with respect to the whole document. The first and last sentence of a paragraph tends to be more important, as the first and last paragraph (e.g., assume a paragraph has *n* sentences, hence the punctuation of each sentence is its position in the paragraph over the value of *n*).
d	Length	It refers to the amount of sentence words (e.g., number of words in the sentence over the value of words in the longest sentence).
e	Nouns	Words that are names of people, places, or named entities that represent a concept (e.g., number of nouns in the sentence over the value of sentence length).
f	List of acronyms	Acronyms also have importance to the weight of a sentence, they can be names.
g	Phrase suggestive	Are suggestive sentences with phrases such as: “In conclusion, this article, this report develop the objective, among others” (e.g., the value is 1 if term *i* appears at least once in the sentence; otherwise, 0).
h	Words influential	A word list for a particular domain can be set.
i	Words theme	These words express the topics discussed in the document (e.g., number of words theme in the sentence over the value of sentence length).
j	Numeric information	Sentences with numerical information are important (e.g., the value of numerical information is 1 if term *i* appears at least once in the sentence; otherwise, 0).
k	Style of words	Words with emphasis (bold, italic, and underline), or, in capital letters, are also important.
l	Pronoun	Pronouns are not included in important characteristics, unless they are accompanied by a corresponding noun.
m	Cohesion sentence-sentence	Calculate the likeness of a sentence with the other and add a value of similarity (e.g., euclidean distance).
n	Cohesion award centroid	For each sentence s is computed vector representing the centroid of the document, which is the arithmetic mean of the values of the corresponding coordinates of all sentences of the document. After that, it calculates the similarity between the centroid and each of the sentences, getting a gross amount for each sentence (e.g., cosine similarity between two sentences, Jaccard coefficient).
o	Nonessential information	Are words that occur often in the beginning of sentences, such as “because due furthermore generally normal.” This is a binary feature where the sentence takes the true value if it has at least one such discourse markers, and false otherwise.
p	Discourse analysis	Information on the level of discourse in a text is also a good feature. Analyzing the discourse is possible to identify the general structure of the text and then remove phrases that are peripheral until you get the main message (e.g., a chain score may be determined by the number and weight of the relations between chain members).

Table 3
Fuzzy systems in fuzzy summarization works.

id*	Membership function	Fuzzy system	Datasets	Evaluation$^+$	Performance
1.	Bell	Genetic algorithms	Documents	P, R and F1	0.72 (F1)
2.	Grades	Coreference algorithm	DUC 2003–2006	ROUGE	0.40 (F1)
3.	Gaussian	Mamdani	TOEFL tests	Human	0.77 (Quality)
4.	Gaussian	Mamdani	DUC 2002	ROUGE	0.47 (F1)
5.	Triangular	Mamdani	DUC 2002	ROUGE	0.47 (F1)
6.	Triangular	Genetic algorithms	TeMário	ROUGE	0.74 (R)
7.	Trapezoidal	Mamdani	DUC 2002	ROUGE	0.45 (F1)
8.	Triangular	Mamdani	DUC 2003	ROUGE	0.60 (P)
9.	Triangular	Genetic algorithms and SRL	DUC 2002	ROUGE	0.47 (F1)
10	Triangular	Mamdani	Science articles	Accuracy	0.46 (P)
11.	Trapezoidal	Mamdani	DUC 2002	ROUGE	0.48 (F1)
12.	Triangular	Mamdani	Legal documents	P, R and F1	0.46 (F1)
13.	Gaussian	Mamdani	DUC 2002	ROUGE	0.33 (F1)
14.	Triangular	Mamdani	Legal documents	P, R and F1	0.26 (F1)
15.	Trapezoidal	Mamdani	DUC 2002	ROUGE	0.48 (F1)

* as in Table 1, SRL = Semantic role labeling, $^+$ P = Precision, R = Recall, F1 = F-measure

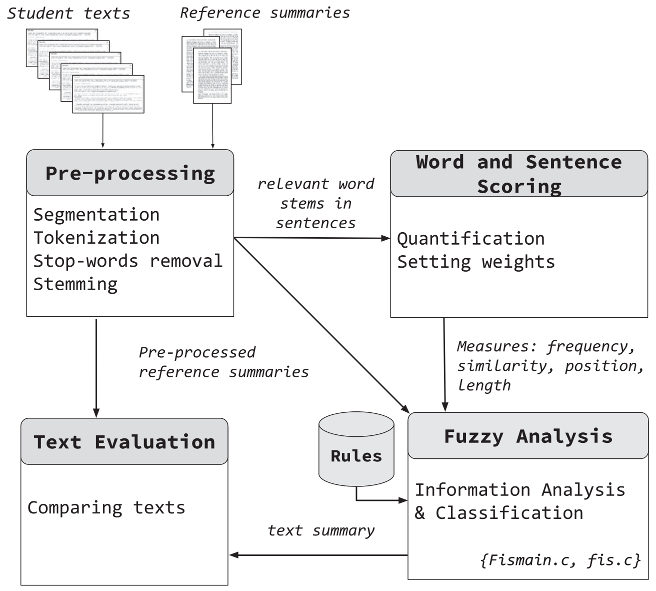
Table 3 presents some details of the summarization systems used by each work listed in Table 1. Notice that linear membership functions (e.g., triangular, trapezoidal) are far more common than non-linear ones (e.g., Gaussian, Bell). This is because the mathematical operations needed by linear membership functions are simpler (as they cope with shapes described by linear equations), and consequently faster to calculate. Fuzzy inference methods are classified in direct methods and indirect methods. Direct methods, such as Mamdani is the most commonly used. Indirect methods are more complex.

4. Proposed method

Our method employs measures founded on fuzzy logics for ATS, including concrete or objective attributes (features observable in a text) and metrics for more abstract, higher-level, or somewhat subjective attributes (fuzzy metric) (Schofield, 2005). Fig. 1 provides an overview of the proposed method specified as a process. The method is composed of four major tasks executed in subsequent stages: (i) Pre-processing, (ii) Word and (iii) Sentence Scoring, (iv) Fuzzy analysis and (v) Text Evaluation.

4.1. Pre-processing

The pre-processing module includes: the identification of sentences *segmentation*; breaking the continuous stream of characters in words or terms, also called tokens *tokenization*; the elimination of words that do not add relevant information such as, articles, prepositions, adverbs, numbers, pronouns and punctuation *stop-word removal*; and a linguistic normalization in order to obtain the radical of each word *stemming*. In English, for example, the tokens: **lies, lying** and **lie** can be transformed to the stem **lie** after removing the prefix and suffix. The pre-processing is an important step, as it reduces the amount of information to analyze and can contribute for the next steps to generate better results. Tools like Pre-Text (Matsubara, Martins, & Monard, 2003) and NLTK (Bird, 2006)



An excerpt from a student text:

```
Introduction: relating study time of high school students
to course to which they wish to provide entrance exam is a
challenge. This research has as objective to analyze
whether students who wish to take entrance exams for the
busiest courses have an average of greater study to...
```

An excerpt from a reference summary:

```
Introduction: the general objective of this research is to
apply/discovery/analyze/evaluate data about...
The sample size has been calculated to detect the
prevalence...
```

Fig. 1. Proposed process.

```
[1] => introduct relat studi time school student cours
wish provid entranc exam challeng
[2] => research object analyz student wish entranc
exam busiest cours averag studi...
```

Fig. 2. Relevant word stems in sentence.

can be used for this module. Fig. 2 provides an overview of the pre-processing output for two sentences of an input text.

4.2. Word and sentence scoring

The word and sentence scoring module extracts the features that express the relevance of the words and sentences - *quantification*; and assigns weights based on measures - *setting weights*. In this module, the text goes through several steps in order to prepare it for the next module, the fuzzy analysis.

4.2.1. Measures

The first step in the word and sentence scoring module is to identify the textual features that can be taken into account when determining the importance of sentences. The scoring technique uses statistical or empirical methods for the summary considering textual features such as the position and length of the sentence, word frequency, and the distance between sentences. Through the analysis of textual features, it is possible to derive measures that recalls the relevance of a sentence. For this work, the textual features are defined as any characteristic of the text that can be observed and quantified by a measure.

To detail what happens in the word and sentence scoring module and assist in understanding the measures used, in what follows a formalization of the problem similar to the document representation in the vector space model (VSM).

A text T can be represented by a set of sentences $T = \{S_1, S_2, \dots, S_j\}$, where j is the number of sentences, and each $S_k \in T (1 \le k \le j)$ is a set of i terms (words), $S_k = \{t_1, t_2, \dots, t_i\}$. In the literature on text summarization there are different ways to measure a source of information. The measures used in our method are identified by m_A, having $A = \{frequency, similarity, position, length\}$ as the features observed and referring to the equations (2|4|5|6), respectively. In our work, we use the set of characteristics in A because these are the most used in the works and research in several areas involving ATS in the last thirteen years (Ferreira et al., 2013).

The *frequency* feature assumes that the main idea of a text can be expressed by a set of frequent words, and in this case, the sentence that presents more of these words should be considered important. The *similarity* refers to the similarity between the sentence words and words that were identified as concepts. The *position* refers to the positioning of the sentence with respect to paragraph, with respect to a section of text or with respect to the text as a whole. The *length* refers to the length of the sentence, too long or short sentence is generally not relevant. We have one equation to each feature. Thus, the measures used in the analysis stage are $m_{fre}, m_{sim}, m_{pos}$ and m_{len} referring to the equations (2),(4),(5) and (6), respectively. Measures m_{fre}, m_{sim} are defined by criteria of similarity measure which originate from *tf-idf* and in the case of text summarization is named *tf-isf* (term frequency - inverse sentence frequency).

$$w_{ij} = tf_{ij} * isf = \frac{f_{ij}}{\sum_{k=1}^{n} f_{kj}} * log\frac{n}{|n_i|} \qquad (1)$$

$$m_{fre}(S_j) = \frac{\sum_{i=1}^{k} w_i(S_j)}{Max(\sum_{i=1}^{k} w_i(S_j))}, \quad j = 1, \dots, n \qquad (2)$$

In Eq. (1), n is the total number of sentences, n_i is number of sentences in which word i occurs and k is the number of words in sentence j (Abbasi-ghalehtaki et al., 2016). The weight w_{ij} is given by frequency tf_{ij} multiplied by the measure of the general significance of the term *isf*. The f_{ij} is the number of occurrences of the term t_i in a set of sentences S_j. The *isf* is defined as the logarithm of the ratio of the total number of sentences N and the number of sentences containing the term n_i. The frequency is normalized by *isf* to prevent a bias in long texts. At last, the score of each word w_{ij} in S_j is computed by Eq. (2).

$$sim(C_o, S_j) = \frac{\sum_{i=1}^{k} w_{ij} w_{io}}{\sqrt{\sum_{i=1}^{k} w_{ij}^2 \cdot \sum_{i=1}^{k} w_{io}^2}}, \qquad (3)$$

$$j, o = 1, \dots, n$$

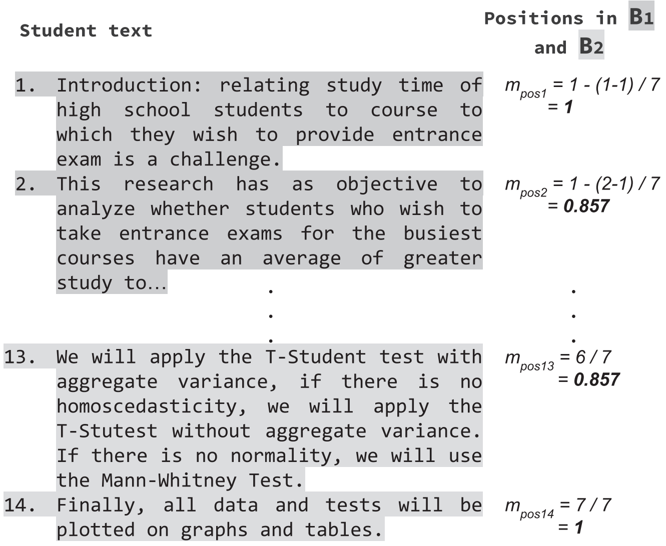
$$m_{sim}(S_j) = sim(C_o, S_j) \qquad (4)$$

In Eqs. (3) and (4), m_{sim} is a value of similarity between words in the sentence S_j and words that were identified as concepts $C_o = \{c_1, c_2, \dots, c_o\}$, also known as biased words (Rautray & Balabantaray, 2017).

$$m_{pos}\left(p_{ij}, n_{S_j}\right) = \begin{cases} 1 - \dfrac{p_{ij} - 1}{n_{S_j}} & \text{if } S_j \in B_1 \\ \dfrac{p_{ij}}{n_{S_j}} & \text{if } S_j \in B_2 \end{cases} \qquad (5)$$

$$p_{ij} = 1, \dots, n_{S_j}$$

In Eq. (5), the position score of sentences m_{pos} is defined by the ratio between the position of each sentence p_{ij} and the number of



Student text

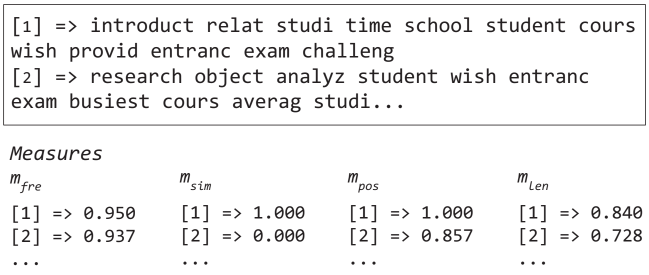
Positions in B1 and B2

1. Introduction: relating study time of high school students to course to which they wish to provide entrance exam is a challenge. — $m_{pos1} = 1 - (1-1) / 7 = \mathbf{1}$
2. This research has as objective to analyze whether students who wish to take entrance exams for the busiest courses have an average of greater study to... — $m_{pos2} = 1 - (2-1) / 7 = \mathbf{0.857}$

.
.
.

13. We will apply the T-Student test with aggregate variance, if there is no homoscedasticity, we will apply the T-Stutest without aggregate variance. If there is no normality, we will use the Mann-Whitney Test. — $m_{pos13} = 6 / 7 = \mathbf{0.857}$
14. Finally, all data and tests will be plotted on graphs and tables. — $m_{pos14} = 7 / 7 = \mathbf{1}$

Fig. 3. Example of position equation.



```
[1] => introduct relat studi time school student cours
wish provid entranc exam challeng
[2] => research object analyz student wish entranc
exam busiest cours averag studi...
```

```
Measures
m_fre           m_sim           m_pos           m_len
[1] => 0.950    [1] => 1.000    [1] => 1.000    [1] => 0.840
[2] => 0.937    [2] => 0.000    [2] => 0.857    [2] => 0.728
...             ...             ...             ...
```

Fig. 4. Measure results.

sentences n_{S_j} as a whole B. Sentences are divided into two subsets B_1 and B_2, which are reviewed individually to assign greater importance to the first and last sentences of text (Fig. 3).

The m_{len} defines an equation to calculate the relevance of a sentence in relation to its length. The value of m_{len} is calculated by the natural logarithm of the standard score. In Eq. (6), $\bar{t}$ is the mean of the terms considering all sentences from the text, t is number of terms in sentence S_j and σ is the standard deviation. The results of measuring weight m_{len} increases as the number of terms of the sentence is closer to the mean $\bar{t}$. According to the text structure and the measure used, too long sentences may be privileged at the expense of others, but this does not occur with the m_{len}.

$$m_{len}(S_j) = ln\left(\bar{t} - \left|\frac{\bar{t} - t(S_j)}{\sigma}\right|\right), \quad j = 1, \dots, n \tag{6}$$

All measures presented are normalized between 0 and 1 for improving the fuzzy analysis process and be treated with equal proportion. Fig. 4 shows the setting weights for each sentence.

4.3. Fuzzy analysis

The fuzzy module does the informativeness classification. We can say that the dataset itself is its own most informative description, and any other summary implies a loss of information. Thus, informativeness (Yager, Ford, & Cañas, 1991) is a measure that represent to what extent a particular summary is informative (Boran et al., 2016). After all measures are computed, the scores are used to model words through the fuzzy set theory (Zadeh, 1999). Formally, a fuzzy set on X, denoted by A, is defined as $A = \{\langle x, \mu_A(x)\rangle | x \in X\}$, where $\mu_A(x)$ is the membership function of x. The α-cut and α^+-cut of A are the crisp sets $A_\alpha = \{x \in X | \mu_A(x) \geqslant \alpha\}$ and $A_\alpha^+ = \{x \in X | \mu_A(x) > \alpha\}$, where α is a number in the interval [0,1] (Boran, Akay, & Yager, 2014). Let $O = \{o_1, o_2, \dots, o_n\}$ defined as a set of n objects, $V = \{v_1, v_2, \dots, v_l\}$ as a set of attributes, and $X_i = \{i = 1, 2, \dots, l\}$ be the domain of v_i. Then, $v_i^n \equiv v_i(o_n) \in X_i$ is the value of the *ith* attribute for the *nth* object.

The type of linguistic summary rule we use are *"if/then"* fuzzy rules. The fuzzy system rule base comprises a set of fuzzy rules R where the *ith* rule is defined as R^i: *If* m_1 *is* v_1^i *and* m_2 *is* v_2^i *and* ... *and* m_N *is* v_N^i *then* y *is* Y^i, where v_N^i and Y^i are linguistic values combined by the operator. The first part of the rule "m_N *is* v_N^i" is called the antecedent or premise, while the second part of the rule "y *is* Y^i" is called the consequent or conclusion.

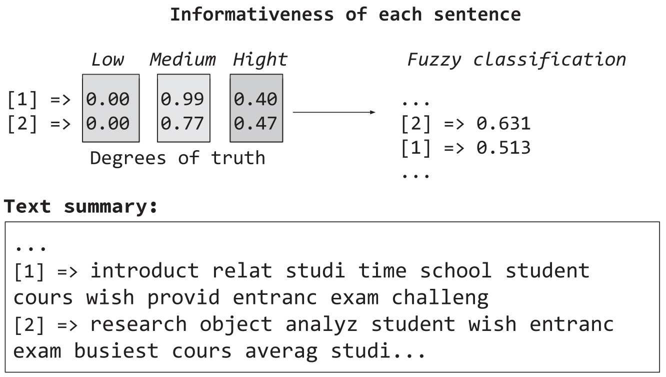
In the fuzzy module the rule antecedents are concatenated by *and* connectives that involve the operator *MIN* and all outputs of the rules are grouped by the *MAX* method. We produced 27 rules (by the combination all input variables) in the rule base and the fuzzy sets were represented with the bell membership function. The parameter values of the bell function were predicted with manual experimentation and a priori knowledge. Different measures are used in different problems and no uniform measure can be used in all kinds of fuzzy environments (Xia & Guo, 2004). The inputs and outputs of the rules are shown in Table 4.

The reason we choose the bell membership function is because the number of fuzzy rules is small and the value of the fuzzy output score of sentence is unrepeatable. Note that the fuzzy module output is a metric over the space of texts that produces values in the range [0, 1]. There is an equivalence between the triangle inequality and fuzzy points to measure the distances. Fig. 5 illustrates the fuzzy classification that represents the informativeness of a sentence. The sentence [2] is more informative than [1] and both may be in text summary.

Table 4
Fuzzy modeling.

Linguistic variables (Type)	Linguistic terms set	Parameters of bell membership functions [a b c]
x^* (input)	Low	[0.241 2.040 0.046]
	Medium	[0.173 2.810 0.462]
	Hight	[0.387 3.118 1.030]
m_{fre} (input)	Low	[0.244 2.410 0.008]
	Medium	[0.159 2.675 0.411]
	Hight	[0.318 3.465 0.889]
m_{sim} (input)	Low	[0.258 2.409 0.008]
	Medium	[0.186 1.983 0.456]
	Hight	[0.318 2.814 0.952]
Informativeness (output)	Low	[0.253 4.160 0.029]
	Medium	[0.195 3.117 0.470]
	Hight	[0.305 4.560 0.975]

*Regression model (Section 5.1.3).



```
...
[1] => introduct relat studi time school student
cours wish provid entranc exam challeng
[2] => research object analyz student wish entranc
exam busiest cours averag studi...
```

Fig. 5. Fuzzy analysis.

4.4. Evaluation

In our evaluation setup, the text summary is compared with reference summaries created by an expert (human) or gold standard. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) was used to evaluate the similarity between the text summary and the reference summaries. ROUGE measures quality by counting overlapping units such as n-grams, word sequences and word pairs between the automatic summary and the reference summary (Lin & Hovy, 2003). ROUGE results produce three measures: recall, precision and f-measure.

The precision is the number of unigrams that co-occur in the text summary and reference summaries divided by the total number of unigrams into reference summaries. The recall is similar the precision but is divided by the total number of unigrams in the text summary. Both are combined into a single measure, f-measure (the weighted harmonic mean of precision and recall). Precision and recall range from 0 to 1. When the precision score is 1, all n-grams in the text summary are present in the reference summaries.

5. Experiments

This section reports experiments for evaluating the performance of our fuzzy method compared with other summarizers. The Baseline, Score and Model methods were also implemented to use in the evaluation process as a benchmark against the fuzzy method.

5.1. Compared methods

5.1.1. Baseline

Our baseline is a typical method for generating summaries, where the first n sentences in text summary are selected for the summary according to the compression rate. Despite its simple construction, this procedure provides a relatively strong baseline for the performance of any text summarization method (Nenkova & McKeown, 2012).

5.1.2. Score

The Score method produces a traditional way summary. It selects the highest scoring sentences S_j given by the sum of the measures m_f. Where $f = \{freq, sim, pos, len\}$.

$$score(S_j) = \frac{\sum_{j=1}^{4} m_f(S_j)}{Max\left(\sum_{j=1}^{4} m_f(S_j)\right)} \tag{7}$$

5.1.3. Model

The Model method selects the sentences the same way as the previous method, but uses the m_{pos} and m_{len} measures in a corresponding regression model. To estimate the regression model was used a sample of 79 randomly selected texts (0.05 significance level) of a dataset already used in other experiments. The relationship between variables (amount of sentence terms t_{S_j} and score of sentence position m_{pos}) and dependent variable (x) was the following regression model: $x = -0.084 + 0.008t_{S_j} + 2.344m_{pos}$. The best fit parameters were estimated using the least-square method. The model test showed a coefficient of determination $R^2 = 0.954$, standard error of 0.15 and p-value < 0.01.

5.1.4. Fuzzy

Our fuzzy method employs three input variables: m_{pos} + m_{len} (estimated with regression model), m_{fre} and m_{sim} to compose a fuzzy metric. The selected sentences should represent a informativeness classified in output variable as High/Medium.

In our work, the Fuzzy Logic Toolbox was used to design the fuzzy module (Sivanandam, Sumathi, Deepa et al., 2007). This toolbox is a collection of functions implementing a framework on the MATLAB for creating and editing fuzzy inference systems. It provides build stand-alone C programs that can integrate with others systems. The type of inference used in our method is Mamdani and the method to compute the output is the center of gravity. The center of gravity method and Mamdani inference system are both very simple and very fast.

5.1.5. Sentence

The Sentence method selects the sentences based on the frequency of words. It first chooses the sentence considered the most important of the original text and then selects the sentences that have some similarity to it. This sentence score is by tf-isf. This method was developed in the GistSumm system (Balage Filho, Pardo, & Nunes, 2007).

5.2. Dataset and gold standard

A collection of 63 texts in Portuguese collected in a task realized in a VLE[1] is used for the construction of a dataset. This dataset has 1392 sentences, 31,326 words, 592 average words per text and 358 of standard deviation.

The VLE in question is a tool used in blended learning in the statistic courses by undergraduate students in Engineering from our university. The task selected is about the preparation of the statistical survey analysis and requires only one writing text.

The reference models of this task (reference summaries) were created by the course professor based on the desired key concepts and learning objectives. The professor has prepared three responses (gold standards) that are used in the evaluation module. It was also told the professor that the maximum word limit adopted for the reference models did not exceed 50% of the average word text of dataset.

The result of the processing of reference models was presented to the professor who pointed out the relevance for the thirteen key concepts: variables, sample, level of significance, confidence intervals, sample size, association representative, quantitative (QT), qualitative (QL), search, correlation. We used the key concepts to compute the similarity between concepts and words of the sentences. The processing for the statistic dataset as well as identifying the main key concepts was held in the Sketch Engine[2] tool.

5.3. Results and discussion

To perform the experiments the dataset was partitioned into two subsets: one with 40% of the data for training and another

[1] http://sestatnet.ufsc.br/.

[2] https://the.sketchengine.co.uk.

Table 5
Performance comparison.

Method	Size of summary - 20%				Size of summary - 30%				Size of summary - 40%			
	P	R	F1	CI for F1	P	R	F1	CI for F1	P	R	F1	CI for F1
Baseline	0.284	0.390	0.328	0.305-0.349	0.284	0.390	0.328	0.305-0.349	0.284	0.391	0.329	0.305-0.349
Score	0.406	0.382	0.393	0.364-0.414	0.361	0.491	0.415	0.387-0.437	0.314	0.547	0.399	0.375-0.422
Model	0.405	0.383	0.393	0.365-0.414	0.360	0.492	0.415	0.386-0.437	**0.316**	0.550	0.401	0.375-0.428
Fuzzy	**0.417**	**0.398**	**0.406**	**0.369–0.436**	**0.366**	**0.496**	**0.421**	**0.389–0.450**	0.315	**0.556**	**0.402**	**0.375–0.430**
Sentence	0.373	0.275	0.316	0.277-0.350	0.352	0.383	0.366	0.330-0.394	0.310	0.478	0.376	0.345-0.399

P = Precision, R = Recall, F1 = F-measure

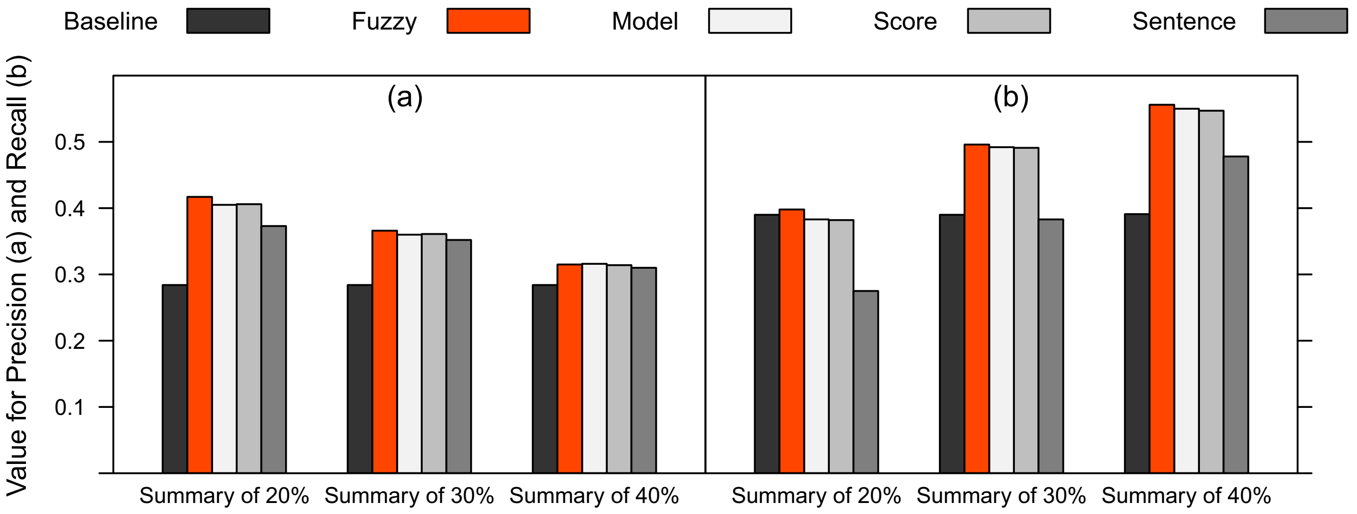


Fig. 6. Precision (a) and Recall (b).

one with 60% for evaluation. The training subset was used to develop the method, analyze the text structure and check for possible inconsistencies in the pre-processing step. Since the evaluation subset was used to measure the efficiency of the proposed method.

In general, the evaluation of a text summarization method requires analysis of the result of the summarization process, the summary. In order to minimize the work and time spent with the summary evaluation, the evaluation conducted in this work is of the intrinsic type. The evaluation is performed automatically, without human judgement and the summaries are evaluated separately.

Due do the small size of the evaluation subset just the ROUGE-1 metric was selected to evaluate the informativeness of summaries. Table 5 illustrates the results of unigrams co-occurrence among summaries produced by five summarization methods with the size of 20%, 30% and 40%. Precision indicates how close the tasks summary is of the reference tasks and recall indicates how much the reference tasks information is in the summary of the task. The f-measure represent the similarity between the tasks summary and the reference models, where 0 indicates that the tasks summary is different from the reference models, while 1 indicates the maximum proximity between them. Table 5 highlights the best results with respect to precision, recall and f-measure for each size of summary and methods evaluated.

In the tasks summary with the size of 20% and 30%, the Fuzzy method presented the best averages of precision (0.417 and 0.366), recall (0.398 and 0.496) and f-measure (0.406 and 0.421). In the tasks summary with the size of 40%, the Model method showed the best average of precision (0.361) and the Fuzzy method the best averages of recall (0.556) and f-measure (0.402). Fig. 6 presents the results of the qualitative evaluation. The summaries with size of 20% achieved the best results for precision and the size of 40% the best results for recall. Considering the dataset, the results are reasonable because the student's tasks are not well structured. Students are not concerned about the structure of the text. Thus, metrics that consider suggestive phrases and significant words can be used to improve the performance. Probably, the precision and recall could be higher if small texts were used in theses experiments.

The results concerning the application of ROUGE-1 metric shows that the Fuzzy method has the best values. Fig. 7 illustrates the classification of methods according to performance confidence intervals (CI). The CI and f-measure of the methods show that the systems cannot be considered as similar. The Sentence method performance is close to that of the Fuzzy method when the size of the summaries is greater, while the Baseline method remains the farthest. Analyzing the intervals between Fuzzy method and Sentence we can see how the Fuzzy method is more efficient than Sentence (86% in summary with size of 30% and 48% in summary with size of 40%, respectively, of the Fuzzy method over the Sentence. This is because the Sentence method is based word frequency, what improves the performance when the summary has more words. The dataset used in the evaluation does not present a clear textual structure what influenced the performance of the Sentence method. The Sentence method may have better results with structured texts (Balage Filho et al., 2007).

Meanwhile, the results of the Score and Model methods are very close. Thus our proposed correlation features used in the Model method solves the optimization problem about fuzzy rules and obtains better performance. This suggests that some features in text can be grouped in a kind of good feature for summarization and helps to reduce the measures useful.

The Baseline method just selects the first text sentences. The poor performance of the Baseline method in relation to the other methods was expected because it is simple and benchmarking with other methods.

Therefore, although the difference in performance of the methods is considerable, the precision, recall and f-measure rates may be increased with the following suggestions: a) dataset size - in-

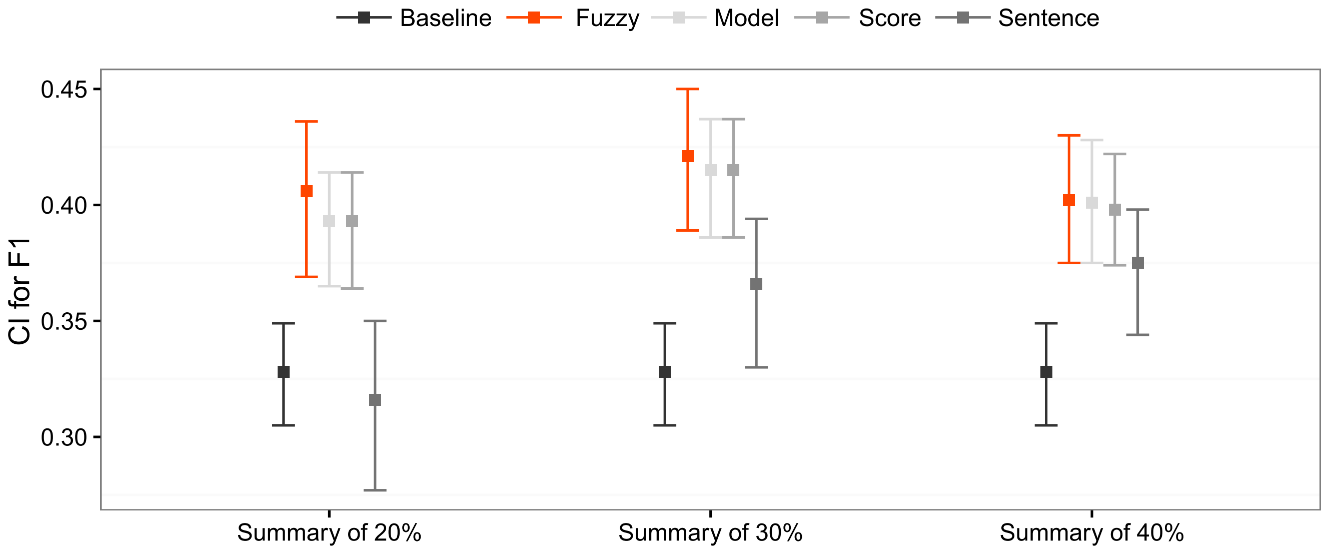


Fig. 7. Confidence intervals (CI) for F1.

creasing the number of texts in dataset tends to reduce the differences between the averages and improve outcomes; b) the text structure - texts of the tasks applied in the VLE are usually unstructured, and in many cases provide information out of context (these problems can be reduced by implementing text normalization rules); c) the number of words in the students answers - summaries with the size of 20% were the most harmed in the evaluation because the standard deviation of 358 average words per text can be considered high. We will deeply explore this observation in the future.

6. Conclusions and future works

This paper presented a method that uses fuzzy metrics to determine the most informative sentences in the text, and then compare and classify text written in response to learning tasks with reference responses for these tasks. Fuzzy logic was used because it is a widely recommended approach for applications that handle uncertain information (e.g., ambiguity in linguistic terms) (Das, 2013; Ross, 2010). We show that fuzzy logic associated with text summarization is able to represent the main concepts in natural language on students assessment.

The major contributions of this paper are:

1. description and quantification of measures founded on fuzzy logics for text summarization; and a model for reducing the number of measures necessary for summarizing the texts;
2. a method that uses those text preprocessing techniques to improve performance of text comparison, assessment, and classification;
3. test of the proposal in the context of Computer-Assisted Assessment (CAA) in an VLE;

In fuzzy summarization the quantity of measures used to identify the main concepts can be a problem since the greater the number of measures, the higher the number of fuzzy rules to express knowledge. As we have observed in the literature review the use of a high number of measure in fuzzy summarization systems (shown in Tables 1 and 2), we tested our method with correlated measures. Two different textual characteristics were used to compose one measure by employing a model obtained through multiple linear regression. We reduced the number of input variables in the fuzzy inference system which implies the decrease of fuzzy rules without optimization.

In this work, it is understood that the summary of a free-text answer is designed to inform assessors and students the global communicative intent of the original text. It is an aid material for learning, as pointed out by previous works (shown in Section 3). However, more comprehensive and systematic studies are needed to evaluate its use in the assessment process.

The evaluation of the proposed method demonstrates that the selected text summarization measure satisfies the objective of a CAA method with superior performance to the methods considered in related works. The summaries produced by the proposed method presented the main key concepts of reference models defined by the professor. However, to implement the proposed method in a VLE as the only form of assessment other comparative studies are needed.

The overall performance of our method depends crucially on the efficiency of the measures and the text evaluation. In the text evaluation unigrams, answers containing few words provide more accurate results than those with many words. Therefore, to improve the performance of our method, it is necessary to deepen the analysis with semantic verification of the terms (e.g., utilizing lexical bases to identify word synonyms to improve the matching of synonym words between the summary and reference models). The implementation of the proposed method in a VLE can contribute to the automation of AAC.

The fuzzy summarization not only improves informativeness of the summary but also helps to identify the concepts that are suitable to guide study by teachers and students. We investigate one solution for this essays correlating different features in a model. We are proving that the correlating features can be an effective alternative to reduce data dimensionality for summarizers systems that use many measures in their design.

Acknowledgments

This work was conducted during a scholarship supported by the International Cooperation Program CAPES at the Pompeu Fabra University. Financed by CAPES Brazilian Federal Agency for Support and Evaluation of Graduate Education within the Ministry of Education of Brazil to Fabio Bif Goularte, PDSE - Sandwich Doctoral Programme, process: 88881.135663/2016-01. Horacio Saggion is (partly) supported by the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502) and by the TUNER project (TIN2015-65308-C5-5-R, MINECO/FEDER, UE).

References

Abbasi-ghalehtaki, R., Khotanlou, H., & Esmaeilpour, M. (2016). Fuzzy evolutionary cellular learning automata model for text summarization. *Swarm and Evolutionary Computation, 30*, 11–26.

Abdi, A., Idris, N., Alguliev, R. M., & Aliguliyev, R. M. (2015). Automatic summarization assessment through a combination of semantic and syntactic information for intelligent educational systems. *Information Processing & Management, 51*(4), 340–358.

Aggarwal, C. C., & Zhai, C. (2012). *Mining text data*. Springer Science & Business Media.

Alencar, F. E. S., Magalhães, R. M., & Diniz, F. A. (2013). Um sistema para o gerenciamento e correção de avaliações objetivas em dispositivos móveis. In *Proceedings of international conference on engineering and computer education: 8* (pp. 128–132).

Almahy, I., & Salim, N. (2014). Web discussion summarization: Study review. In *Proceedings of the first international conference on advanced data and information engineering (daeng-2013)* (pp. 649–656). Springer.

de Ávila, R. L., & Soares, J. M. (2013). Textual preprocessing techniques and comparison algorithms to support the correction of questions: experiments, analyzes and contributions. In *Proceedings of the brazilian symposium of informatics in education: 24* (p. 727).

Balage Filho, P. P., Pardo, T. A. S., & Nunes, M. d. G. V. (2007). *Sumarização automática de textos científicos: Estudo de caso com o sistema gistsumm*. ICMC-USP.

Baralis, E., Cagliero, L., & Farinetti, L. (2015). Generation and evaluation of summaries of academic teaching materials. In *Computer software and applications conference (compsac), 2015 ieee 39th annual: 2* (pp. 881–886). IEEE.

Basu, S., Jacobs, C., & Vanderwende, L. (2013). Powergrading: A clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics, 1*, 391–402.

Binwahlan, M. S., Salim, N., & Suanmali, L. (2010). Fuzzy swarm diversity hybrid model for text summarization. *Information Processing & Management, 46*(5), 571–588.

Bird, S. (2006). Nltk: the natural language toolkit. In *Proceedings of the coling/acl on interactive presentation sessions* (pp. 69–72). Association for Computational Linguistics.

Boran, F. E., Akay, D., & Yager, R. R. (2014). A probabilistic framework for interval type-2 fuzzy linguistic summarization. *Fuzzy Systems, IEEE Transactions on, 22*(6), 1640–1653.

Boran, F. E., Akay, D., & Yager, R. R. (2016). An overview of methods for linguistic summarization with fuzzy sets. *Expert Systems with Applications, 61*, 356–377.

Brew, C., & Leacock, C. (2013). Automated short answer scoring. In *Handbook of automated essay evaluation: Current applications and new directions* (p. 136).

Brown, S., Bull, J., & Race, P. (2013). *Computer-assisted assessment of students*. Routledge.

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education, 25*(1), 60–117.

Cambria, E., & White, B. (2014). Jumping nlp curves: A review of natural language processing research [review article]. *IEEE Computational Intelligence Magazine, 9*(2), 48–57.

Carbonaro, A. (2010). Towards an automatic forum summarization to support tutoring. In *Technology enhanced learning. quality of teaching and educational reform* (pp. 141–147). Springer.

Carvalho, J. P., Batista, F., & Coheur, L. (2012). A critical survey on the use of fuzzy sets in speech and natural language processing. In *Fuzzy systems (fuzz-ieee), 2012 ieee international conference on* (pp. 1–8). IEEE.

Chang, M.-M. (2005). Applying self-regulated learning strategies in a web-based instructionan investigation of motivation perception. *Computer Assisted Language Learning, 18*(3), 217–230.

Chen, C.-M., Lee, H.-M., & Chen, Y.-H. (2005). Personalized e-learning system using item response theory. *Computers & Education, 44*(3), 237–255.

Das, S. (2013). Pattern recognition using the fuzzy c-means technique. *International Journal of Energy, Information and Communications, 4*(1), 1–14.

DUC (2003). Workshop on text summarization. *NIST*. Edmonton, Canada: HLT-NAACL. http://duc.nist.gov/pubs.html.

DUC (2006). Workshop on text summarization. *NIST*. New York Marriott at the Brooklyn Bridge Brooklyn, New York, USA: HLT-NAACL. http://duc.nist.gov/pubs.html.

Ferreira, R., de Souza Cabral, L., Lins, R. D., e Silva, G. P., Freitas, F., Cavalcanti, G. D., et al. (2013). Assessing sentence scoring techniques for extractive text summarization. *Expert systems with applications, 40*(14), 5755–5764.

Goularte, F. B., Wilges, B., Nassar, S. M., & Cislaghi, R. (2014). Metrics of automatic text summarization tasks in virtual learning environment. In *Proceedings of the brazilian symposium of informatics in education: 25* (p. 752).

Hannah, M. E., Geetha, T., & Mukherjee, S. (2011). Automatic extractive text summarization based on fuzzy logic: A sentence oriented approach. In *Swarm, evolutionary, and memetic computing* (pp. 530–538). Springer.

Haque, Md. A. & Rahman, T. (2014). Sentiment analysis by using fuzzy logic. arXiv preprint arXiv:1403.3185,.

He, Y., Hui, S. C., & Quan, T. T. (2009). Automatic summary assessment for intelligent tutoring systems. *Computers & Education, 53*(3), 890–899.

Heilman, M., & Madnani, N. (2015). The impact of training data on automated short answer scoring performance. *Silver Sponsor*, 81–85.

Jordan, S., & Mitchell, T. (2009). E-assessment for learning? the potential of short--answer free-text questions with tailored feedback. *British Journal of Educational Technology, 40*(2), 371–385.

Jurafsky, D., & Martin, J. H. (2009). Speech and language processing: An introduction to natural language processing. *Computational Linguistics and Speech Recognition, 2nd Ed.*.

Karanikolas, N. N. (2010). Computer assisted assessment (caa) of free-text: Literature review and the specification of an alternative caa system. In *2010 19th ieee international workshops on enabling technologies: Infrastructures for collaborative enterprises* (pp. 116–118). IEEE.

Kearns, L. R. (2012). Student assessment in online learning: Challenges and effective practices. *Journal of Online Learning and Teaching, 8*(3), 198.

Kiani, A., & Akbarzadeh, M. R. (2006). Automatic text summarization using hybrid fuzzy ga-gp. In *Fuzzy systems, 2006 ieee international conference on* (pp. 977–983). IEEE.

Kitchenham, B. (2004). *Procedures for performing systematic reviews* (33, pp. 1–26). Keele, UK, Keele University.

Kumar, Y. J., Salim, N., Abuobieda, A., & Albaham, A. T. (2014). Multi document summarization based on news components using fuzzy cross-document relations. *Applied Soft Computing, 21*, 265–279.

Kyoomarsi, F., Khosravi, H., Eslami, E., & Davoudi, M. (2010). Extraction-based text summarization using fuzzy analysis. *Iranian Journal of Fuzzy Systems, 7*(3), 15–32.

Kyoomarsi, F., Khosravi, H., Eslami, E., Dehkordy, P. K., & Tajoddin, A. (2008). Optimizing text summarization based on fuzzy logic. In *Seventh ieee/acis international conference on computer and information science* (pp. 347–352). IEEE.

Kyoomarsi, F., Rahmani, F., Eslami, E., Dehkordy, P., & Tajoddin, A. (2011). Text summarization based on cellular automata. *International conference on information communication and management*.

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities, 37*(4), 389–405.

Leite, D., & Rino, L. H. (2009). A genetic fuzzy automatic text summarizer. In *Anais do csbc 2009: 1* (pp. 779–788). SBC.

Lin, C.-Y., & Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 conference of the north american chapter of the association for computational linguistics on human language technology-volume 1* (pp. 71–78). Association for Computational Linguistics.

Lloret, E., & Palomar, M. (2012). Text summarisation in progress: A literature review. *Artificial Intelligence Review, 37*(1), 1–41.

Matsubara, E. T., Martins, C. A., & Monard, M. C. (2003). Pretext: A pre-processing text tool using the bag-of-words approach. *Technical Re-port, 209*.

Megala, S. S., Kavitha, A., & Marimuthu, A. (2014). Enriching text summarization using fuzzy logic. *International Journal of Computer Science and Information Technologies, 5*(1).

Megala, S. S., Kavitha, A., & Marimuthu, A. (2015). Text summarization system using fuzzy logic and conditional random field algorithm. *International Journal of Computer Science and Information Technologies, 1*(5), 863–867.

Mohler, M., Bunescu, R., & Mihalcea, R. (2011). Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1* (pp. 752–762). Association for Computational Linguistics.

Morais, A. M., & Arajo, J. M. F. R. (2013). Educational data mining for support e-learning teacher based on decision tree. In *Proceedings of the conferncia iadis ibero-americana www/internet* (pp. 141–148). IADIS.

Nenkova, A., & McKeown, K. (2012). A survey of text summarization techniques. In *Mining text data* (pp. 43–76). Springer.

Noorbehbahani, F., & Kardan, A. A. (2011). The automatic assessment of free text answers using a modified bleu algorithm. *Computers & Education, 56*(2), 337–345.

Palmer, K., & Richardson, P. (2003). On-line assessment and free-response input-a pedagogic and technical model for squaring the circle. In *Proceedings of the 7th computer assisted assessment conference* (pp. 289–300).

Pascual-Nieto, I., Santos, O. C., Perez-Marin, D., & Boticario, J. G. (2011). Extending computer assisted assessment systems with natural language processing, user modeling, and recommendations based on human computer interaction and data mining. In *Ijcai proceedings-international joint conference on artificial intelligence: 22* (p. 2519). Citeseer.

Pérez-Marín, D., Pascual-Nieto, I., & Rodríguez, P. (2009). Computer-assisted assessment of free-text answers. *The Knowledge Engineering Review, 24*(04), 353–374.

Pope III, F., Shirvani, R. A., Rwebangira, M. R., Chouikha, M., Taylor, A., & Ramirez, A. A., et al. (2016). Automatic detection of small groups of persons, influential members, relations and hierarchy in written conversations using fuzzy logic. arXiv preprint arXiv:1610.01720,.

Pope III, F. D. (2016). *Fuzzy logic clustering algorithm to natural language processing on the usage of the second person pronoun:" you"*. Howard University Ph.D. thesis..

Radev, D. R., Hovy, E., & McKeown, K. (2002). Introduction to the special issue on summarization. *Computational Linguistics, 28*(4), 399–408.

Ramos-Soto, A., Bugarín, A., & Barro, S. (2016). On the role of linguistic descriptions of data in the building of natural language generation systems. *Fuzzy Sets and Systems, 285*, 31–51.

Ramos-Soto, A., Tintarev, N., De Oliveira, R., Reiter, E., & Van Deemter, K. (2016). Natural language generation and fuzzy sets: An exploratory study on geographical referring expression generation. In *Fuzzy systems (fuzz-ieee), 2016 ieee international conference on* (pp. 587–594). IEEE.

Rautray, R., & Balabantaray, R. C. (2017). An evolutionary framework for multi document summarization using cuckoo search approach: Mdscsa. *Applied Computing and Informatics*.

Rodrigues, F., & Araújo, L. (2012). Automatic assessment of short free text answers.. In *Csedu (2)* (pp. 50–57).

Rodrigues, F., & Oliveira, P. (2014). A system for formative assessment and monitoring of students' progress. *Computers & Education, 76*, 30–41.

Rodrigues, R. C. R., & Rocha, F. E. L. (2011). Webcmtool: Um ambiente web para facilitar a avaliao da aprendizagem baseado em mapas conceituais e ontologias de domnio. In *Proceedings of the conferncia iadis ibero-americana www/internet* (pp. 219–226). IADIS.

Ross, T. (2010). *Fuzzy logic with engineering applications*. John Wiley & Sons.

Saggion, H., & Poibeau, T. (2013). Automatic text summarization: Past, present and future. In *Multi-source, multilingual information extraction and summarization* (pp. 3–21). Springer.

Sankarasubramaniam, Y., Ramanathan, K., & Ghosh, S. (2014). Text summarization using wikipedia. *Information Processing & Management, 50*(3), 443–461.

Saraswathi, S., Hemamalini, M., Janani, S., & Priyadharshini, V. (2011). Multi-document text summarization in e-learning system for operating system domain. In *Advances in computing and communications* (pp. 175–186). Springer.

Schofield, J. (2005). The statistically unreliable nature of lines of code. *CrossTalk Apr*.

Selvi, P., & Bnerjee, A. (2010). Automatic short-answer grading system (asags). arXiv preprint arXiv:1011.1742,.

Shrivastav, H., & Hiltz, S. R. (2013). Information overload in technology-based education: a meta-analysis. In *Proceedings of the amcis*. AMCIS.

Sim, G., Holifield, P., & Brown, M. (2004). Implementation of computer assisted assessment: lessons from the literature. *Research in Learning Technology, 12*(3).

Sivanandam, S. N., Sumathi, S., & Deepa, S. N. (2007). *Introduction to fuzzy logic using MATLAB*: 1. Springer.

Suanmali, L., Binwahlan, M. S., & Salim, N. (2009). Sentence features fusion for text summarization using fuzzy logic. In *Hybrid intelligent systems, 2009. his'09. ninth international conference on: 1* (pp. 142–146). IEEE.

Suanmali, L., Salim, N., & Binwahlan, M. S. (2009). Fuzzy logic based method for improving text summarization. arXiv preprint arXiv:0906.4690,.

Suanmali, L., Salim, N., & Binwahlan, M. S. (2011). Fuzzy genetic semantic based text summarization. In *Dependable, autonomic and secure computing (dasc), 2011 ieee ninth international conference on* (pp. 1184–1191). IEEE.

TOEFL (2017). Test of english as a foreign language.

Wilbik, A., & Keller, J. M. (2012). A distance metric for a space of linguistic summaries. *Fuzzy Sets and Systems, 208*, 79–94.

Witte, R., & Bergler, S. (2007). Fuzzy clustering for topic analysis and summarization of document collections. In *Advances in artificial intelligence* (pp. 476–488). Springer.

Xia, Z.-Q., & Guo, F.-F. (2004). Fuzzy metric spaces. *Journal of Applied Mathematics and Computing, 16*(1–2), 371–381.

Yager, R., Ford, K., & Cañas, A. (1991). An approach to the linguistic summarization of data. *Uncertainty in Knowledge Bases*, 456–468.

Yang, G., Chen, N.-S., Kinshuk, Sutinen, E., Anderson, T., & Wen, D. (2013). The effectiveness of automatic text summarization in mobile learning contexts. *Computers & Education, 68*, 233–243.

Yang, G., Wen, D., Kinshuk, Chen, N. S., & Sutinen, E. (2012). Personalized text content summarizer for mobile learning: An automatic text summarization system with relevance based language model. In *Technology for education (t4e), 2012 ieee fourth international conference on* (pp. 90–97). IEEE.

Zadeh, L. A. (1999). Fuzzy sets as a basis for a theory of possibility. *Fuzzy sets and systems, 100*, 9–34.