Pamukkale Univ Muh Bilim Derg, 27(2), 139-150, 2021

Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi

Pamukkale University Journal of Engineering Sciences

Certainty factor model in paraphrase detection

Eşanlatım tespitinde eminlik faktörü modeli

Senem KUMOVA METİN[1*], *Bahar KARAOĞLAN*[2], *Tarık KIŞLA*[3], *Katira SOLEYMANZADEH*[4]

[1]Department of Software Engineering, Faculty of Engineering, İzmir University of Economics, İzmir, Turkey.
senem.kumova@ieu.edu.tr

[2]International Computer Institute, Ege University, İzmir, Turkey.
bahar.karaoglan@ege.edu.tr, katira.sole@gmail.com

[3]Department of Computer Education and Instructional Technology, Faculty of Education, Ege University, İzmir, Turkey.
tarik.kisla@ege.edu.tr

Received/Geliş Tarihi: 08.10.2019 Revision/Düzeltme Tarihi: 24.04.2020 doi: 10.5505/pajes.2020.75350
Accepted/Kabul Tarihi: 25.04.2020 Research Article/Araştırma Makalesi

Abstract

In this paper, we address the problem of uncertainty management in identification of paraphrase sentence pairs. Paraphrase sentences are simply sets/pairs of sentences that express the same facts and/or opinions using different words or order of words. We propose the use of certainty factor (CF) model in paraphrase detection. A set of succeeding paraphrase detection features (generic and distance based features) is built by filtering and this set is used as evidences in CF model. The CF model is evaluated by F1 and accuracy measures on Microsoft Research Paraphrase corpus. The results are compared to the well-known Bayesian reasoning. The experimental results showed that CF model is an alternating paraphrase detection method to Bayes model.

Keywords: Paraphrase, Paraphrase detection, Certainty factor, Evidence, Evidence selection.

Öz

Bu makalede, eşanlatımlı cümle çiftlerinin belirlenmesindeki belirsizlik problemi üzerinde durulmuştur. Eşanlatım cümleleri basitçe aynı olay ve/veya fikri farklı sözcük veya sözcüklerin farklı dizilişleri ile ifade eden cümle çiftleri/kümeleridir. Çalışmada eşanlatım tespitinde eminlik faktörü (EF) modelinin kullanılması önerilmiştir. EF modelinde kullanılmak üzere filtreleme yöntemi ile eşanlatım tespitinde başarılı olan öznitelikler (jenerik ve uzaklık tabanlı öznitelikler) belirlenmiş ve bu öznitelikler kümesi EF modelinde deliller olarak kullanılmıştır. EF modeli Microsoft Eşanlatım derlemi üzerinde F1 ve doğruluk ölçekleri ile sınanmıştır. Yöntemin başarımı Bayes karar verme yaklaşımı ile kıyaslanmıştır. Deney sonuçları EF modelinin eşanlatım tespitinde Bayes modeline bir alternatif yöntem olduğunu göstermiştir.

Anahtar kelimeler: Eşanlatım, Eşanlatım tespiti, Eminlik faktörü, Delil, Delil seçimi.

1 Introduction

Text similarity measurement underlies the major language understanding and processing tasks like spelling checking, text classification (concept identification, emotion analysis, new event detection and tracking, etc.), summarization, machine translation and many more. In this study, we propose certainty factor (CF), which is used in expert systems, as a metric to measure text similarity, within the scope of paraphrase identification.

The paraphrase pairs of text are described as two passages of text where the same meaning is to be given to the reader or the listener. The same sentence might be understood differently by different people. From this point of view, we see language understanding as an expert task and applying expert approaches may be a remedy. In expert systems, certainty factor model is introduced as an alternative to Bayesian reasoning to cope with the problems where the uncertainty exists. The theory is firstly proposed by Shortliffe and Buchanan [1] in MYCIN (an expert system in diagnosis and therapy of blood infections and meningitis) due to the lack of reliable statistical data in domain and mathematically inconsistent and/or illogical expressions of experts on the strength of their beliefs.

In this study, we present CF model as an alternative to Bayesian reasoning considering paraphrase detection as an expert system problem where the required information to decide on the type of the text pair may be incomplete, inconsistent or uncertain. The general CF model requires pre-determined evidences, rules, human-expert's degree of belief/disbelief to the evidences/rules and a structure to accumulate the whole set of rule-evidence pairs. In the paraphrase identification, we propose the CF model that consists of 3 main processes: *evidence selection*, *rule formulation* and *rule accumulation*. The evidences are selected from a set of 17 features that are categorized in two: generic features (e.g. sentence length ratio, word overlap ratio, word ordering ratio, common word group ratio) and distance-based features (e.g. Jaccard distance, Euclidean distance). In rule formulation process, several parameters such as the value-range of evidences, the degree of belief/disbelief to the evidences and the rules are determined. Finally, the rules are fired in order similar to the standard CF model that will be detailed in section 3.2.

The proposed CF model in paraphrase identification is realized by utilizing the renowned paraphrase corpus of Microsoft Research (MSRP) [2] that is stated to be a standard resource in paraphrase identification studies [3]. The evaluation is performed by F1 and accuracy measures.

The main contribution of the study is bringing an expert system look on paraphrase identification problem. The experimental results revealed that the proposed CF model, which is a rule-based expert system model, is promising in determination of

*Corresponding author/Yazışılan Yazar

paraphrase pairs when compared to the traditional Bayesian reasoning model. In addition, it is showed that in paraphrase identification task, entropy based measures may be used as an alternative to the human-experts beliefs in order to set parameters of the reasoning models.

The paper is organized as following. We first review related work in Section 2. The mathematical background on proposed method and Bayesian reasoning is presented in Section 3. Section 4, 5 and 6 give the overall methodology, experimental results and conclusion respectively.

2 Related work

The earliest text similarity detection studies were mainly on information retrieval area where the relevant documents to user queries were to be detected [4]. Following these studies, the text similarity is used in a variety of different areas such as text classification, word sense disambiguation [5], summarization [6] and automatic assessment of machine translation [7].

Identification of paraphrase sentence pairs bases on measuring semantic similarity between two texts considering some syntactic or semantic features. The identification methods mainly depend on machine learning techniques where it is possible to assess the combined impact of different features. In Table 1, a number of different references on paraphrase detection where MSRP corpus is utilized are listed together with the methods and/or the features that are employed, the type of the machine learning algorithm and the classification performance results of those approaches.

In this study, some of the features presented in the works given in Table 1 are employed and the experiments are run on the same corpus to enable comparable results. Below, the works in Table 1 will be briefly explained.

In an earlier study on MSRP corpus, Zhang and Patrick [8] transformed sentence pairs of MSRP corpus to a generic and simpler form that is introduced as the canonicalized text. The canonicalized texts are given as inputs to a decision tree that employs lexical matching features such as longest common subsequence, edit distance for supervised learning process. In a similar effort in paraphrase classification, in [9], the utility of machine translation evaluation methods such as BLEU [10] and NIST [11], are investigated and a new classification method is proposed. In a more recent work, [12], machine translation metrics are re-examined, a meta-classifier that considers the weighted probability estimates of three classifiers is trained.

Kozareva and Montoyo [13] considered paraphrase identification as a classification task and used lexical and semantic features in supervised methods (e.g. support vector machines, k-nearest neighbour and maximum entropy) to classify the data set in two classes as paraphrase and non-paraphrase. In the study, semantic features that are extracted from WordNet [14],[15], lexical features such as longest common subsequence that are used in a variety of studies are utilized. In [16], it is proposed to use an enhanced set of similar lexical features together with semantic heuristics in machine learning methods.

Rus et al. in [17] proposed a method based on lexico-syntactic graph-subsumption that uses word orderings, synonym and antonym information. The synonym and antonym information is extracted from WordNet [14],[15] and the linguistic information is represented in a graph structure. The paraphrasing is detected considering the existence of subsumption relation between the graphs of the sentences in the regarding pair. In [18], unlike the majority of previous studies, the main goal is stated as making paraphrasing judgement based on the significance of dissimilarity between the sentences instead of similarity

Table 1. A number of paraphrase identification studies utilizing *MSRP* corpus.

Reference	Methods/features	Type	Accuracy	*F1*
Zhang and Patrick [8]	Text canonicalization	supervised	0.703	0.795
Mihalcea, Corley, and Strapparava [7]	Word-to-word similarity features	unsupervised	0.703	0.813
Rus et al. [17]	Lexico-Syntactic graph subsumption	unsupervised	0.706	0.805
Qiu, Kan, and Chua [18]	Dissimilarity classification	supervised	0.720	0.816
Islam and Inkpen [20]	Combination of semantic and syntactical features	unsupervised	0.726	0.813
Fernando and Stevenson [3]	Wordnet measure and vector based similarity	unsupervised	0.741	0.824
Ul-Qayyum and Altaf [16]	Semantic heuristic features	supervised	0.747	0.818
Finch, Hwang, and Sumita [9]	Machine translation methods	supervised	0.750	0.827
Wan et al. [22]	Dependency-based features	supervised	0.756	0.830
Kozareva and Montoyo [13]	Lexical and semantic similarity features	supervised	0.766	0.796
Socher, Huang and Pennington [21]	Dynamic pooling and unfolding recursive auto-encoders	supervised	0.768	0.836
Madnani, Tetreault and Chodorow [12]	Machine translation metrics	supervised	0.774	0.841
Wang et al. [23]	Sentence similarity learning by lexical decomposition and composition	supervised	0.784	0.847
He et al. [24]	Multi-perspective convolutional neural networks and structured similarity layer	supervised	0.786	0.847
Cheng and Kartsaklis [25]	Recursive neural networks using syntax-aware multi-sense word embeddings	supervised	0.786	0.853
Filice et al. [26]	Combination of convolution kernels and similarity scores	supervised	0.791	0.852

The proposed method requires two phases. In the first phase, the common information nuggets or individual semantic content units in the sentences are defined. It is assumed that if the pair is a paraphrase pair then the sentences must share some amount of these nuggets/units. Secondly, uncommon nuggets are found and they are classified as significant or not by an SVM.

Fernando and Stevenson in [3] offered the use of a similarity matrix in paraphrase identification and experimented on MSRP corpus. In this approach, similarity between the sentences a and b that are represented by binary vectors (with elements equal to 1 if a word is present and 0 otherwise), $\vec{a}$ and $\vec{b}$, can be computed using the following formula:

$$sim(\vec{a},\vec{b}) = \frac{\vec{a}W \cdot \vec{b}}{|\vec{a}||\vec{b}|} \quad (1)$$

where W is the matrix containing the information about the similarity of word pairs. W matrix is populated by six WordNet similarity metrics (e.g. Lesk [19]). Fernando and Stevenson [3] stated that their approach performs better than previously published methods.

In literature, it is observed that many of the researchers utilized word-based (word-to-word) similarity methods in paraphrase identification. For example, a word-based similarity method that uses the features such as semantic word similarity and a modified and normalized version of the longest common subsequence is proposed in [20]. One other word-based similarity approach is presented in [7]. The knowledge and corpus-based metrics such as WordNet based similarity, latent semantic analysis and point-wise mutual information are employed to identify the paraphrase pairs in [7]. The metrics are combined by a function that considers the word similarity. The sentences that produced similarity values higher than the predefined threshold value (= 0.5) are classified as paraphrased pair.

In [21], the texts are stored in a tree-based structure and a recursive auto-encoder is used to measure similarity features in an unsupervised manner and the texts with different lengths are made comparable by dynamic pooling.

In paraphrase recognition, Wan et al. [22] employed syntactical features that are extracted from dependency trees grounding on the idea that the dependency trees of paraphrase/similar sentences must have also similar alignments. In this study, features extracted from trees are used together with machine translation methods.

In more recent works, similar to the proposed solutions to other problems in natural language processing field, different types of neural networks (convolutional or recursive) [23]-[25], and/or combinations of neural networks [26] and/or vector space representations such as word embeddings are being used to increase performance in paraphrase identification though their black box nature and computational burden.

3 Mathematical background

In this section we give brief explanation regarding the mathematical instruments on which we base our methodology. Bayesian reasoning is taken as the baseline for performance assessment of the proposed method, CF model, in decision making to handle uncertainty with an expert approach. Entropy based measures, information gain and information gain ratio, are employed in selection of effective evidences on the classification of sentence pairs as paraphrase or not and the value ranges of the evidences that fit to classes.

3.1 Bayesian Reasoning

In Bayes decision theory, it is stated that the probability of an event may change after it has been learned that some other event has occurred. The new probability is called the conditional probability of the event H given that event E is true. In hypothesis testing, event H represents the hypothesis and event E is accepted as an evidence for the regarding hypothesis. The conditional probability is formulated as

$$P(H \mid E) = \frac{P(E \mid H) \times P(H)}{P(E|H) \times P(H) + P(E|\neg H) \times P(\neg H)} \quad (2)$$

where $P(H)$ is the probability of hypothesis H being true. $P(\neg H)$ is the probability of H being false. $P(E|H)$ represents the probability of evidence E to be observed given H is true, and $P(E|\neg H)$ is the conditional probability of evidence E given that H is false. In cases where the uncertainty on H is reduced by observing multiple independent evidences E_1, E_2, E_3 ... E_n, the conditional probability of H expands to

$$P(H \mid E_1 E_2 \dots E_n) = \frac{\prod_{i=1}^{n} P(E_i \mid H) \; xP(H)}{\prod_{i=1}^{n} P(E_i \mid H) \; xP(H) + \prod_{i=1}^{n} P(E_i \mid \neg H) \; xP(\neg H)} \quad (3)$$

In rule-based expert systems that employ Bayesian reasoning, the rules in knowledge base are written in the following form:

IF E_i is true { LS, LN}

THEN H is true {prior $P(H)$}

where LS is likelihood of sufficiency and LN is likelihood of necessity. Further information on LS and LN can be found in [27].

In identification of paraphrase sentence pairs, H is defined as the hypothesis that states "The sentence pair is a paraphrase pair" and $\neg H$ is the hypothesis that is "The sentence pair is not a paraphrase pair". The text similarity features are employed as evidences that trigger the change in probability of both hypotheses. For each evidence, a pair of rules; one rule for H and one rule for $\neg H$; is defined as the example given in Figure 1. The rules are fired one by one and the resulting $P(H)$ and $P(\neg H)$ values are obtained. Finally, if $P(H) \leq P(\neg H)$ the sentence pair is classified as paraphrase and vice versa.

Rule 1:
IF Two sentences involve at least 3 words in common {LS=0.8 LN= 0.2}
THEN The sentence pair is a paraphrase pair {*prior* 0.2}
Rule 2:
IF Two sentences don't involve at least 3 words in common {LS=0.6 LN= 0.5}
THEN The sentence pair is not a paraphrase pair {*prior* 0.8}

Figure 1. A pair of rules in Bayesian reasoning (LS, LN and prior probability values are given randomly).

3.2 Certainty factor model

The rule-based expert systems handle uncertainty in decision problems by the help of two notions: experience and the expertise. The classical approach in rule-based systems considering those notions is the Bayesian reasoning. In Bayesian reasoning, conditional probabilities are employed to handle uncertain cases and simply degree of probability to an outcome is measured. One of the major problems in Bayesian

reasoning is that when some evidence E is observed, the belief in hypothesis H to be true is represented by $P(H \mid E)$). The belief in the opposite hypothesis H' is formulated simply as $P(H' \mid E) = 1 - P(H \mid E)$ though in real life problems there may be cases where $P(H' \mid E) \neq 1 - P(H \mid E)$. In such cases, an alternating approach to classical Bayesian reasoning is required. Shortliffe and Buchanan [1] proposed certainty factor model in order to handle such uncertainties in practice.

In certainty factor model, the concept of certainty factor (cf) is proposed as a number to measure the expert's belief, which ranges between -1 and 1. There exist three main cf values in the model. The first of cf values is. cf_{rule} is used to represent the degree of belief in hypothesis when the evidence is observed. The second cf value, cf_{evidence}, gives the degree of belief in the evidence. A positive cf value represents a degree of belief and a negative value shows a degree of disbelief. That is to say, $cf = 1$ means a complete belief and $cf = -1$ vice versa.

In certainty factor theory, the knowledge base includes the rules that have the following syntax:

IF Evidence E is true

THEN Hypothesis H is true { cf_{rule} }

where cf_{rule} represents belief in hypothesis H given that evidence E has occurred. cf_{rule} value is formulated as follows:

$$cf_{rule} = \frac{MB(H,E) - MD(H,E)}{1 - min[MB(H,E), MD(H,E)]} \tag{4}$$

where is $MD(H,E)$ measure of disbelief and $MB(H,E)$ is measure of belief. Measure of belief is the degree to which belief in hypothesis would be increased if evidence E is observed. Measure of disbelief is the degree to which disbelief in hypothesis would be increased by observing the evidence [28]. $MD(H,E)$ and $MB(H,E)$ that ranges between 0 and 1 are given as

$$MB(H,E) = \begin{cases} 1 & if\ P(H) = 1 \\ \dfrac{max[p(H|E), P(H)] - P(H)}{max[1,0] - P(H)} & otherwise \end{cases}$$

$$MD(H,E) = \begin{cases} 1 & if\ P(H) = 0 \\ \dfrac{min[P(H|E), P(H)] - P(H)}{min[1,0] - P(H)} & otherwise \end{cases}$$

where $P(H)$ is the prior probability of hypothesis H being true and $P(H|E)$ is the probability that hypothesis H is true given evidence E.

In cases where the expert's belief in evidence is also uncertain, the net certainty for a single rule, cf_{net} (the third cf value), is calculated by multiplying the certainty factor of the evidence $cf_{evidence}$ and the certainty factor of the rule cf_{rule}.

$$cf_{net} = cf(H,E) = cf_{evidence} \times cf_{rule} \tag{5}$$

For rules where multiple evidences that are combined by "AND" or "OR" statements, the net certainty of the hypothesis/rule is calculated considering the whole set of evidences.

For conjunctive rules such as

IF <evidence E_1>

AND <evidence E_2>

AND <evidence E_3>

…

AND <evidence E_n>

THEN <hypothesis> $\{cf_{rule}\}$

The net certainty is established as follows

$$cf_{net} = cf(H.E_1 \cap E_2 \cap E_3 \cap \dots \cap E_n) = min\left[cf_{evidence\_1}.cf_{evidence\_2} \dots cf_{evidence\_n}\right] \times cf_{rule} \tag{6}$$

For disjunctive rules such as

IF <evidence E_1>

OR <evidence E_2>

OR <evidence E_3>

…

OR <evidence E_n>

THEN <hypothesis> $\{cf_{rule}\}$

The certainty of the hypothesis is given as

$$cf_{net} = cf(H.E_1 \cup E_2 \cup E_3 \cup \dots \cup E_n) = max\left[cf_{evidence\_1}.cf_{evidence\_2} \dots cf_{evidence\_n}\right] \times cf_{rule} \tag{7}$$

In CF model, the accumulation of the rules on the same hypothesis is performed by merging the individual net certainty factors of the rules. Suppose that the knowledge base includes following two rules:

Rule 1: IF "A is X"

THEN H $\{cf_{rule\_1}\}$

Rule 2: IF "B is Y"

THEN H $\{cf_{rule\_2}\}$

Firing the first rule, we obtain cf_{net1} value of hypothesis H when evidence "A is X" is observed, $cf_{net1} = cf_{evidence}$("A is X") $\times$ cf_{rule_1}. Similarly second rule is fired when "B is Y" is true, the certainty factor value is $cf_{net2} = cf_{evidence}$("B is Y") $\times$ cf_{rule_2}. The combined certainty factor value is obtained by the following equation.

$$cf_{net\ 1+2} = \begin{cases} if\ cf_{net1} > 0\ and\ cf_{net2} > 0 \\ cf_{net1} + cf_{net2} \times (1 - cf_{net1}) \\ \\ if\ cf_{net1} < 0\ \ or\ cf_{net2} < 0 \\ \dfrac{cf_{net1} + cf_{net2}}{1 - min[|cf_{net1}|.|cf_{net2}|]} \\ \\ if\ cf_{net1} < 0\ and\ cf_{net2} < 0 \\ cf_{net1} + cf_{net2} \times\ (1 + cf_{net1}) \end{cases} \tag{8}$$

Similar to Bayesian reasoning, in CF model, it is accepted that there exists two hypotheses to test in order to identify paraphrase/non-paraphrase sentence pairs by employing text similarity features as evidences. The first hypothesis is that the given sentence pair is a paraphrase pair. The second is that the given pair includes non-paraphrase sentences. To exemplify, assume that the hypothesis is "Given sentence pair is a paraphrase pair" and the evidences are listed as;

E_1 : The number of words that are observed in both sentences is greater than 2,
E_2 : The sentences include same named entities,
E_3 : The sentences have same number of words.

where the certainty factors of evidences in order are $cf_{evidence\_1} = 0.3$, $cf_{evidence\_2} = 0.13$, $cf_{evidence\_3} = 0.15$. In this example, the rules may be stated as

Rule 1: IF The number of words that are observed in both sentences is greater than 2

THEN Given sentence pair is a paraphrase pair

$\{cf_{rule\_1} = 0.70\}$

Rule 2: IF The sentences include same named entities

THEN Given sentence pair is a paraphrase pair

$\{cf_{rule\_2} = 0.40\}$

Rule 3: IF The sentences include words with opposite meanings.

THEN Given sentence pair is a paraphrase pair

$\{ cf_{rule\_3} = -0.60\}$

The cf values of first two rules in our example present that these evidences when observed increase the belief in hypothesis. On the other hand, the negative certainty value given in Rule 3 means that when observed this evidence decreases the belief in the same hypothesis.

Assuming that the evidences "The number of words that are observed in both sentences is greater than 2" and "The sentences include same named entities" are observed/true, the certainty value of the regarding hypothesis is calculated by firing these rules one by one. When Rule 1 is fired the net certainty value is obtained as cf_{net1} =cf(E_1, "Given sentence pair is a paraphrase pair")=0.30x0.70=0.21. The net certainty factor when Rule 2 is fired is cf_{net2} =cf(E_1, "Given sentence pair is a paraphrase pair")=0.13x0.4=0.052. Both cf_{net1} and cf_{net2} are greater than zero as a result the combined certainty value of Rule 1 and Rule 2 is calculated as

$$cf_{net1+2} = cf(cf_{net1}cf_{net2}) = cf_{net1} + cf_{net2} \times (1 - cf_{net1}) = 0.21 + 0.052 \times (1 - 0.21) = 0.251$$

meaning that if first two evidences are observed the belief in hypothesis to be true is 0.251. The last rule has a negative certainty value. cf_{net3} = cf(E_3, "Given sentence pair is a paraphrase pair")= 0.15 x (-0.60)= -0.09 that decreases the belief to the hypothesis. Merging this negative impact to previous combined certainty value

$$cf_{net1+2+3} = cf(cf_{net1+2}cf_{net3}) = \frac{cf_{net1+2} + cf_{net3}}{1 - min[|cf_{net1+2}|.|cf_{net3}|]} = \frac{0.251 + (-0.09)}{1 - min[|0.251|.|-0.09|]} = 0.177$$

is obtained. The resulting net cf value that is close to zero may be interpreted as a weak belief to the hypothesis to be true after considering all regarding evidences.

3.3 Entropy based measures

Mitchell [29] defines information gain (IG) and gain ratio (GR) as measures of the effectiveness of an attribute/feature in classifying training data in decision trees. We employed these measures in two folds. Firstly, both measures are used as attribute evaluators in evidence selection. Secondly, they are employed to determine the value-ranges that classify the data.

IG and GR are determined by well-known notion of entropy. In information theory, entropy is a measure that represents the amount of uncertainty/disorder of samples in a given data set. For example, if all the samples in data set belong to a different class, the uncertainty/disorder reaches to its maximum value. Entropy is defined as follows in dataset S in which n different classes of samples exist.

$$H(S) = -\sum_{i=1}^{n} p_i \, log_2 \, p_i \tag{9}$$

where p_i is the proportion of samples that belongs to the class *i*. Information gain is the reduction of uncertainty in samples based on a specific feature . This is why; as the information gain gets higher the uncertainty gets lower supporting the effective classification. Information gain is calculated as follows

$$IG(S, f) = H(S) - H(S|f) = H(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} H(S_i) \tag{10}$$

where S is the dataset. $H(S|f)$ is the entropy measured given the feature *f* and S_i is the subset *i* that includes samples of class *i*.

Gain ratio (GR) is the ratio of information gain to feature's entropy value. Assuming S is class and f is the regarding feature. GR is determined as follows

$$GR(S, f) = \frac{IG(S, f)}{H(f)} \tag{11}$$

4 Proposed method

In deciding whether a given sentence pair is a paraphrase pair or not, variety of text similarity features may be employed and their joint contribution may be accumulated by several methods. In this study, we propose to formulate rules that accept the outcomes of selected text similarity features as evidences and accumulate the belief/disbelief in paraphrasing by the certainty factor model.

The stages of the proposed method, depicted in Figure 2, may be defined briefly as follows:

1. Evidence Selection: The similarity features that succeed in distinguishing paraphrase and non-paraphrase pairs are selected as evidences,
2. Rule Formulation: CF model requires the propagation of a list of IF-THEN-ELSE rules to decide whether the sentence pair is paraphrase or not. In rule formulation process, for each evidence, a decision rule must be built for both hypotheses: 1)"Given sentence pair is a paraphrase pair". 2) "Given sentence pair is a non-paraphrase pair". In order to generate the rule for a specific evidence-hypothesis pair, firstly the evidence value-range of the hypothesis must be set. The notion of value-range is accepted to be the range where the hypothesis is being strongly supported when the pair's evidence value falls in this range. Secondly, $cf_{evidence}$ must be set based on the expert's belief/disbelief on the given evidence. And finally, a rule for each evidence-hypothesis pair must be formulated by measuring cf_{rule} based on the equations given in section 3.2,

3. Rule Accumulation: At this stage rules are fired for each hypothesis with the evidences collected from the sentence pair and the final decision is made by using the formulas presented in section 3.2.

In the following subsections, firstly the similarity features (evidence candidates) will be presented and then the stages in proposed reasoning system will be defined in detail.

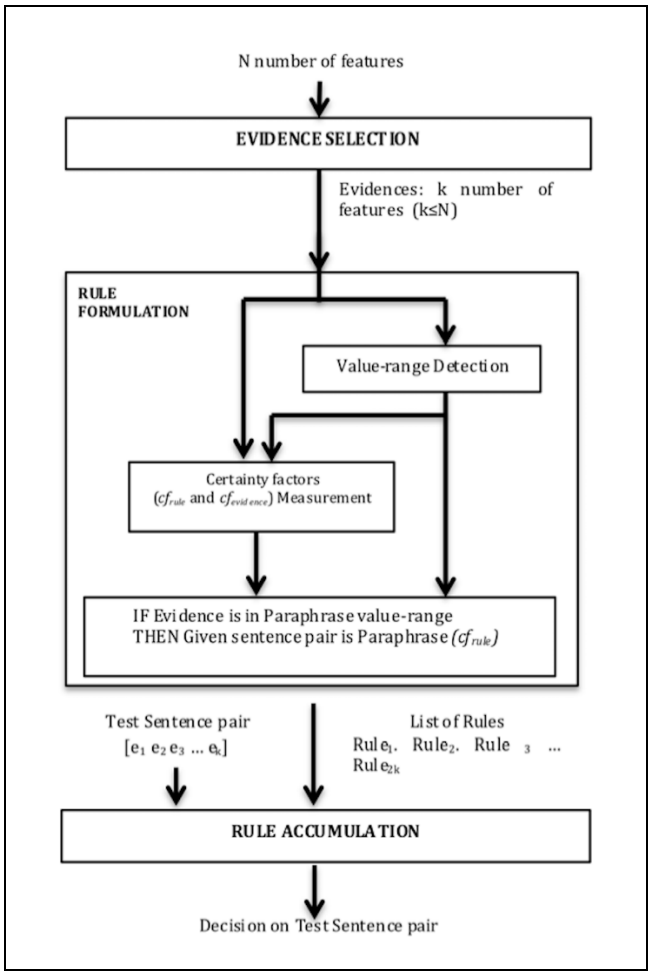

Figure 2. Flow chart of the proposed CF method.

4.1 Similarity features: Evidence candidates

In this study, sentence similarity features that are accepted as evidence candidates are categorized in two groups: generic syntactical features and distance-based features.

The first category of features, generic syntactical features, produces a value in the range [0 1] for each sentence pair. Values closer to 1 indicate higher probability for sentence pair to be paraphrases and values closer to 0 indicate higher probability for sentence pair to be non-paraphrase. The generic features considered in the study are sentence length ratio (LS), matching word ratio (MW), matching POS (Part of Speech) ratio (MW_POS), common word group ratio (MB), common POS group ratio (MB_POS), word ordering ratio (OW) and POS ordering ratio (OW_POS).

Sentence length ratio (LS) is measured by determining the number of words in sentences. The number of words in sentence is accepted as the sentence length. The length of the shorter sentence in sentence pair is divided by the length of the longer sentence in order to obtain sentence length ratio. LS value ranges between 0 and 1 theoretically. LS reaches to its maximum value for the pair that include sentences that have the same number of words.

Matching word ratio (MW) is a feature that indicates the similarity in terms of constituting words in sentences in given sentence pair. The assumption behind this feature is that if two sentences have some words in common, they tend to be paraphrases of each other. MW is calculated by dividing the number of words that occur in both sentences by the number of different words in sentence pair. The feature gets its maximum value, 1, if sentences in pair hold exactly same

words. The minimum MW value is zero in case where there is not a single word that is used in both sentences.

MW is modified to POS overlap ratio (MW_POS) by employing part of speech tags instead of the words. Thus, not only the word overlaps but also the overlaps on part of speeches may be considered in identification of paraphrase pairs. Similar to MW, the range of MW_POS is [0 1]. It gets the value 1 for a complete overlap and 0 for vice versa.

Common word group ratio, matching blocks (MB) is the feature that quantifies the contribution of common word groups to the sentence similarity [30]. It is accepted that in paraphrase pairs, the same word sequences are observed in both sentences. MB is calculated by determining the longest sequences of words that occur in both sentences as follows:

$$MB = \sum_{i=1}^{n} \frac{(LB_i)^2}{L_1 \cdot L_2} \tag{12}$$

Where LB_i is the number of words in i[th] common word sequence. L_1 and L_2 are sentence lengths in pair in terms of their word counts. The same procedure is followed to calculate.

POS group ratio (MB_POS) except that in MB_POS part of speech tags are considered on behalf of words in MB. It is expected that if the MB_POS is close to its maximum value (1), the sentences are paraphrases since they contain same part of speech groups. In case where MB_POS=0, the sentences do not have any common part of speech tag groups, supporting the hypothesis "Given sentence pair is non-paraphrase pair".

Word ordering ratio (OW) measures how similar the order of the words is in given sentences. It is believed that if the words are observed in same order or in almost same order in sentences, the probability of pair being paraphrase increases [18]. In order to attain word-ordering ratio, for each common word in pair, the difference in word position, *PD*, is to be calculated. For the words that are observed only in one of the sentences, *PD* value is accepted to be V where V is the total number of different words in pair. OW of the given pair is obtained as follows:

$$OW = 1 - \sum_{i=1}^{V} \frac{|PD_i|}{V^2} \tag{13}$$

OW ranges between 0 and 1. If the sentences are composed of same words in same positions, the value is 1 and if the sentences do not have any common-words, OW gets the value 0. To exemplify, in Figure 3, the OW is measured for the sample sentence pair: "But Gelinas says only six have been fully re-evaluated" and "Ms. Gelinas said only 1.5 per cent of those have been fully re-evaluated."

POS ordering ratio (OW_POS) is the feature that indicates how similar the order of the part of speech tags is in given sentences. The feature employs the OW equation on part of speech tags to measure the similarity.

The category of distance-based features involves renowned sentence similarity metrics of cosine, Jaccard, Hamming, Chebychev and Sumo distance, as formulated in Table 2, In Table 2, x_s and x_t are the representative vectors of the first and second sentence respectively.

The vectors are built by occurrence frequency values of composing words/tokens in sentence pair. Figure 4 gives the representative vectors. $x_s = [1\ 1\ 0\ 1\ 1\ 1\ 0\ 0\ 1\ 0\ 1\ 0\ 1\ 1\ 0\ 0]$ and $x_t = [1\ 0\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 0\ 0\ 1\ 1]$. of the sentences "But Gelinas says only six have been fully re-evaluated" and "Ms. Gelinas said only 1.5 per cent of those have been fully re-evaluated.", respectively. In Figure 4, f_y represents the occurrence frequency of the word/token y in the regarding sentence. For example, $f_{"but"} = 1$ in first sentence since the word/token "but" is observed only once in this sentence and $f_{"but"} = 0$ for the second sentence where "but" is never used.

To exemplify the use of representative vectors to measure distance-based features, we will calculate the Hamming distance of the previous sentence pair: "But Gelinas says only six have been fully re-evaluated" and "Ms. Gelinas said only 1.5 per cent of those have been fully re-evaluated.". Hamming distance employs the number of words/tokens that have different occurrence frequencies in two sentences. Simply to count this type of words/tokens in given sentence pair, x_t is subtracted from x_s to obtain the difference vector $[0\ 1 - 1\ 0\ 0\ 0 - 1 - 1\ 0 - 1\ 0 - 1\ 1\ 1 - 1 - 1]$. Each value that is not a zero in difference vector means that the regarding word/token is observed with different number of occurrences in given two sentences. Counting the values other than zero in difference vector, we obtain 10 for the example sentence pair. Dividing this value by n=15 (n is the length of the representative vector) Hamming distance is measured as 0.67. Similar procedures are applied for all distance-based features.

Word Order	1	2	3	4	5	6	7	8	9	10	11	12	13
Sentence 1	But	Gelinas	says	only	six	have	been	fully	re-evaluated.				
Sentence 2	Ms.	Gelinas	said	only	1.5	per	cent	of	those	have	been	fully	re-evaluated.

V=16

$PD_{"but"}$=16 $PD_{"gelinas"}$=2-2=0 $PD_{"says"}$=16 $PD_{"only"}$=4-4=0
$PD_{"six"}$=16 $PD_{"have"}$=6-10=-4 $PD_{"been"}$=7-11=-4 $PD_{"fuly"}$=8-12=4
$PD_{"re\text{-}evaluated"}$=9-13=-4 $PD_{"Ms."}$=16 $PD_{"said"}$=16 $PD_{"1.5"}$=16
$PD_{"per."}$=16 $PD_{"cent"}$=16 $PD_{"of"}$=16 $PD_{"those"}$=16

$$OW = 1 - \frac{|16| + |0| + |16| + |0| + |16| + |-4| + |-4| + |-4| + |-4| + |16| + |16| + |16| + |16| + |16| + |16| + |16|}{16^2}$$

$$OW = 1 - \frac{176}{256} = 0.31$$

Figure 3. Word ordering ratio of a sample sentence pair.

Table 2. Distance-based features.

Distance-based feature	Equation
Chebyshev Distance	$d_{st} = max_j\{\lvert x_{sj} - x_{tj}\rvert\}$
Hamming Distance	$d_{st} = (\#(x_s \neq x_t)/n)$
Jaccard Distance	$d_{st} = 1 - \frac{\sum_i \min(x_{si}.x_{ti})}{\sum_i \max(x_{si}.x_{ti})}$
Cosine Distance	$d_{st} = 1 - \frac{x_s x_t}{\sqrt{(x_s x_s)(x_t x_t)}}$
Sumo Distance [31]	$\alpha.\beta \in [0.1]$ $d_{st} = \begin{cases} \log_2 \frac{\lvert x_s\rvert}{\lvert x_s \cap x_t\rvert} + \beta \log_2 \frac{\lvert x_t\rvert}{\lvert x_s \cap x_t\rvert} & if\ \log_2 \frac{\lvert x_s\rvert}{\lvert x_s \cap x_t\rvert} + \beta \log_2 \frac{\lvert x_t\rvert}{\lvert x_s \cap x_t\rvert} < 1 \\ e^{-k * \log_2 \frac{\lvert x_s\rvert}{\lvert x_s \cap x_t\rvert} + \beta \log_2 \frac{\lvert x_t\rvert}{\lvert x_s \cap x_t\rvert}} & otherwise \end{cases}$

	f_{been}	f_{but}	f_{cent}	f_{fully}	$f_{gelinas}$	f_{have}	f_{ms}	f_{of}	f_{only}	f_{per}	$f_{re\text{-}evaluated}$	f_{said}	f_{says}	f_{six}	f_{those}	$f_{1.5}$	
x_s=[	1	1	0	1	1	1	0	0	1	0	1	0	1	1	0	0	]
x_t=[	1	0	1	1	1	1	1	1	1	1	1	1	0	0	1	1	]

Figure 4. Representative vectors x_s and x_t of the sentences "But Gelinas says only six have been fully re-evaluated" and "Ms. Gelinas said only 1.5 per cent of those have been fully re-evaluated.", respectively.

Each distance metric generates a value in a predefined range for each sentence pair in the corpus. It is observed that frequently the distance values of paraphrase pairs are lower than the values of non-paraphrase pairs. The metrics are utilized for both the stemmed and surface form of the sentence pairs resulting with ten different features: cosine distance of stemmed pair (C_ST), cosine distance of surface formed pair (C_SU), Jaccard distance of stemmed pair (J_ST), Jaccard distance of surface formed pair (J_SU), Hamming distance of stemmed pair (H_ST), Hamming distance of surface formed pair (H_SU), Chebyshev distance of stemmed pair (CH_ST), Chebyshev distance of surface formed pair (CH_SU), Sumo distance of stemmed pair (S_ST), Sumo distance of surface formed pair (S_SU).

4.2 Evidence selection

In classification problems, feature selection is defined as a pre-process commonly reducing the number of features in order to simplify the classification models, shorten the training times, detect succeeding features and understanding the data set. The feature selection methods are categorized in three: filtering methods, wrappers and embedded methods [32]. The wrappers aim to identify the most effective subset of features in classification by evaluating the performances employing well-known classification methods. Filtering methods employ a feature evaluator (e.g. information gain, gain ratio) to evaluate the classification performance of features individually. In filtering, a ranked list of features is provided that enables the comparison of features. The last category, embedded methods, both wrappers and filtering methods may be employed.

In this study, we proposed the use of feature selection methods in order to select evidences from the given set of text similarity features. Briefly, in our approach, accepting the paraphrase detection problem as a classification problem, the features that are highlighted to be effective in classification by feature selection methods are used as evidences. In evidence selection, as outlined in Figure 5, we employed filtering. Two feature evaluators are utilized in filtering: gain ratio and chi-square.

In gain ratio filtering, the worth of each feature is measured by gain ratio value and the features are sorted in descending order. In the sorted list L_{gain}, the features holding lower ranks (e.g. first, second) are accepted to be more successful compared to others.

Input: An initial set of 17 text similarity features
- Run first evaluator: Sort features based on gain ratio measure
- Store sorted features in list L_{gain}
- Run second evaluator: Sort features based on chi-square measure
- Store sorted features in list L_{chi}
- Loop while there exist a feature in L_{gain}
  - Select next feature *f* in list L_{gain}
  - Find the feature *f* in list L_{chi}
  - Calculate and record average rank of feature *f*

  Average rank= (rank(L_{gain}*ff*) + rank(L_{chi}*ff*))/2
  - Store *f* and its average rank in list *L*
- Sort list *L* according to the average ranks in ascending order

Output: The sorted list of features: *L*
The top-most *N* elements in list *L* are used as evidences.

Figure 5. Evidence selection algorithm.

The chi-square evaluator computes the worth of a feature by the value of the chi-squared statistic with respect to the class. Simply, the evaluator sorts the given features and the features that are mostly related to class information hold the lower ranks in sorted list L_{chi} of features. The top most features in list L_{chi} are accepted to be most successful features in distinguishing paraphrase pairs from non-paraphrase pairs.

Table 3 gives the resulting ranks of features that are obtained by the use of WEKA machine learning tool [33]. In Table 3, the features are sorted in increasing order according to the average of ranks that are obtained by two evaluators. For example, C_ST is ranked as 9th and 1st best classifying feature for the gain ratio and chi-square respectively. Thus, the average rank of C_ST is ((9+1))/2=5. In order to determine features that fail in classification, average rankings may be considered. The reliability on average rankings, in other words the agreement among the raters, is measured by Kendall-Tau statistics [34].

Kendall-Tau ranges between -1 and 1 where -1 is interpreted as no agreement and 1 as a complete agreement among raters. The resulting Kendall-Tau is calculated as -0.0294 (two sided p-value = 0.9) meaning that the agreement among the raters is not such strong to automatically select the features according to the average rankings. This directed us to measure the change in classification performance with an empirical approach. We measured the performance of paraphrase detection methods, employing best N features as evidences based on the average rankings where N ranges from 3 to 17.

Table 3. The features ranked by filtering methods.

Feature	*Gain Ratio*	*Chi-Square*	*Average Rank*
MW	*7*	*2*	*4.5*
C_ST	*9*	*1*	*5.0*
OW	*6*	*4*	*5.0*
S_ST	*10*	*3*	*6.5*
H_SU	*1*	*12*	*6.5*
H_ST	*4*	*11*	*7.5*
J_SU	*2*	*13*	*7.5*
S_SU	*12*	*5*	*8.5*
C_SU	*11*	*6*	*8.5*
J_ST	*3*	*14*	*8.5*
CH_ST	*5*	*16*	*10.5*
MW_POS	*14*	*7*	*10.5*
MB	*13*	*9*	*11.0*
OW_POS	*15*	*8*	*11.5*
CH_SU	*8*	*17*	*12.5*
MB_POS	*16*	*10*	*13.0*
LS	*17*	*15*	*16.0*

4.3 Rule formulation

The reasoning system in this study requires two rules for each evidence, one for the hypothesis "Given sentence pair is a paraphrase pair" and one for the opposite hypothesis "Given sentence pair is not a paraphrase pair". We formulated the rule pair for evidence E as:

IF The value of evidence E is in range [a b]

THEN Given sentence pair is paraphrase $\{cf_{rule\_paraphrase}\}$

IF The value of evidence E is in range [c d]

THEN Given sentence pair is not paraphrase $\{cf_{rule\_non\_paraphrase}\}$

In order to generate/define the rule pair, three parameters

- The range [a b] and [c d] (named as value-range in following sections)
- $cf_{evidence}$ values that show the degree of belief/disbelief to the evidences ("The evidence value is in range [a b] and "The evidence value is in range [c d]")
- cf_{rule} values that show the degree of belief/disbelief to hypotheses given the evidences

must be known.

In the following subsections, the proposed approaches to obtain those parameters from the training set are presented in detail. The result of rule formulation is a collection of rules where half is owned by the hypothesis " Given sentence pair is a paraphrase pair" and the other half belong to the opposite hypothesis.

4.3.1 Determining Value-ranges of Evidences

In identification of paraphrase sentence pairs, for each evidence an evidence value that is actually a similarity score in a predefined range is calculated for the sentence pairs. If the evidence value of the given pair falls in the value-range that belongs to the paraphrase pairs, the degree of belief to paraphrasing increases for the regarding pair and vice versa.

In this study, we propose to set the value-range $[a\ b\ ]$ that strongly supports the hypothesis "Given pair is a paraphrase pair" and to use the range $\neg[a\ b\ ]$ for the opposing hypothesis in order to build the rule pair for regarding evidence.

IF The value of evidence E for given sentence pair is in range $[a\ b\ ]$

THEN Given sentence pair is a paraphrase pair

IF The value of evidence E for given sentence pair is in range $\neg[a\ b\ ]$

THEN Given sentence pair is not paraphrase pair

For each evidence, the value-range assignment process begins with normalizing the evidence scores to [0 1] in the training set. Following, the value a is set to zero and increased by 0.1 increments till one (a=0.1, 0.2, 0.3, … 0.8, 0.9, 1). For each a, all b values that satisfies a<b and b∈[0 1] are calculated and alternative value-ranges are generated for the regarding value. For example when a=0.4, alternative [a b] value-ranges are [0.4 0.5], [0.4 0.6], [0.4 0.7], [0.4 0.8], [0.4 0.9], [0.4 1].

The most successful value-range in distinguishing paraphrase pairs from non-paraphrase pairs is determined by two methods: information gain and gain ratio. The information gain is measured for each value-range by utilizing training set. The value-range that gives the highest score is assigned as the value-range $[a\ b\ ]$ for the regarding evidence. The same procedure is applied by measuring gain ratio and gain ratio value-ranges are obtained for all evidences. Further details on determining value-ranges may be found in [35].

4.3.2 Certainty factors (cf_{rule} and $cf_{evidence}$) Measurement

In CF model, two certainty factor values are required to formulate the rules. Though the proposed CF model enables domain experts to decide on those values, in our experiments, we employed statistical methods in order to provide stable comparable results to Bayesian reasoning.

The first certainty factor is cf_{rule} that represents the belief/disbelief on the hypothesis given that evidence is observed. cf_{rule} value is calculated by the equations, given in section 3.2 that combine MB and MD metrics. The required probability of the hypothesis $P(H)$ is the ratio of number of samples that hypothesis is observed to be true to total number of samples in the training set. The conditional probability of hypothesis given the evidence $P(H \mid E)$ is the ratio of samples where both hypothesis and evidence are observed to the samples that evidence is true.

The second certainty factor is $cf_{evidence}$ that indicates the degree of belief/disbelief to the evidence. $cf_{evidence}$ in our experiments is calculated as

$$cf_{evidence} = \frac{\#\ of\ Samples\ that\ both\ Hypothesis\ and\ Evidence\ are\ true}{\#\ of\ Samples\ that\ Evidence\ is\ true} \quad (14)$$

4.4 Rule accumulation

The evidences directed us to define 17 rules for the hypothesis "Given pair is paraphrase pair" and equal number of rules for the opposite hypothesis. In this stage, for a given sentence pair whose evidence values are already known, the rules are fired one by one. The accumulated cf value is accepted as the belief value for the regarding hypothesis. The final belief values of two hypotheses are compared and the hypothesis that has a higher degree of belief is accepted to be the resulting decision.

5 Experimental results

The data set in our experiments is constructed from 5670 sentence pairs from MSRP corpus where 3807 (67%) pairs are paraphrase pairs and 1863 (33%) are non-paraphrase pairs. In the evaluation of CF and Bayesian reasoning approaches, F1 and accuracy measures are considered. F1 measure combines well-known measures of precision (P) and recall (R) and is formulated as follows

$$P = \frac{TP}{TP + FP} \quad (15)$$

$$\mathrm{R} = \frac{TP}{TP + FN} \quad (16)$$

$$F1 = \frac{P \cdot R}{P + R} \quad (17)$$

where TP is the number of pairs that are both classified as and annotated in corpus as paraphrase, and FP is the number of pairs that are classified as paraphrase but annotated as non-paraphrase in corpus. FN refers to the pairs that are annotated as paraphrase in corpus but assigned to non-paraphrase class by the classifiers. Accuracy is formulated as

$$A = \frac{TP + TN}{TP + TN + FP + FN} \quad (18)$$

Where TN is the number of pairs that are classified as non-paraphrase but annotated in corpus as paraphrase.

The evaluation tests are performed in 5-fold basis both for Bayesian reasoning and CF methods. Table 4 and 5 give average values of F1, accuracy (A) together with the standard deviation on F1 (S_F1) and accuracy (S_A) values for tests where threshold values are obtained by information gain and gain ratio, respectively. F1(%) and A(%) columns present the increase in F1 and accuracy values when compared to the performance of whole evidence set. The shaded cells in Table 4

and 5 present the maximum evaluation scores. For example, the maximum F1 scores are observed when most succeeding 3 evidences (MW, C_ST, OW) are employed in Bayes method both in Table 4 & 5.

The experimental evaluation revealed the following outcomes:

1. It is observed that employing gain ratio measure in determination of threshold value pairs (value-ranges) generates higher evaluation scores compared to information gain measure,
2. The highest F1 scores 0.810 and 0.808 (respectively for Bayes and CF methods) are provided by 3 best evidence where gain ratio is employed in determination of threshold values,
3. The accuracy measure results show that the subsets of evidence where size>4 succeed for both Bayes and CF methods when value-ranges are measured by gain ratio,
4. Considering accuracy measure, it is seen that maximum score is 0.697 and it may be obtained by application of both Bayes and CF methods,
5. Overall examination of the evaluation scores shows that no method is consistently outperforming the other. Thus. CF model is observed to be a good alternative to traditional Bayes method when evidence selection is performed.

6 Conclusion

Seeing the decision on paraphrasing as an expert problem, here, we propose the use of certainty factor as a remedy. In this respect annotated sets of sentences from the well-known MSRP corpus are scrutinized to find the evidences that may reveal the paraphrasing status of the sentence pairs. Generic and distance based similarity features are exploited as the evidence base. Filtering is applied to find the best discriminating features, which are named as evidences; among the paraphrase and non-paraphrase pairs and the regarding value-ranges are decided via gain ratio and information gain measures.

F1 and accuracy metrics are used to evaluate the performance of the model and the results are compared to the well-known Bayesian reasoning. The experimental results showed that CF model can be an alternating paraphrase detection method to Bayes model and previously proposed methods of supervised and unsupervised learning. As a further work, we plan to tune the parameters such as cf_{rule} and $cf_{evidence}$ of CF model by the help of human-experts in order to improve the performance.

7 Author contribution statements

In the scope of this study, Senem KUMOVA METİN contributed to formation of the idea and the design, coducting of the analyses, literature review, Bahar KARAOĞLAN contibuted to formation of the idea and the design, Tarık KIŞLA contributed to formation of the idea and the design, collecting the data and conducting the analyses and Katira SOLEYMANZADEH contributed to formation of the idea and the design, coducting of the analyses, literature review.

8 Ethics committee approval and conflict of interest statement

There is no need to obtain permission from the ethics committee for the article prepared.

There is no conflict of interest with any person / institution in the article prepared.

Table 4. The evaluation results of CF and Bayes methods (Value-ranges are obtained by Information Gain).

	BAYES						*CF*					
Number of Evidences	*F1*	*A*	*F1 (%)*	*A (%)*	*S_F1*	*S_A*	*F1*	*A*	*F1 (%)*	*A (%)*	*S_F1*	*S_A*
3	0.741	0.677	0.014	0.009	0.005	0.008	0.741	0.677	0.004	0.002	0.005	0.008
4	0.735	0.675	0.006	0.006	0.007	0.011	0.736	0.675	-0.003	0.000	0.007	0.011
5	0.735	0.675	0.006	0.006	0.007	0.011	0.736	0.675	-0.003	0.000	0.007	0.011
6	0.724	0.667	-0.009	-0.006	0.007	0.012	0.725	0.667	-0.018	-0.012	0.007	0.012
7	0.722	0.665	-0.013	-0.009	0.007	0.012	0.725	0.667	-0.018	-0.012	0.007	0.012
8	0.727	0.669	-0.006	-0.003	0.007	0.010	0.736	0.675	-0.002	0.000	0.007	0.010
9	0.725	0.668	-0.009	-0.005	0.007	0.010	0.730	0.671	-0.010	-0.007	0.007	0.010
10	0.725	0.668	-0.009	-0.005	0.006	0.010	0.730	0.671	-0.011	-0.006	0.006	0.010
All Evidences	0.731	0.671	-	-	0.007	0.011	0.738	0.675	-	-	0.007	0.011

Table 5. The evaluation results of CF and Bayes methods (Value-ranges are obtained by Gain Ratio).

	BAYES						*CF*					
Number of Evidences	*F1*	*A*	*F1 (%)*	*A (%)*	*S_F1*	*S_A*	*F1*	*A*	*F1 (%)*	*A (%)*	*S_F1*	*S_A*
3	0.810	0.690	0.026	-0.002	0.005	0.008	0.808	0.690	0.047	0.019	0.005	0.008
4	0.806	0.691	0.022	-0.001	0.007	0.011	0.807	0.697	0.045	0.028	0.008	0.013
5	0.806	0.693	0.022	0.002	0.007	0.011	0.807	0.697	0.046	0.028	0.008	0.013
6	0.807	0.697	0.024	0.007	0.007	0.012	0.807	0.697	0.046	0.028	0.007	0.012
7	0.807	0.697	0.024	0.007	0.007	0.012	0.807	0.697	0.046	0.028	0.007	0.012
8	0.805	0.695	0.020	0.005	0.007	0.010	0.805	0.695	0.042	0.026	0.006	0.010
9	0.805	0.695	0.020	0.005	0.007	0.010	0.805	0.695	0.042	0.026	0.006	0.010
10	0.805	0.695	0.020	0.004	0.006	0.010	0.804	0.695	0.042	0.025	0.006	0.010
All Evidences	0.789	0.692	-	-	0.007	0.011	0.772	0.678	-	-	0.012	0.015

9 References

[1] Shortliffe EH, Buchanan BG. "A model of inexact reasoning in medicine" *Mathematical Biosciences.* 23(3-4), 351-379, 1975.

[2] Dolan B, Quirk C, Brockett C. "Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources". *20th International Conference Computational Linguistic (COLING '04)*, Geneva, Switzerland, 23-27 August 2004.

[3] Fernando S, Stevenson M. "A Semantic Similarity Approach to Paraphrase Detection". Proceedings of the 11th annual research colloquium of the UK special interest group for computational linguistics (CLUK 2008), Oxford, United Kingdom, 1-3 March 2008.

[4] Salton G, Lesk ME. "Computer Evaluation of Indexing and Text Processing". *Journal of the ACM (JACM)*, 15(1), 8-36, 1968.

[5] Schütze H. "Automatic word sense discrimination". *Computational Linguistic.* 24(1), 97-123, 1998.

[6] Lin CY Hovy E. "The potential and limitations of automatic sentence extraction for summarization". *Proceedings of the HLT-NAACL 03 Text Summarization Workshop*, Edmonton, AB, Canada, 31 May- 3 June 2003.

[7] Mihalcea R, Corley C, Strapparava C. "Corpus-based and knowledge-based measures of text semantic similarity". *Proceeding 21st Conference Artifical Intelligence*, Boston, Massachusetts, USA, 16-20 July 2006.

[8] Zhang Y, Patrick J. "Paraphrase Identification by Text Canonicalization". *Proceedings of the Australasian Language Technology Workshop*, Sydney, Australia, 9-11 December 2005.

[9] Finch A, Hwang YS, Sumita E. "Using Machine Translation Evaluation Techniques to Determine Sentence-level Semantic Equivalence". *The Third International Workshop on Paraphrasing (IWP 2005)*, Jeju Island, Korea, 14 October 2005.

[10] Papineni K, Roukos S, Ward T, Zhu W. "BLEU: a method for automatic evaluation of machine translation". *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, 7-12 July 2002.

[11] Doddington G. "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics". *Proceedings of the Second İnternational Conference on Human Language Technology Research*, San Diego, California, USA, 24-27 March 2002.

[12] Madnani N, Tetreault J, Chodorow M. "Re-Examining Machine Translation Metrics for Paraphrase Identification". *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT '12)*, Montreal, Canada, 3-8 June 2012.

[13] Kozareva Z, Montoyo A. "Paraphrase identification on the basis of supervised machine learning techniques". *International Conference on Natural Language Processing (FinTAL 2006)*, Turku, Finland, 23-25 August 2006.

[14] Miller GA. "WordNet: a lexical database for English". *Communications of the ACM*, 38(11), 39-41, 1995.

[15] Fellbaum C. *WordNet: An Electronic Lexical Database*. 1st ed. Cambridge, Massachusetts, USA, MIT Press, 1998.

[16] Ul-Qayyum Z, Altaf W. "Paraphrase identification using semantic heuristic features". *Research Journal of Applied Sciences, Engineering and Technology*, 4(22), 4894-4904, 2012.

[17] Rus V, McCarthy PMM, Lintean MC, McNamara DS, Graesser AC. "Paraphrase ıdentification with lexico-syntactic graph subsumption". *Twenty-First International Florida Artifical Intelligence Reseach Society Conference (FLAIRS '08)*, Florida, USA, 15-17 May 2008.

[18] Qiu L, Kan MY, Chua TS. "Paraphrase recognition via dissimilarity significance classification". *Conference on Empirical Methods in Natural Language Processing (EMNLP '06)*, Sydney, Australia, 22-23 July 2006.

[19] Banerjee S, Pedersen T. "Extended gloss overlaps as a measure of semantic relatedness". *IJCAI International Joint Conference on Artificial Intelligence*, Acapulco, Mexico, 9-15 August 2003.

[20] Islam A, Inkpen D. "Semantic text similarity using corpus-based word similarity and string similarity". *ACM Transactions on Knowledge Discovery from Data*, 2(2), 1-25, 2008.

[21] Socher R, Huang E, Pennington J. "Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection". *Advances in Neural Information Processing Systems*, Granada, Spain, 12-14 December 2011.

[22] Wan S, Dras M, Dale R, Paris C. "Using Dependency-Based Features to Take the 'Para-farce' out of Paraphrase". *Proceedings of the Australasian Language Technology Workshop*, Sydney, Australia, 30 November- 1 December 2006.

[23] Wang Z, Mi H, Ittycheriah A. "Sentence similarity learning by lexical decomposition and composition". *COLING 2016-26th International Conference on Computational Linguistics,* Osaka, Japan, 11-16 December 2016.

[24] He H, Gimpel K, Lin J. "Multi-perspective sentence similarity modeling with convolutional neural networks". *EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 17-21 September 2015.

[25] Cheng J, Kartsaklis D. "Syntax-aware multi-sense word embeddings for deep compositional models of meaning". *EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 17-21 September 2015.

[26] Filice S, Da San Martino G, Moschitti A. "Structural representations for learning relations between pairs of texts". *ACL-IJCNLP 2015-53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, Beijing, China, 26-31 July 2013.

[27] Dwivedi A, Mishra D, Kalra PK. "Handling uncertainties-using probability theory to possibility theory" *Mag. IIT Kanpur*, 7(3), 1-12, 2006.

[28] Negnevitsky M. *Artificial Intelligence: A guide to Intelligent Systems.* 2nd ed. Essex, England, Pearson Education, 2005.

[29] Mitchell TM. *Machine learning.* Boston, USA, McGraw-Hill, 1997.

[30] Kışla T, Karaoğlan B, Metin SK. "Extracting the features of similarity in short texts". *IEEE 23th Signal Processing and Communications Applications Conference*, Malatya, Turkey, 16-19 May 2015.

[31] Cordeiro J, Dias G, Brazdil P. "A Metric for Paraphrase Detection". International Multi-Conference on Computing in the Global Information Technology (ICCGI'07), Guadeloupe, French Caribbean, 4-9 March 2007.

[32] Guyon I, Elisseeff A. "An ıntroduction to variable and feature selection". *Journal of Machine Learning Research*, 3(3), 1157-1182, 2003.

[33] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. "The WEKA data mining software- an update". *SIGKDD Explorations Newsletter*, 11(1), 10-18, 2009.

[34] Kendall MG, Smith BB. "The problem of *m* rankings". *Annals Mathematical. Statistics*, 10(3), 275-287, 1939.

[35] Kumova Metin S, Karaoglan B, Kısla T. "Attribute value-range detection in identification of paraphrase sentence pairs". *24th Signal Processing and Communication Application Conference (SIU)*, Zonguldak, Turkey, 16-19 May 2016