Expert Systems With Applications
Manuscript Draft

Manuscript Number: ESWA-D-17-00087R1

Title: Certainty Factor Model in Paraphrase Detection

Article Type: Full length article

Keywords: Paraphrase; paraphrase detection; certainty factor; evidence; evidence selection

Highlights

- Certainty Factor (*CF*) model is used in paraphrase detection.
- A set of 17 paraphrase detection features is considered as evidence candidates in *CF* model.
- The feature set is reduced by filtering to obtain a set of evidences.
- *CF* model is promising in determination of paraphrase pairs when compared to the traditional Bayes model.

Certainty Factor Model in Paraphrase Detection

Senem Kumova-Metin (Corresponding Author)[1], Bahar Karaoğlan[2], Tarık Kışla[3], Katira Soleymanzadeh[4]

[1]Izmir University of Economics, Faculty of Engineering, No.156, 35330, Balçova-İzmir, Turkey, Email: senem.kumova@ieu.edu.tr
[2]Ege University, International Computer Institute, Bornova-İzmir, Turkey, Email: bahar.karaoglan@ege.edu.tr
[3]Ege University, Department of Computer Education and Instructional Technology, Bornova-İzmir, Turkey, Email: tarik.kisla@ege.edu.tr
[4]Ege University, International Computer Institute, Bornova-İzmir, Turkey, Email: katira.sole@gmail.com

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 (Dolan, Quirk, & Brockett, 2004) 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.

Highlights

- Certainty Factor (*CF*) model is used in paraphrase detection.
- A set of 17 paraphrase detection features is considered as evidence candidates in *CF* model.
- The feature set is reduced by filtering to obtain a set of evidences.
- *CF* model is promising in determination of paraphrase pairs when compared to the traditional Bayes model.

Keywords: Paraphrase, paraphrase detection, certainty factor, evidence, evidence selection

POINT-TO-POINT RESPONSES: REVIEWER #1

First of all we want to mention that we are very thankful for your valuable, affirmative and encouraging comments. We revised our paper according to them and we believe that our paper is improved so much. Below, you may find our point-to-point responses for your comments.

Reviewer #1: In this paper the authors propose to use the certainty factor model in paraphrase detection. The idea appears to be quite promising. The authors implement and explain the concept, but in its current form the paper cannot be recommended to publication and has to be seriously revised

1. **The paper has to be reorganized. Apparently, it makes sense to concentrate all needed mathematical notions (the entropy, information gains, the binary classification features and so on) in a separate section, something like mathematical background.**

   The paper is reorganized. The revised paper is organized as follows:

   Abstract
   1. Introduction
   2. Related Work
   3. Background Information
      3.1. Bayesian Reasoning
      3.2. Certainty Factor Model
      3.3. Information-based metrics
   4. Proposed Reasoning Methodology
      4.1. Similarity Features: Evidence Candidates
      4.2. Evidence Selection
      4.3. Rule Formulation
         4.3.1. Determining Value-ranges of Evidences
         4.3.2. Certainty factors (cf_{rule} and $cf_{evidence}$) Measurement
      4.4. Rule Accumulation
   5. Experimental results
   6. Conclusion

   The mathematical notions (e.g. information gain, entropy), Bayesian reasoning and certainty factor model is presented in section "3. Background Information"

2. **Processes of the feature selection procedure or of the Paraphrase Detection procedure has to be represented in a pseudocode fashion.**
   Feature selection is renamed as "Evidence Selection". The evidence selection is represented in pseudocode fashion in Figure is presented in Figure 3 (Section 4.2).

3. **Abuse of notation (for example, for cf) has to be discarded.**
   We are really thankful for this and the following comment. We believe that these comments improved our paper a lot.

   We rewrote the proposed methodology section. The term *cf* is used to represent three different notions in *CF* theory. We named them as
   - $cf_{evidence}$: the degree of belief/disbelief to evidence
   - cf_{rule} : the degree of belief/disbelief to the hypothesis to be true when evidence is observed
   - cf_{net} : the overall belief to the hypothesis when the rules are accumulated

4. **There are a lot good examples explaining the notions, but they are presented in inaccurate form. The calculation must be explained more clearly.**
   The methodology section is reworded and divided into 4 subsections. All the mathematical information is given in a separate section.

5. **What is min[1,0] or max[1,0]? (abuse??)**
   Regarding expressions are removed from the paper and expressed as statements.

6. **Estimation of the certainty factors of evidences is not understandable**
   The estimation of *cf* values are reworded and given in section 4.3.2.

7. **The text has to be corrected. For example, what does a sentence "The lower the rank the higher is the expectation for the feature to classify the data effectively" mean?**
   The text is corrected, we did out best to remove all typing/grammatical mistakes.

8. **The 5-fold cross validation is mentioned several times. It is completely unclear how it was done.**
   The 5-fold cross validation is explained in Experimental Results section (page 23).

9. **It is more acceptable to present the conclusion after the experimental results.**
   The conclusion is given after the experimental results.

POINT-TO-POINT RESPONSES: REVIEWER #2

First of all, we want to mention that we are very thankful for your valuable, encouraging and affirmative comments on our study. We revised our paper based on your comments and we believe that it improved a lot. We tried to do our best to fulfill the requirements and remove all mistakes.

Reviewer #2: In this paper, the authors address the problem of uncertainty management in identification of paraphrase sentence pairs. There are several issues needed to be addressed:

The revised paper is organized as follows:

Abstract
7. Introduction
8. Related Work
9. Background Information
   9.1. Bayesian Reasoning
   9.2. Certainty Factor Model
   9.3. Information-based metrics
10. Proposed Reasoning Methodology
   10.1. Similarity Features: Evidence Candidates
   10.2. Evidence Selection
   10.3. Rule Formulation
      10.3.1. Determining Value-ranges of Evidences
      10.3.2. Certainty factors (cf_{rule} and $cf_{evidence}$) Measurement
   10.4. Rule Accumulation
11. Experimental results
12. Conclusion

Below, you may find point-to-point responses for your comments.

1) **The related work is not well written. The authors have just listed several works without any logic connection and do not illustrate the advantages and disadvantages about the work.**
   Related work section is reorganized in revised paper. Since there exists a wide range of studies in paraphrase identification, we just provide prominent studies that employ MSRP corpus in order to enable the performance comparison with our proposed method. The previous methods are briefly explained in related work section and a comprehensive resource for further discussion and detailed information is provided to the readers. In addition in experimental results section we provided our comment to compare the proposed method and the previous studies.

2) **The highlight is not clear in this manuscript as I cannot see why the authors choose certainty factor model for paraphrase detection. This is important as this the main contribution of the manuscript. The authors should use some sentences to illustrate the advantages of certainty factor model for paraphrase detection using the literature review or some simple example.**
   We thank very much to this valuable comment. We believe that this comment triggered us to explain the contribution of proposed model and improved our paper so much.

   The introduction section is reworded to mention why CF model is more advantageous compared to Bayesian model when uncertainty exists. Briefly, in Bayesian reasoning, ithe probabilities of hypothesis "H" and the opposite hypothesis "notH" when some evidence is given are not independent ($P(H' \mid E) = 1 - P(H \mid E)$). But in real world

problems, such as paraphrase identification, we cannot state that opposite hypotheses are dependent or independent. CF model contrary to Bayesian model, enables to assign independent probabilities to opposite hypotheses. This fact encouraged us to build CF model for paraphrase identification.

3) **The presentation of the algorithm is not clear. I cannot see the flowchart of proposed method about how to handle the problem of paraphrase detection. I guess the author should at least give a table or figure to show the basic steps of the proposed method for handling paraphrase detection.**
The proposed method is presented by a flow chart in Figure 1 (Introduction section) in revised paper. And the proposed methodology section is reorganized to explain the stages in flowchart.

4) **The simulation results seem to be not sufficient as lack of compared work. The authors should compare with some simulation results of other state-of-the-art method listed in the related work.**
The related work section includes some prominent paraphrase identification studies in which MSRP corpus is employed. In this study, we also employed MSRP corpus actually to enable the comparison with the previous studies. In the earlier version of our paper this fact was not mentioned. In revised version, we clearly explained this fact and in experimental results section we comment on performance results of proposed model and previous methods.

1. Introduction

Paraphrasing is the restatement of facts or opinions in a given passage of text, keeping the original meaning. Simply, it is performed by reading the passage and rewording it without copying the original author's style or wording. Based on the definition of paraphrasing, the paraphrase pairs of text are described as two passages of text where the same meaning is to be given to the reader.

Since the identification of paraphrase text pairs becomes increasingly prominent in various areas such as plagiarism detection, summarization, and machine translation; various approaches are proposed to solve this problem. In the majority of current works, the paraphrase identification is simplified as the task of reasoning on the given text pairs as paraphrase or non-paraphrase considering various text similarity features. Though there exists several text similarity features to be employed, the performance of studies is limited to the quality/content of the corpus where feature values are obtained. If the required information to decide on the type of the text pair is incomplete, inconsistent, uncertain or all three in the corpus, the decision may be wrong or even decision may not be made.

The rule-based expert systems handle uncertainty in decision problems by the help of two notions: experience and the expertise (Negnevitsky, 2005). 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 problem in Bayesian reasoning is that when some evidence E is observed, the belief in hypothesis H to be true is represented by $P(H \mid E)$ and the belief in the opposite hypothesis, H', is formulated 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, certainty factor model may be used (Dwivedi, Mishra, & Kalra, 2006). In certainty factor model, instead of conditional probabilities, the measure of belief/disbelief to the hypothesis given the evidence is employed (Heckerman, 1992). The model enables the assignment of independent belief values to the hypothesis H and the opposite hypothesis H' given the evidence E. It is also possible to have no statement on belief when there exists no evidence in *CF* model. In other words, in *CF* model, it is possible to separate belief, disbelief and ignorance. On contrary, Bayesian reasoning requires assigning probabilities even if no information is available.

In this study, we proposed the use of *CF* model in paraphrase identification since that the reliable statistical data are unavailable due to the limitations on corpora and the vagueness in paraphrasing rules that results with inability to handle exceptional cases where no rule is applicable. The general *CF* model requires pre-determined evidences, rules, 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 a *CF* model, depicted in Figure 1, where text similarity features that are highlighted to being successful in paraphrase detection by feature selection methods are selected as evidences. In our experiments, the evidences are extracted 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). To generate rules for

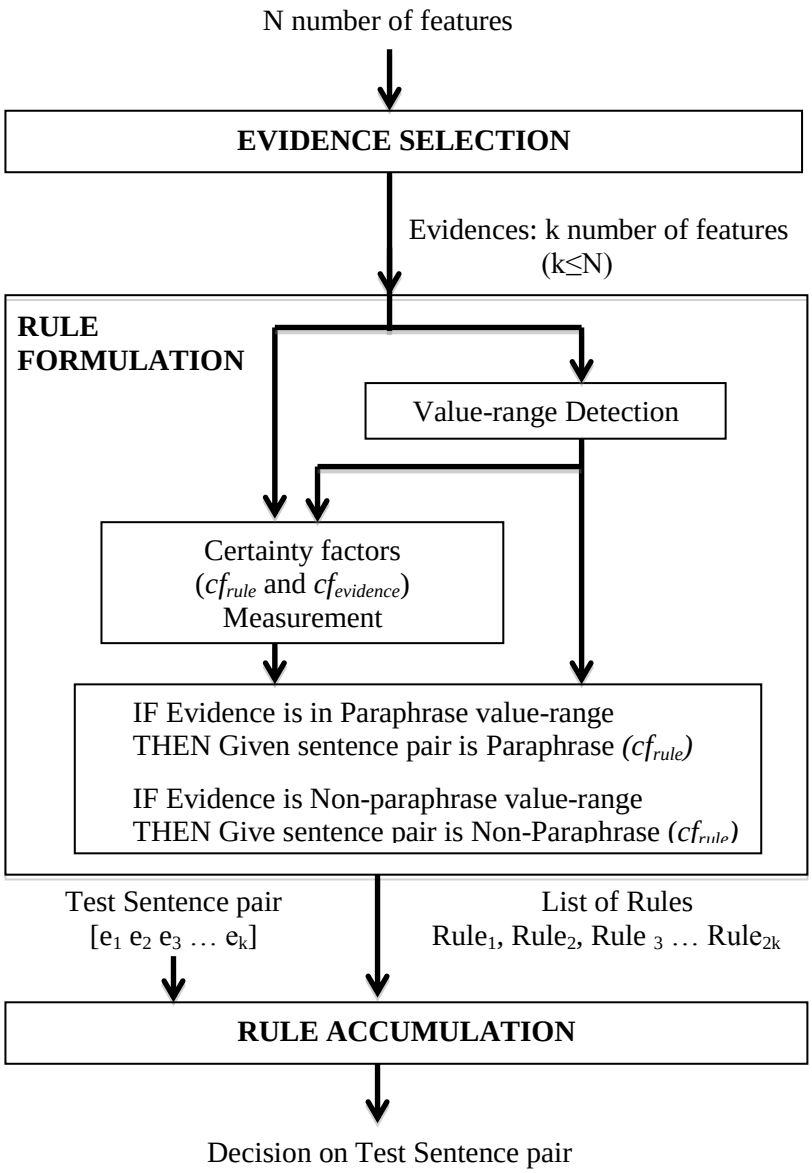

Figure 1. Flow chart of the proposed *CF* method

evidences, the value-range of paraphrases/non-paraphrases and the degree of belief/disbelief to the evidences and rules are determined. We employed two alternative measures (information gain and gain ratio) in determination of regarding evidence value-ranges and the belief/disbelief to the evidences/rules are measured from the training set. The rule accumulation is performed similar to the general *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*) (Dolan et al., 2004). The evaluation is performed by *F1* and accuracy measures. It is observed that the *CF* model is promising in determination of paraphrase pairs when compared to the traditional Bayesian reasoning model.

The paper is organized as following. We first review related work in Section 2. The background information on proposed and alternative reasoning method are presented

in Section 3. Section 4 and 5 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 (Salton & Lesk, 1968). Following these studies, the text similarity is used in a variety of different areas such as text classification, word sense disambiguation (Lesk, 1986; Schütze, 1998) summarization (C.-Y. Lin & Hovy, 2003) and automatic assessment of machine translation (Mihalcea, Corley, & Strapparava, 2006).

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.

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

Reference	**Methods/features**	**Type**	**Accuracy**	***F1***
Zhang and Patrick (2005)	Text canonicalization	supervised	0.703	0.795
Mihalcea, Corley, and Strapparava (2006)	Word-to-word similarity features	unsupervised	0.703	0.813
(Rus, McCarthy, Lintean, McNamara, & Graesser, 2008)	Lexico-Syntactic graph subsumption	unsupervised	0.706	0.805
Qiu, Kan, and Chua (2006)	Dissimilarity classification	supervised	0.720	0.816
Islam and Inkpen (2008)	Combination of semantic and syntactical features	unsupervised	0.726	0.813
Blacoe and Lapata (2012)	Semantic spaces from word clustering	supervised	0.730	0.823
Fernando and Stevenson (2008)	Wordnet measure and vector based similarity	unsupervised	0.741	0.824
Ul-Qayyum and Altaf (2012)	Semantic heuristic features	supervised	0.747	0.818
Finch, Hwang, and Sumita (2005)	Machine translation methods	supervised	0.750	0.827
Wan et al. (2006)	Dependency-based features	supervised	0.756	0.830
Kozareva and Montoyo (2006)	Lexical and semantic similarity features	supervised	0.766	0.796
Socher, Huang, and Pennington (2011)	Dynamic pooling and unfolding recursive auto-encoders	supervised	0.768	0.836
Madnani, Tetreault, and Chodorow (2012)	Machine translation metrics	supervised	0.774	0.841

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 the comparable results. Below, the works in Table 1 will be briefly explained. For further related

discussion and detailed information on the paraphrasing and the regarding methods, Androutsopoulos & Malakasiotis's study (2010) may be seen.

In an earlier study on *MSRP* corpus, Zhang and Patrick (2005) 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, Finch, Hwang, and Sumita (2005) investigated the utility of machine translation evaluation methods such as BLEU (Papineni, Roukos, Ward, & Zhu, 2002), NIST (Doddington, 2002), WER (Su, Wu, & Chang, 1992) and PER(Tillmann, Vogel, Ney, & Zubiaga, 1997) and proposed a PER-based classification method. In a more recent work, Madnani, Tetreault, and Chodorow (2012) re-examined the machine translation metrics, a meta-classifier that considers the weighted probability estimates of three classifiers is trained. The proposed method is stated to be the best performing system ever reported on the *MSRP* corpus when compared to all previously published work.

Kozareva and Montoyo (2006) 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 (Fellbaum, 1998; Miller, 1995), lexical features such as longest common subsequence that are used in a variety of studies are utilized. Zia and Ul-Qayyum and Altaf (2012) proposed to use an enhanced set of similar lexical features together with semantic heuristics in machine learning methods.

Rus et al. (2008) 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 (Fellbaum, 1998; Miller, 1995) 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 the study of Qiu, Kan, and Chua (2006) unlike the majority of studies in literature, the main goal is stated as making paraphrasing judgement based on the significance of dissimilarity between the sentences instead of similarity. 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 (2008) 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:

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

where W is the matrix containing the information about the similarity of word pairs. W matrix is populated by 6 WordNet similarity metrics (Lesk (Banerjee & Pedersen, 2003), Lch (Leacock & Chodorow, 1998), Wup (Wu & Palmer, 1994) Res (Resnik, 1995), Lin (D. Lin, 1998) ve Jcn (Jiang & Conrath, 1997)). Fernando and Stevenson (2008) 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 by Islam and Inkpen (2008). One other word-based similarity approach is presented by Mihalcea, Corley, and Strapparava (2006). 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. 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 paraphrase pair.

Socher, Huang, and Pennington (2011) introduced a differentiating approach where both word-based similarity and multiword-based similarity are considered. In this study, 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. (2006) 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.

3. Background Information

In probability theory, it is assumed that the exact knowledge is available, though in real life problems the knowledge is commonly uncertain and/or unsuitable to reach a perfectly reliable solution. The lack of the exact knowledge that results with uncertainty may be due to the weak implications, imprecise language, unknown data and/or combining the views of different experts (Negnevitsky, 2005). Two of major decision models that handle uncertainty in problems are Bayesian reasoning and certainty factor model. In this section, firstly background information on Bayesian reasoning and certainty factor will be provided. Following, information based metrics that are employed in construction of reasoning models will be explained briefly.

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

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

where $P(H)$ is the probability of hypothesis H being true, $P(\neg H)$ is the probability of H being false, $P(E \mid H)$ represents the probability of evidence E to be observed given H is true, and $\mathrm{P}(E|\neg H)$ 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 $E_1, E_2, E_3 \ldots E_n$, the conditional probability of H expands to

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

Based on the Bayes theory 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 (likelihood of sufficiency) represents a measure of belief in hypothesis H given evidence E_i is present, LN (likelihood of necessity) is a measure of disbelief to H if evidence E_i is missing (Negnevitsky, 2005). Likelihood of sufficiency (LS) and likelihood of necessity (LN) are defined as

$$LS = \frac{P(Ei \mid H)}{P(Ei|\neg H)} \qquad \text{and} \qquad LN = \frac{P(\neg Ei \mid H)}{P(\neg Ei|\neg H)}$$

where $P(Ei \mid H)$ and $P(Ei|\neg H)$ are the conditional probabilities of ith evidence given H being true and given H being false respectively. $P(\neg Ei|\neg H)$ is the probability of ith evidence being not observed given H is false. In knowledge bases, the prior probability of the hypothesis $P(H)$ that is provided by the experts is represented by prior odds $O(H)$:

$$O(H) = \frac{p(H)}{1 - p(H)}$$

Thus, we attain the posterior probabilities as follows

$$P(H \mid Ei) = \frac{LS \ \times \ O(H)}{1 + LS \ \times \ O(H)}$$

and

$$P(H \mid \neg Ei) = \frac{LN \ \times \ O(H)}{1 + LN \ \times \ O(H)}$$

The Bayesian accumulation of evidences $E_1, E_2, E_3 \ldots E_n$, is performed by firing the rules where the regarding evidence is observed. The conditional probability value $P(H \mid Ei)$ or $P(H \mid \neg Ei)$ is used as the prior odds of the next rule with evidence E_{i+1}. Merging all the evidences by Bayesian accumulation, $P(H)$ value is obtained.

In identification of paraphrase sentence pairs, H is the hypothesis that states “The sentence pair is a paraphrase pair” and text similarity features are evidences that

trigger the change in probability of the hypothesis. The same procedure is followed for the opposite hypothesis ($\neg H$) that states that the sentences are non-paraphrase. Comparing the resulting values of $P(H)$ and $P(\neg H)$ it may be decided that the given sentence pair is a paraphrase pair if $P(H) > P(\neg H)$ and non-paraphrase pair vice versa.

3.2. Certainty Factor Model

Certainty factor theory 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 (1975) 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, Shortliffe and Buchanan (1975) introduced certainty factor (*cf*), a number to measure the expert's belief, which ranges between -1 and 1. The *cf* value is used to represent the degree of belief in hypothesis when the evidence is observed (cf_{rule}) and the degree of belief in the evidence ($cf_{evidence}$). A positive *cf* value represents a degree of belief and a negative value 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:

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

where *MD(H,E)* is measure of belief and *MB(H,E)* is measure of disbelief. 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 (Negnevitsky, 2005). *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}, is calculated by multiplying the certainty factor of the evidence, $cf_{evidence}$ and the certainty factor of the rule, cf_{rule}.

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

For rules where multiple evidences that are combined by "AND" or "OR" statements exist, 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

$$cf_{net} = cf(H, E_1 \cap E_2 \cap E_3 \cap \ldots \cap E_n) = min\left[cf_{evidence\_1}, cf_{evidence\_2}, cf_{evidence\_3} \ldots cf_{evidence\_n}\right] \times cf_{rule}$$

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

$$cf_{net} = cf(H, E_1 \cup E_2 \cup E_3 \cup \ldots \cup E_n) = max\left[cf_{evidence\_1}, cf_{evidence\_2}, cf_{evidence\_3} \ldots cf_{evidence\_n}\right] \times cf_{rule}$$

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.

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

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 $cf_{net2}=cf(E_2$, "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. Information-based metrics

Information gain and gain ratio metrics that ground on information theory are employed in different processes (e.g. evidence selection and value-range detection) in the proposed model. Both metrics 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

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

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 (Mitchell, 1997). This is why; as the information gain gets higher the uncertainty gets lower supporting the effective classification. Information gain is calculated as follows

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

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

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

4. Proposed Reasoning Methodology

The decision whether a given sentence pair is a paraphrase pair or not, may be made based on the degree of similarity between the given sentences. In such a decision, a variety of text similarity features may be employed and their joint contribution may be measured by several methods. In this study, we propose to formulate rules that

accept the selected text similarity features as evidences, and accumulate the belief/disbelief on paraphrase pairs by the certainty factor model.

The stages of the proposed method (depicted in Figure 1) 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 on paraphrase/non-paraphrase pairs.

   In rule formulation process, for each evidence, a decision rule must be built for both hypotheses ("Given sentence pair is a paraphrase pair" and "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 determined. 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 determined 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:** This stage is in which the pre-determined rules are fired one by one for the given hypothesis and the sentence pair. The *CF* model, defined in section 3.2, is employed once for the hypothesis "Given sentence pair is a paraphrase pair" and once for the hypothesis "Given sentence pair is a non-paraphrase pair", the resulting belief values are compared and the hypothesis that generates the higher belief value is returned as the decision.

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

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, produce a value in the range [0 1] for each sentence pair. In this group of features, it is common to observe higher feature values, closer to the upper limit, for paraphrase pairs and lower values for non-paraphrase pairs. 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 short sentence is divided by the length of the long 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 (Kışla, Karaoğlan, & Metin, 2015). 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:

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

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*) and *MB* 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 (Islam & Inkpen, 2008). 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:

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

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 2, 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.”

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

POS ordering ratio (*OW_POS*) is the feature that indicates measures 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 composed of occurrence frequency values of words in sentences.

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 (Cordeiro, Dias, & Brazdil, 2007)	$\alpha, \beta \in [0,1]$ $d_{st} = \begin{cases} \log_2 \frac{\lvert x_s\rvert}{\lvert x_s \cap x_t\rvert} + \beta \log_2 \frac{\lvert x_t\rvert}{\lvert x_s \cap x_t\rvert} & if\ \log_2 \frac{\lvert x_s\rvert}{\lvert x_s \cap x_t\rvert} + \beta \log_2 \frac{\lvert x_t\rvert}{\lvert x_s \cap x_t\rvert} < 1 \\ e^{-k * \log_2 \frac{\lvert x_s\rvert}{\lvert x_s \cap x_t\rvert} + \beta \log_2 \frac{\lvert x_t\rvert}{\lvert x_s \cap x_t\rvert}} & otherwise \end{cases}$

The distance metrics generate 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 (Guyon & Elisseeff, 2003). 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 3, we employed filtering. Two feature evaluators are utilized in filtering: gain ratio and chi-square.

- **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}, *f*) + rank(L_{chi}, *f*))/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 3. Evidence selection algorithm

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.

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 (Hall et al., 2009). 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 (Kendall & Smith, 1939). 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* is 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 given below:

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 $\neg[a\ b]$ [1]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 $b \in [0\ 1]$ are calculated and alternative value-ranges are generated for regarding a 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

[1] $\neg[a\ b]$ represents the set/range of values that are not in range $[a\ b]$ (NOT$[a\ b]$)

regarding evidence. The same procedure is applied by measuring gain ratio and gain ratio value-ranges are obtained for all evidences.

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

In *CF* model, two certainty factors are required to formulate the rules. Though the proposed *CF* model enables domain experts to decide on those certainty factors, 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 $P(H|E)$ 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

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

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 (Dolan et al., 2004) 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

$$P = \frac{TP}{TP + FP}$$
$$R = \frac{TP}{TP + FN}$$
$$F1 = \frac{P \cdot R}{P + R}$$

where *TP* is the number of pairs that are both classified as and annotated in corpus as paraphrase, and *FP* are pairs that are number of pairs that are classified as paraphrase but non-paraphrase in corpus. And *FN* are the pairs that are annotated as paraphrase in

corpus but assigned to non-paraphrase class by the classifiers. Accuracy is formulated as

$$A = \frac{TP + TN}{TP + TN + FP + FN}$$

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[2] 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 succeding 3 evidences are employed in Bayes method both in Table 4 & 5.

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

[2] The data set is splitted to 5 equally sized subsets that each subset includes equal proportions of positive and negative samples with the whole set. Each subset is employed once as testing set and the remaining sets are merged to be used in training. The experiments are run totally 5 times and the evalution values are averaged to be reported.

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. Moreover, when compared to the methods (mentioned in Section 2) employing same corpus, it is examined that *CF* method beats many of them based on *F1* measure.

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 metrics.

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 is 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.

Acknowledgement
This work is carried under the grant of TÜBİTAK – The Scientific and Technological Research Council of Turkey to Project No: 114E126. Using Certainty Factor Approach and Creating Paraphrase Corpus for Measuring Similarity of Short Turkish Texts and Ege University Scientific Research Council Project No 2015/BİL/034, Developing a Paraphrase Corpus for Turkish Short Text Similarity Studies.

References

Androutsopoulos, I., & Malakasiotis, P. (2010). A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research, 38*, 135–187. http://doi.org/10.1613/jair.2985

Banerjee, S., & Pedersen, T. (2003). Extended gloss overlaps as a measure of semantic relatedness. In *IJCAI International Joint Conference on Artificial Intelligence* (pp. 805–810).

Blacoe, W., & Lapata, M. (2012). A Comparison of Vector-based Representations for Semantic Composition. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '12)*, (July), 546–556. Retrieved from http://dl.acm.org/citation.cfm?id=2391011

Cordeiro, J., Dias, G., & Brazdil, P. (2007). A Metric for Paraphrase Detection. *2007 International Multi-Conference on Computing in the Global Information Technology (ICCGI'07)*. http://doi.org/10.1109/ICCGI.2007.4

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *Proceedings of the Second International Conference on …*, 138–145. http://doi.org/10.3115/1289189.1289273

Dolan, B., Quirk, C., & Brockett, C. (2004). Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. *Proceedings of the 20th International Conference on Computational Linguistics (COLING '04)*, Article No. 350. http://doi.org/10.3115/1220355.1220406

Dwivedi, A., Mishra, D., & Kalra, P. K. (2006). Handling Uncertainties—Using Probability Theory to Possibility Theory. *The Magazine of IIT Kanpur*, *7*(3), 1–12.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. *MIT Press, Cambridge, London, England* (Vol. 71). http://doi.org/10.1139/h11-025

Fernando, S., & Stevenson, M. (2008). 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)*, 45–52. http://doi.org/10.1.1.144.4680

Finch, A., Hwang, Y.-S., & Sumita, E. (2005). Using Machine Translation Evaluation Techniques to Determine Sentence-level Semantic Equivalence. In *Proceedings of the Third International Workshop on Paraphrasing (IWP 2005)* (pp. 17–24). Retrieved from http://acl.ldc.upenn.edu/I/I05/I05-5003.pdf

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research (JMLR)*, *3*(3), 1157–1182. http://doi.org/10.1016/j.aca.2011.07.027

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software. *SIGKDD Explorations Newsletter*, *11*(1), 10. http://doi.org/10.1145/1656274.1656278

Heckerman, D. (1992). The certainty-factor model. In *Encyclopedia of Artificial Intelligence*.

Islam, A., & Inkpen, D. (2008). Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data*, *2*(2), 1–25. http://doi.org/10.1145/1376815.1376819

Jiang, J. J., & Conrath, D. W. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *Proceedings of International Conference Research on Computational Linguistics*, 19–33. http://doi.org/10.1.1.269.3598

Kendall, M. G., & Smith, B. B. (1939). The Problem of *m* Rankings. *The Annals of Mathematical Statistics*, *10*(3), 275–287. http://doi.org/10.1214/aoms/1177732186

Kışla, T., Karaoğlan, B., & Metin, S. K. (2015). Extracting the Features of Similarity in Short Texts. In *IEEE 23th Signal Processing And Communications*

Applications Conference (pp. 180–183).
Kozareva, Z., & Montoyo, A. (2006). Paraphrase Identification on the Basis of Supervised Machine Learning Techniques. *Advances in Natural Language Processing: 5th International Conference on NLP (FinTAL 2006)*, 524–533. http://doi.org/10.1007/11816508_52
Leacock, C., & Chodorow, M. (1998). Combining Local Context and WordNet Similarity for Word Sense Identification. *WordNet: An Electronic Lexical Database.*, (JANUARY 1998), 265–283. http://doi.org/citeulike-article-id:1259480
Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries. In *Proceedings of the 5th annual international conference on Systems documentation - SIGDOC '86* (pp. 24–26). http://doi.org/10.1145/318723.318728
Lin, C.-Y., & Hovy, E. (2003). The Potential and Limitations of Automatic Sentence Extraction for Summarization. *Proceedings of the {HLT}-{NAACL} 03 on {Text} {Summarization} {Workshop} - {Volume} 5*, 73–80. http://doi.org/10.3115/1119467.1119477
Lin, D. (1998). An Information-Theoretic Definition of Similarity. *Proceedings of ICML*, 296–304. http://doi.org/10.1.1.55.1832
Madnani, N., Tetreault, J., & Chodorow, M. (2012). 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)*, 182–190. http://doi.org/10.1.1.374.2895
Mihalcea, R., Corley, C., & Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. *Proceedings of the 21st National Conference on Artificial Intelligence*, *1*, 775–780. http://doi.org/10.1.1.65.3690
Miller, G. a. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39–41. http://doi.org/10.1145/219717.219748
Mitchell, T. M. (1997). *Machine Learning*. *Machine Learning* (Vol. 1). http://doi.org/10.1007/BF00116892
Negnevitsky, M. (2005). *Artificial intelligence: a guide to intelligent systems*. Pearson Education.
Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. *… of the 40Th Annual Meeting on …*, (July), 311–318. http://doi.org/10.3115/1073083.1073135
Qiu, L., Kan, M.-Y., & Chua, T.-S. (2006). Paraphrase Recognition via Dissimilarity Significance Classification. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP '06)*, (July), 18–26. http://doi.org/10.3115/1610075.1610079
Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *Roceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1 - IJCAI'95*, *1*, 6. http://doi.org/10.1.1.55.5277
Rus, V., McCarthy, P. M. M., Lintean, M. C., McNamara, D. S., & Graesser, A. C. (2008). Paraphrase Identification with Lexico-Syntactic Graph Subsumption. *Proceedings of the Twenty-First International Florida Artificial Intelligence Research Society Conference (FLAIRS '08)*, 201–206. Retrieved from http://www.aaai.org/Papers/FLAIRS/2008/FLAIRS08-051.pdf
Salton, G., & Lesk, M. E. (1968). Computer Evaluation of Indexing and Text Processing. *Journal of the ACM*, *15*(1), 8–36.

http://doi.org/10.1145/321439.321441
Schütze, H. (1998). Automatic Word Sense Discrimination. *Computational Linguistics*, *24*(1), 97–123.
Shortliffe, E. H., & Buchanan, B. G. (1975). A model of inexact reasoning in medicine. *Mathematical Biosciences*, *23*(3–4), 351–379. http://doi.org/10.1016/0025-5564(75)90047-4
Socher, R., Huang, E., & Pennington, J. (2011). Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection. *Advances in Neural Information Processing Systems*, 801–809. Retrieved from http://machinelearning.wustl.edu/mlpapers/paper_files/NIPS2011_0538.pdf%5Cnhttps://papers.nips.cc/paper/4204-dynamic-pooling-and-unfolding-recursive-autoencoders-for-paraphrase-detection.pdf
Su, K. Y., Wu, M. W., & Chang, J. S. (1992). A new quantitative quality measure for machine translation systems. In *COLING* (pp. 433–439).
Tillmann, C., Vogel, S., Ney, H., & Zubiaga, A. (1997). A DP-based Search Using Monotone Alignments in Statistical Translation. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics* (pp. 289–296). http://doi.org/10.3115/976909.979654
Ul-Qayyum, Z., & Altaf, W. (2012). Paraphrase identification using semantic heuristic features. *Research Journal of Applied Sciences, Engineering and Technology*, *4*(22), 4894–4904.
Wan, S., Dras, M., Dale, R., & Paris, C. (2006). Using Dependency-Based Features to Take the "Para-farce" out of Paraphrase. *Proceedings of the Australasian Language Technology Workshop (ALTW 2006)*, (2005), 131–138. Retrieved from https://www.aclweb.org/anthology-new/U/U06/U06-1.pdf#page=139
Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. *32nd Annual Meeting on Association for Computational Linguistics*, 133–138. http://doi.org/10.3115/981732.981751
Zhang, Y., & Patrick, J. (2005). Paraphrase Identification by Text Canonicalization. *Proceedings of the Australasian Language Technology Workshop*, (December 2005), 160–166. Retrieved from http://www.academia.edu/download/31114316/10.1.1.92.730.pdf#page=174