Evaluating Short Answer Using Text Similarity Measures

Bennouar Djamel
Informatics' Department
Buira University
Bouira ,Algeria
dbennouar@gmail.com

Djellatou Rachda
Informatics' Department
Saad Dahlab University
Blida,Algeria
Rachda.djellatou@gmail.com

Abstract— **Our study is to develop a semi-automatic correction tool for the evaluation of students in exams online. Our problem is the comparison of the texts of which it is the response of the student and the other is the response model of the teacher. The texts are introduced using a computer, tablet or Smartphone, calculated results are in real time to a remote server. the results in this case must be relevant and obtained very quickly. We proceeded as track of study the similarity between texts.**

Keywords—E-Learning ;short answer;Text similarity

I. INTRODUCTION

To evaluate the similarity between two documents is a problem in many disciplines such as text data analysis, information research or text data knowledge extraction. In each area, the similarity is used for different purposes.

-In textual data Analysis (ADT): the similarities are used for the description and data mining.

-in Information research: the assessment of the similarity between document and query is used to identify the relevant document in relation to the information needs expressed by users

-Text mining -in the similarity are used to produce synthetic representation of large materials collection [1]

II. TEXT SIMILARITY MEASURES

1. Syntactic Similarity

Is to compare two string by measuring their similarity using algorithmic approaches. Among the similarity measure is cited distance lavenshtein [2]; the dice coefficient [3], the index of jackard [4], the cosine [5]etc ..

1.1. String-Based Similarity**[5]**

The measurement is based on the string sequence , and the characters composition for comparing the similarity.
A chain measure is an indicator that measures the similarity or dissimilarity (distance) between two text strings for matching string or the comparison,it is decomposed into two category measures : Character-based distance measures andTerm-based distance measures .

1.1.1. Character-based distance measures:

1.1.1.1. Hamming distance

Hamming distance, d (z1; z2) between two words of the same length z1 and z2 (. Ie of two tuples in the most general case) is the number of symbols (ie d. positions) where z1 and z2 are different. [6] The algorithm of Needleman and Wunsch is based on this distance.
example:

- The Hamming distance between 101010 and 111010 is 1 for these two words differ only in the second position.
- The Hamming distance between 1010 and 0101 is 4 for these two words differ in all positions.
- The Hamming distance between 1010 and 111010 is not defined (and will not be considered).

1.1.1.2. The distance Needleman- Wunsch

The Needleman-Wunsch algorithm is an algorithm that performs a maximum overall alignment of two strings. It is commonly used in bioinformatics to align protein sequences or nucleotides. The algorithm was presented in 1970 by Saul Needleman and Christian Wunsch in their article "A general method applicable to the search for Similarities in the aminoacid sequence of two proteins." [2]

1.1.1.3. Levenshtien distance

Levenshtein distance giving a mathematical measure of the similarity between two strings. It is equal to the minimum number of characters you have to delete, insert or replace to go from one channel to another. It was proposed by Vladimir Levenshtein in 1965. it is also known under the names of edit distance or time dynamic deformation, including pattern recognition, especially speech recognition [2]

Exemple :

		m	e	i	l	e	n	s	t	e	i	n
	0	1	2	3	4	5	6	7	8	9	10	11
l	1	1	2	3	3	4	5	6	7	8	9	10
e	2	2	1	2	3	3	4	5	6	7	8	9
v	3	3	2	2	3	4	4	5	6	7	8	9
e	4	4	3	3	3	3	4	5	6	6	7	8
n	5	5	4	4	4	4	3	4	5	6	7	7
s	6	6	5	5	5	5	4	3	4	5	6	7
h	7	7	6	6	6	6	5	4	4	5	6	7
t	8	8	7	7	7	7	6	5	4	5	6	7
e	9	9	8	8	8	7	7	6	5	4	5	6
i	10	10	9	8	9	8	8	7	6	5	4	5
n	11	11	10	9	9	9	8	8	7	6	5	4

1.1.1.4. Smith-Waterman Distance

The Smith-Waterman algorithm is a sequence alignment algorithm used especially in bioinformatics. For example it is used to align nucleotide sequences or proteins. This algorithm was invented by Temple F. Smith and Waterman 1981. Michael Smith-Waterman algorithm is an optimal algorithm which gives an alignment corresponding to the highest score possible correspondence between the amino acids or nucleotides of the two sequences. The calculation of this score is based on the use of similarity matrices or substitution matrices. [5]

1.1.1.5. JaroWinkler distance

Jaro-Winkler distance measures the similarity between two strings. This is a variant proposed in 1999 by William E. Winkler, from the distance of Jaro, which is mainly used in the detection of duplicates.

More distance Jaro-Winkler between two strings, the higher they are similar. This is particularly suitable for the treatment of short string such as names or passwords. The result is normalized so as to have a measure between 0 and 1, with zero representing the absence of similarity. [2]

The similarity of two channels C1 and C2précédentes characters is calculated according to the index, or distance, Jaro as follows:

$$S_R = \frac{1}{3}\left(\frac{m}{|l1|} + \frac{m}{|l2|} + \frac{m-t}{m}\right) \quad (1)$$

The Jaro-Winkler index is calculated using the following formula:

$$S_W = S_R + l.p(1-S_R) \quad (2)$$

S is the similarity index of Jaro, the length of the common prefix of two string of R Characters, and p a weighting factor for promoting or not the chaine having a common prefix. The value proposed by Winkler 0,1.La p is the length of the prefix common to the two channels should not exceed 4.

If we take the example in previous similarity indices, we can build with strings C1 and C2in the following table:

		1	2	3	4	5	6	7	8	9	10	11	12
		i	n	f	o	r	m	a	t	i	q	u	e
1	i	1	0	0	0	0	0	0	0	1	0	0	0
2	n	0	1	0	0	0	0	0	0	0	0	0	0
3	f	0	0	1	0	0	0	0	0	0	0	0	0
4	o	0	0	0	1	0	0	0	0	0	0	0	0
5	r	0	0	0	0	1	0	0	0	0	0	0	0
6	m	0	0	0	0	0	1	0	0	0	0	0	0
7	a	0	0	0	0	0	0	1	0	0	0	0	0
8	t	0	0	0	0	0	0	0	1	0	0	0	0
9	i	1	0	0	0	0	0	0	0	1	0	0	0
10	o	0	0	0	1	0	0	0	0	0	0	0	0
11	n	0	1	0	0	0	0	0	0	0	0	0	0

The maximum distance allowed for the corresponding characters here:

$$e = \left(\frac{max(|l1|,|l2|)}{2}\right) - 1 = \frac{12}{2} - 1 = 5 \quad (3)$$

Note that the following characters are distant from each other by more than 5 characters, and will not be considered relevant:

- ✓ The first "i" of "information" and the second "i" in " informatique "
- ✓ The first "i" of " informatique " and the second "i" of "information"
- ✓ The "n" of " informatique " and the second "n" of "information"
- ✓ The "o" of " informatique " and the second "o" of "information"

The corresponding number of characters:

$$M = 9 \quad (4)$$

The respective lengths of the two strings are:

$$|l_1| = 1_2 \quad |l_2| = 1_1$$

The number of transpositions is:

$$t = 0$$

The index of Jaro for both chains is:

$$S_R = \frac{1}{3}\left(\frac{m}{|l1|} + \frac{m}{|l2|} + \frac{m-t}{m}\right) = \frac{1}{3}\left(\frac{9}{12} + \frac{9}{11} + \frac{9-0}{10}\right) = 0.82 \quad (5)$$

For 9 = (length of common prefix of two strings: "informati") and p = 0.1 (default value), the similarity index Jaro-Winkler for both string C1 and C2 will be:

$$S_W = S_R + l.p(1-SR) = 0.82 + 9*0.1(1-0.82) + 0.98 = 0.98 \quad (6)$$

1.1.1.6. The similarity using the N-gram

An n-gram is a sequence of nwords that appear consecutively in the text. An n-gram of size 1 is called an unigram, an n-gram of size 2 is called bigram and an n-gram of size 3 is called a trigram. n-gram co-occurrence measures how well a candidate summary overlaps with a reference summary using a weighted average of variable-length n-gram matches [7]

1.1.1.7. BLEU(Bilingual Evaluation Understudy)

Algorithm created by Papineni et al [8]to evaluate machine translation system from one language to another language [9] using the coverage of n-grams of translation reference [10]; he was used to evaluate google translate and bing (0.31) and (0.29) respectively. [11]

the BLUE algorithm was used with the principle of a comparison with the answer the teacher's model and the student's response with a penalty index in order to increase its performance The main idea of the comparison is to measure the near the candidate reference model named M-BLUE [7] for more performance on this type of algorithm it is necessary to make a given warehouse model reference response from the instructor or teacher; for each student response is necessary to have several typical response to increase the performance of the algorithm and compare the student's response with the most appropriate response.

1.1.2. Term-based distance measures

1.1.2.1. Manhattan distance (city block)

It calculates the distance that would be traveled to move from one data point to another if a gate path is followed. It is the sum of the differences of their corresponding components [5]then the formula is noted as follows:

$$d = \sum_{i=1}^{n} |xi - yi| \qquad (7)$$

1.1.2.2. Cosine similarity

It measures the similarity between two vector by measuring the cosine of the angle between them. [5]

$$Sc = \cos(\theta) = \frac{A.B}{|A||B|} = \frac{\sum_{i=1}^{n} Ai * Bi}{\sqrt{\sum_{i=1}^{n} (Ai)^2} \sqrt{\sum_{i=1}^{n} (Bi)^2}} \qquad (8)$$

The scalar product of two vectors A and B is divided by the product of the two vectors standards. The result of this calculation will always be between 0 and 1.

If we consider here as character strings C1 = " informatique " and C2 = "information" As their union {a, e, f, i, m, n, o, q, r, t, u} we can translate each of the two strings in a vector, considering the number of occurrences of each character contained in a string against the union of the two strings. We obtain:

C1 = (1,1,1,2,1,1,1,1,1,1,1) and C2 = (1,0,1,2,1,2,2,0,1, 1.0) Cosine is thus calculated for channels C1 and C2de characters as follows:

$$lc = \frac{(1*1)+(1*0)+(1*1)+(2*2)+(1*1)+(1*2)+(1*2)+(1*0)+(1*1)+(1*1)+(1*0)}{\sqrt{1^2+1^2+1^2+2^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2} \sqrt{1^2+0^2+1^2+2^2+1^2+2^2+2^2+0^2+1^2+1^2+0^2}}$$

$$lc = 0.41 \qquad (9)$$

1.1.2.3. Dice's coefficient

Dice's coefficient is defined as twice the number of common terms in the compared strings divided by the total number of terms in both strings.[3]

$$S = \frac{2*|X \cap Y|}{|X|+|Y|} \qquad (10)$$

Exemple :

Take the example of two strings "hello" and "hello", which was like the coefficient of Dice. The respective bigrams are {he, el, ll, lo} (ie 4 bigrams) and {al, ll, lo} (ie 3 bigrams). The intersection of the two lots are the values {ll, lo}. Thus, the equation becomes:

$$s = 2*2 /(4+3) = 4/7 = 0,57 \qquad (11)$$

With this value of 0.57 it is possible to determine that these two words are roughly similar. The higher the value is close to 1 and the chains are similar. If the result is 1, it means that the two channels are identical.

1.1.2.4. The Jaccard index

Jaccard distance are two metrics used statistics to compare the similarity and diversity between samples. They are named after the Swiss botanist Paul Jaccard. [4]It is calculated as follows:

$$S_J = \frac{A \cap B}{A \cup B} \qquad (12)$$

If we consider, for example, the two strings:
C1 = " informatique " and C2 = "information" the union of these two chains {a, e, f, i, m, n, o, q, r, t, u} has length 11 and their intersection {a f, i, m, n, o, r, t} of length 8. The Jaccard index is in this case:

$$S_J = \frac{8}{11} = 0.72 \qquad (13)$$

1.1.2.5. *Matching coefficient*

is a very simple approach based vectors that simply counts the number of similar terms (dimensions), to which the two vectors are not zero. [5]

2. Semantic Similarity

Semantic similarity is a concept based on a set of terms are given a metric based on the similarity of their meaning / semantic content Among such similarity measures, we find Resnik, LSA (Latent Semantic Analysis), ESA (Explicit Semantic Analysis)

2.1. Corpus-Based Similarity

The principle of similarity based corpus is to detect the similarity between the word according to information obtained from a large corpus as Wékipédia .there are several technique that uses this principle as LSA Explicit Semantic Analysis (ESA), PointwiseMutual Information- information Retrieval (PMI-IR) and Extraction words similar distributional using co-occurrences [11] Among the most important technique LSA technology

2.1.1. Latent SemanticAnalysis (LSA)

LSA is a statistical model developed by Landauer et al in 1988, which enabled the comparison of word semantic similarity [12]This method allows the construction of thematic classes in the form of weighted words bags made by the vocabulary of a text .dropoff window After several extensions, this model is used to treat multilingual data and extract thematic similarities. [4]

LSA is used by a matrix that captures the words and calculates the frequency of appearance in the text; the matrix is transformed into SVD (Singular Value Decomposition) it is possible to determine the similarity between the words, it is able to produce score results to approximate the results of experts; however; LSA does not consider the word order. . [13][14][15][5]

LSA has been used by different ITS (intelligent tutoring systems) as AutoTutor [16],Intelligent EssayAssessor [17] Summary Street [18]; Apex [19]etc.

2.1.2. Explicit Semantic Analysis (ESA)

Is a measure used to calculate the semantic proximity between two arbitrary texts, technique is based on Wikipedia terms (or texts) as large vectors, each vector entry with the weight of TF-IDF between term and a Wikipedia article. Semantic proximity between two words (or text) is expressed by measuring the cosine between the corresponding vectors [20]

2.1.3. PointwiseMutual Information InformationRetrieval (PMI-IR)

Is a method to calculate the similarity between pairs of words he uses advanced syntax AltaVista search queries to calculate probabilities. Most often cooccur two words near each other on a web page, the higher their PMI-IR Similarity Score is high.

2.1.4. Semantic analysis Saillant

SalientSemanticAnalysis) incorporates a similar semantic abstraction and interpretation of the words as ESA, but it uses the salient concepts collected from encyclopedic knowledge, where a "concept" is a word or phrase with unambiguous concrete meaning, and offers an encyclopedic definition. Salience in this case is determined on the basis of the word is a hyperlink in context, which means that they are highly relevant to the given text.SSA is an example of Vecteur space generalized model (GVSM) where the vectors representing the associations of words-concepts. [20][21]

2.1.5. Random Projection (RP)

is a high-dimensional space projected onto a lower one dimension, using a matrix generated randomly. contrary to LSA, RP is computationally efficient for large corpus, while maintaining accurate similarity vector and giving comparable results. [2]

2.1.6. Best Alignment Strategy

Let Ta and Tb be two text fragments of size a and b respectively. After removing all stop words, we first determine the number of shared terms (ω) between Ta and Tb. Second, we calculate the semantic relatedness of all possible pairings between non-shared terms in Ta and Tb.
We further filter these possible combinations by creating a list ϕ which holds the strongest semantic pairings between the fragments' terms, such that each term can only belong to one and only one pair.

$$Sim(T_a;T_b)=\frac{\left(\omega+\sum_{i=1}^{|\phi|}\phi i\right)*(2\ ab)}{a+b} \qquad (14)$$

where ϕi is the similarity score for the pairing[2]

2.1.7. The KNN algorithm

The KNN algorithm is among the simplest machine learning algorithms. In a classification context of a new observation x, simple basic idea is to vote nearest neighbors of this observation. X The class is determined by the majority class among the k nearest neighbors of observation x. The KNN method is a method based on neighborhood, non parametric; This mean that the algorithm allows a classification without making assumptions on the function y = f (x1, x2, ... xp) which connects the dependent variable to the independent variables[22]

To use the KNN algorithm in automatic correction of tries, the problem is resolved as follows:

The author uses the Text Categorization, then he tries transformed into vector after pre-preparation (adverb, pronoun, adjective), calculates the frequency of each feature and transformed into a vector of the same size for KNN algorithm, all the different functions of the tests make up a vector space. Each vector is expressed by the tf-idf weights. and after computes the similarity between the test trying and learning tests with the cosine formula [23]

2.1.8. Evaluation Based on Topics

Method used to evaluate tries; as he defined topics each of which is noted by a value, the final score is obtained adding all the notes. Such as NAPLAN has 10 function as assessed (spelling scored from 0 to 6 points, 0-5 vocabulary, the idea, ext ..) and each characteristic is rated by 0 a6, at the end an overall rating is attributed by summing all the functions. . [24]

	Criteria				Points
	4	3	2	1	
Level of engagement in class	Student proactively contributes to class by offering ideas and asking questions more than once per class.	Student proactively contributes to class by offering ideas and asking questions once per class.	Student rarely contributes to class by offering ideas and asking questions.	Student never contributes to class by offering ideas and asking questions.	
Listening, questioning and discussing	Respectfully listens, discusses and asks questions and helps direct the group in solving problems.	Respectfully listens, discusses and asks questions.	Has trouble listening with respect, and takes over discussions without letting other people have a turn.	Does not listen with respect, argues with teammates, and does not consider other ideas. Blocks group from reaching agreements.	
Behavior	Student almost never displays disruptive behavior during class discussions and group activities.	Student rarely displays disruptive behavior during class discussions and group activities.	Student occasionally displays disruptive behavior during class discussions and group activities.	Student almost always displays disruptive behavior during class discussions and group activities.	
Preparation	Student is almost always prepared with assignments and required class materials.	Student is usually prepared with assignments and required class materials.	Student is rarely prepared with assignments and required class materials.	Student is almost never prepared with assignments and required class materials.	
Problem-solving	Actively seeks and suggests solutions to problems.	Improves on solutions suggested by other group members.	Does not offer solutions, but is willing to try solutions suggested by other group members.	Does not try to solve problems or help others solve problems.	
Group/partner teamwork	Works to complete all group goals. Always has a positive attitude about the tasks and work of others. All team members contribute equally. Performed all duties of assigned team role.	Usually helps to complete group goals. Usually has a positive attitude about the tasks and work of others. Assisted team members in the finished project. Performed nearly all duties of assigned team role.	Occasionally helps to complete group goals. Sometimes makes fun of the group tasks and work of others. Finished individual task but did not assist team members. Performed some duties of assigned team role.	Does not work well with others and shows no interest in completing group goals. Often makes fun of the work of others and has a negative attitude. Contributed little to group effort. Did not perform duties of assigned team role.	
				Total	

Figure 1: example on Evaluation based topics

2.1.9. NLP

Is the application of Artificial Intelligence for the analysis of natural language. Several techniques used in this area as the parser to find the linguistic structure of a text [25][26]the speech analyzer for analyzing the structure of discourse in the text, lexical similarity measures, to analyze the use of the word in a text [25][27]

III. CONCLUSION

The goal of this state of the art was to view a deferent text similarity measures existing and their efficiency to choose one of them or hybrid combination in our approach to evaluate student in short answer exam and compare results

References

[1] E. Negre, “Comparaison de textes: quelques approches...,” 2013.

[2] M. Conference, S. Task, and S. T. Similarity, ** SEM 2013 : The Second Joint Conference on Lexical and Computational Semantics*, vol. 1. 2013.

[3] “Measures of the Amount of Ecologic Association Between Species Author (s): Lee R . Dice Published by : Ecological Society of America Stable URL : http://www.jstor.org/stable/1932409 .,” vol. 26, no. 3, pp. 297–302, 2013.

[4] U. D. Avignon, E. T. Des, and P. D. E. Vaucluse, “Traduction automatique statistique et adaptation à un domaine spécialisé,” 2011.

[5] W. Gomaa and A. Fahmy, “Short Answer Grading Using String Similarity And Corpus-Based Similarity,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 3, pp. 115–121, 2012.

[6] “The Bell System Technical Journal,” vol. xx, no. 2, 1950.

[7] F. Noorbehbahani and a. a. Kardan, “The automatic assessment of free text answers using a modified BLEU algorithm,” *Comput. Educ.*, vol. 56, no. 2, pp. 337–345, Feb. 2011.

[8] K. Papineni, S. Roukos, T. Ward, and W. Zhu, “B LEU : a Method for Automatic Evaluation of Machine Translation,” no. July, pp. 311–318, 2002.

[9] P. Diana and E. Alfonseca, “Application of the Bleu algorithm for recognising textual entailments,” pp. 1–4, 2001.

[10] S. Barrachina, O. Bender, F. Casacuberta, J. Civera, E. Cubel, A. Lagarda, H. Ney, and E. Vidal, “Statistical Approaches to Computer-Assisted Translation,” no. December 2007, 2008.

[11] W. Hassan and A. Aly, “Automatic scoring for answers to Arabic test questions ☪,” *Comput. Speech Lang.*, 2013.

[12] T. K. Landauer, P. W. Foltz, and D. Laham, “An introduction to latent semantic analysis,” *Discourse Process.*, vol. 25, no. 2–3, pp. 259–284, Jan. 1998.

[13] Y. He, S. C. Hui, and T. T. Quan, “Automatic summary assessment for intelligent tutoring systems,” *Comput. Educ.*, vol. 53, no. 3, pp. 890–899, Nov. 2009.

[14] S. Valenti, F. Neri, and A. Cucchiarelli, “An Overview of Current Research on Automated Essay Grading,” *J. Inf. Technol. Educ.*, vol. 2, pp. 3–118, 2003.

[15] R. Ziai, N. Ott, and D. Meurers, “Short Answer Assessment : Establishing Links Between Research Strands,” pp. 190–200, 2012.

[16] P. Wiemer-hastings, A. C. Graesser, D. Harter, B. Klettke, K. Link, B. Olde, V. Pomeroy, K. Wiemer-, H. Yetman, S. Craig, P. Chipman, M. Ring, and C. Web-, “The Foundations and Architecture of Autotutor.”

[17] P. W. Foltz and D. Laham, “Automated Essay Scoring : Applications to Educational Technology Automated scoring with LSA,” pp. 939–944.

[18] R. Through, T. H. E. Use, O. F. For, N. By, N. O. F. Of, T. Submitted, T. O. The, O. F. The, A. O. F. The, N. Of, O. Fulfillment, R. For, T. H. E. Degree, O. Of, and E. Of, “B.s., u,” 2001.

[19] W. Hou, J. Tsao, S. Li, and L. Chen, “Automatic Assessment of Students ’ Free-Text Answers,” pp. 235–243, 2010.

[20] C. Banea, S. Hassan, M. Mohler, and R. Mihalcea, “UNT : A Supervised Synergistic Approach to Semantic Text Similarity,” pp. 635–642, 2012.

[21] S. Hassan and R. Mihalcea, “Semantic Relatedness Using Salient Semantic Analysis,” pp. 884–889, 2011.

[22] E. Mathieu-dupas, “Algorithme des K plus proches voisins pondérés (WKNN) et Application en diagnostic,” 2010.

[23] L. Bin, L. Jun, Y. Jian-Min, and Z. Qiao-Ming, “Automated Essay Scoring Using the KNN Algorithm,” in *2008 International Conference on Computer Science and Software Engineering*, 2008, vol. 1, pp. 735–738.

[24] A. Fazal, F. K. Hussain, and T. S. Dillon, “An innovative approach for automatically grading spelling in essays using rubric-based scoring,” *J. Comput. Syst. Sci.*, vol. 79, no. 7, pp. 1040–1056, Nov. 2013.

[25] R. Swartz, “AUTOMATED EVALUATION OF ESSAYS AND SHORT ANSWERS Jill Burstein.”

[26] S. Abney, “Part-of-Speech Tagging and Partial Parsing,” pp. 1–23, 1996.

[27] J. Z. Sukkarieh and E. Bolge, “Building a Textual Entailment Suite for the Evaluation of Automatic Content Scoring Technologies,” pp. 3149–3156.