Journal of Theoretical and Applied Information Technology
15th February 2023. Vol.101. No 3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A HEURISTIC RANKING OF DIFFERENT SIMILARITY TECHNIQUES USED FOR EFFECTIVE LANGUAGE TRANSLATION AND PLAGIARISM DETECTION

PELURU JANARDHANA RAO[1], Dr. KUNJAM NAGESWARA RAO[2], Dr. SITARATNAM GOKURUBOYINA[3]

[1]Research Scholar, Department of Computer Science and Systems Engineering, Andhra University College of Engineering, Andhra University, Visakhapatnam, India
[2]Professor, Department of Computer Science and Systems Engineering, Andhra University College of Engineering, Andhra University, Visakhapatnam, India
[3]Research Scientist, Institute of Bioinformatics and Computational Biology (Recognized as SIRO), Visakhapatnam, India

E-mail: [1]peluru.janardhanarao@gmail.com, [2]kunjamnag@gmail.com, [3]sitagokuruboyina@gmail.com

ABSTRACT

With the rapid growth of language translation tools and digital libraries, text documents can be easily translated from one language to other resulting cross-language or multilingual plagiarism. Through this article we are presenting a detailed study in the comparison of multilingual documents for the efficient language translation. Parallel corpus is used to compare multilingual text which is a collection of similar sentences and sentences which are translation of each other. A detailed study is presented in this paper with various methods used in the literature to identify the similarity between French and English languages. A heuristic ranking model was developed to assess the suitability of various string similarity methods for determining language similarity. Through this study we concluded that the Fuzzy-Wuzzy (Partial-ratio) string similarity method out performs in terms of accuracy and Spacy similarity technique finds the similarity between the languages used for translation in less amount of time.

Keywords: *Natural Language Processing, Plagiarism, String Similarity, Fuzzy-Wuzzy, Sequence Matcher, Levenshtein Distance*

1. INTRODUCTION

Natural Language Processing (NLP) is a major field of Artificial Intelligence. The ability to make computers understand human language relies heavily on NLP. NLP employs a variety of techniques, such as text similarity and clustering, to allow machines to recognize and extract patterns from large amounts of text data.

One of the most important NLP methods for evaluating the similarity between two chunks of text based on their meaning is text similarity. Bag of Words, word2vec, and TF-IDF are word embedding techniques that are used to encrypt the text data. The encoding method makes it possible to compare sentences, extract related questions from FAQs, and scan documents in the database. The main steps in identifying text similarity in NLP are Text preprocessing, Feature extraction, Vector similarity, and Decision function.

2. RELATED WORK

Latent semantic indexing is a corpus-based approach used to evaluate text similarity based on the semantic relations among words. Cross-Lingual Latent Semantic Indexing (CL-LSI) [1] is used to find the similarity between Arabic and English documents. The paper compares HAPAX, dictionary-based similarity measures with CL-LSI methods. Monolingual and cross-lingual LSI approaches [9] are compared with the dictionary approach and LSI approaches outperformed the dictionary approach.

Comparable documents [7-10] in Arabic, English and French languages are extracted from Wikipedia and Euro News. A personalized web crawler is used [13] to build comparable corpora from Wikipedia and truly parallel sentences are filtered from comparable sentence pairs.

Comparable documents are identified by using hapax words [7] and then the documents that are paired to the same target document are filtered out using pigeonhole reasoning and cross-lingual information. Comparability of documents is measured using Binary and Cosine comparability measures [8]. Vector space model is used to represent multilingual documents and LSI is used to reduce the dimensionality of VSM.

Statistical machine translation model [6] is used to produce a bag-of-words representation in Cross-lingual information retrieval. Machine translation systems in ACCURAT project [10] are improved by using comparable corpora. Probabilistic method [11] proposed to model cross-lingual semantic similarity in context that depends on latent cross-lingual concepts. Bilingual Word Embeddings Skip-Gram model [12] learns bilingual word embeddings that are based on comparable data. Plagiarism between French and English documents is detected [14] by using Sequence matcher technique. Sequence matcher, Levenshtein distance, Fuzzy-Wuzzy approaches [2-5] are used to retrieve math formulae from text documents.

3. STRING SIMILARITY

The method of determining the degree of correspondence between two strings is known as string similarity. There are several techniques available in the literature that have been used enormously to identify the similarity of strings. In this paper, we have presented the study of similarity between strings using Levenshtein distance, sequence Matcher, Fuzzy – wuzzy (Ratio), Fuzzy–wuzzy (Partial-Ratio), Spacy and Word2Vec techniques.

3.1. Sequence Matcher

Sequence Matcher technique identifies the best ever contiguous matching subsequence that contains no useless elements. After comparing two strings, the Sequence Matcher gives the comparison ratio ranges between 0 and 1. If the comparison ratio of two strings is in between 0.7 to 0.9 then it will be deliberated as keyword and keyword will be stored in the data set. Sequence Matcher objects in python has the following key functions: Set_seq1 (a), Set_seq2 (b), Ratio ().

3.2. Levenshtein Distance

Levenshtein distance is widely used in computational linguistics, bioinformatics, DNA analysis, and molecular biology. The similarity between source string and objective string is evaluated by Levenshtein Distance. Edit distance or Levenshtein distance is mainly used for spell checking, error correction in a program, and measuring the melodies similarities or rhythms in music. Levenshtein distance is primarily used in speech recognition and Plagiarism detection.

3.3. Fuzzy-Wuzzy

Fuzzy string matching also defines an accurately precise String Matching in order to find the string that matches nearly in a given pattern. It is used in many applications including text re-use detection, spell-checking, spam filtering, as well in bioinformatics domain. Fuzzy-Wuzzy library is used to identify the string similarity between two words and gives the ratio between 0 and 1. The words are more similar if the ratio is nearer to 1. The words are irrelevant to each other if the ratio is nearer to 0. The two popular Fuzzy-Wuzzy techniques for finding the similarity between the languages are Fuzzy-Wuzzy (Ratio) technique which uses pure Levenshtein Distance based matching and Fuzzy-Wuzzy (Partial-Ratio) technique in which matching is done based on best substrings.

3.4. Spacy

The Spacy technique makes a prediction by comparing the objects and presents how similar two objects are. Similarity Prediction is beneficial for flagging replicas. Word vectors and Context Sensitive tensors techniques are supported by Spacy to identify the similarity between the words. The similarity between two sentences ranges from 0 to 1, 1 if two sentences are more similar and 0 if two sentences are not similar. There is a possibility to have high similarity value even though sentences have no common words. To remove this high similarity between unmatched sentences text preprocessing is used.

3.5. Word2Vec

Word2Vec is a family of model architectures and optimization for learning word embedding's from large datasets, rather than a single algorithm. Word2Vec-taught embedding does have proved to be effective in a number of downstream natural language processing tasks. Since 2013, Word2vec has been a tool for efficiently creating word embeddings.

3.6. Problem Statement

The similarity between two texts can be found out by semantic similarity. There are two kinds of relations between the documents direct and indirect. The similarity between the documents

always depends on this relationship. Semantic similarity can also be used in Twitter for measuring the semantic association between the texts. Measuring semantic relationship is the main task in NLP. This semantic similarity between the words is the main motivation for this paper.

4. EXPERIMENTAL ANALYSIS AND RESULTS

Experimentation is done with 800 documents consists of French and English text. Accuracy is the first metric used to identify similarity between English and French documents. Time analysis is the second metric used to find the time taken to identify similarities between English and French documents. To reduce the retrieval time, the speeds of RAM and processor plays major role. 4GB RAM and I3 processor are used in this experiment to find the retrieval time.

Table 1: The Ranking of the models in the literature based on Accuracy and Time.

Math Retrieval Technique	Description	Accuracy Ranking	Time Ranking
Sequence Matcher	Sequence matcher's main goal is to identify the best ever contiguous matching subsequence with no unnecessary items.	5	4
Levenshtien Distance	The Levenshtein Distance is used to determine how similar the source and objective strings are.	2	3
Fuzzy-Wuzzy	String Matching, which is the method of identifying strings that roughly fit a given pattern, is also known as fuzzy string matching.	1	2
Spacy	By comparing the sets, the Spacy technique makes a prediction and shows how close two objects are.	3	1
Word2Vec	Word2Vec (W2V) takes a text corpus as input and returns a vector representation for each word.	4	5

Table 1 explains the rank assigned to the proposed models and the models in the literature based on the performance metric accuracy and time. Fuzzy-Wuzzy outperforms well both in terms of accuracy and time. The proposed models were nearly tested on 800 documents. Table 2 to table 17 presents the number French words, French Synonyms, English words, English Synonyms in each documents and the accuracy of translation with all the models discussed and the documents are mapped in one-one mappings.

Table 2: Similarity between English (Active) and French (Active) languages using Sequence Matcher.

Sample	Sequence Matcher (English Active + French Active)					
	French Word Count	French Synonyms Count	French + English Count	English Word Count	Final Left words	Accuracy
1	208	1545	8942	170	37	78.03
2	268	1688	9404	214	53	74.23
3	353	2271	11234	291	56	80.15
4	401	2227	10772	340	84	74.29
5	448	2689	12329	373	85	76.21
6	295	1775	9658	247	68	71.46
7	332	2194	11307	289	86	70.24

Table 3: Similarity between English (Active) and French (Passive) languages using Sequence Matcher.

Sample	**Sequence Matcher** Sequence Matcher (English Active + French Passive)					
	French Word Count	French Synonyms Count	French + English Count	English Word Count	Final Left words	Accuracy
1	218	1481	8780	170	36	71.94
2	282	1954	10424	214	60	70.02
3	380	2348	11618	291	65	76.66
4	431	2431	11369	340	90	73.12
5	413	2687	12562	373	107	70.31
6	302	1658	9001	247	75	69.63
7	343	2414	11955	289	97	66.43

Table 4: Similarity between English (Passive) and French (Active) languages using Sequence Matcher.

Sample	Sequence Matcher (English Passive + French Active)					
	French Word Count	French Synonyms Count	French + English Count	English Word Count	Final Left words	Accuracy
1	208	1545	8942	187	37	73.86
2	268	1688	9404	230	60	73.31
3	353	2271	11234	320	66	78.37
4	401	2227	10772	375	92	73.46
5	448	2689	12329	405	85	78.01
6	295	1775	9658	261	74	71.64
7	332	2194	11307	301	88	70.76

Table 5: Similarity between English (Passive) and French (Passive) languages using Sequence Matcher.

Sample	Sequence Matcher (English Passive + French Passive)					
	French Word Count	French Synonyms Count	French + English Count	English Word Count	Final Left words	Accuracy
1	218	1481	8780	187	39	73.19
2	282	1954	10424	230	57	74.71
3	380	2348	11618	320	63	80.11
4	431	2431	11369	375	78	78.20
5	413	2687	12562	373	86	77.76
6	302	1658	9001	261	70	73.18
7	343	2414	11955	301	81	73.08

Table 6: Similarity between English (Active) and French (Active) languages using Levenshtein Distance.

Sample	Levenshtein Distance (English Active + French Active)					
	French Word Count	French Synonyms Count	French + English Count	English Word Count	Final Left words	Accuracy
1	208	1545	8942	170	17	98.23
2	268	1688	9404	214	4	98.13
3	353	2271	11234	291	17	99.31
4	401	2227	10772	340	13	96.76
5	448	2689	12329	373	17	96.24
6	295	1775	9658	247	13	97.16
7	332	2194	11307	289	7	94.80

Table 7: Similarity between English (Passive) and French (Active) languages using Levenshtein Distance.

Sample	Levenshtein Distance (English Active + French Passive)					
	French Word Count	French Synonyms Count	French + English Count	English Word Count	Final Left words	Accuracy
1	218	1481	8780	170	26	98.23
2	282	1954	10424	214	3	98.59
3	380	2348	11618	291	10	99.31
4	431	2431	11369	340	20	96.47
5	413	2687	12562	373	26	94.63
6	302	1658	9001	247	20	97.16
7	343	2414	11955	289	6	95.15

Table 8: Similarity between English (Active) and French (Passive) languages using Levenshtein Distance.

Sample	Levenshtein Distance (English Passive + French Active)					
	French Word Count	French Synonyms Count	French + English Count	English Word Count	Final Left words	Accuracy
1	208	1545	8942	187	18	97.32
2	268	1688	9404	230	3	98.69
3	353	2271	11234	320	9	98.12
4	401	2227	10772	375	19	98.13
5	448	2689	12329	405	16	97.28
6	295	1775	9658	261	19	97.31
7	332	2194	11307	301	10	96.01

Table 9: Similarity between English (Passive) and French (Passive) languages using Levenshtein Distance.

Sample	Levenshtein Distance (English Passive + French Passive)					
	French Word Count	French Synonyms Count	French + English Count	English Word Count	Final Left words	Accuracy
1	218	1481	8780	187	23	97.32
2	282	1954	10424	230	2	99.13
3	380	2348	11618	320	19	98.12
4	431	2431	11369	375	10	97.86
5	413	2687	12562	373	12	96.79
6	302	1658	9001	261	10	97.31
7	343	2414	11955	301	11	96.67

Table 10: Similarity between English (Active) and French (Active) languages using Fuzzy-Wuzzy (Partial-Ratio).

Sample	Fuzzy- Wuzzy (Partial-Ratio) (English Active + French Active)					
	French Word Count	French Synonyms Count	French + English Count	English Word Count	Final Left words	Accuracy
1	208	1545	8942	170	2	100
2	268	1688	9404	214	0	100
3	353	2271	11234	291	0	100
4	401	2227	10772	340	1	100
5	448	2689	12329	373	1	100
6	295	1775	9658	247	0	99.59
7	332	2194	11307	289	1	100

Table 11: Similarity between English (Passive) and French (Active) languages with Fuzzy-Wuzzy (Partial-Ratio).

Sample	Fuzzy- Wuzzy (Partial-Ratio) (English Active + French Passive)					
	French Word Count	French Synonyms Count	French + English Count	English Word Count	Final Left words	Accuracy
1	218	1481	8780	170	1	100
2	282	1954	10424	214	0	100
3	380	2348	11618	291	1	100
4	431	2431	11369	340	1	100
5	413	2687	12562	373	0	100
6	295	1775	9658	261	1	99.59
7	332	2194	11307	301	1	100

Table 12: Similarity between English (Active) and French (Passive) languages with Fuzzy-Wuzzy (Partial-Ratio).

Sample	Fuzzy- Wuzzy (Partial-Ratio) (English Passive + French Active)					
	French Word Count	French Synonyms Count	French + English Count	English Word Count	Final Left words	Accuracy
1	208	1545	8942	187	2	100
2	268	1688	9404	230	0	100
3	353	2271	11234	320	1	100
4	401	2227	10772	375	1	100
5	448	2689	12329	405	0	100
6	295	1775	9658	261	1	99.61
7	332	2194	11307	301	1	100

Table 13: Similarity between English (Passive) and French (Passive) languages with Fuzzy-Wuzzy(Partial-Ratio).

Sample	Fuzzy- Wuzzy (Partial-Ratio) (English Passive + French Passive)					
	French Word Count	French Synonyms Count	French + English Count	English Word Count	Final Left words	Accuracy
1	218	1481	8780	187	3	100
2	282	1954	10424	230	1	100
3	380	2348	11618	320	0	100
4	431	2431	11369	375	0	100
5	413	2687	12562	373	0	100
6	302	1658	9001	261	1	99.61
7	343	2414	11955	301	1	100

Table 14: Similarity between English (Active) and French (Active) languages using Spacy.

Sample	Spacy (English Active + French Active)					
	French Word Count	French Synonyms Count	French + English Count	English Word Count	Final Left words	Accuracy
1	208	1545	8942	170	4	90.76
2	268	1688	9404	214	4	91.50
3	353	2271	11234	291	4	91.72
4	401	2227	10772	340	2	92.86
5	448	2689	12329	373	4	92.56
6	295	1775	9658	247	9	92.44
7	332	2194	11307	289	4	91.73

Table 15: Similarity between English (Passive) and French (Active) languages using Spacy.

Sample	Spacy (English Active + French Passive)					
	French Word Count	French Synonyms Count	French + English Count	English Word Count	Final Left words	Accuracy
1	218	1481	8780	170	6	90.87
2	282	1954	10424	214	6	91.20
3	380	2348	11618	291	7	91.48
4	431	2431	11369	340	4	93.07
5	413	2687	12562	373	6	92.30
6	302	1658	9001	247	6	92.94
7	343	2414	11955	289	9	91.45

Table 16: Similarity between English (Active) and French (Passive) languages using Spacy.

Sample	Spacy (English Passive + French Active)					
	French Word Count	French Synonyms Count	French + English Count	English Word Count	Final Left words	Accuracy
1	208	1545	8942	187	4	92.88
2	268	1688	9404	230	4	94.01
3	353	2271	11234	320	5	93.39
4	401	2227	10772	375	2	94.69
5	448	2689	12329	405	4	93.99
6	295	1775	9658	261	7	93.93
7	332	2194	11307	301	6	93.35

Table 17: Similarity between English (Passive) and French (Passive) languages using Spacy.

Sample	Spacy (English Passive + French Passive)					
	French Word Count	French Synonyms Count	French + English Count	English Word Count	Final Left words	Accuracy
1	218	1481	8780	187	5	92.99
2	282	1954	10424	230	5	93.78
3	380	2348	11618	320	3	93.22
4	431	2431	11369	375	1	94.95
5	413	2687	12562	373	5	93.78
6	302	1658	9001	261	5	94.51
7	343	2414	11955	301	7	93.15

Table 18 and 19 represents the accuracy with models discussed in one-to-many mapping of documents and Tables 20-22 presents the results with many-to-many mapping of documents. It is observed from the Tables 18 to 22 that Fuzzy-Wuzzy (Partial-Ratio) outperforms the remaining techniques in terms of accuracy. Table 23 deals with time taken to find the similarities between documents with different document sizes. It is observed from the table that Spacy similarity technique identifies the similarity between the documents in less time compared to the remaining techniques.

Table 18: Accuracy with Similarity techniques in one-to-many mode.

French Sample		English Sample		ACCURACY				
				Sequence-Matcher	Levenshtein-Distance	Fuzzy-Wuzzy (Partial-Ratio)	Spacy	Word2Vec
1	Active	1	Active	78.23	98.23	99.41	90.67	78.23
1		2		60.28	97.66	100	91.13	60.28
1		3		67.01	98.62	99.65	92.28	67.35
1		4		60.58	96.76	100	92.99	60.58
1		5		54.95	93.02	99.19	92.46	54.95
1	Passive	1	Active	72.94	98.23	99.41	90.87	72.94
1		2		61.21	96.72	100	91.21	61.21
1		3		67.01	98.62	99.65	92.37	67.35
1		4		59.11	96.17	100	93.05	59.11
1		5		53.61	93.29	99.19	92.60	53.61
1	Active	1	Passive	74.86	97.32	99.46	92.88	74.86
1		2		66.95	98.26	100	93.73	66.95
1		3		70.93	97.81	99.68	94.07	71.25
1		4		65.86	98.13	99.73	95.00	65.86
1		5		62.22	94.81	99.50	94.19	62.22
1	Passive	1	Passive	73.79	97.32	100	92.99	73.79
1		2		63.91	97.39	99.56	93.77	63.91
1		3		69.37	97.81	99.68	94.16	69.68
1		4		62.13	97.6	100	95.09	62.13
1		5		61.97	94.81	99.75	92.99	61.97

Table 19: Accuracy with Similarity techniques in one-to-many mode.

French Sample		English Sample		ACCURACY				
				Sequence-Matcher	Levenshtein-Distance	Fuzzy-Wuzzy (Partial-Ratio)	Spacy	Word2Vec
2	Active	1	Active	67.05	97.64	100	90.74	68.25
2		2		75.23	98.13	100	91.50	75.70
2		3		65.97	98.28	99.65	92.30	66.66
2		4		61.17	96.17	100	93.17	61.17
2		5		56.56	93.03	99.73	92.45	56.83
2	Passive	1	Active	68.23	97.64	100	90.24	69.41
2		2		71.02	98.59	100	91.20	71.49
2		3		68.04	98.62	99.65	91.96	68.72
2		4		62.05	96.17	100	93.05	62.05
2		5		55.76	92.76	99.73	92.30	56.03
2	Active	1	Passive	66.31	97.32	97.32	92.84	67.37
2		2		73.91	98.69	98.69	94.01	74.34
2		3		68.43	97.5	97.5	94.07	69.06
2		4		64.266	97.6	97.6	95.11	64.53
2		5		60.49	94.56	94.56	94.18	60.74
2	Passive	1	Passive	66.84	97.86	100	92.44	67.91
2		2		75.21	99.13	99.56	93.78	75.65
2		3		72.18	97.81	99.68	93.73	72.81
2		4		66.13	97.6	100	94.97	66.40
2		5		62.71	95.06	99.75	94.02	62.96

Table 20: Accuracy with levenshtein and Sequence matcher Similarity techniques in many-to-many mode.

French pair	English pair	Levenshtein - Distance	French pair	English pair	Sequence - Matcher
[5 10]	[3 6]	97.43	[3 10]	[1 9]	74.03
[4 10]	[3 8]	98.08	[3 7]	[3 6]	76.33
[4 7]	[3 9]	98.48	[2 7]	[2 7]	73.72
[2 10]	[3 9]	98.10	[3 9]	[3 10]	75.34
[5 9]	[2 6]	97.72	[1 8]	[1 6]	71.35
[1 6]	[1 6]	97.53	[4 6]	[3 6]	74.75
[4 8]	[1 9]	97.72	[5 9]	[5 6]	75.80
[4 8]	[2 9]	97.68	[1 10]	[2 9]	72.00
[3 10]	[2 9]	97.89	[5 7]	[5 9]	77.65
[5 10]	[2 9]	97.89	[1 7]	[3 9]	73.86

Table 21: Accuracy with Spacy and Word2Vec Similarity techniques in many-to-many mode.

French pair	English pair	Spacy	French pair	English pair	Word2Vec
[4 6]	[4 6]	93.17	[4 9]	[4 6]	75.22
[1 10]	[4 6]	93.30	[3 10]	[3 9]	77.27
[2 9]	[4 8]	93.27	[3 8]	[4 9]	73.5
[1 8]	[5 8]	93.26	[5 7]	[5 8]	74.34
[4 8]	[4 9]	93.10	[3 8]	[3 7]	73.51
[4 10]	[4 6]	93.06	[2 6]	[3 6]	73.27
[1 9]	[5 10]	92.20	[1 6]	[1 6]	77.77
[1 7]	[4 6]	93.49	[4 10]	[3 10]	76.04
[2 9]	[4 6]	93.09	[5 7]	[5 6]	77.22
[5 10]	[5 8]	92.92	[5 8]	[3 9]	75.37

Table 22: Accuracy with Fuzzy-Wuzzy (Ratio) and Fuzzy-Wuzzy (Partial-Ratio) Similarity techniques in many-to-many mode.

French pair	English pair	Fuzzy-Wuzzy (Ratio)	French pair	English pair	Fuzzy-Wuzzy (Partal-Ratio)
[3 10]	[3 8]	93.19	[4 9]	[1 6]	100
[1 6]	[1 6]	94.07	[5 9]	[3 10]	100
[2 7]	[3 6]	93.29	[5 9]	[3 7]	100
[5 6]	[2 10]	93.82	[1 9]	[1 8]	100
[1 8]	[2 9]	94.10	[2 9]	[4 10]	100
[3 10]	[3 9]	94.69	[5 9]	[1 7]	100
[5 10]	[2 9]	93.68	[3 8]	[2 7]	100
[3 7]	[3 8]	93.36	[4 6]	[3 8]	100
[2 6]	[1 6]	92.83	[3 8]	[4 7]	100
[4 10]	[3 10]	94.09	[3 9]	[1 7]	100

Table 23: Time for finding the similarity between languages with Similarity techniques.

Sample Size	TIME					
	Sequence matcher	Levenshtein Distance	Fuzzy-Wuzzy (ratio)	Fuzzy-Wuzzy (Partial-Ratio)	Spacy	Word2Vec
[Size 3KB]	9	8	6	5	5	12
[Size 6KB]	16	14	12	9	8	22
[Size 9KB]	24	19	14	14	10	34
[Size 15KB]	36	28	23	22	16	53
[Size 18KB]	43	36	27	26	19	64
[Size 21KB]	50	42	33	33	22	72
[Size 30KB]	66	58	47	46	33	101
[Size 33KB]	72	64	52	51	36	110
[Size 36KB]	76	70	56	55	38	121
[Size 45KB]	93	84	72	71	46	152

5. CONCLUSION

In this paper we have discussed heuristic ranking for finding the best model for language similarity between French and English documents. The results presented in this paper with some string similarity techniques like Spacy similarity technique, sequence matcher, Fuzzy-Wuzzy (partial ratio), and Levenshtein distance techniques. The performance of the proposed techniques compared in terms of Accuracy and time. The accuracy is more and translation time is less for Fuzzy- Wuzzy (Partial-Ratio) compared to all the models discussed.

REFERENCES:

[1] David Langlois, Motaz Saad, Kamel Smaïli. "Alignment of comparable documents: comparison of similarity measures on French-English-Arabic data". Natural Language Engineering, Cambridge University Press (CUP), 2018, 24 (5), pp.677-694. 10.1017/S1351324918000232. hal-01819710.

[2] G AppaRao, K VenkataRao, P V G D Prasad Reddy and T Lava Kumar, "An Efficient Procedure for Characteristic mining of Mathematical Formulas from Document", International Journal of Engineering Science and Technology (IJEST), Mar 2018, Vol. 10 No.03, pp. 152-157.

[3] G. Appa Rao, G. Srinivas, K.Venkata Rao, P.V.G.D. Prasad Reddy, "Characteristic mining of Mathematical Formulas from Document - A Comparative Study on Sequence Matcher and Levenshtein Distance procedure," International Journal of Computer Sciences and Engineering, Vol.6, Issue.4, pp.400-404, 2018.

[4] G AppaRao, G Srinivas, K VenkataRao and P V G D Prasad Reddy, "A partial ratio and ratio based fuzzy-wuzzy procedure for characteristic mining of mathematical formulas from documents", IJSC- ICTACT Journal on Soft Computing, July 2018, Vol 8, Issue 4, pp. 1728-1732

[5] K.N.Brahmaji Rao, G.Srinivas, P.V.G.D Prasad Reddy, T.surendra "A Heuristic ranking of different Characteristic mining based Mathematical Formulae retrieval models", Volume-9 Issue-1, October 2019.

[6] Felix Hieber and Stefan Riezler. 2015. "Bag-of-Words Forced Decoding for Cross-Lingual Information Retrieval". In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1172–1182, Denver, Colorado. Association for Computational Linguistics.

[7] Emmanuel Morin, Amir Hazem, Florian Boudin, and Elizaveta Loginova-Clouet. 2015. LINA: "Identifying Comparable Documents from Wikipedia. In Proceedings of the Eighth Workshop on Building and Using Comparable Corpora", pages 88–91, Beijing, China. Association for Computational Linguistics.

[8] Motaz Saad, David Langlois, Kamel Smaïli, "Extracting Comparable Articles from Wikipedia and Measuring their Comparabilities", Procedia - Social and Behavioral Sciences, Volume 95, 2013, Pages 40-47, ISSN 1877-0428, https://doi.org/10.1016/j.sbspro.2013.10.620.

[9] Saad, M., Langlois, D., and Smaïli, K. 2014. "Cross-lingual semantic similarity measure for comparable articles". In Proceedings of the Advances in Natural Language Processing – 9th International Conference on NLP (PolTAL 2014), Warsaw, Poland, Springer International Publishing, https://doi.org/10.1007/978-3-319-10888-9_11.

[10] Skadina, I., Aker, A., Mastropavlos, N., Su, F., Tufis, D., Verlic, M., Vasiljevs, A., Babych, B., Clough, P., Gaizauskas, R., Glaros, N., Paramita , M.L., and Pinnis, M. 2012. "Collecting and using comparable corpora for statistical machine translation". In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey, European Language Resources Association (ELRA), pp. 438–445.

[11] Ivan Vulić and Marie-Francine Moens. 2014. "Probabilistic Models of Cross-Lingual Semantic Similarity in Context Based on Latent Cross-Lingual Concepts Induced from Comparable Data". In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 349–362, Doha, Qatar. Association for Computational Linguistics.

[12] Ivan Vulić and Marie-Francine Moens. 2015. "Monolingual and Cross-Lingual Information Retrieval Models Based on (Bilingual) Word Embeddings". In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15). Association for Computing Machinery, New York, NY, USA, 363–372. https://doi.org/10.1145/2766462.2767752.

[13] Krzysztof Wołk, Krzysztof Marasek, "Building Subject-aligned Comparable Corpora and Mining it for Truly Parallel Sentence Pairs", Procedia Technology, Volume 18, 2014, Pages 126-132, ISSN 2212-0173, https://doi.org/10.1016/j.protcy.2014.11.024.

[14] Sree Ram Kiran Nag, M., Srinivas, G., Venkata Rao, K., Vakkalanka, S., Nagendram, S. (2022). "An Efficient Procedure for Identifying the Similarity Between French and English Languages with Sequence Matcher Technique". Lecture Notes on Data Engineering and Communications Technologies, vol 86. Springer, Singapore. https://doi.org/10.1007/978-981-16-5685-9_4.