Comparison between Surface-based and Dependency-based Relation Extraction Approaches for Automatic Generation of Multiple-Choice Questions

Naveed Afzal[1] and Abdullah Bawakid[2]

***Abstract*— Multiple Choice Questions (MCQs) are frequently used as an assessment tool in various e-Learning applications. In this paper, we compare two systems for automatic generation of multiple-choice question (MCQs) that are based on semantic relations. Both systems used an unsupervised approach for relation extraction to be applied in the context of automatic generation of MCQs. Both approaches aim to identify the most important semantic relations in a document without assigning explicit labels to them in order to ensure broad coverage, unrestricted to predefined types of relations. One system is based on surface-based semantic relations while other utilizes dependency-based semantic relations. The surface-based MCQ system extract semantic relations between named entities in a text via Information Extraction methodologies and automatically generate questions from extracted semantic relations while the dependency-based MCQ system extract semantic relations between named entities by employing a dependency-based tree model. Our findings indicate that the dependency-based MCQ system performs better than the surface-based MCQ system.**

***Keywords*— Biomedical informatics, Electronic learning, Data mining, Natural language processing**

I. INTRODUCTION

E-Learning in the last two decades has seen an exceptional growth and now many organisations and educational institutes employ e-Learning applications for training and testing of their staff and students respectively. The continuous development in the area of information technology and increasing use of the internet has resulted in a huge global market and rapid growth for e-Learning and its applications. One of the most popular e-Learning applications is MCQ tests that are frequently used for objective assessment. MCQ tests provide an effective and efficient measure of test-taker's performance and feedback test results to learners. MCQs are straightforward to conduct and in many disciplines instructors use MCQs as a preferred assessment tool and it is estimated that 45% - 67% student assessments utilise MCQs [1].

In the literature (see, e.g., [2]) the structure of a multiple choice question is described as follows. A multiple choice question is known as an *item*. The part of text which states the question is called *the stem* while the set of possible answers (correct and incorrect) are called *options*. The correct answer is called the *key* while incorrect answers are called *distractors*. Figure 1 shows an example of a multiple choice question. The work done in the area of automatic generation of MCQs does not have a long history. A detailed overview of various existing approaches has been presented by [3 and 4].

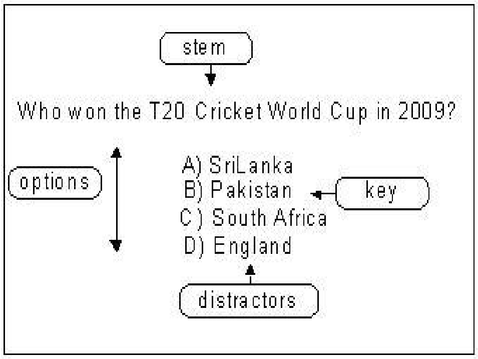

Fig. 1: An example of MCQ

II. SURFACE-BASED MCQ SYSTEM

Afzal and Pekar [5] presented a surface-based MCQ system that extract semantic rather than syntactic relations between key concepts in a given text by using Information Extraction (IE) methodologies. Questions were automatically generated from these extracted semantic relations using a certain set of rules while distractors were automatically generated using a distributional similarity measure (see [3] for complete description of the system).

[1] Naveed Afzal, Department of Biomedical Informatics, Mayo Clinic Rochester, MN, USA, 55901, (Phone: 507-513-2894, Email: afzal.naveed@mayo.edu)

[2] Abdullah Bawakid, Faculty of Computing and Information Technology, University of Jeddah, Jeddah, Saudi Arabia, (Email: abawakid@uj.edu.sa)

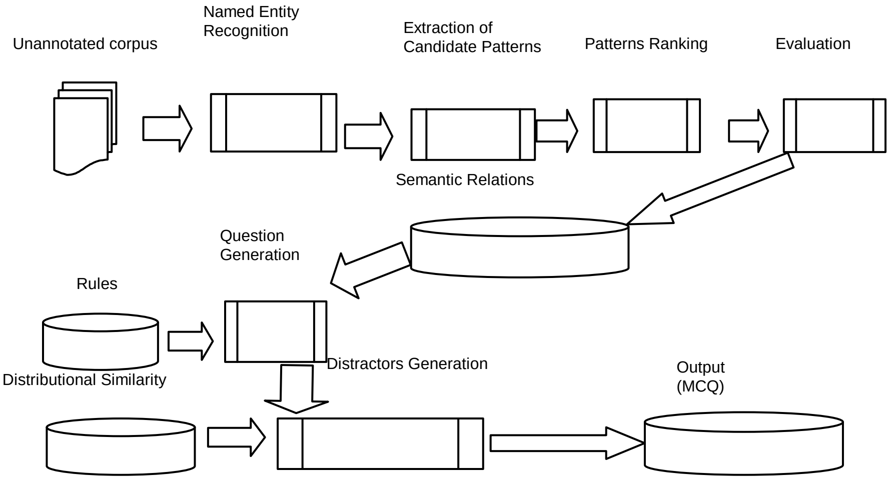

Fig. 2: System Architecture

The IE component consist of two main phases, in the first phase unannotated text is first processed by Named Entity Recogniser (NER) and in the second phase candidates patterns are extracted, ranked according to their domain relevance by using various information theoretic concepts and statistical tests of association and in the final stage extracted patterns are intrinsically evaluated in terms of precision, recall and F-score. The initial experimental results of IE components are reported in [5]. Two different measures were used for selecting ranked patterns: score-thresholding and rank-thresholding. The system used GENIA corpus as the domain corpus and the British National Corpus (BNC) as a general corpus. In the intrinsic evaluation phase, GENIA EVENT Annotation corpus [6] was used. In the automatic question generation phase the extracted semantic relations were automatically transformed into questions by employing certain set of rules while the distractors were automatically generated using a distributional similarity measure.

III. DEPENDENCY-BASED MCQ SYSTEM

Afzal and Mitkov [4] presented a dependency-based MCQs system that uses a dependency tree model to extract semantic relations from a given text. The dependency tree model was chosen because dependency trees are regarded as a suitable basis for semantic patterns acquisition as they abstract away from the surface structure to represent relations between elements (entities) of a sentence. In a dependency tree a pattern is defined as a path in the dependency tree passing through zero or more intermediate nodes within a dependency tree [7]. An insight of usefulness of the dependency patterns was provided by [8] in their work as they revealed that dependency parsers have the advantage of generating analyses which abstract away from the surface realisation of text to a greater extent than phrase structure grammars tend to, resulting in semantic information being more accessible in the representation of the text which can be useful for IE.

The dependency-based MCQ system follow the same system architecture that followed by surface-based MCQ system. The IE component of the dependency-based MCQ system is discussed in detail in [9].

IV. IE COMPONENT COMPARISON

In the IE component of surface-based MCQ system [5] discussed three different surface type patterns (e.g., untagged word patterns, PoS-tagged word patterns and verb-centred patterns) along with prepositions and their experimental results revealed that the verb-centred pattern type along with prepositions performed better than compared to other pattern types and moreover inclusion of prepositions provide useful insight into extracted semantic relations. Among various ranking methods they found that CHI and NMI are the best performing ranking methods. CHI is the best performing ranking method in terms of precision scores but recall scores are very low while using NMI they were able to attain much better recall scores. Moreover, the score-thresholding measure performs better than the rank-thresholding.

In the IE component of dependency-based MCQ system [9] explored dependency-based pattern approach and there too they found that overall CHI and NMI are the best performing ranking methods while the score-thresholding ranking measure outperforms the rank-thresholding.

In this section, we compare the precision scores obtained by using the best performing ranking methods (NMI and CHI) for the dependency-based patterns with the surface-based verb-centred patterns along with prepositions for the GENIA corpus. Figure 3 shows the comparison of precision scores obtained using NMI ranking method for GENIA corpus between the dependency-based patterns and the surface-based verb-centred patterns along with prepositions.

Figure 3 shows that the NMI ranking method in dependency-based patterns is able to achieve higher precision scores compare with the NMI ranking method in surface-based verb-centred patterns while Figure 4 shows the same comparison but using CHI ranking method.

Figure 4 also shows that precision scores attained by the dependency-based approach are higher than the scores attained by the surface-based approach.

Overall, the results achieved from Figures 3 and 4 revealed that the dependency-based patterns outperform the best performing surface-based pattern type (verb-centred along with prepositions) in terms of precision scores.

Moreover, the dependency-based approach provided more coverage compared to the surface-based approach. The dependency-based approach enabled us to extract semantic relations that the surface-based approach was unable to extract as it abstract away from different surface realisations of semantic relations. The surface-based approach was able to extract much more effectively those semantic relations that involved PROTEIN and DNA named entities but it was unable to extract a few semantic relations that involved the following named entities (CELL_LINE, CELL_TYPE and RNA) while the dependency-based approach was able to extract these effectively. For example:

[V/express] (subj[CELL_LINE] + obj[RNA])

[V/activate] (p[CELL_LINE] + p[CELL_LINE])

[V/show] (subj[CELL_TYPE] + obj[expression] + prep[of] + P[RNA])

[V/enhance] (a[RNA] + obj[transcription] + prep[in] + p[CELL_LINE])

[V/inhibit] (a[RNA] + obj[transcription] + prep[in] + p[CELL_LINE])

[V/mediate] (obj[transcription] + prep[of] + p[DNA] + prep[in] + p[CELL_LINE])

Our detailed analysis has revealed that the IE component of dependency-based approach is much more effective in extracting semantic relations than the IE component of surface-based approach.

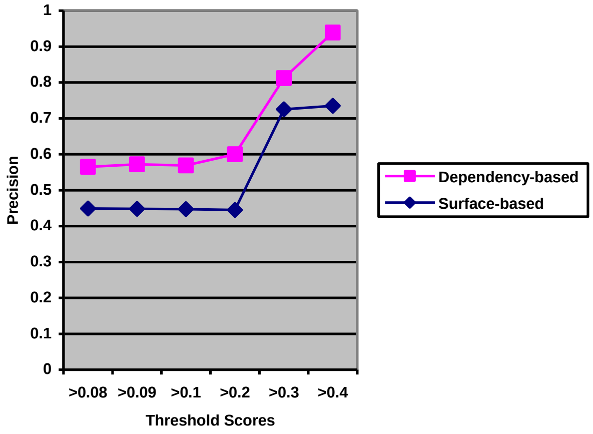

Fig. 3: Comparison of precision scores using NMI for GENIA corpus between dependency-based and surface-based patterns

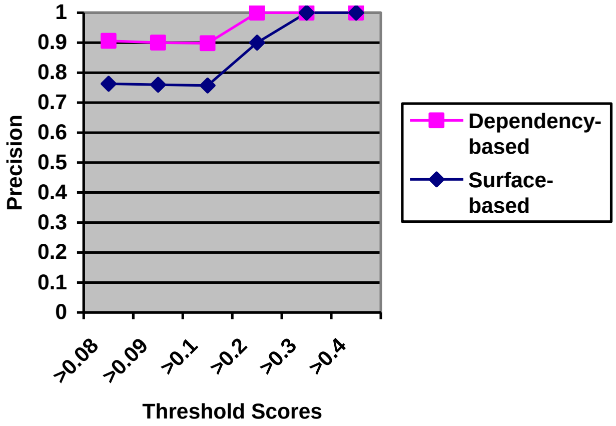

Fig. 4: Comparison of precision scores using CHI for GENIA corpus between dependency-based and surface-based patterns

V. EXTRINSIC EVALUATION COMPARISONS

In the previous section, we have compared IE component of both surface-based and dependency-based MCQs systems. In this section, we will perform a comparison between extrinsic evaluation results of both systems. Both systems: surface-based and dependency-based; were evaluated as a whole in a user-centred fashion. In both systems, the quality of automatically generated MCQs was evaluated by human evaluators that were experts in a biomedical domain. Both systems were evaluated in terms of their robustness, effectiveness and efficiency. From a given dataset, surface-based MCQ system automatically generated 80 MCQs while on the same dataset 52 MCQs were automatically generated by dependency-based MCQ system. Both MCQs systems were extrinsically evaluated by two biomedical experts on the basis of following criteria: readability, usefulness of semantic relation, relevance, acceptability and overall MCQ usability. (For more details please see Afzal 2015, Afzal and Mitkov 2014). Table 1 shows the results obtained after the evaluation of both MCQs systems where QR, DR, USR, QRelv, DRelv, QA, DA and MCQ Usability represents Question Readability, Distractors Readability, Question Relevance, Distractors Relevance, Question Acceptability, Distractors Acceptability and Overall MCQ Usability respectively.

In order to compare the evaluation results (Table 1) of both MCQ systems, we take average scores of all the categories for each MCQ system and compare them. Figure 5 shows the comparison between two MCQ systems.

The results from Figure 5 show that MCQs generated using the dependency-based approach achieve better results during extrinsic evaluation in terms of question readability, usefulness of semantic relation, question and distractors relevance, question and distractors acceptability and overall usability of MCQ. These results are better compared with the extrinsic evaluation results of surface-based MCQs system respectively. In terms of overall MCQ usability, the extrinsic evaluation results show that in surface-based MCQ system 35% of MCQ items were considered directly usable, 30% needed minor revisions and 14% needed major revisions while 21% MCQ items were deemed unusable. In case of dependency-based MCQ system, we found that 65% of MCQ items were considered directly usable, 23% needed minor revisions and 6% needed major revisions while 6% of MCQ items were unusable.

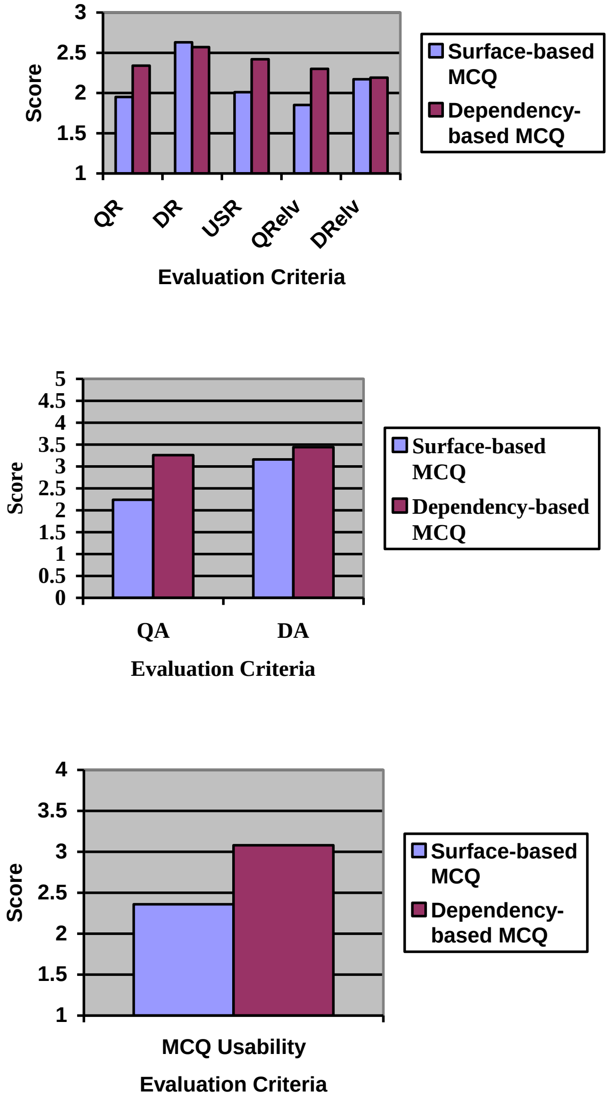

Fig. 5: Comparison between surface-based and dependency-based MCQ systems

VI. DISCUSSIONS

We used Kappa statistics [10] in order to measure the agreement between the two evaluators. Kappa statistics are a quite useful and popular quantitative measure that is used to measure the agreement between evaluators. The Kappa coefficient between evaluators is defined as:

$$K = \frac{P_A - P_E}{1 - P_E}$$

where P_A is the times evaluators agree and P_E is the proportion of times that we would expect the evaluators to agree by chance. K = 1 when there is a complete agreement among the evaluators while K = 0 when there is no agreement. The interpretation of the Kappa score is very important and an example of a commonly used scale is presented in Table 2 [10].

TABLE 1: INTERPRETATION OF KAPPA SCORES

Kappa Score	Agreement
<0.20	Poor
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Good
0.81 – 1.00	Excellent

As in extrinsic evaluation, both the evaluators evaluated both systems according to the aforementioned criteria in Section 5. We measured the agreement between the evaluators by using Kappa score which is shown in Table 2.

TABLE 2: KAPPA SCORES

Evaluation Criteria	Kappa Score (Surface-based MCQ)	Kappa Score (Dependency-based MCQ)
Question Readability	0.29	0.31
Distractors Readability	0.08	-0.13
Usefulness of Semantic Relation	0.21	0.42
Question Relevance	0.27	0.22
Distractors Relevance	0.29	0.31
Question Acceptability	0.27	0.26
Distractors Acceptability	0.12	0.10
Overall MCQ usability	0.25	0.23

The average Kappa score is 0.27 which is fair according to Table 2 but not very high due to various different sub-categories present in the extrinsic evaluation.

We used weighted Kappa [11] to measure the agreement across major sub-categories in which there is a meaningful difference.

For example, in question readability there was three sub-categories: ‘Clear’, ‘Rather Clear’ and ‘Incomprehensible’. In this case we may not care whether one evaluator chooses question readability as ’Clear’ while another evaluator chooses ‘Rather Clear’ in regards to the same question. We might care however if one evaluator chooses question readability as ‘Clear’ while another evaluator chooses question readability for the same question meaning it is recorded as ‘Incomprehensible’. In weighted Kappa, we assigned a score of 1 when both of the evaluators agree while a score of 0.5 is assigned when one evaluator chooses the question readability of a question as ‘Clear’ while the other evaluator chooses it as ‘Rather Clear’. We used a similar sort of criteria during distractors readability, usefulness of semantic relation, question relevance and distractors relevance.

TABLE 3: EVALUATION RESULTS OF SURFACE-BASED AND DEPENDENCY-BASED MCQ SYSTEMS

	QR (1-3)	**DR (1-3)**	**USR (1-3)**	**QRelv (1-3)**	**DRelv (1-3)**	**QA (0-5)**	**DA (0-5)**	**MCQ Usability (1-4)**
Surface-based MCQs System								
Evaluator1	2.15	2.96	2.14	2.04	2.24	2.53	3.04	2.61
Evaluator2	1.74	2.29	1.88	1.66	2.10	1.95	3.28	2.11
Average	1.95	2.63	2.01	1.85	2.17	2.24	3.16	2.36
Dependency-based MCQs System								
Evaluator1	2.42	2.98	2.38	2.37	2.31	3.25	3.73	3.37
Evaluator2	2.25	2.15	2.46	2.23	2.06	3.27	3.15	2.79
Average	2.34	2.57	2.42	2.30	2.19	3.26	3.44	3.08

In questions and distractors acceptability, we assigned an agreement score of 1 when both evaluators agree completely while a score of 0.5 was assigned when both of the evaluators choose questions and distractors acceptability between ‘0’ and ‘2’. A score of 0.5 was also assigned when both of the evaluators choose questions and distractors acceptability between ‘3’ and ‘5’. In overall MCQ usability, we assigned a score of 1 when both of the evaluators agreed and a score of 0.5 was assigned when one of the evaluator assigned an MCQ as ‘Directly Usable’ while the other evaluators marked the same MCQ as ‘Needs Minor Revision’. An agreement score of 0.5 was assigned when an MCQ was assigned by one of the evaluator as ‘Needs Major Revision’ while the other evaluator marked the same MCQ as ‘Unusable’. Table 4 shows the results obtained using weighted Kappa.

The results in Table 4 show that the use of weighted Kappa has increased the agreement between the two evaluators from fair to moderate. The agreement between the two evaluators is not very high. Because of this we are not looking at average scores between the two evaluators but instead we analyse the scores assigned by each evaluator separately.

One of the main reasons for not having high agreement score between the two evaluators is that these MCQs are generated from a part of the GENIA EVENT corpus which is very different to an instructional text or teaching material. As mentioned earlier, the GENIA EVENT corpus consists of MEDLINE abstracts so due to that some automatically generated MCQs are ambiguous or lacks context. For example in an MCQ, one evaluator classified the question readability as ‘Clear’ and the same MCQ is classified as ‘Rather Clear’ by the other evaluator due to the lack of context. This can be explained from the following example:

TABLE 4: WEIGHTED KAPPA SCORE

Evaluation Criteria	**Kappa Score (Surface-based MCQ)**	**Kappa Score (Dependency-based MCQ)**
Question Readability	0.44	0.44
Distractors Readability	0.48	0.37
Usefulness of Semantic Relation	0.37	0.51
Question Relevance	0.43	0.42
Distractors Relevance	0.48	0.54
Question Acceptability	0.46	0.45
Distractors Acceptability	0.39	0.39
Overall MCQ usability	0.43	0.41

Sentence: *Conversely inhibition of NF-kappaB confers a tenfold increase in glucocorticoid mediated apoptosis establishing that NF-kappaB also functions as an antiapoptotic factor.*

The following question was automatically generated from the aforementioned sentence:

Which protein also functions as an antiapoptotic factor?

According to the feedback of one evaluator this question is ambiguous and needs more context as there are hundreds of apoptotic factors and so there is a possibility of more than one right answer for this question. Similarly NF-Kappa B protein refers to a family of several proteins rather than one protein only so context is also important in automatically generating good quality MCQs. Moreover, sometimes the GENIA named entity tagger’s inability to recognize the boundaries of a named entity also resulted in MCQ where the answer of a particular question is partially given in the question. This can be elaborated from the following example:

Sentence: *The B cell-specific nuclear factor OTF-2 positively regulates transcription of the human class II transplantation gene DRA.*

The following question was automatically generated from the aforementioned sentence:

Which protein OTF-2 positively regulates transcription of the human class II transplantation gene DRA?

According to the evaluator’s feedback the answer of the question is partially given in the question and the actual question should be:

Which protein positively regulates transcription of the human class II transplantation gene DRA?

But due to the GENIA tagger’s inability to recognize some named entity boundaries our system was unable to automatically generate the correct question.

In order to test the significance of the difference between two sets of (surface-based and dependency-based) MCQ systems we used the Chi-Square test, which being a non-parametric statistical test, is suitable as we cannot assume a normal distribution of evaluator scores. In carrying out the test, we compared two sets of scores assigned by one evaluator: the scores assigned to MCT items generated with the surface-based method and those assigned to MCT items generated with the dependency-based method. Table 5 shows the p-values of Chi-Square test obtained from using the evaluation scores provided by the two evaluators.

TABLE 5: P-Values OF CHI-SQUARE

Evaluation Criteria	**p-values of Chi-Square Test**	
	Evaluator 1	**Evaluator 2**
Question Readability	0.1912	**0.0011**
Distractors Readability	0.5496	0.4249
Usefulness of Semantic Relation	0.2737	**0.0002**
Question Relevance	0.0855	**0.0004**
Distractors Relevance	0.1244	0.7022
Question Acceptability	0.1449	**0.0028**
Distractors Acceptability	0.0715	0.4123
Overall MCQ Usability	**0.0026**	**0.0010**

In Table 5, where there is a statistical significant difference (at the level of $p < 0.05$), between surface-based and dependency-based MCQ systems, the number is shown in bold. Both evaluators agreed during the extrinsic evaluation that the dependency-based MCQ system is better than the surface-based MCQ system in terms of overall MCQ usability. This has been proved by the p-values of Chi-Square (Table 5). Indeed there is a statistical difference between surface-based and dependency-based MCQ systems in terms of overall MCQ usability. The MCQs generated by the dependency-based system are more usable than the MCQs generated by the surface-based system.

VII. CONCLUSIONS

In this paper, we have compared the design and implementation of two unsupervised semantic-based systems for MCQ. Both systems attempt to identify the most important semantic relations in a document without assigning explicit labels to them to maximize coverage by having a range of unrestricted to predefined types of relations.

One of the two covered systems is based on surface-based semantic relations while other utilizes dependency-based semantic relations. The surface-based MCQ system extracts semantic relations between named entities in a text via Information Extraction methodologies and automatically generate questions from extracted semantic relations while the dependency-based MCQ system extract semantic relations between named entities by employing a dependency-based tree model. Our findings indicate that the dependency-based MCQ system performs better than the surface-based MCQ system.

REFERENCES

[1] W.E. Becker and M. Watts, "Teaching methods in U.S. and undergraduate economics courses" Journal of Economics Education, 32(3), 2001, pp. 269 – 279.

[2] G. Isaacs, "Multiple choice testing: A guide to the writing of multiple choice tests and to their analysis" Campbell town: HERDSA, 1994.

[3] N. Afzal, "Automatic Generation of Multiple Choice Questions using Surface-based Semantic Relations". (To appear in) International Journal of Computational Linguistics (IJCL) 6 (3), 2015.

[4] N. Afzal and R. Mitkov, "Automatic Generation of Multiple Choice Questions using Dependency-based Semantic Relations". Soft Computing. Volume 18, Issue 7, 2014 pp. 1269-1281. DOI: 10.1007/s00500-013-1141-4

[5] N. Afzal and V. Pekar, "Unsupervised relation extraction for automatic generation of multiple-choice questions", In Proceedings of the Recent Advances in Natural Language Processing (RANLP-2009). Borovets, Bulgaria, 2009, pp. 1-5.

[6] J-D. Kim, T. Ohta and J. Tsujii, "Corpus Annotation for Mining Biomedical Events from Literature", BMC Bioinformatics, 2008.

[7] K. Sudo, S. Sekine and R. Grishman, "An Improved Extraction Pattern Representation Model for Automatic IE Pattern Acquisition". In Proc. of the 41st Annual Meeting of ACL-03, Sapporo, Japan, 2003, pp. 224–231.

[8] M. Stevenson and M. Greenwood, "Dependency Pattern Models for Information Extraction," Research on Language and Computation 2009.

[9] N. Afzal, R. Mitkov and A. Farzindar, "Unsupervised relation extraction using dependency trees for automatic generation of multiple-choice questions". In Proceedings of the C. Butz and P. Lingras (Eds.): Canadian Artificial Intelligence, LNAI 6657. Newfoundland and Labrador, Canada: Springer, Heidelberg, 2011, pp. 32-43.

[10] J. Cohen, "A coefficient of agreement for nominal scales", Educational and Psychological Measurement, 20(1), 1960, pp. 37-46

[11] J. Cohen, "Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit," Psychological Bulletin, 70(4), 1968, pp. 213-220.