Automatic Generation of Multiple Choice Questions using Surface-based Semantic Relations

Naveed Afzal *afzal.naveed@mayo.edu*
Division of Biomedical Statistics and Informatics
Department of Health Sciences Research, Mayo Clinic
Rochester, MN, USA, 55901

Abstract

Multiple Choice Questions (MCQs) are a popular large-scale assessment tool. MCQs make it much easier for test-takers to take tests and for examiners to interpret their results; however, they are very expensive to compile manually, and they often need to be produced on a large scale and within short iterative cycles. We examine the problem of automated MCQ generation with the help of unsupervised Relation Extraction, a technique used in a number of related Natural Language Processing problems. Unsupervised Relation Extraction aims to identify the most important named entities and terminology in a document and then recognize semantic relations between them, without any prior knowledge as to the semantic types of the relations or their specific linguistic realization. We investigated a number of relation extraction patterns and tested a number of assumptions about linguistic expression of semantic relations between named entities. Our findings indicate that an optimized configuration of our MCQ generation system is capable of achieving high precision rates, which are much more important than recall in the automatic generation of MCQs. Its enhancement with linguistic knowledge further helps to produce significantly better patterns. We furthermore carried out a user-centric evaluation of the system, where subject domain experts from biomedical domain evaluated automatically generated MCQ items in terms of readability, usefulness of semantic relations, relevance, acceptability of questions and distractors and overall MCQ usability. The results of this evaluation make it possible for us to draw conclusions about the utility of the approach in practical e-Learning applications.

Keywords: E-Learning, Automatic Assessment, Educational Assessment, Natural Language Processing, Information Extraction, Unsupervised Relation Extraction, Multiple Choice Questions Generation, Biomedical Domain.

1. INTRODUCTION

Multiple Choice Questions (MCQs), also known as multiple-choice tests are a popular form of objective assessment in which a user selects one answer from a set of alternatives (distractors) for a given question. MCQs are straightforward to conduct and instantaneously provide an effective measure of the test-takers' performance and feedback test results to the learner. In many disciplines instructors use MCQs as a preferred assessment tool and it is estimated that between 45% and 67% of student assessments utilize MCQs [7].

In this paper, we present a new approach to automatic MCQs generation, where we first identify important concepts, as well as the relationships between them in the input texts. In order to achieve this, we study unsupervised Information Extraction methods with the purpose of discovering the most significant concepts and relations in the domain text, without any prior knowledge of their types or their exemplar instances (seeds). Information Extraction (IE) is an important problem in many information access applications. The goal is to identify instances of specific semantic relations between named entities of interest in the text. Named Entities (NEs) are generally noun phrases in the unstructured text e.g. names of persons, posts, locations and organizations, while relationships between two or more NEs are described in a pre-defined way

e.g. “interact with” is a relationship between two biological objects (proteins). We will employ this approach for the automatic generation of MCQ task, where it will be used to find relations and NEs in educational texts that are important for testing students’ familiarity with key facts contained in the texts. In order to achieve this, we need an IE method that has a high precision and at the same time works with unrestricted semantic types of relations (i.e. without reliance on seeds). Recall is of secondary importance to precision.

The use of unsupervised IE for MCQ generation offers a number of important advantages. First, because the approach finds significant semantic relations between concepts, rather than just individual concepts, it is able to capture a wider range of important facts contained in instructional texts and does so with greater accuracy, eventually achieving greater quality of MCQs. Second, in contrast to approaches that make use of manually encoded extraction rules, seed patterns or annotated examples, our approach has a potentially unrestricted coverage, as it does not target any pre-defined types of semantic relations. Third, our unsupervised approach to MCQ generation makes it suitable to be applied in situations where manually annotated text is unavailable or is very expensive to create, which is a common scenario in many e-Learning applications.

To validate this approach we employed two modes of evaluation. In the intrinsic evaluation, we examined the ability of the method to extract the most relevant semantic relations from a text by comparing automatically extracted relations with a gold standard – manually annotated relations contained in a publicly available domain corpus. In the extrinsic evaluation, domain experts were asked to judge the quality of the final MCQ items that our system generated in terms of readability, relevance, and overall usability of questions and distractors. The results of the extrinsic evaluation make it possible for us to draw conclusions about the practical utility of the use of unsupervised IE methods for MCQ generation.

2. RELATED WORK

Even though NLP has made significant progress in recent years, NLP methods, and the area of automatic generation of MCQs in particular, have started being used in e-Learning applications only very recently. One of the most comprehensive study in this area was published by [31, 32], who presented a computer-aided system for the automatic generation of multiple choice question items. Their system mainly consists of three parts: term extraction, stem generation and distractors selection. The system used a linguistic textbook in order to generate MCQs and found that 57% of automatically generated MCQs were judged worthy of keeping as test items, of which 94% required some level of post-editing. The main disadvantage of this system is its reliance on the syntactic structure of sentences to produce MCQs that produces questions from sentences, which have SVO, or SV structure. Moreover, the identification of key terms in a sentence is also an issue as identification of irrelevant concepts (key terms) results in unusable stem generation. Karamanis et al. [25] conducted a pilot study to use the [32] system in a medical domain and their results revealed that some questions were simply too vague or too basic to be employed as MCQ in a medical domain. They concluded that further research is needed regarding question quality and usability criteria. Skalban [38] presented a detailed analysis of the [32] system and highlighted the shortcomings it faced. Her work identified critical errors in the system including key term error, generic item error and subordinate source clause error. Her work also revealed that key term error created the most unusable MCQs.

Brown et al. [8] used an approach that evaluated the knowledge of students by automatically generating test items for vocabulary assessment. Their system produced six different types of questions for vocabulary assessment by making use of WordNet. The six different types of questions include definition, synonym, antonym, hypernym, hyponym and cloze questions. The cloze question requires the use of a target word in a specific context. The approach presented in this paper relied heavily on WordNet and is unable to produce any questions for words that are not present in WordNet.

Chen et al. [10] presented an approach for the semi-automatic generation of grammar test items by employing NLP techniques. Their approach was based on manually designed patterns, which were further used to find authentic sentences from the Web and were then transformed into grammatical test items. Distractors were also obtained from the Web with some modifications in manually designed patterns e.g. changing part of speech, adding, deleting, replacing or reordering of words. The experimental results of this approach revealed that 77% of the generated MCQs were regarded as worthy (i.e. can be used directly or needed only minor revision). The disadvantage of this approach is that it requires a considerable amount of effort and knowledge to manually design patterns that can later be employed by the system to generate grammatical test items.

A semi-automatic system to assist teachers in order to produce cloze tests based on online news articles was presented by [22]. In cloze tests, questions are generated by removing one or more words from a passage and the test takers have to fill in the missing words. The system focuses on multiple-choice fill-in-the-blank tests and generates two types of distractors: vocabulary distractors and grammar distractors. User evaluation reveals that 80% of the generated items were deemed suitable.

Papasalouros et al. [34] presented a system for automatic generation of MCQs, which makes use of domain ontologies. Ontology defines a common vocabulary for agents (including people) who need to share information in a domain. Ontologies include machine-interpretable definitions of basic concepts in the domain and relations among them. In order to generate MCQs, this paper utilized three different strategies: class-based strategies (based on hierarchies), property-based strategies (based on roles between individuals) and terminology-based strategies. The MCQs generated by this approach were evaluated in terms of quality, syntactic correctness and a number of questions were produced for different domain specific ontologies. The experimental results revealed that not all questions produced are syntactically correct and in order to overcome this problem more sophisticated Natural Language Generation (NLG) techniques are required.

Most of the previous approaches to automatically generating MCQs have been used for vocabulary and grammatical assessments of English. The main drawback of these approaches is that generated MCQs are very simple, easy to answer and mostly based on recalling facts, so the main challenge is to automatically generate MCQs which will allow the examiner/instructor to evaluate test takers not only on the superficial memorization of facts but also on higher levels of cognition.

3. OUR APPROACH

This paper tries to solve the problem by extracting semantic rather than surface-level or syntactic relations, between key concepts in a text via IE methodologies and then generating questions from such semantic relations. The research work presented in this paper is the extension of the work done by [3]. We carried out our experiments using a biomedical domain, as it is more complex as compared to other domains [5].

There is a large body of research dedicated to the problem of extracting relations from texts of various domains. Most previous work focused on supervised methods and tried to both extract relations and assign labels describing their semantic types. As a rule, these approaches required a manually annotated corpus, which is very laborious and time-consuming to produce. Semi-supervised approaches rely on seed patterns and/or examples of specific types of relations [4, 6, 39 and 9]. Unsupervised approaches do not rely on any hand-labelled training data or seed examples. Most of the unsupervised learning algorithms use clustering. Examples of unsupervised learning algorithms applied in IE include [20, 36, 37, 16 and 14]. In the biomedical domain, most approaches were supervised and relied on regular expressions to learn patterns [13] while semi-supervised approaches exploited pre-defined seed patterns and cue words [23, 29]. Relation Extraction in the biomedical domain has been addressed primarily with supervised

approaches or those based on manually written extraction rules, which are rather inadequate in scenarios where relation types of interest are not known in advance.

Our assumption for Relation Extraction is that it is between NEs stated in the same sentence and that presence or absence of relation is independent of the text prior to or succeeding the sentence. Our system (Figure 1) consists of three main components: IE, question generation and distractors generation. In an IE component, unannotated text is first processed by NER (Section 4) and then candidate patterns are extracted from the text (Section 5). The candidate patterns are ranked according to their domain relevance and we then intrinsically evaluate the candidate patterns in terms of precision, recall and F-score (Section 8). In automatic question generation components (Section 9), these extracted semantic relations are automatically transformed into questions by employing a certain set of rules while in automatic distractors generation (Section 10), distractors are generated using a distributional similarity measure.

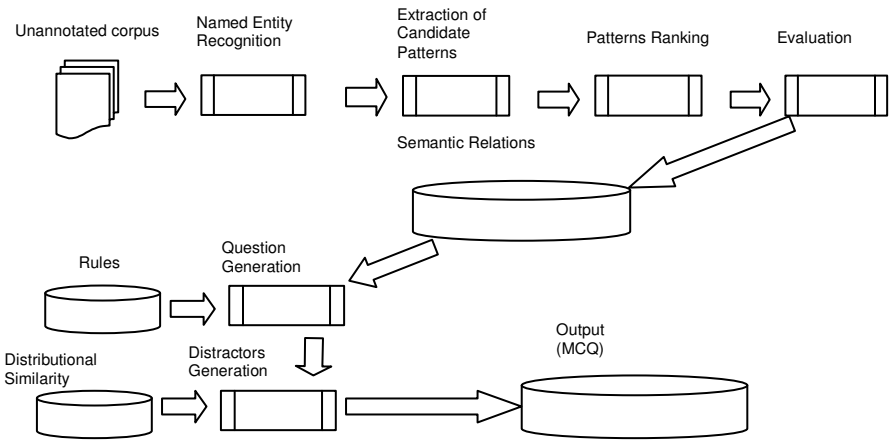

FIGURE 1: System Architecture.

4. NAMED ENTITY RECOGNITION (NER)

NER is an integral part of any IE system as it identifies NEs present in a text. Biomedical NER is generally considered more difficult than other domains such as newswire text. Biomedical NEs are expressed in various linguistic forms such as abbreviations, plurals, compound, coordination, cascades, acronyms and apposition. Sentences in such texts are syntactically complex as the subsequent Relation Extraction phase depends upon the correct identification of the NEs and correct analysis of linguistic constructions expressing relations between them [47].

There are huge numbers of NEs in the biomedical domain and new ones are constantly added [45] which means that neither dictionaries nor the training data approach will be sufficiently comprehensive for NER. Moreover, Grover et al. [19] presented a report, investigating the suitability of current NLP resources for syntactic and semantic analysis for the biomedical domain.

The GENIA NER [44, 45] is a specific tool designed for biomedical texts; the NE tagger is designed to recognize mainly the following NEs: protein, DNA, RNA, cell_type and cell_line. Table 1 shows the performance of GENIA NER[1]. The GENIA NER also provide us Part-of-Speech (PoS) information along with base form of a word.

[1] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

Entity Type	Precision	Recall	F-score
Protein	65.82	81.41	72.79
DNA	65.64	66.76	66.20
RNA	60.45	68.64	64.29
Cell Type	56.12	59.60	57.81
Cell Line	78.51	70.54	74.31
Overall	67.45	75.78	71.37

TABLE 1: GENIA NER Performance.

5. EXTRACTION OF CANDIDATE PATTERNS

Our general approach to the discovery of interesting extraction patterns consists of two main stages: (i) the construction of potential patterns from an unannotated domain corpus and (ii) their relevance ranking.

Once the training corpus has been tagged with the GENIA tagger, the process of pattern building takes place. Its goal is to identify which NEs are likely to be semantically related to each other.

The procedure for constructing candidate patterns is based on the idea that important semantic relations are expressed with the help of recurrent linguistic constructions, and these constructions can be recognized by examining sequences of content words (nouns, verbs, adjectives and adverbs) appearing between NEs along with prepositions. Prepositions are used to express relations of place, direction, time or possessions. Semantic patterns are widely used in the area of IE. As in IE, we are interested in extraction of semantic classes of objects (NEs), relationships among these NEs and relations in which these entities participate. To find such constructions, we impose a limit on the number of content words intervening between the two NEs. We experimented with different thresholds and finally settled on a minimum of one content word and a maximum of three content words to be extracted between two NEs. The reason for introducing this condition is that if there are no content words between two NEs then, although some relation might exist between them, it is likely to be a very abstract grammatical relation. On the other hand, if there are too many content words intervening between two NEs, then it is likely they are not related at all. We build patterns using this approach and store each pattern along with its frequency in a database. In extracted patterns, lexical items are represented in lowercase while semantic classes are capitalized. For example in the pattern “PROTEIN encode PROTEIN”, here encode is a lexical item while PROTEIN is a semantic class.

We carried out our experiments using three different pattern types:

- Untagged word patterns
- PoS-tagged word patterns
- Verb-centered patterns

Untagged word patterns consist of NEs and their intervening content words along with prepositions. Some examples of untagged word patterns from the GENIA corpus along with their frequencies are shown in Table 2.

Patterns	Frequency
DNA contain DNA	22
PROTEIN expression in CELL_TYPE	14
PROTEIN induce PROTEIN	13
PROTEIN bind to DNA	12

TABLE 2: Untagged word patterns along with their frequencies.

The motive for choosing these different types of surface patterns is that verbs typically express semantic relations between nouns that are used as their arguments. PoS-tagged word patterns contain the PoS of each content word along with prepositions also as shown in Table 3.

Patterns	Frequency
DNA contain_v DNA	22
PROTEIN activate_v PROTEIN	19
PROTEIN express_v PROTEIN	18
PROTEIN expression_n in_i CELL_TYPE	16

TABLE 3: PoS-tagged word patterns along with their frequencies.

In Verb-centered patterns, the presence of a verb is compulsory in each pattern. Table 4 shows a few examples of verb-centered patterns. We require the presence of a verb in the verb-based patterns, as verbs are the main predicative class of words, expressing specific semantic relations between two named entities.

Patterns	Frequency
DNA contain_v DNA	22
DNA activate_v PROTEIN	19
CELL_TYPE express_v PROTEIN	18
PROTEIN encode_v PROTEIN	13

TABLE 4: Verb-centered word patterns along with their frequencies

Moreover, in the pattern building phase, the patterns containing the passive form of the verb like:

PROTEIN be_v express_v CELL_TYPE

are converted into the active voice form of the verb like:

CELL_TYPE express_v PROTEIN

Because such patterns were taken to express a similar semantic relation between NEs, passive to active conversion was carried out in order to relieve the problem of data sparseness: it helped to increase the frequency of unique patterns and reduce the total number of patterns. For the same reason, negation expressions (not, does not, etc.) were also removed from the patterns as they express a semantic relation between NEs equivalent to one expressed in patterns where a negation particle is absent. In addition, patterns containing only stop-words were also filtered out.

6. PATTERN RANKING

After candidate patterns have been constructed, the next step is to rank the patterns based on their significance in the domain corpus. The ranking methods we use require a general corpus that serves as a source of examples of pattern use in domain-independent texts. To extract candidates from the general corpus, we treated every noun as a potential NE holder and the candidate construction procedure described above was applied to find potential patterns in the general corpus. In order to score candidate patterns for domain-relevance, we measure the strength of association of a pattern with the domain corpus as opposed to the general corpus. The patterns are scored using the following methods for measuring the association between a pattern and the domain corpus: Information Gain (IG), Information Gain Ratio (IGR), Mutual Information (MI), Normalized Mutual Information (NMI), Log-likelihood (LL) and Chi-Square (CHI). These association measures were included in the study as they have different theoretical principles behind them: IG, IGR, MI and NMI are information-theoretic concepts while LL and CHI are statistical tests of association. Some of these ranking methods have been used in

classification of words according to their meanings [35] but to our knowledge, this approach is the first one to explore these ranking methods to rank IE patterns.

In addition to these six measures, we introduce a meta-ranking method that combines the scores produced by several individual association measures (apart from MI), in order to leverage agreement between different association measures and downplay idiosyncrasies of individual ones. Apart from the aforementioned pattern ranking methods, we also used most frequently used pattern ranking method: tf-idf in our experiments. All these pattern-ranking methods are discussed in details in [2]. We used a method based on setting a threshold on the association score below which the candidate patterns are discarded (henceforth, score-thresholding measure).

7. INTRINSIC EVALUATION

We used intrinsic evaluation to evaluate the quality of IE component of our MCQ system. We used the GENIA corpus as the domain corpus while the British National Corpus (BNC) was used as a general corpus. The GENIA corpus consists of 2,000 abstracts extracted from the MEDLINE containing 18,477 sentences. In the evaluation phase, GENIA EVENT Annotation corpus is used [27] and it is quite similar to GENIA corpus consisting of MEDLINE abstracts. It consists of 9,372 sentences. We measured precision, recall and F-score and to test the statistical significance of differences in the results of different methods and configurations, we used a paired t-test, having randomly divided the evaluation corpus into 20 subsets of equal size; each subset containing 461 sentences on average. We collected precision, recall and F-score for each of these subsets and then using paired t-test, we found statistical significance between different surface pattern types and between different ranking methods using score-thresholding measure.

7.1 Results

The numbers of untagged word patterns extracted from each corpus are GENIA 10093, BNC 991004 and GENIA EVENT 4854. Figure 2 shows results of the score-thresholding measure for untagged word patterns. Here we are considering only those threshold values, which enable us to attain high precision scores (see Table 9 at the end of the paper for complete results in terms of precision, recall and F-score).

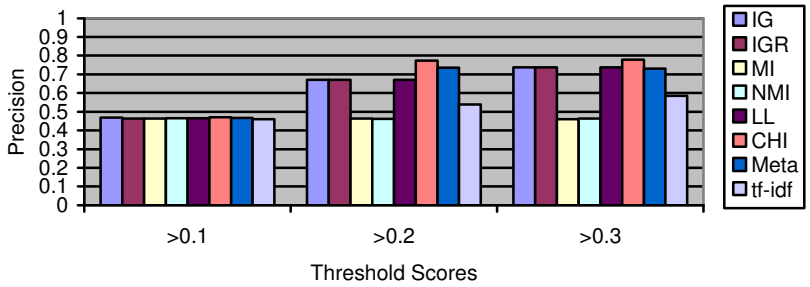

FIGURE 2: Score-thresholding results for untagged word patterns.

We carried out a similar set of experiments using PoS-tagged word patterns. The numbers of PoS-tagged word patterns extracted from each corpus are GENIA 9237, BNC 840057 and GENIA EVENT 4446. Figure 3, shows results of the score-thresholding measure for PoS-tagged word patterns (see Table 10 at the end of the paper for complete results in terms of precision, recall and F-score).

The numbers of verb-centred word patterns extracted from each corpus are GENIA 6645, BNC 598948 and GENIA EVENT 3271. Figure 4 shows results of score-thresholding measures for verb-centred word patterns respectively (see Table 11 at the end of the paper for complete results in terms of precision, recall and F-score).

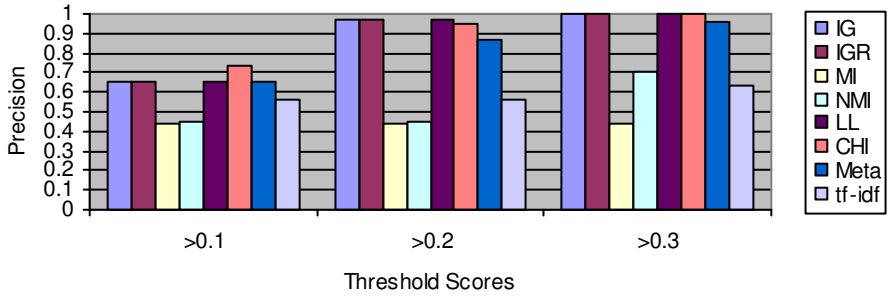

FIGURE 3: Score-thresholding results for PoS-tagged word patterns.

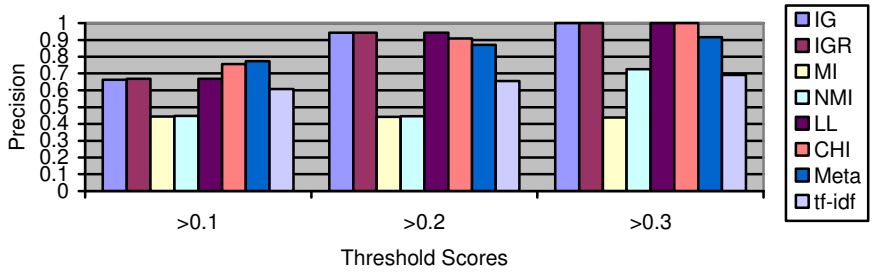

FIGURE 4: Score-thresholding results for verb-centered word patterns.

In all experiments, we found that IG, IGR and LL achieved quite similar results while CHI Meta and NMI are the best performing ranking methods while MI is the worst in terms of precision scores. The tf-idf ranking method performed better than MI on all occasions but it does not really apply to our work as our corpus consists of those documents that describe relevant domain information only as compared to the corpus used by [41]. Even though CHI and Meta ranking methods attained higher precision scores, recall scores are very low. One reason for having a low recall is due to the small size of GENIA corpus. This can be remedied by using a large corpus because a large corpus will produce a much greater number of patterns and increase the recall. In score-thresholding, it is possible to optimize for high precision (up to 100%), though recall and F-score is generally quite low. MCQ applications rely on the production of good questions rather than the production of all possible questions, so high precision plays a vital role in such applications. We explored three surface pattern types and found that verb-centered and PoS-tagged pattern types are better than untagged word patterns. Figure 5 shows the precision scores for the best performing ranking methods (CHI and NMI) in the score-thresholding measure.

Verb-centered patterns work well, because verbs are known to express semantic relations between NEs to the verb; PoS-tagged word patterns add important semantic information into the pattern and possibly disambiguate words appearing in the pattern. In order to find out whether the differences between the three patterns types are statistically significant, we carried out the paired t-test. We found that there is no statistically significant difference between PoS-tagged word patterns and verb-centred patterns. Apart from IG, IGR and LL there is a statistically significant difference between all the ranking methods of untagged word patterns and PoS-tagged word patterns, untagged word patterns and verb-centered patterns respectively. In terms of F-score, verb-centered word patterns achieved a higher F-score as compared to other pattern types.

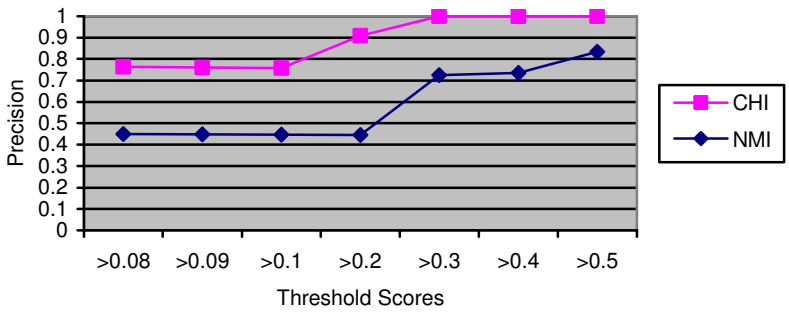

FIGURE 5: Best performing ranking methods.

8. QUESTION GENERATION

This component of our system transforms extracted semantic relations into questions automatically by employing a certain set of rules. The questions automatically generated by our approach are more accurate as it automatically generates questions from important concepts present in the given domain by relying on the semantic relations. Our approach for automatic generation of questions depends upon accurate output of the NE tagger and the parser.

Patterns	Questions Examples
SC1 verb SC2 DNA contain_v DNA	Which DNA contains cis elements? Which DNA is contained by inducible promoter?
SC1 verb preposition SC2 CELL_TYPE culture_v with_i PROTEIN	Which cell_type is cultured with IL-4?
SC1 verb adjective SC2 CELL_TYPE express_v several_j PROTEIN	Which cell_type expresses several low molecular weight transmembrane adaptor proteins?
SC1 verb verb SC2 CELL_TYPE exhibit_v enhance_v PROTEIN	Which cell_type exhibits enhance IL-2?
SC1 adverb verb SC2 PROTEIN efficiently_a activate_v DNA	Which DNA is efficiently activated by Oct2?
SC1 verb preposition SC2 PROTEIN bind_v to_t DNA	Which protein binds to ribosomal protein gene promoters?
SC1 verb noun preposition SC2 CELL_LINE confirm_v importance_n of_i PROTEIN	Which cell_line confirms importance of NF-kappa B?
SC1 verb preposition adjective SC2 CELL_TYPE derive_v from_i adherent_j CELL_TYPE	Which cell_type derives from adherent PBMC?
SC1 verb preposition noun preposition SC2 CELL_TYPE result_v in_i activation_n of_i PROTEIN	Which cell_type results in activation of TNF-alpha?
SC1 adverb verb noun preposition SC2 CELL_LINE specifically_a induce_v transcription_n from_i DNA	Which cell_line specifically induces transcription from interleukin-2 enhancer?

TABLE 6: Examples of extracted patterns along with automatically generated questions.

In order to automatically generate questions, we first assume that the user has supplied a set of documents on which students will be tested. We will refer to this set of documents as “evaluation corpus” (e.g. in this research, we used a small subset of GENIA EVENT Annotation corpus as an evaluation corpus). We select semantic patterns, attaining higher precision/ higher F-score at certain score thresholds using the score-thresholding measure. We extract surface-based semantic patterns from the evaluation corpus and try to match these patterns with the semantic patterns learned from the GENIA corpus and when a match is found; we extract the whole sentence from the evaluation corpus and then automatically transform the extracted pattern into a question. This process can be illustrated by the following example:

Pattern: DNA contain_v DNA

Step 1: Identify instantiations of a pattern in the evaluation corpus; this involves finding the template (in the above example, the verb 'contain') and the slot filler (two specifics DNA's in the above example). We then have the aforementioned pattern being matched in the evaluation corpus and the relevant sentence is extracted from it.

Thus, the gamma 3 ECS is an inducible promoter containing cis elements that critically mediate CD40L and IL-4-triggered transcriptional activation of the human C gamma 3 gene.

Step 2: The part of the extracted sentence that contains the template together with slot fillers is tagged by <QP> and </QP> tags as shown below:

Thus, the <DNA> gamma 3 ECS </DNA> is an <QP> <DNA> inducible promoter </DNA> containing <DNA> cis elements </DNA> </QP> that critically mediate <protein> CD40L </protein> and IL-4-triggered transcriptional activation of the <DNA> human C gamma 3 gene </DNA>.

Step 3: In this step, we extract semantic tags and actual names from the extracted sentence by employing Machinese Syntax parser (Tapanainen and Järvinen, 1997). After parsing, the extracted semantic pattern is transformed into the following question:

Which DNA contains cis elements?

In order to automatically generate questions from the aforementioned extracted semantic patterns, we developed a certain set of rules. Table 6 shows those rules, which are based on semantic classes (NEs), and part-of-speech (PoS) information present in a pattern. We employ verb-centered patterns for question generation as the presence of a verb between two NE generally represent a meaningful semantic relation between them. During the automatic generation of questions, we also employed a list of irregular verbs in order to produce past participle form of irregular verbs. Table 5 contains some of the examples of patterns and their respective automatically generated questions. Here SC represents the Semantic Class (e.g. NEs). All these rules are domain-independent and only rely on the presence of semantic classes and PoS information between these semantic classes.

We are able to automatically generate only one type of questions (Which questions) regarding named entities present in a semantic relation. Our approach is not capable of automatically generating different types of questions (e.g. Why, How and What questions), and in order to do that one has to look at various NLG techniques. This would be beyond the scope of this paper.

9. DISTRACTORS GENERATION

Our approach relies on a distributional similarity measure to automatically generate distractors. This component of our system is discussed in detail in [2].

10. EXTRINSIC EVALUATION

The real application users have a vital role to play in the extrinsic or user-centered evaluation process. In this section, we will evaluate the MCQ system as a whole in a user-centered fashion. The evaluation used in our approach is mainly concerned with the adequate and appropriate generation of MCQs as well as the amount of human intervention required. In other words, we want to evaluate our system in terms of its robustness and efficiency.

The extrinsic evaluations of our system used the same criteria that was used by [2]. We evaluated MCQ's in terms of their reliability, usefulness of semantic relation, relevance, acceptability and overall MCQ usability (See [2] for further details). We found that NMI and CHI are the best performing ranking methods (Section 8). CHI achieved very high precision scores but

recall scores are very low while in NMI recall scores are relatively better than CHI. Due to this reason, during the extrinsic evaluation phase we employ NMI and selected a score-thresholding (score> 0.01) for NMI as it gives a maximum F-score of 54%. We generated 80 MCQs using a small subset of GENIA EVENT Annotation corpus using NMI score > 0.01.

In the extrinsic evaluation, two biomedical experts (both post-doc) were asked to evaluate MCQs according to the aforementioned criteria. Figure 6 shows the screenshot of the interface used during the extrinsic evaluation of automatically generated MCQs. Both evaluators were vastly experienced, one evaluator's main area of research focuses on isolation, characterizing and growing stem cells from Keloid and Dupuytren's disease and is currently working at Plastics and Reconstructive Surgery Research while the other biomedical expert is a bio-curator with a PhD in molecular biology and is currently working for the Hugo Gene Nomenclature Committee (HGNC). Both evaluators were asked to give a scoring value for the readability of questions and distractors from 1 (incomprehensible) to 3 (clear) usefulness of semantic relation from 1 (incomprehensible) to 3 (clear), question and distractors relevance from 1 (not relevant) to 3 (very relevant), question and distractors acceptability from 0 (unacceptable) to 5 (acceptable) and overall MCQ usability from 1 (unusable) to 4 (directly usable).

Extrinsic evaluation results of overall MCQ usability show that 35% of MCQ items were considered directly usable, 30% needed minor revisions and 14% needed major revisions while 21% MCQ items were deemed unusable by the evaluators. Table 7 shows the results obtained for a surface-based MCQ system 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.

	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)**
Evaluator 1	2.15	2.96	2.14	2.04	2.24	2.53	3.04	2.61
Evaluator 2	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

TABLE 7: Extrinsic evaluation results.

11. DISCUSSION

We used weighted Kappa [12] to measure the agreement across major sub-categories in which there is a meaningful difference. K = 1 when there is a complete agreement among the evaluators while K = 0 when there is no agreement. For example, in question readability we had 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 if one evaluator chooses question readability as 'Clear' while another evaluator chooses question readability for the same question meaning as 'Incomprehensible'. In weighted Kappa, we assigned a score of 1 when both of the evaluators agree; a score of 0.5 is assigned when one evaluator chooses the question readability of a question as 'Clear' while the other evaluator choose it as 'Rather Clear'. We used a similar sort of criteria during distractors readability, usefulness of semantic relation, question relevance and distractors relevance. In question 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 question and distractors acceptability between '0' and '2'. A score of 0.5 was also assigned when both of the evaluators choose question 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 evaluators 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 one of the evaluator as 'Needs Major

Revision' assigned an MCQ while the other evaluator marked the same MCQ as 'Unusable'. Table 8 show the Kappa score:

Evaluation Criteria	Kappa Score
Question Readability	0.44
Distractors Readability	0.48
Usefulness of Semantic Relation	0.37
Question Relevance	0.43
Distractors Relevance	0.48
Question Acceptability	0.46
Distractors Acceptability	0.39
Overall MCQ usability	0.43

TABLE 8: Kappa Score.

Due to various sub-categories, we are only able to attain a moderate agreement between the two evaluators. One of the main reasons for not having high agreement scores 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. The GENIA EVENT corpus consists of MEDLINE abstracts so due to that some automatically generated MCQs are ambiguous or lacks context.

In 2010, First Question Generation Shared Task Evaluation Challenge (QGSTEC) also used a similar sort of evaluation criteria where they evaluated the automatically generated questions in terms of relevance, question type, syntactic correctness and fluency, ambiguity and variety.

Moreover, according to our knowledge, this is the first system that have used IE in the context of automatic generation of MCQs. Due to which there is no direct comparison possible with other approaches. We have compared this approach with our dependency based approach [2] that also used IE and results of that comparison are discussed in detail in [1].

12. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented an approach for automatic generation of MCQs based on unsupervised surface-based semantic relations. Our approach consisted of three main components: in the first component, we used IE methodologies to extract semantic relations and in the second component, we automatically generated questions using these semantic relations. In the third component, distractors were automatically generated using a distributional similarity measure.

We examined three different types of surface patterns each implementing different assumptions about linguistic expression of semantic relations between named entities. We explored different information-theoretic and statistical measures to rank candidate semantic patterns by domain relevance as well as meta-ranking (a method that combines multiple pattern-ranking methods). The experimental results revealed that the CHI and NMI ranking methods obtained higher precision than the other ranking methods.

These extracted semantic relations allowed us to automatically generate better quality questions by focusing on the important concepts present in a given text as questions are automatically generated using these semantic relations. We used a certain set of rules based on named entities and part-of-speech information to automatically generate questions from these semantic patterns. The plausible distractors were automatically generated by using a distributional similarity measure. Distributional similarity is known to adequately model the semantic similarity between lexical expressions and it is used quite frequently in many NLP applications. Distributional similarity measures are corpus-driven and have a broad coverage compared with the thesaurus-based methods that have a limited coverage.

We extrinsically evaluated the whole MCQ system in terms of question and distractor readability, usefulness of semantic relation, relevance, acceptability of question and distractor and overall MCQ usability. Two domain experts evaluated the system according to the aforementioned criteria and the results revealed that our approach is able to automatically generate good quality MCQs. In the future, we are planning to carry out extrinsic evaluation using item response theory [18] as conducted by [32] and compare our results with their approach on the same dataset.

13. REFERENCES

[1] N. Afzal and A. Bawakid, "Comparison between Surface-based and Dependency-based Relation Extraction Approaches for Automatic Generation of Multiple-Choice Questions". IJMSE, Volume 6, Issue 8, 2015.

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

[3] N. Afzal and V.Pekar, "Unsupervised Relation Extraction for Automatic Generation of Multiple-Choice Questions". In Proc. of RANLP'2009 14-16 September 2009. Borovets, Bulgaria.

[4] E. Agichtein and L. Gravano, "Snowball: Extracting relations from large plain text collections". In Proc. of the 5th ACM International Conference on Digital Libraries, 2000.

[5] S. Ananiadou and J. McNaught eds. "Text Mining for Biology and Biomedicine", Artech House, 2006.

[6] R.K. Ando and T. Zhang, "A high-performance semi-supervised learning method for text chunking". In Proc. of the 43rd Annual Meeting on Association for Computational Linguistics (ACL-05). Association for Computational Linguistics, pp. 1-9, 2005.

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

[8] J. Brown, G. Frishkoff and M. Eskenazi, "Automatic question generation for vocabulary assessment". In Proc. of HLT/EMNLP. Vancouver, B.C. 2005.

[9] R. Bunescu and R. Mooney, "Learning to extract relations from the web using minimal supervision". In Proc. of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07). Prague, Czech Republic, 2007.

[10] C.-Y. Chen, H.-C. Liou and J.S. Chang, "FAST- An automatic generation system for grammar tests". In Proc. of COLING/ACL Interactive Presentation Sessions, Sydney, Australia, 2006.

[11] A.M. Cohen and W.R. Hersh, "A survey of current work in biomedical text mining". Briefings in Bioinformatics, pp. 57-71, 2005.

[12] J. Cohen, "Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit". Psychological Bulletin, 1968.

[13] D. P. Corney, D. Jones, B. Buxton and W. Langdon, "BioRAT: Extracting biological information from full-length papers". Bioinformatics, pp. 3206-3213, 2004.

[14] K. Eichler, H. Hemsen and G. Neumann, "Unsupervised relation extraction from web documents". In Proc. of the 6th International Language Resources and Evaluation (LREC-08). Marrakech, Morocco, 2008.

[15] G. Erkan, A. Ozgur and D. R. Radev, "Semi-supervised classification for extracting protein interaction sentences using dependency parsing". In Proc. of CoNLL-EMNLP, 2007.

[16] O. Etzioni, M. Banko, S. Soderland and D. S. Weld, "Open information extraction from the web". Communications of the ACM, 51(12), pp.68-74, 2008.

[17] M. Greenwood, M. Stevenson, Y. Guo, H. Harkema and A. Roberts, "Automatically acquiring a linguistically motivated genic interaction extraction system". In Proc. of the 4th Learning Language in Logic Workshop, Bonn, Germany, 2005.

[18] N. Gronlund, "Constructing Achievement Tests". New York, USA: Prentice Hall, 1982.

[19] C. Grover, A. Lascarides and M. Lapata, "A comparison of parsing technologies for the biomedical domain". Natural Language Engineering 11 (1), pp. 27 -65, 2005.

[20] T. Hasegawa, S. Sekine and R. Grishman, "Discovering relations among named entities from large corpora". In Proc. of ACL'04, 2004.

[21] A. Hoshino and H. Nakagawa, "A real-time multiple-choice question generation for language testing – A preliminary study". In Proc. of the 43rd ACL'05 2nd Workshop on Building Educational Applications Using Natural Language Processing, pp.17-20, 2005.

[22] A. Hoshino and H. Nakagawa "Assisting cloze test making with a web application". In Proc. of Society for Information Technology and Teacher Education International Conference. Chesapeake, VA, 2007.

[23] M. Huang, X. Zhu, G. D. Payan, K. Qu and M. Li, "Discovering patterns to extract protein-protein interactions from full biomedical texts". Bioinformatics, pp. 3604-3612, 2004.

[24] D. Jurafsky and J. H. Martin, "Speech and Language Processing". Second Edition. Prentice Hall, 2008.

[25] N. Karamanis, L. A. Ha and R. Mitkov, "Generating multiple-choice test items from medical text: A pilot study". In Proc. of the 4th International Natural Language Generation Conference, (July), pp.111-113, 2006.

[26] S. Katrenko and P. Adriaans, "Learning relations from biomedical corpora using dependency trees". In Proc. of the 1st International Workshop on Knowledge Discovery and Emergent Complexity in Bioinformatics, Ghent, pp. 61–80, 2006.

[27] J-D. Kim, T. Ohta and J. Tsujii, "Corpus annotation for mining biomedical events from literature", BMC Bioinformatics, 2008.

[28] D. Lin and P. Pantel, "Concept discovery from text". In Proc. of Conference on CL'02. pp. 577-583. Taipei, Taiwan, 2002.

[29] C. Manning and H. Schütze, "Foundations of Statistical Natural Language Processing". The MIT Press, Cambridge, U.S. 1999.

[30] E. P. Martin, E. Bremer, G. Guerin, M-C. DeSesa and O. Jouve, "Analysis of protein/protein interactions through biomedical literature: Text mining of abstracts vs. Text mining of full text articles". Berlin: Springer-Verlag, pp. 96-108, 2004.

[31] R. Mitkov and L. A. An, "Computer-aided generation of multiple-choice tests". In Proc. of the HLT/NAACL 2003 Workshop on Building educational applications using Natural Language Processing, 17-22. Edmonton, Canada, 2003.

[32] R. Mitkov, L. A. Ha and N. Karamanis, “A computer-aided environment for generating multiple-choice test items”. Natural Language Engineering 12(2). Cambridge University Press, pp. 177-194, 2006.

[33] T. Ono, H. Hishigaki, A. Tanigami and T. Takagi, “Automated extraction of information on protein–protein interactions from the biological literature”. Bioinformatics, pp. 155-161, 2001.

[34] A. Papasalouros, K. Kanaris and K. Konstantinos, “Automatic generation of multiple choice questions from domain ontologies”. In Proc. of IADIS International Conference e-Learning, 2008.

[35] V. Pekar, M. Krkoska and S. Staab, “Feature weighting for co-occurrence-based classification of words”. In Proc. of the 20th International Conference on Computational Linguistics (COLING-04). Geneva, Switzerland, pp. 799-805, 2004.

[36] S. Sekine, “On-demand information extraction”. In Proc. of the COLING/ACL, 2006.

[37] Y. Shinyama, and S. Sekine, “Preemptive information extraction using unrestricted relation discovery”. In Proc. of the HLT Conference of the North American Chapter of the ACL. New York, pp. 304-311, 2006.

[38] Y. Skalban, “Improving the output of a multiple-choice test generator: Analysis and proposals”. University of Wolverhampton, 2009.

[39] M. Stevenson and M. Greenwood, “A semantic approach to IE pattern induction”. In Proc. of ACL’05, pages 379-386, 2005.

[40] M. Stevenson and M. Greenwood, “Dependency pattern models for information extraction”. Research on Language and Computation, 2009.

[41] 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, pp. 224–231, Sapporo, Japan, 2003.

[42] I. Szpektor, H. Tanev, I. Dagan and B. Coppola, “Scaling Web-based acquisition of entailment relations”. In Proc. of EMNLP-04, Barcelona, Spain, 2004.

[43] P. Tapanainen and T. Järvinen, “A non-projective dependency parser”. In Proc. of the 5th Conference on Applied Natural Language Processing, pages 64–74, Washington, 1997.

[44] Y. Tsuruoka, Y. Tateishi, J-D. Kim, T. Ohta, J. McNaught, S. Ananiadou and J.Tsujii, “Developing a robust PoS tagger for biomedical text”. Advances in Informatics – 10th Panhellenic Conference on Informatics, LNCS 3746, pp. 382-392, 2005.

[45] Y. Tsuruoka and J. Tsujii, “Bidirectional inference with the easiest-first strategy for tagging sequence data”. Proc. of HLT/EMNLP, pp. 467-474, 2005.’

[46] J. Wilbur, L. Smith and T. Tanabe, “BioCreative 2. Gene mention task. Proc. of the 2nd Bio-Creative Challenge Workshop pp. 7-16, 2007.

[47] G. Zhou, J. Su, D. Shen and C. Tan, “Recognizing name in biomedical texts: A machine learning approach”. Bioinformatics, pp. 1178-1190, 2004.

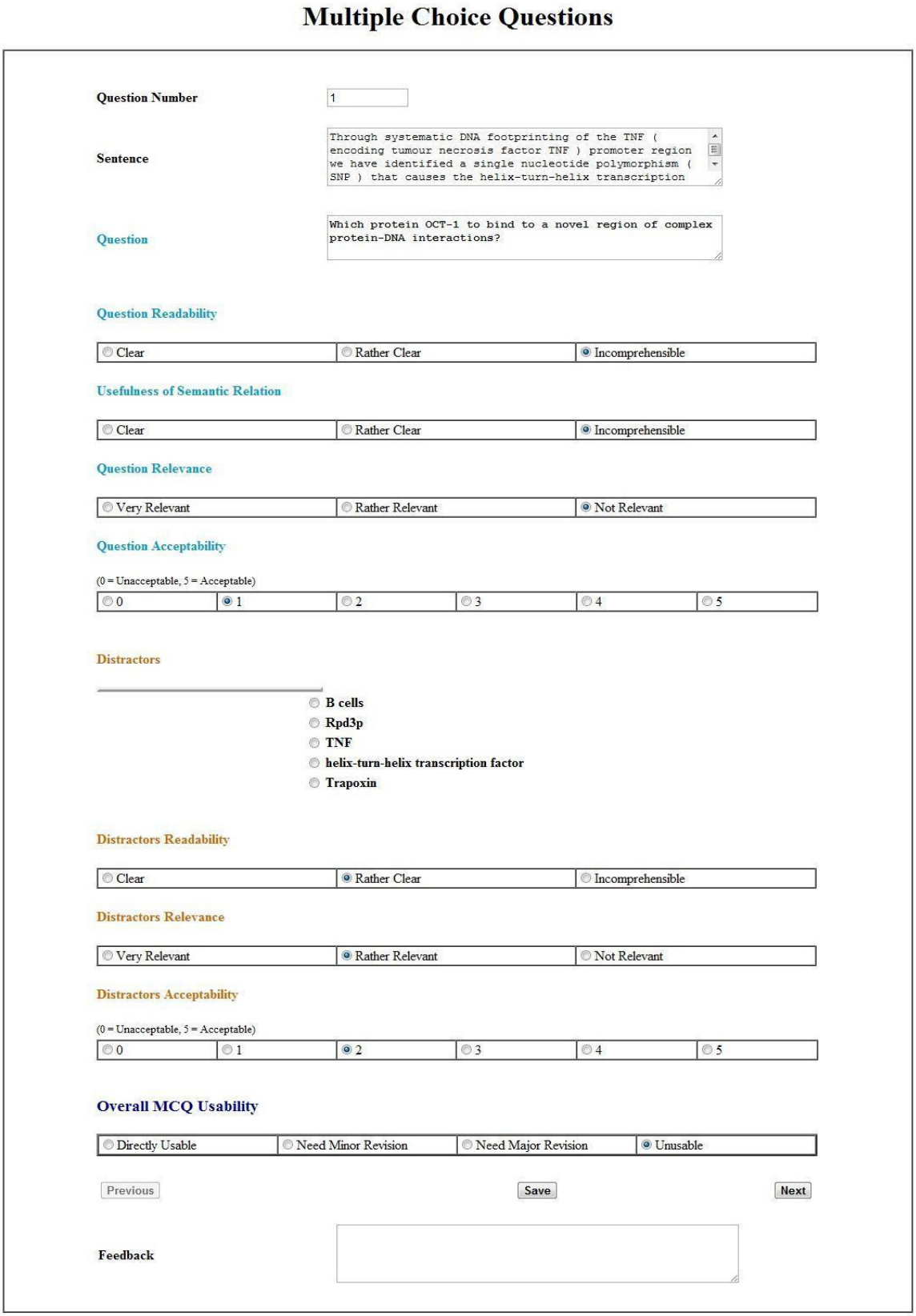

FIGURE 6: Screenshot of extrinsic evaluation interface.

Ranking Methods	Untagged Word Patterns					
	P	**R**	**F-score**	**P**	**R**	**F-score**
Threshold score > 0.01				*Threshold score > 0.1*		
IG	0.449	0.693	0.545	0.469	0.189	0.269
IGR	0.448	0.759	0.563	0.464	0.280	0.349
MI	0.444	0.776	0.564	0.464	0.454	0.459
NMI	0.449	0.728	0.555	0.465	0.341	0.393
LL	0.448	0.759	0.563	0.465	0.280	0.349
CHI	**0.466**	0.488	0.477	**0.470**	0.172	0.252
Meta	0.450	0.728	0.556	0.466	0.275	0.346
tf-idf	0.435	0.684	0.532	0.460	0.184	0.262
Threshold score > 0.02				*Threshold score > 0.2*		
IG	0.449	0.693	0.545	0.670	0.028	0.054
IGR	0.450	0.716	0.552	0.670	0.028	0.054
MI	0.452	0.700	0.549	0.464	0.314	0.374
NMI	0.454	0.657	0.537	0.462	0.260	0.333
LL	0.450	0.716	0.552	0.670	0.028	0.054
CHI	**0.470**	0.405	0.435	**0.773**	0.018	0.034
Meta	0.449	0.692	0.545	0.735	0.023	0.046
tf-idf	0.433	0.650	0.520	0.539	0.074	0.129
Threshold score > 0.03				*Threshold score > 0.3*		
IG	0.449	0.693	0.545	0.738	0.012	0.024
IGR	0.457	0.577	0.510	0.738	0.012	0.024
MI	0.453	0.653	0.535	0.460	0.280	0.348
NMI	0.463	0.522	0.491	0.463	0.166	0.244
LL	0.457	0.577	0.510	0.738	0.012	0.024
CHI	**0.468**	0.345	0.398	**0.778**	0.012	0.023
Meta	0.459	0.536	0.495	0.730	0.017	0.033
tf-idf	0.414	0.487	0.448	0.586	0.046	0.084

TABLE 9: Score-thresholding results for untagged word patterns.

Ranking Methods	PoS-tagged Word Patterns					
	P	**R**	**F-score**	**P**	**R**	**F-score**
Threshold score > 0.01				*Threshold score > 0.1*		
IG	0.439	0.615	0.512	0.653	0.029	0.055
IGR	0.440	0.583	0.501	0.649	0.028	0.054
MI	0.444	0.696	0.542	0.439	0.407	0.423
NMI	0.444	0.648	0.527	0.444	0.312	0.367
LL	0.440	0.583	0.501	0.649	0.028	0.054
CHI	**0.447**	0.342	0.387	**0.737**	0.016	0.031
Meta	0.440	0.610	0.511	0.649	0.031	0.059
tf-idf	0.388	0.559	0.458	0.559	0.045	0.083
Threshold score > 0.02				*Threshold score > 0.2*		
IG	0.447	0.379	0.410	0.970	0.007	0.014
IGR	0.443	0.449	0.446	0.970	0.007	0.014
MI	0.442	0.623	0.517	0.441	0.306	0.361
NMI	0.443	0.538	0.486	0.446	0.194	0.271
LL	0.443	0.449	0.446	**0.970**	0.007	0.014
CHI	**0.450**	0.229	0.303	0.952	0.004	0.009
Meta	0.437	0.442	0.439	0.870	0.009	0.018
tf-idf	0.391	0.538	0.453	0.577	0.025	0.048
Threshold score > 0.03				*Threshold score > 0.3*		
IG	0.447	0.379	0.410	1.000	0.003	0.007
IGR	0.450	0.339	0.386	1.000	0.003	0.007
MI	0.439	0.560	0.492	0.436	0.264	0.329
NMI	0.441	0.446	0.444	0.703	0.018	0.034
LL	0.450	0.339	0.387	1.000	0.003	0.007
CHI	**0.452**	0.200	0.277	**1.000**	0.002	0.005
Meta	0.448	0.362	0.401	0.955	0.005	0.009
tf-idf	0.399	0.456	0.425	0.637	0.016	0.032

TABLE 10: Score-thresholding results for PoS-tagged word patterns.

Ranking Methods	Verb-centred Word Patterns					
	P	**R**	**F-score**	**P**	**R**	**F-score**
Threshold score > 0.01				*Threshold score > 0.1*		
IG	0.447	0.675	0.538	0.662	0.026	0.051
IGR	0.447	0.622	0.520	0.667	0.027	0.052
MI	0.452	0.700	0.550	0.444	0.403	0.423
NMI	0.450	0.659	0.535	0.447	0.309	0.365
LL	0.447	0.622	0.520	0.667	0.027	0.052
CHI	**0.452**	0.345	0.391	0.757	0.016	0.032
Meta	0.448	0.612	0.517	**0.774**	0.025	0.049
tf-idf	0.409	0.609	0.489	0.608	0.027	0.051
Threshold score > 0.02				*Threshold score > 0.2*		
IG	0.453	0.412	0.432	0.944	0.005	0.010
IGR	0.447	0.444	0.446	0.944	0.005	0.010
MI	0.449	0.634	0.525	0.443	0.303	0.359
NMI	0.450	0.532	0.487	0.445	0.201	0.277
LL	0.447	0.444	0.446	**0.944**	0.005	0.010
CHI	**0.452**	0.238	0.312	0.909	0.003	0.006
Meta	0.448	0.437	0.442	0.871	0.008	0.016
tf-idf	0.411	0.554	0.472	0.654	0.016	0.032
Threshold score > 0.03				*Threshold score > 0.3*		
IG	0.453	0.412	0.432	1.000	0.002	0.005
IGR	0.454	0.356	0.399	1.000	0.002	0.005
MI	0.447	0.558	0.496	0.438	0.274	0.337
NMI	0.448	0.443	0.445	0.725	0.018	0.035
LL	0.454	0.356	0.399	1.000	0.002	0.005
CHI	**0.451**	0.206	0.283	**1.000**	0.002	0.004
Meta	0.451	0.408	0.428	0.917	0.003	0.007
tf-idf	0.426	0.424	0.425	0.690	0.009	0.018

TABLE 11: Score-thresholding results for verb-centered word patterns.