# Automatic Cloze-Questions Generation

**Annamaneni Narendra, Manish Agarwal and Rakshit shah**
LTRC, IIIT-Hyderabad, India
{narendra.annamaneni|manish_agrwal|rakshit_shah}@students.iiit.ac.in

## Abstract

Cloze questions are questions containing sentences with one or more blanks and multiple choices listed to pick an answer from. In this work, we present an automatic Cloze Question Generation (CQG) system that generates a list of important cloze questions given an English article. Our system is divided into three modules: sentence selection, keyword selection and distractor selection. We also present evaluation guidelines to evaluate CQG systems. Using these guidelines three evaluators report an average score of 3.18 (out of 4) on Cricket World Cup 2011 data.

## 1 Introduction

Multiple choice questions (MCQs) have been proved efficient to judge students' knowledge. Manual construction of such questions, however, is a time-consuming and labour-intensive task. Cloze questions (CQs) are fill-in-the-blank questions, where a sentence is given with one or more blanks in it with four alternatives to fill those blanks. As opposed to MCQs where one has to generate the WH style question, CQs use a sentence with blanks to form a question. The sentence could be picked from a document on the topic avoiding the need to generate a WH style question. As a result, automatic CQG has received a lot of research attention recently.

1. *Zaheer Khan opened his account with three consecutive maidens in the world-cup final.*
*(a) Zaheer Khan (b) Lasith Malinga (c) Praveen Kumar (d) Munaf Patel*

In the above example CQ, the underlined word (referred to as *keyword*) *Zaheer Khan* is blanked out in the sentence and four alternatives are given. In area of cloze questions, (Sumita et. al., 2005; Lee and Seneff, 2007; Lin et. al., 2007; Pino et. al., 2009; Smith et. al., 2010) have mostly worked in the domain of English language learning. Cloze questions have been generated to test students knowledge of English in using the correct verbs (Sumita et. al., 2005), prepositions (Lee and Seneff, 2007) and adjectives (Lin et. al., 2007) in sentences. Pino et. al. (2009) and Smith et. al. (2010) have generated questions to teach and evaluate student's vocabulary. Agarwal and Mannem (2011) have generated factual cloze questions from a biology text book through heuristically weighted features. They do not use any external knowledge and rely only on information present in the document to generate the CQs with distractors. This restricts the possibilities during distractor selection and leads to poor distractors.

In this work, we present an end-to-end automatic cloze question generating system which adopts a semi-structured approach to generate CQs by making use of a knowledge base extracted from a Cricket [1] portal. Also, unlike previous approaches we add context to the question sentence in the process of creating a CQ. This is done to disambiguate the question and avoid cases where there are multiple answers for a question. In Example 1, we have disambiguated the question by adding context *in the world-cup final*. Such a CQG system can be used in a variety of applications such as quizzing systems, trivia games, assigning fan ratings on social networks by posing game related questions etc.

Automatic evaluation of a CQG system is a very difficult task; all the previous systems have been evaluated manually. But even for the manual evaluation, one needs specific guidelines to evaluate fac-

[1] A popular game played in commonwealth countries such as Australia, England, India, Pakistan etc..

*Proceedings of Recent Advances in Natural Language Processing*, pages 511–515,
Hissar, Bulgaria, 7-13 September 2013.

tual CQs when compared to those that are used in language learning scenario. To the best of our knowledge there are no previously published guidelines for this task. In this paper, we also present guidelines to evaluate automatically generated factual CQs.

## 2 Approach

Our system takes news reports on Cricket matches as input and gives factual CQs as output using a knowledge base on Cricket players and officials collected from the web.

Given a document, the system goes through three stages to generate the cloze questions. In the first stage, informative and relevant sentences are selected and in the second stage, keywords (or words/phrases to be questioned on) are identified in the selected sentence. Distractors (or answer alternatives) for the keyword in the question sentence are chosen in the final stage.

The Stanford CoreNLP tool kit is used for tokenization, POS tagging (Toutanova et. al, 2003), NER (Finkel et. al, 2005), parsing (Klein et. al, 2003) and coreference resolution (Lee et. al, 2011) of sentences in the input documents.

### 2.1 Sentence Selection

In sentence selection, relevant and informative sentences from a given input article are picked to be the question sentences in cloze questions.

Agarwal and Mannem (2011) uses many summarization features for sentence selection based on heuristic weights. But for this task it is difficult to decide the correct relative weights for each feature without any training data. So our system directly uses a summarizer for selection of important sentences. There are few abstractive summarizers but they perform very poorly, (Michael et. al., 1999) for example. So our system uses an extractive summarizer, MEAD [2] to select important sentences. Top 10 percent of the ranked sentences from the summarizer's output are chosen to generate cloze questions.

### 2.2 Keywords Selection

This step of the process is selection of words in the selected sentence that can be blanked out. These words are referred to as the keywords in the sentence. For a good factual CQ, a keyword should be the word/phrase/clause that tests the knowledge of the user from the content of the article. This keyword shouldn't be too trivial and neither should be too obscure. For example, in an article on Obama, Obama would make a bad keyword.

The system first collects all the potential keywords from a sentence in a list and then prunes this list on the basis of observations described later in this section.

Unlike the previous works in this area, our system is not bound to select only one token keyword or to select only nouns and adjectives as a keyword. In our work, a keyword could be a Named Entity (person, number, location, organization or date) (NE), a pronoun (that comes at beginning of a sentence so that its referent is not present in that sentence) or a constituent (selected using the parse tree). In Example 2, the selected keyword is a noun phrase, *carrom ball*.

2. *R Ashwin used his <u>carrom ball</u> to remove the potentially explosive Kirk Edwards in Cricket World Cup 2011.*

#### 2.2.1 Observations

According to our data analysis we have some observations to prune the list that are described below.

- **Relevant tokens should be present in the keyword** There must be few other tokens in a keyword other than stop words[3], common words[4] and topic words [5]. We observed that words given by the TopicS tool are trivial to be keywords as they are easy to predict.

- **Prepositions** The preposition at the beginning of the keyword is an important clue with respect to what the author is looking to check. So, we keep it as a part of the question sentence rather than blank it out as the keyword. We also prune the keywords containing one or more prepositions as they more often than not make the question unanswerable and sometimes introduce a possibility for multiple answers to such questions.

[2] MEAD is a publicly available toolkit for multi-lingual summarization and evaluation. The toolkit implements multiple summarization algorithms (at arbitrary compression rates) such as position-based, Centroid[RJB00], TF*IDF, and query-based methods (http://www.summarization.com/mead)

[3] In computing, stop words are words which are filtered out prior to, or after processing of natural language data (text). `http://armandbrahaj.blog.al/2009/04/14/list-of-english-stop-words/`

[4] Most common words in English taken from `http://en.wikipedia.org/wiki/Most\_common\_words\_in\_English`.

[5] Topics (words) which the article talks about. We used the TopicS tool (Lin and Hovy, 2000)

We also use the observations, presented by (Agarwal and Mannem, 2011) in their keyword selection step, such as, a keyword must not repeat in the sentence again and its term frequency should not be high, a keyword should not be the entire sentence, etc. We use the score given by the TopicS tool to filter the keywords with high frequency.

The above criteria reduces the potential keywords' list by a significant amount. Among the rest of the keywords, our system gives preference to NE (persons, location, organization, numbers and dates (in order)), noun phrases, verb phrases in order. To preserve the overall quality of a set of generated questions, system checks that any answer should not be present in other questions. In case of a tie term frequency is used.

## 3 Distractor Selection

The previous two stages (*sentence selection* and *keyword selection*) are not domain specific in nature i.e. they work fine irrespective of the dataset and domain chosen. But the same is not true for *distractor selection* because the quality of distractors largely depends on the domain. We have performed experiments and presented the results on the domain Cricket. Consider Example 3.

3. *Sehwag had hit a boundary from the first ball of six of India's previous eight innings in Cricket World Cup 2011.*
   *(a) Ponting (b) Sehwag (c) Zaheer (d) Marsh*

In *Example 3*, although all the distractors are of the domain of Cricket, the distractors are not good enough to create confusion. We have some clues in the given sentence that can be exploited to provide distractors that pose a greater challenge to the students: (i) Someone hitting a boundary on the first ball must be a Top-order batsman and (ii) *India* in the sentence implies that the batsman is from Indian team. But out of the three distractors, one is an Indian bowler (*Zaheer*) and the other two are Australian Top-order batsmen (*Ponting* and *Marsh*). Hence answer of the question can easily be chosen which is *Sehwag*.

| Player's name | Team | Playing Role | Batting Style | Bowling Style |
|---|---|---|---|---|
| Sachin Ramesh Tendulkar | India | Top-order batsman | Right hand | Right-arm, Off Break |
| Zaheer Khan | India | Fast bowler | Right hand | Left-arm, Faster |
| Virendra Sehwag | India | Top-order batsman | Right hand | Right-arm, Off Break |
| Ricky Ponting | Australia | Top-order batsman | Right hand | - |

Table 1: Knowledge Base

To present more meaningful and useful distractors, the stage is domain dependent and also uses a knowledge base. The system extracts clues from the sentences to present meaningful distractors. The knowledge base is collected by crawling players' pages available at `http://www.espncricinfo.com`. Each page has a variety of information about the player such as name, playing style, birth date, playing role, major teams etc. This information is widely used to make better choices through out the system. Sample rows and columns from the database of players are shown in the Table 1. The Distractors are selected such that none of them already occur in the question sentence.

For the Cricket domain, the system takes only the NEs as keywords. So if a keyword's NE Tag is location/number/date/organization, then system selects three distractors from the database randomly. But in case when the NE tag is a person's name, three distractors are selected based on (i) the properties of the keyword and (ii) the clues in the question sentence. The distractor selection method is shown in Figure 1.

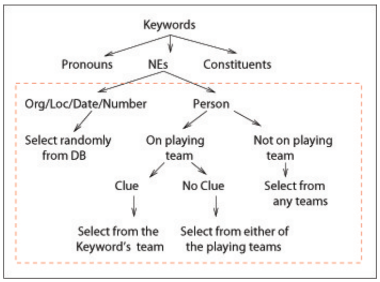

Figure 1: Distractor Selection Method

In case of a person's name *team name*, *playing role*, *batting style* and *bowling style* are the features of a keyword (Table 1). The system looks for clues in the sentence such as team names and other player names. According to the features and clues extracted by the system, three distractors are chosen either from the same team as that of the keyword or from both playing teams or from any team playing in the tournament. Distractors are selected such that none of them already occur in the question sentence. Remainder of this section describes different strategies incorporated in order to handle different cases.

### 3.1 Select distractors from a single team

The presence of a team name or of a team player of any of the two playing teams is a direct clue for selecting the distractors from the team of the keyword. It does not matter that the team name is of

| Score | Sentence | | Keyword | Distractor |
|---|---|---|---|---|
| 4 | Very informative | Very relevant | Question worthy | Three are useful |
| 3 | Informative | Relevant | Question worthy but span is wrong | Two are useful |
| 2 | Remotely informative | Remotely relevant | Question worthy but not the best | One is useful |
| 1 | Not at all informative | Not at all relevant | Not at all question worthy | None is useful |

Table 2: Evaluation Guidelines

the player which is our keyword or of the team he is playing against as long as it is either of these two. Consider *Example 3* and *Example 4*.

4. ***MS Dhoni*** *trumped a poetic century from Mahela Jayawardene to pull off the highest run-chase ever achieved in a World Cup final.*
   *(a) Kumar Sangakkara (b) Upul Tharanga*
   *(c) Mahela Jayawardene (d) Chamara Silva*

In *Example 3*, the system finds explicitly *India*, the team name whereas in *Example 4*, the system finds a player of the opponent team, *MS Dhoni*. In both these cases, the distractors are selected from the team that the keyword belongs to.

### 3.2 Select distractors from both the teams

We observed that we could choose distractors from either of the teams if there are no features indicating a particular playing team and the keyword is from one of the two teams. So the system can select three distractors from any of the two playing teams, which is a larger source to select the distractors.

In *Example 1*, there are no features indicating that the distractors should all belong to either team *India* or team *Sri Lanka* knowing that the world cup final was played between India and Sri Lanka. So, we can select distractors from both the teams in such cases.

### 3.3 Select distractors from any team

If the keyword in a question does not belong to either of the teams then it could be a name of an umpire or a player from the other teams. In case of an umpire, we randomly select three umpires from the list of umpires for that tournament. And in case of a player that belongs to neither of the teams playing the match, we randomly pick three players with the same *playing role* as that of the keyword from any team, doesn't matter playing or not.

## 4 Evaluation Guidelines and Results

Automatic evaluation of any CQG system is difficult for two reasons i) agreeing on standard evaluation data is difficult ii) there is no one particular set of CQs that is correct. Most question generation systems hence rely on manual evaluation. However, there are no specific guidelines for the manual evaluation either. In this paper, we also present evaluation guidelines for a CQG system that we believe are suitable for the task. The proposed evaluation guidelines are shown in Table 2.

Evaluation is done in three phases: (i) Evaluation of selected sentences, (ii) Evaluation of selected keywords and (iii) Evaluation of selected distractors. The evaluation of the selected sentences is done using two metrics, namely, informativeness and relevance. Merging the two metrics into one can mislead because a sentence might be informative but not relevant and vice versa. In such a case, assigning a score of three for one possibility and two to the other will not do justice to the system. The keywords are evaluated for their question worthiness and correctness of their span. Finally, the distractors are evaluated for their usability (i.e. the score is the number of distractors that are useful). A distractor is useful if it can't be discounted easily through simple elimination techniques.

The overall score for every cloze question is calculated by taking the average of all the four metrics for a question. The overall score on the entire data is the mean of scores of each question.

| **Evaluator** | | **4** | **3** | **2** | **1** |
|---|---|---|---|---|---|
| Eval-1 | Informativeness | 8 | 10 | 3 | 1 |
| | Relevance | 4 | 15 | 3 | 0 |
| | Keywords | 16 | 0 | 5 | 1 |
| | Distractors | 11 | 4 | 4 | 3 |
| Eval-2 | Informativeness | 13 | 7 | 2 | 0 |
| | Relevance | 9 | 11 | 2 | 0 |
| | Keywords | 7 | 0 | 15 | 0 |
| | Distractors | 6 | 14 | 1 | 1 |
| Eval-3 | Informativeness | 9 | 9 | 4 | 0 |
| | Relevance | 8 | 10 | 4 | 0 |
| | Keywords | 7 | 0 | 15 | 0 |
| | Distractors | 14 | 5 | 3 | 0 |

Table 3: Results (Eval: Evaluator)

Cloze questions generated from news reports on two Cricket World Cup 2011 matches were used for evaluation. 22 questions (10+12) were generated and evaluated by three different evaluators using the above mentioned guidelines. The results are listed in Table 3. The overall accuracy of our system is 3.15 (Eval-1), 3.14 (Eval-2) and 3.26 (Eval-3) out of 4. The accuracy of the distractors is 3.05 (Eval-1), 3.14 ((Eval-2) and 3.5 (Eval-3) out of 4.

## 5 Conclusion & Future Work

This paper proposed the automatic generation of *Multiple Choice Questions*(MCQs). The proposed method generates MCQs using summarisation tool ,TopicS tool and knowledge base from the web.We have proposed a novel approach for distractor selection using knowledge base for the specific domain.The proposed constraints for the distractor selection makes questions effective.We have proposed the evaluation guidelines to evaluate multiple choice questions at three stages.

We believe that there is still much room for improvement.Firstly distractor selection proposal was done for specific domain ,these constraints can be generalised to any domain. Proposed evaluation guidelines do evaluation question by question only.The overall performance of the system,taking into account the entire document is not performed .This is left for future work.

## References

Chin-Yew Lin and Eduard Hovy 2000 *The automated acquisition of topic signatures for text summarization.* In Proceedings of COLING 2000.

Dan Klein and Christopher D. Manning 2003 *Accurate Unlexicalized Parsing.* Proceedings of the 41st Meeting of the Association for Computational Linguistics, pp. 423-430.

Eiichiro Sumita, Fumiaki Sugaya, and Seiichi Yamamoto. 2005 *Measuring Non-native Speakers Proficiency of English by Using a Test with Automatically-Generated Fill-in-the-Blank Questions.* 2nd Wkshop on Building Educational Applications using NLP, Ann Arbor

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky. 2011 *Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task.* In Proceedings of the CoNLL-2011 Shared Task, 2011.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning 2005 *Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling.* Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370

John Lee and Stephanie Seneff. 2007 *Automatic Generation of Cloze Items for Prepositions.* CiteSeerX - Scientific Literature Digital Library and Search Engine [http://citeseerx.ist.psu.edu/oai2] (United States).

Juan Pino, Michael Heilman and Maxine Eskenazi. 2009 *A Selection Strategy to Improve Cloze Question Quality* Wkshop on Intelligent Tutoring Systems for Ill-Defined Domains. 9th Int. Conf. on ITS.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer 2003 *Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network* In Proceedings of HLT-NAACL 2003, pp. 252-259.

Lin, Y. C., Sung, L. C., Chen and M. C.: 2007 *An Automatic Multiple-Choice Question Generation Scheme for English Adjective Understanding* CCE 2007 Workshop Proc. of Modeling, Management and Generation of Problems / Questions in eLearning, pp. 137-142.

Manish Agarwal and Prashanth Mannem 2011 *Automatic Gap-fill Question Generation from Text Book.* In the proceeding of, The 6th Workshop on Innovative Use of NLP for Building Educational Applications, ACL-HLT 2011.

Michael J.Witbrock and Vibhu O. Mittal. 1999 *Ultra-Summarization: A Statistical Approach to Generating Highly Condensed Non-Extractive Summaries.* In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.

Simon Smith, P.V.S Avinesh and Adam Kilgarriff. 2010 *Gap-fill Tests for Language Learners: Corpus-Driven Item Generation.*