Automatic Cloze Generation based on Cross-document Information Extraction

Wen-Pin Lin and Heng Ji
Computer Science Department
Queens College and Graduate Center
City University of New York
New York, NY 11367, USA
hengji@cs.qc.cuny.edu

Abstract

Most of the previous work on cloze prediction focused on grammar testing such as preposition generation. In this paper we explore a different problem on knowledge testing by designing a system to automatically generate blanks for cloze test. We present a cross-document Information Extraction (IE) driven approach, and compare its performance on two different tasks with and without background documents respectively. Experimental results show that compared to traditional approaches, our method can significantly save time for educators in designing high-quality cloze tests.

Keywords—cloze generation, cross-document information extraction

1 Introduction

A cloze test is a traditional exercise consisting of a portion of text with certain words removed, where the student is asked to replace the missing words (Taylor, 1953). Cloze exercises have a variety of uses in e-learning, including testing and developing reading, vocabulary, grammar, or listening skills. Cloze generation has been an effective educational instrument for the following two purposes: (1) measuring language proficiency for a second/foreign language learner (Oller, 1973); (2) measuring reading comprehension (e.g., Fotos, 1991; Jonz, 1991). Cloze test scores were proven to correlate highly with other language test scores. However, it’s time consuming for the teachers to manually make blanks and evaluate answers.

Automatic cloze prediction methods could greatly help the students to capture and organize the main content (what/who/when/where/how/why questions), and also provide baselines for the teachers to design questions. It has been widely assumed by language professionals that cloze tests are testing an underlying "grammar of expectancy", the capability of individuals to synthesize and analyze sequential linguistic elements in realistic contexts of use. Therefore most of the previous research on automatic cloze generation focused on the above purpose (1), namely on evaluating grammatical blanks based on simple techniques, such as every n-th word (e.g. 6th or 7th word) or part-of-speech tagging. In this paper we focus on a more chal-
lenging task in order to address the purpose (2), namely to generate semantically informative and salient blanks. We apply state-of-the-art cross-document Information Extraction (IE) techniques (Ji et al., 2009) to detect important facts from each learning article, and incorporate additional features including background knowledge for cloze generation. The novel contributions of this paper are as follows:

- the first attempt to generate cloze based on IE techniques;
- imitate human learning by exploiting background knowledge in cloze generation;
- investigate the impact of our approach on different learning levels;
- propose a new evaluation metric of cloze prediction based on browsing cost.

2 Task Definition and Baseline Systems

The subjects for cloze test are usually chosen based on the target learning groups. For example, the cloze test articles for high school students focus a lot on the biography facts of historical figures and scientific facts, while elementary school materials focus on basic world knowledge such as geographical facts. Two examples are presented as follows.

[Cloze Test for Elementary School Students]
There are 12 months in a year.
The months are January, February, March, April, May, June, July, August, September, October, November, and December.
There are about four weeks in each month and 52 weeks in a year.
There are seven days in a week.
The days are Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, and Saturday.
New Year's Day is January 1st.

[Cloze Test for High School Students]
John Adams (1735 - 1826) was the second President of the United States of America.
He was President from 1797 until 1801. His Vice-President was Thomas Jefferson.
John Adams was born in Quincy, Massachusetts. His father was a farmer.
Adams was a delegate to both the First and Second Continental Congresses, and helped write the Declaration of Independence. In 1789, Adams was elected the first Vice-President of the US, serving two terms under President George Washington.
Adams was elected President in 1797, barely beating Thomas Jefferson, who became his Vice-President. John Adams was the first President to live in the White House; his family moved there in 1800.
John Adams died on July 4, 1826, the 50th anniversary of the signing of the Declaration of Independence. Thomas Jefferson had died earlier that same day. They were the only two signers of the Declaration of Independence that were elected President of the USA.

Usually a collection of candidate words, called “word bank” is provided to the students to fill in the blanks.
Many blanks can be covered by IE output. The goal of our task is to generate the word bank automatically using IE approach. There are generally two different types of cloze tests: (1) using background documents; and (2) not using background documents. The mode (1) aims to imitate the procedure of human learning – a student learns a long article from the book and is asked to take a quiz on a shorter summary after the class. In this paper we will conduct experiments in both settings and compare the results.

For comparison we will apply the following baseline approaches for cloze generation.

- Extract all content words as blanks; in this paper we apply the Stanford English part-of-speech tagger (Toutanova and Manning, 2000) to extract all nouns, verbs and adjectives;
- Extract all prepositions as blanks using the part-of-speech tagger;
- Extract words randomly;
- Extract every-seventh words;
- Extract top N most frequent words.

3 IE based Cloze Generation

As we can see most of the baseline approaches are based on shallow processing, so we need deeper understanding on the texts in order to generate cloze in a more reliable way. Although the candidate blanks generated by teachers vary a lot based on different topics and learning levels, most of them involve facts (entities, time expressions, relations, events, etc.) which can be detected by IE techniques.

3.1 IE Approach Overview

We apply our English cross-document IE system (Ji et al., 2009) to extract facts from texts. They were developed for the NIST Automatic Content Extraction Program (ACE2005). ACE defined 7 types of entities (persons(PER), geo-political entities(GPE), organization(ORG), facilities(FAC), weapons(WEA) and vehicles(VEH)), 18 types of relations (e.g. “a town some 50 miles south of Salzburg” indicates a located relation.), and 33 distinct types of relatively ‘dynamic’ events (e.g. “Barry Diller on Wednesday quit as chief of Vivendi Universal Entertainment.” indicates a “personnel-start” event). The IE pipeline includes name tagging, nominal mention tagging, coreference resolution, time expression extraction and normalization, relation extraction and event extraction. Most of these components are learned based on Maximum Entropy Models incorporating diverse features from lexical processing, part-of-speech tagging, syntactic parsing, dependency parsing, semantic role labeling and domain knowledge. We produce the head words of entity mentions, relation and event arguments, context words in relation mentions, time and value expressions, and event trigger words as the blank candidates.

3.2 Inference Constraints

It’s noteworthy that there are other characteristics about cloze generation that need to be taken into account. Ideally the answer of a blank cannot be inferred easily from the words in the same article. For example, if “John Adams” appears as the central topic of an article, then it’s not appropriate to remove one instance of “Adams” as the blank. Therefore, selecting informative words themselves is not sufficient. After we extract candidates from IE output, we apply the following filtering steps in order to match this constraint.
• If the number of mentions (other name strings, noun phrases or pronouns) referring to a name is larger than 8, delete this candidate name;
• Delete all pronouns, stop words and suffix words from the candidate set;
• If an event trigger has a lot of synonyms in the same article, delete it from the candidate set;
• For any event mention, only keep the trigger word and head words of event arguments;
• For any relation mention: only keep head words of arguments and intervening words;
• If there are background documents, remove any candidate blanks that appear fewer than 6 times in the collection of the target document and background documents.

4 Experimental Results

4.1 Data and Evaluation Metric

We derive our test data from the following two sources:

• Online data set: 59 documents from the e-learning website (http://www.enchantedlearning.com/), including 894 tokens in the word bank and 14618 tokens in the corpus. The word bank was provided by the website.
• Offline data set: 12 summary documents, with each document associated with 10 background documents from the NIST TAC2009 Summarization task (Dang and Owczarzak, 2009). The summary documents include 5175 tokens, and the word bank includes 155 tokens. Each word bank was annotated by two annotators in parallel and adjudicated at the end.

For each document our system will generate the same number of blanks as the word bank. We apply the Browsing Cost measure in (Ji et al., 2009) to evaluate our approach of cloze generation:

\[
Browsing \ Cost \ (i) = \text{the number of incorrect blanks that a user must examine before finding i correct blanks (which match the word bank)}.
\]

4.2 Overall Results

Figure 1 summarizes the browsing cost results for cloze prediction on the online data set. Figure 1 indicates that among all the baselines, the content word based approach achieved the best results. It’s not surprising that the preposition based approach performs poorly because our cloze test does not specifically target at grammar checking. The baselines based random selection and every-seven-word performed almost equally.

23 correct answers are extracted by IE but not by the content word based method, including 18 numbers, 3 prepositions, 1 adverb and 1 pronoun. Since the content words only include nouns, verbs and adjectives, all of the other words that have different part-of-speech tags are missed by the content word deletion baseline. But some important pronouns and prepositions are part of entity mentions, and some numbers are part of time expressions or values. Therefore our IE-driven approach can successfully cover all of them. For example, 13 numbers in-
dicate years which are considered as important facts by the teachers during cloze design. For example, in the following text, two numbers are chosen as blanks:

In 1947, Robinson played his first major league baseball game (he played for the Brooklyn Dodgers in an exhibition game against the New York Yankees).

... 
Robinson was born in the year 1919 in Cairo, Georgia.

Figure 1. Browsing Cost of Cloze Prediction for Online Data

The IE approach itself did not perform better than the content word based approach in terms of browsing cost. However, after applying the filtering steps described in section 3.2, our method achieved much better results than any baseline – a user needs to browse much fewer incorrect blanks before seeing any number of correct blanks. In total IE with filtering can cover 53% of the correct answers. Each filtering step can filter many incorrect answers. Especially the frequency based filtering steps successfully removed 321 frequent but incorrect candidates. For example, in the following text, obviously “George Washington Carver” is the central topic of the article and IE approach can identify all instances. But they cannot be selected as blank candidates because the students are expected to learn other facts about this person. Our frequency based filtering steps can remove these candidates, because the name “Carver” appears 8 times and there are 7 pronouns referring to it.
George Washington Carver (1865?-1943) was an American scientist, educator, humanitarian, and former slave. Carver developed hundreds of products from peanuts, sweet potatoes, pecans, and soybeans. His discoveries greatly improved the agricultural output and the health of Southern farmers. Before this, the only main crop in the South was cotton. The products that Carver invented included a rubber substitute, adhesives, foodstuffs, dyes, pigments, and many other products. Carver was born in the state of Missouri and was sickly as a child. He was orphaned when he was young, and was brought up by Moses and Susan Carver on their farm. He began school at age 12 and later attended Simpson College in Indianola, Iowa, where he was the first black student. He transferred to Iowa Agricultural College to study science, earning a Bachelor of Science degree (in 1894) and a Master of Science degree in bacterial botany and agriculture (in 1897). He then became the first black faculty member at that college. Booker T. Washington convinced Carver to teach at the Tuskegee Normal and Industrial Institute for Negroes (now called Tuskegee University) in Alabama, USA, where Carver headed the agricultural department for 50 years. Carver donated his life savings to a fund designed to encourage agricultural research.

4.3 Impact of Using Background Documents

In Figure 2 we show the impact of using background documents in the IE driven cloze prediction approach. We can see in general the background documents provided positive gains in improving the quality of cloze generation. In order to check the robustness of using background document, we conducted the Wilcoxon Matched-Pairs Signed-Ranks Test to compare these two curves in Figure 2 for all the points. The results show that we can reject the hypothesis that the improvements using background documents were random at a 98.4% confidence level. To conclude it is important to imitate the procedure of student learning in automatic cloze prediction. If a word is salient and informative in the background documents, it’s likely to appear as a blank in the cloze test because it may be some knowledge that the teachers expect the students to learn and memorize. For example, after we remove all the event trigger candidates that appear fewer than 6 times in the collection of the target document and background documents, we can remove 69 incorrect answers while only lose 2 correct blanks.
4.4 Impact of Learning Levels

The IE-driven approach generally performed better than all the baselines on all the learning materials we have collected. But we have noticed that our approach is more advantageous on materials from advanced levels such as high schools. Figure 3 shows the results on learning materials from elemental schools. It indicates that although IE-driven approach can save a lot of browsing costs for the top 20 correct blanks, it tends to reach an upper-bound on recall. In contrast, the content word based approach is able to generate much more correct blank candidates. The main reason is that the learning materials from low levels don’t involve sophisticated knowledge that can be covered by entities, relations and events. For example, in the following article for elementary school students, IE does not produce any facts while the content word based approach can cover all the blank candidates:

*Apples are a type of widely-cultivated fruit that grows on trees.*
*An apple tree can grow to over 35 feet tall.*
*Each spring, an apple tree produces pink and white flowers.*
*After a blossom has been pollinated (fertilized), an apple develops.*
*Inside each apple are small, brown seeds, which can grow into new apple trees.*
*Each fall, apple trees lose their leaves - they are deciduous.*
5 Related Work

Lam et al. (1992) concluded that the random word deletion and preposition deletion methods are suited to measuring the perceptual process of English language use. Sachs et al. (1997) further proved that content word and preposition word deletion methods can produce more reliable tests than the random approaches. (Chen et al., 2006) used manually-designed patterns to generate testing questions. In this paper we took a further step and proposed an IE-driven approach that can achieve much better results than these traditional cloze generation methods. Coniam (1997) also proved that if a word appears very frequently in an article, it’s likely to indicate the central topic and thus not likely to be chosen as a blank candidate. In our previous work (Ji and Grishman, 2008) we demonstrated background documents can be exploited to improve IE performance.

A lot of other work focused on using cloze tests for grammar checking. Most of the features used for these previous methods were derived from part-of-speech tagging and parsing. For example, (Lee and Seneff, 2007) proposed two methods, based on collocations and on non-native English corpora, to generate distractors for prepositions. (Sumita et al., 2005) used part-of-speech tagging based cloze generation as the first step for question generation. Some systems (e.g. Mitkov and Ha, 2003; Goto et al., 2009) can generate multiple choices to test the students’ abilities at filling in the blanks.
6 Conclusion and Future Work

We described a novel approach of using cross-document IE techniques to predict blanks in cloze tests, and demonstrated how filtering methods and background documents can be exploited to enhance the performance. This approach is able to capture the semantic content embedded in learning materials, especially those for high school students, and produce high quality blanks. Most of the previous work on cloze generation focused on grammar testing, we hope our work can provide a shared task definition and a new evaluation metric for other researchers who are interested in studying the cloze tests for evaluating reading comprehension. In the future we intend to extend our work to automatic answer generation and scoring for cloze tests. We will also gather more labeled data so that we can adapt our IE system to cover broader types of domain-specific facts (e.g. extended named entity types such as “Dinosaurs” in the biology domain).

Acknowledgement

This work was supported by the U.S. Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053, the U.S. NSF CAREER Award under Grant IIS-0953149, Google, Inc., DARPA GALE Program, CUNY Research Enhancement Program, PSC-CUNY Research Program, Faculty Publication Program and GRTI Program. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References


John Sachs, Peter Tung and Raymond Lam. 1997. How to construct a cloze test: Lessons from testing measurement theory models. *Perspectives* (City University of Hong Kong) Vol. 9 (1997 Spring)

