A Selection Strategy to Improve Cloze Question Quality
提高完形填空问题质量的选择策略

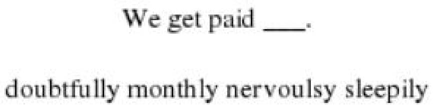
Abstract. We present a strategy to improve the quality of automatically generated cloze and open cloze questions which are used by the REAP tutoring system for assessment in the ill-defined domain of English as a Second Language vocabulary learning. Cloze and open cloze questions are fill-in-the-blank questions with and without multiple choice, respectively. The REAP intelligent tutoring system [1] uses cloze questions as part of its assessment and practice module. First we describe a baseline technique to generate cloze questions which uses sample sentences from WordNet [2]. We then show how we can refine this technique with linguistically motivated features to generate better cloze and open cloze questions. A group of English as a Second Language teachers evaluated the quality of the cloze questions generated by both techniques. They also evaluated how well-defined the context of the open cloze questions was. The baseline technique produced high quality cloze questions 40% of the time, while the new strategy produced high quality cloze questions 66% of the time. We also compared our approach to manually authored open cloze questions.

摘要。我们提出了一种策略来提高自动生成的完形填空的质量和开放的完形填空问题，这些问题被REAP辅导系统用于评估英语作为第二语言词汇学习的不明确领域。完形填空和打开完形填空问题分别是有和没有多项选择的填空问题。REAP智能辅导系统[1]使用完形填空问题作为其评估和实践模块的一部分。首先，我们描述了一种生成完形填空问题的基线技术，该问题使用了来自WordNet的样本句子[2]。然后，我们将展示如何利用语言驱动的功能改进此技术，以生成更好的完形填空和打开完形填空问题。一组英语作为第二语言教师评估了两种技术产生的完形填空问题的质量。他们还评估了开放式完形填空问题的背景是如何明确定义的。基线技术在40％的时间内产生了高质量的完形填空问题，而新策略在66％的时间内产生了高质量的完形填空问题。我们还将我们的方法与手动创作的开放式完形填空问题进行了比较。

1 Introduction
1 简介

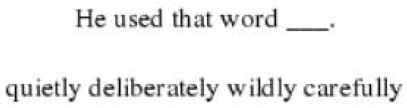
This paper describes a strategy to generate cloze (fill-in-the-blank) and open cloze (without multiple choice) questions, which are part of the assessment and practice module in the REAP system [1]. REAP is an intelligent tutoring system for English as a Second Language (ESL) vocabulary learning that provides a student with authentic documents retrieved from the web that contain target vocabulary words. These documents go through text quality filtering and are annotated for readability level [3] and topic. The system uses these annotations to match documents to the student’s level and interests. After each reading the system provides a series of practice exercises for focus words that were in that reading. Following the practice session, the system updates the learner model for the student based on his or her performance. In the following reading session, the system searches for texts according to the updated model, in order to give the student individualized practice. The REAP system uses several types of questions, for example synonym or related word questions [4], and in particular cloze questions. A cloze question consists of a stem, which is a sentence with a target word removed, and of distractors, which are words semantically or grammatically related to the target word. Examples of good and bad quality cloze questions are shown in Fig. 1 and 2. Brown et al. [5] successfully generated several different types of questions, for example synonym and cloze questions, by using WordNet [2]. However, teachers whose students were using the tutor judged the overall quality of the cloze questions to be insufficient to be used in their class. As a result, teachers generated the cloze items manually and without authoring support, which is time-consuming. This paper presents a strategy to improve the quality of automatically-generated cloze and open cloze questions.

本文描述了一种生成完形填空（填空）和开放完形填空（无多选）问题的策略，这些问题是 REAP 系统中评估和实践模块的一部分[1]。 REAP 是一种智能辅导系统，用于英语作为第二语言（ESL）词汇学习，为学生提供从网络中检索到的包含目标词汇单词的真实文档。这些文档通过文本质量过滤，并注释了可读性级别[3]和主题。系统使用这些注释将文档与学生的级别和兴趣相匹配。每次阅读后，系统都会为阅读中的焦点词提供一系列练习练习。在练习课程之后，系统根据学生的表现更新学生的学习者模型。在下面的阅读环节中，系统根据更新的模型搜索文本，以便为学生提供个性化的练习。 REAP 系统使用几种类型的问题，例如同义词或相关单词问题[4]，特别是完形填空问题。完形填空问题包括一个词干，它是一个删除了目标词的句子，以及一个干扰词，它们是与目标词在语义上或语法上相关的词。好的和坏质量的完形填空问题的例子如图 1 和 2 所示.Brown et al。 [5]通过使用 WordNet [2]成功地生成了几种不同类型的问题，例如同义词和完形填空问题。然而，学生使用导师的教师认为完形填空问题的整体质量不足以在课堂上使用。结果，教师手动生成完形遮挡项目而没有创作支持，这非常耗时。本文提出了一种提高自动生成的完形填空和开放完形填空问题质量的策略。


We get paid ___.

doubtfully monthly nervoulsy sleepily

Fig. 1. A rare case of a short sentence with a sufficiently well-defined context for the target word.


He used that word ___.

quietly deliberately wildly carefully

Fig. 2. A short sentence with multiple answers due to an ill-defined context.

Vocabulary learning can be considered an ill-defined domain due to the fact that there is no absolute measure of vocabulary knowledge. The exact nature of individual word knowledge is an unresolved research question, with various possible theories [6 - 8]. Although there exist ways of encoding lexical knowledge such as WordNet, Latent Semantic Analysis [9], and thesauri which can in some sense be viewed as models, there is no clear consensus on how to model lexical information. Even when assuming that an individual word is known, it is still challenging to make inferences about other, possibly related, words, because words appear in very different contexts [12]. On the contrary, well-structured domains have fully-developed cognitive models which have been applied successfully in intelligent tutoring. Examples include the work of VanLehn et al. [13] in physics and Anderson and Reiser in computer programming [10]. Also, many practice tasks for vocabulary learning, such as writing a sentence containing a vocabulary word (sentence production), lack single, definite correct answers. In fact, the sentence production problem can have an infinite number of correct answers: words can be combined in novel ways since human language is productive.
由于没有词汇知识的绝对量度，词汇学习可被视为不明确的领域。单个单词知识的确切性质是一个尚未解决的研究问题，有各种可能的理论[6-8]。虽然存在编码词汇知识的方法，如 WordNet，潜在语义分析[9]，以及在某种意义上可以被视为模型的叙词表，但对于如何对词汇信息建模没有明确的共识。即使假设单个单词是已知的，对其他可能相关的单词进行推断仍然具有挑战性，因为单词出现在非常不同的语境中[12]。相反，结构良好的领域具有完全开发的认知模型，已成功应用于智能辅导。例子包括 VanLehn 等人的工作。[13]在物理学和 Anderson 和 Reiser 的计算机编程[10]。此外，词汇学习的许多练习任务，例如编写包含词汇单词（句子产生）的句子，缺乏单一，明确的正确答案。实际上，句子生成问题可以有无数正确答案：由于人类语言是有效的，因此可以以新颖的方式组合单词。

Although sentence production is a very good exercise for vocabulary learning and provides a rich measure of the knowledge of a word, it is very hard to assess automatically, precisely because it has many solutions. One could imagine having a model solution and compare to it a student’s solution as in the entity relationship modelling tutor KERMIT [11]. Again this is hard to do automatically because the student’s solution could be very far, for example semantically, from the model and yet correct. On the other hand, the reverse problem, posed by cloze questions, of completing a sentence with a missing word is less openended but still valuable for practice and assessment. One advantage is that cloze questions can be scored automatically. Even though the process of grading is simpler, cloze questions can still assess various aspects of word knowledge. Nation divides the concept of word knowledge into receptive knowledge and productive knowledge [6]. He lists ten criteria for

receptive knowledge and nine criteria for productive knowledge. Multiple-choice cloze questions appear to involve five of the categories for receptive knowledge: students have to identify the written form of the target word and of the distractors. They also have to know at least one meaning of the word, namely the meaning of the word in a given context. They can make use of related words if such words are found in the stem. Finally, they need to recognize the correct grammatical usage of the word. However, there is no test of morphological knowledge. Students can also answer the question if they know the meaning of the word in one specific context only. Finally, they do not have to check if the usage (register, frequency of use) of the word fits in the stem. Furthermore, open cloze questions also involve several categories of productive knowledge. To answer an open cloze question, students need to produce a word expressing a certain meaning in the given context. This word could be related to, or form collocations, that is pairs of frequently co-occurring words, with certain words in the stem, it should form a grammatically correct sentence, it should be spelled correctly and its level of formality should be chosen carefully. Therefore, open cloze questions also assess productive knowledge. Having a reliable assessment of individual word knowledge allows the system to know which words it should focus on.
虽然句子产生是一种非常好的词汇学习练习，并且提供了丰富的单词知识测量，但很难自动评估，正是因为它有很多解决方案。可以想象有一个模型解决方案，并与实体关系建模导师KERMIT [11]中的学生解决方案进行比较。同样，这很难自动完成，因为学生的解决方案可能非常远，例如在语义上，从模型而且是正确的。另一方面，由完形填空问题构成的完成一个缺少单词的句子的相反问题不太开放，但对于练习和评估仍然有价值。一个优点是完整性问题可以自动评分。即使分级过程更简单，但完形填空问题仍然可以评估单词知识的各个方面。国家将词汇知识的概念分为接受性知识和生产性知识[6]。他列出了十项接受性知识标准和九项生产性知识标准。多项选择完形填空问题似乎涉及接受性知识的五个类别：学生必须识别目标词和干扰者的书面形式。他们还必须至少知道这个词的一个含义，即在给定语境中该词的含义。如果在词干中找到这样的词，他们可以使用相关词。最后，他们需要认识到这个词的正确语法用法。但是，没有形态学知识的测试。如果学生只在一个特定的语境中知道单词的含义，他们也可以回答这个问题。最后，他们不必检查单词的用法（注册，使用频率）是否适合词干。此外，开放式完形填空问题还涉及几类生产性知识。要回答一个开放的完形填空问题，学生需要在给定的语境中产生表达某种意义的词。这个词可能与或者形成搭配，即经常共同出现的词对，词干中的某些词，它应该形成一个语法正确的句子，它应该拼写正确，并且应该仔细选择其形式的程度。因此，开放式完形填空问题也会评估生产性知识。对单个单词知识进行可靠的评估可以使系统知道应该关注哪些单词。

Existing work on cloze question generation such as that of Liu et. al [14] has focused on lexical items regardless of their part-of-speech. Lee and Seneff [15] focused on generating cloze questions for prepositions with

a technique based on collocations. Others have generated cloze questions for common expressions. For instance, Higgins [16] generated cloze questions for structures such as “not only the” that assess grammatical skills rather than vocabulary skills. Hoshino and Nagakawa [17] generated questions for both vocabulary and grammatical patterns. Mitkov et al. [18] generated cloze questions about concepts found in a document by converting a declarative sentence into an interrogative sentence.
关于填充问题生成的现有工作，例如Liu等人的工作。 al [14]专注于词汇项目而不管他们的词性。 Lee和Seneff [15]专注于使用基于搭配的技术为介词生成完形填空问题。其他人已经为常见表达式生成完形填空问题。例如，希金斯[16]为结构创建了完形填空问题，例如评估语法技能而不是词汇技能的“不仅仅是”。 Hoshino和Nagakawa [17]产生了词汇和语法模式的问题。 Mitkov等人。 [18]通过将陈述句转换为疑问句，生成了关于文档中发现的概念的完形填空问题。

This work first focused on generating cloze questions for adverbs and was then extended to other parts of speech. Adverbs are considered to stand between open class and closed class words1 , which is why our strategy should be extensible to any part of speech. Additionally, it seems that adverbs with the suffix -ly (for example “clearly”), which are the most frequent kind of adverb, can be easily replaced in a sentence by other adverbs of the same kind without producing a grammatically or semantically incorrect sentence, if this sentence does not sufficiently narrow the context. As a consequence, it is important to avoid having several possible answers for cloze questions on adverbs.
这项工作首先集中在为副词生成完形填空问题，然后扩展到其他词性。 副词被认为是在开放阶级和封闭阶级词之间1，这就是为什么我们的策略应该可以扩展到任何词性。 另外，似乎带有后缀的副词（例如“清楚”），这是最常见的副词，可以在句子中被同类的其他副词轻易替换，而不会产生语法或语义不正确的句子 ，如果这句话没有充分缩小背景。 因此，避免对副词上的完形填空问题有几个可能的答案是很重要的。

This paper concentrates on the quality of the stem and the quality of the distractors. Sumita et al. [19] generate distractors thanks to a thesaurus; if the original sentence where the correct answer is replaced by a distractor gets more than zero hit on a search engine, the distractor is discarded. In [14], the authors use a technique based on word sense disambiguation to retrieve sentences from a corpus containing a target word with a particular sense. Their strategy also uses collocations to select suitable distractors. Our strategy makes use of collocations too by applying it to both stems and distractors in order to ensure their quality.
本文主要关注杆的质量和干扰物的质量。 Sumita等人。 [19]由于词库而产生干扰物; 如果正确答案被干扰者替换的原始句子在搜索引擎上获得的命中数大于零，则会丢弃干扰者。 在[14]中，作者使用基于词义消歧的技术从包含具有

特定意义的目标词的语料库中检索句子。 他们的策略还使用搭配来选择合适的干扰者。 我们的策略也通过将其应用于茎和干扰物来利用搭配，以确保其质量。

2 Cloze Question Generation
2 完形填空问题

We first describe the baseline technique based on WordNet sample sentences, which allowed us to investigate linguistic features for good cloze questions. WordNet is a lexical database in which nouns, verbs, adjectives and adverbs are grouped in synonym sets, or synsets. Synsets have a semantic relationship such as synonym, antonym, etc. For each synset, a definition and, most of the time, sample sentences, are provided. The latter are used to produce the stems. We want to generate a cloze question for an adverb w. We assign its most frequent synset as an adverb to w. This synset may have sample sentences, either involving the target word w itself or synonyms in the same synset. If sentences involving the target word are available, we select them. After this first selection, if there are still several possible sentences, the longest one is preferred, because longer sample sentences in WordNet tend to have richer contexts. The distractors are chosen randomly from a list of adverbs. As a result, they have the same part of speech as the target word which is recommended by Haladyna et al. [20] as a way of avoiding obviously wrong answers. Furthermore, Hensler and Beck [21] have shown that higher proficiency readers among students from grade 1 to 6 can use part of speech as a clue to answer a cloze question. Since the REAP system is used by ESL learners who are able to use parts of speech in their native language, we believe they may be able to use them in English as well. Thus all the distractors have the same part of speech.

我们首先描述基于WordNet样本句子的基线技术，这使我们能够研究语法特征以获得良好的完形填空问题。 WordNet是一个词汇数据库，其中名词，动词，形容词和副词分组在同义词集或同义词集中。同义词具有语义关系，例如同义词，反义词等。对于每个同义词集，提供定义以及大多数时间的样本句子。后者用于生产茎。我们想为副词w生成完形填空问题。我们将其最常用的synset指定为w的副词。该同义词集可以具有样本句子，其涉及目标词w本身或同一同义词集中的同义词。如果涉及目标词的句子可用，我们选择它们。在第一次选择之后，如果仍有几个可能的句子，则最长的句子是首选，因为WordNet中较长的样本句子往往具有更丰富的上下文。干扰者是从副词列表中随机选择的。因此，它们与Haladyna等人推荐的目标词具有相同的词性。 [20]作为一种避免明显错误答案的方法。此外，Hensler和Beck [21]已经证明，1至6年级学生的熟练程度较高的读者可以使用词性作为回答完形填空问题的线索。由于REAP系统由能够以其母语使用词性的ESL学习者使用，我们相信他们也可以使用英语。因此，所有干扰者都具有相同的词性。

2.1 Stem Selection

2.1 茎选择

Similar to the baseline, our strategy aims to select the most suitable sentences from a set of sentences containing the target word. However, for both for the stem and the distractors, our selection criteria are more fine-grained and we expect to produce a better quality output.
与基线类似,我们的策略旨在从包含目标词的一组句子中选择最合适的句子。然而，对于杆和干扰物，我们的选择标准更精细，我们期望产生更好的质量输出。

First, to apply a selection strategy we need to choose from several sample sentences per word. However, WordNet has zero or one sample sentence per word in any given synset. Therefore, we used the Cambridge Advanced Learner’s Dictionary (CALD) which has several sample sentences for each sense of a word. We retained the same selection criterion as for the baseline, namely the length of the sentence, and added new linguistically relevant criteria. Our approach employs the following selection criteria: complexity, well-defined context, grammaticality and length.
首先，要应用选择策略，我们需要从每个单词的几个样本句子中进行选择。 但是,WordNet 在任何给定的 synset 中每个单词都有一个或一个样本句子。 因此，我们使用了剑桥高级学习词典（CALD）,它对每个单词的含义都有几个样本句子。我们保留了与基线相同的选择标准,即句子的长度,并增加了新的语言相关标准。我们的方法采用以下选择标准：复杂性，明确定义的上下文，语法和长度。

Each sample sentence was given a weighted score combining these four criteria. We assessed the complexity of a sentence by parsing it with the Stanford parser [22, 23] and counting the resulting number of clauses. The Stanford parser uses the Penn Treebank syntactic tag set described in [24]. By clause we mean the sequences annotated with the following tags: S (simple declarative clause), SBAR (clause introduced by subordinating conjunction), SBARQ(direct question introduce by wh-word or wh-phrase) and SINV (declarative sentence with subject-auxiliary inversion). We chose this selection criterion through analysis of a dataset of high quality manually-generated cloze questions. We noticed that high quality stems often consist of two clauses, one clause involving the target word, and the other clause specifying the context, as in the following sentence: “We didn’t get much information from the first report, but subsequent reports were much more helpful.” (the target word is italicized). We believe that more clauses tend to make the context more well-defined.
每个样本句子都给出了结合这四个标准的加权分数。我们通过使用斯坦福解析器[22,23]解析它并计算得到的子句数来评估句子的复杂性。 Stanford 解析器使用[24]中描述的 Penn Treebank 语法标记集。 By 子句是指用以下标记注释的序列：S（简单声明性子句），SBAR（由从属连接引入的子句），SBARQ（由 wh-word 或 wh-phrase 引入的直接问题）和 SINV（具有主语的陈述句子）辅助倒置）。

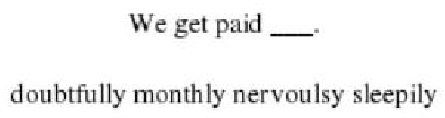
我们通过分析高质量手动生成的完形填空问题的数据集来选择此选择标准。我们注意到高质量的词根通常由两个条款组成，一个条款涉及目标词，另一个条款指定上下文，如下面的句子：“我们没有从第一次报告中获得太多信息，但随后的报告是更有帮助。“（目标词是斜体）。我们认为更多的条款往往会使背景更加明确。

The context of a sentence is considered to be well-defined if it requires the presence of the target word in the sentence and rejects the presence of any another word. A way to assess how well-defined the context is in a sentence with respect to a target word is to sum the collocation scores between the target word and the other words in the sentence. Collocations are pairs, or sometimes larger sets, of words that frequently co-occur, often despite the absence of clear semantic requirements. An example is that tea is much more likely to be called “strong” than “powerful”. Conversely, a car is more likely to be “powerful” than “strong”. An “argument”, however, can be either. The strength of the context around a word is in some sense determined by the presence of very informative collocations. For example, the sentence “I drank a cup of strong (blank) with lemon and sugar” has a very well-defined context for “tea” because of the collocations with “strong”, “lemon”, “sugar”, and “drink”. We followed the method described by Manning and Schutze [25] to estimate a set of collocations for each target word. ¨ We used a corpus consisting of approximately 100,000 texts appropriate for ESL learners, which are a part of the REAP database of potential readings. The system calculated the frequency of co-occurrence of content words by counting co-occurences within a window of two adjacent words. It then identified salient collocations by calculating a likelihood ratio for each possible pair. For this last step, other metrics such as point-wise mutual information and Pearson’s chisquare test were also tried and produced similar estimates of collocations.

如果句子的上下文要求句子中存在目标词并拒绝任何其他词的存在，则认为句子的上下文是明确定义的。评估关于目标词在句子中如何明确定义上下文的方法是将目标词与句子中的其他词之间的搭配分数相加。虽然没有明确的语义要求，但搭配是经常共同出现的词汇，有时甚至是更大的集合。一个例子是，茶更有可能被称为“强大”而非“强大”。相反，汽车更有可能“强大”而不是“强大”。然而，“论证”也可以。围绕一个词的上下文的强度在某种意义上是由非常有信息的搭配的存在决定的。例如，句子“我喝了一杯带有柠檬和糖的强烈（空白）”，因为与“强烈”，“柠檬”，“糖”和“糖”搭配，因此“茶”具有非常明确的背景。喝”。我们按照 Manning 和 Schutze [25]描述的方法来估计每个目标词的一组搭配。 ¨我们使用了一个语料库，该语料库包含大约 100,000 个适合 ESL 学习者的文本，这些文本是 REAP 潜在读数数据库的一部分。系统通过计算两个相邻单词的窗口内的共同出现来计算内容单词的共现频率。然后，它通过计算每个可能对的似然比来识别显着的搭配。对于最后一步，还尝试了其他指标，例如逐点互信息和 Pearson 的 chisquare 测试，并产生了类似的搭配估计。

We also used the Stanford parser to assess the grammaticality of the sentence. Each parsed sentence was assigned a score corresponding to the probabilistic context-free grammar score. Longer sentences generally have poorer scores and thus we normalized this score with the length of the sentence2 . Although the parser is applied to any sentence, even ungrammatical ones, the latter receive a lower score than grammatical sentences. Knight and Marcu [26] use a similar technique to estimate the probability of a short sentence in a noisy channel model for sentence compression.
我们还使用斯坦福解析器来评估句子的语法性。 为每个解析的句子分配对应于概率上下文无关语法分数的分数。 较长的句子通常具有较差的分数，因此我们将该分数与句子的长度进行归一化2。 虽然解析器适用于任何句子，即使是不合语法的句子，但后者的得分低于语法句子。 Knight和Marcu[26]使用类似的技术来估计用于句子压缩的嘈杂信道模型中短句子的概率。

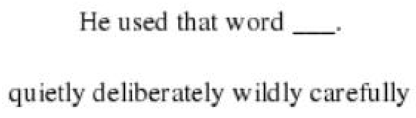
Finally, we used the sentence length as a quality criterion. Indeed, we noticed that the sentences generated by the baseline technique were too short (6 tokens on average) to provide enough context for the target word to be guessed (see Fig. 2), although there were some rare exceptions (see Fig. 1).
最后，我们使用句子长度作为质量标准。 实际上，我们注意到基线技术产生的句子太短（平均6个令牌），无法为猜测的目标词提供足够的上下文（见图2），尽管有一些罕见的例外（见图1））。



We get paid ___.

doubtfully monthly nervoulsy sleepily

Fig. 1. A rare case of a short sentence with a sufficiently well-defined context for the target word.



He used that word ___.

quietly deliberately wildly carefully

Fig. 2. A short sentence with multiple answers due to an ill-defined context.

The weights for each criterion were determined manually by the authors by examining their effects on performance on training data. We randomly chose 30 adverbs as training data and modified the weights so as to obtain the best sample sentences for these adverbs. Our criteria to judge the best samples were the same criteria used by ESL teachers for their assessment of the cloze questions (see Sect. 3). Our final weights are 1 for length, 8 for complexity, 0.3 for grammaticality and 0.2 for the collocation criteria3 . In addition, we filtered out misspelled sentences with the Jazzy spell checker [27]. Although this is not very relevant for the Cambridge Dictionary, it is relevant when using a large corpus of text. 每个标准的权重由作者通过检查它们对训练数据的性能的影响来手动确定。 我们随机选择 30 个副词作为训练数据并修改权重，以获得这些副词的最佳样本句子。我们判断最佳样本的标准与 ESL 教师用于评估完形填空问题的标准相同(见第 3 节）。 我们的最终权重是长度为 1，复杂度为 8，语法为 0.3，搭配标准为 0.2。 另外，我们使用 Jazzy 拼写检查器过滤掉拼写错误的句子[27]。 虽然这与剑桥词典不太相关，但在使用大量文本时它是相关的。

2.2 Distractor Selection

2.2 牵引器选择

The quality of distractors was also scored. In order to produce a score for each distractor, we replaced the target word with a distractor in the sentence, then rescored the sentence. Only the grammaticality and collocation criteria measured how well the distractors fit in the sentence, both from a syntactic and from a semantic point of view. We selected the distractors with the highest scores, that is the ones that fit best in the sentence. Thus we prevented them from being obviously wrong answers. However we also wanted to avoid having several possible answers, which would happen if the distractors fit too well in the sentence. We therefore selected distractors that were semantically “far enough” from the target word. To compute semantic similarity between two words, we used Patwardhan and Pedersen’ s method [28]. In this method, two words w1 and w2 are associated with their definition in WordNet. Each word d of the definition is associated with a first order context vector, computed by counting the co-occurrences of d with other words in a corpus4 . Then, computing the resultant (i.e. the sum) of these context vectors produces a second order context vector, which represents the meaning of the word. Finally the dot product of the second order context vectors associated with w1 and w2 give the semantic similarity between w1 and w2. This method, unlike several other methods based on the WordNet hierarchy, handles all parts-of-speech, not just verbs and nouns.
干扰者的质量也得分。为了给每个分心者产生一个分数，我们用句子中的分心者替换了目标词，然后重新判断了句子。只有语法和搭配标准从句法和语义的角度衡量了干扰者在句子中的适应程度。我们选择了得分最高的干扰者，即在句子中最适合的干扰者。因此，我们阻止他们成为明显错误的答案。然而，我们也想避免有几个可能的答案，如果分心者在句子中适合得太好就会发生这种情况。因此，我们选择了与目标词在语义上“足够远”的干扰物。为了计算两个单词之间的语义相似性，我们使用了 Patwardhan 和 Pedersen 的方法[28]。在这种方法中，两个单词 w1 和 w2 与它们在 WordNet 中的定义相关联。定义的每个单词 d 与一阶上下文向量相关联，通过计算与语料库 4 中的其他单词的共同出现来计算。然后，计算这些上下文向量的结果（即总和）产生二阶上下文向量，其表示该单词的含义。最后，与 w1 和 w2 相关联的二阶上下文向量的点积给出了 w1 和 w2 之间的语义相似性。与基于 WordNet 层次结构的其他几种方法不同，此方法处理所有词性，

而不仅仅是动词和名词。
Using distractors that are semantically different from the target word does not guarantee that they will not fit in the sentence. Figure 2 shows that unrelated words can fit in the same sentence. Therefore, this technique will work with a sentence that has few possible answers.
使用与目标词在语义上不同的干扰物并不能保证它们不适合句子。 图 2 显示不相关的单词可以放在同一个句子中。 因此，这种技术将适用于几乎没有答案的句子。

2.3 Stem selection using several dictionaries and a raw text corpus
2.3 使用多个词典和原始文本语料库进行词干选择

We also applied the technique for stem selection to several dictionaries and a raw text corpus, with all parts of speech included. A corpus provides many sentences per word, therefore producing many good quality cloze questions per word. However, unlike dictionaries where sentences are carefully crafted, a large text corpus will contain many useless sentences that the algorithm ought to discard. We did not generate distractors. This test was mainly designed to evaluate the quality of the sentences independently of the quality of the distractors.
我们还将词干选择技术应用于几个词典和原始文本语料库，包括所有词性。 语料库每个单词提供许多句子,因此每个单词产生许多高质量的完形填空问题。然而，与精心制作句子的词典不同，大型文本语料库将包含算法应该丢弃的许多无用句子。 我们没有产生干扰者。 该测试主要是为了评估句子的质量，而不考虑干扰者的质量。

Dictionaries In order to compare sentences from dictionaries and from raw text corpus, we extracted the sample sentences provided by several dictionaries 5 for each word in the Academic Word List [29].
字典 为了比较字典和原始文本语料库中的句子，我们为学术单词列表中的每个单词提取了几个词典 5 提供的样本句子[29]。

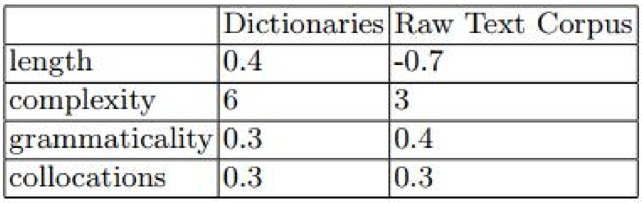
Corpus preprocessing A ten million word subset of the REAP documents, which are gathered from the Web, was filtered for text quality by running a partof-speech (POS) tagger on each document and computing the cosine similarity between the vectors of POS trigrams in the document and a corpus of English literature, known to be grammatical and of high quality. HTML tags of the documents were stripped out. The raw text was chunked into sentences using the sentence detector of OpenNLP toolkit [30]. Only the sentences containing words from the Academic Word List were retained.
语料库预处理 通过在每个文档上运行部分语音（POS）标记器并计算文档中 POS 三元组向量之间的余弦相似性，从 Web 收集的 REAP 文档的一千万字的子集被过滤以获得文本质量。 英语文学语料库，已知具有语法和高质量。 文件的 HTML 标签被删除了。 使用 OpenNLP 工具包的句子检测器将原始文本分成句子[30]。

只保留包含学术词汇表中单词的句子。

Parameters We used two different sets of weights for dictionaries and for the raw text corpus, shown in Tab. 1. In dictionaries, sentences tend to be too short, therefore the length weight is positive. On the contrary, sentences found in raw text are very long. To avoid long sentences, the length weight is negative. This way, we expect to find a trade-off between well-defined context and length. Furthermore, sentences of more than 30 words were discarded as not suitable for a practice exercise. A nonlinear function of the length could also have discarded sentences that are too short or too long. Dictionary sentences often lack a verb, which is why the complexity weight is high. The grammaticality weight is slightly higher for raw text because dictionary sentences are usually well-formed.

参数 我们在字典和原始文本语料库中使用了两组不同的权重,如 Tab1 所示。在词典中，句子往往太短，因此长度权重是正的。 相反，原始文本中的句子很长。为避免长句，长度权重为负。 这样，我们期望在明确定义的上下文和长度之间找到折衷。 此外，超过 30 个单词的句子被丢弃，因为不适合练习。 长度的非线性函数也可以丢弃太短或太长的句子。 字典句子通常缺少动词，这就是复杂性权重高的原因。 原始文本的语法重量略高，因为字典句子通常是格式良好的。

Table 1. Different choice of weights for dictionaries and raw text corpus



	Dictionaries	Raw Text Corpus
length	0.4	-0.7
complexity	6	3
grammaticality	0.3	0.4
collocations	0.3	0.3

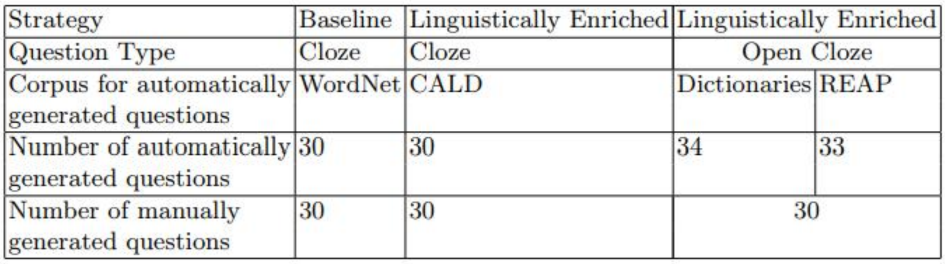
3 Evaluation

3 评估

A series of three experiments were conducted. In each of them, five ESL teachers, two of whom are native speakers, assessed the quality of a set of questions. This set consisted of manually-generated and automatically-generated questions displayed in random order. Open cloze questions were generated to measure how well-defined the context was independently of the choice of distractors.The experiment settings are detailed in Tab. 2.

进行了一系列的三个实验。 在他们每个人中，有五位 ESL 教师，其中两位是母语人士，评估了一组问题的质量。 该集合包括以随机顺序显示的手动生成和自动生成的问题。 生成开放式完形填空问题以测量上下文与受干扰者的选择无关的定义。实验设置在表格 2 中详细说明。

Table 2. Experiment Settings



Strategy	Baseline	Linguistically Enriched	Linguistically Enriched	
Question Type	Cloze	Cloze	Open Cloze	
Corpus for automatically generated questions	WordNet	CALD	Dictionaries	REAP
Number of automatically generated questions	30	30	34	33
Number of manually generated questions	30	30	30	

For each cloze question, the teachers answered a series of questions, illustrated in Fig. 3. The questions targeted specific reasons regarding the suitability of cloze questions.
对于每个完形填空问题，教师回答了一系列问题，如图 3 所示。这些问题针对完形填空问题的适用性的具体原因。

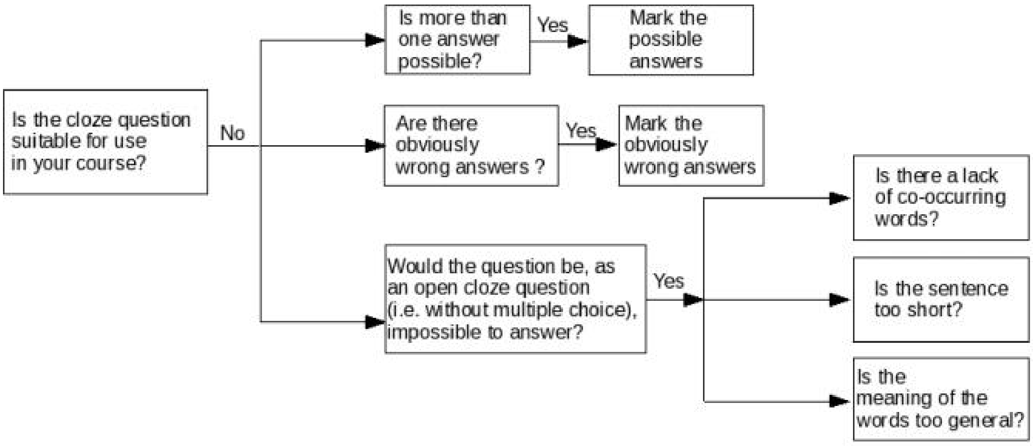

Is the cloze question suitable for use in your course?
No
Is more than one answer possible?
Yes
Mark the possible answers
Are there obviously wrong answers ?
Yes
Mark the obviously wrong answers
Would the question be, as an open cloze question (i.e. without multiple choice), impossible to answer?
Yes
Is there a lack of co-occurring words?
Is the sentence too short?
Is the meaning of the words too general?

Fig. 3. Flow chart of the cloze question evaluation process

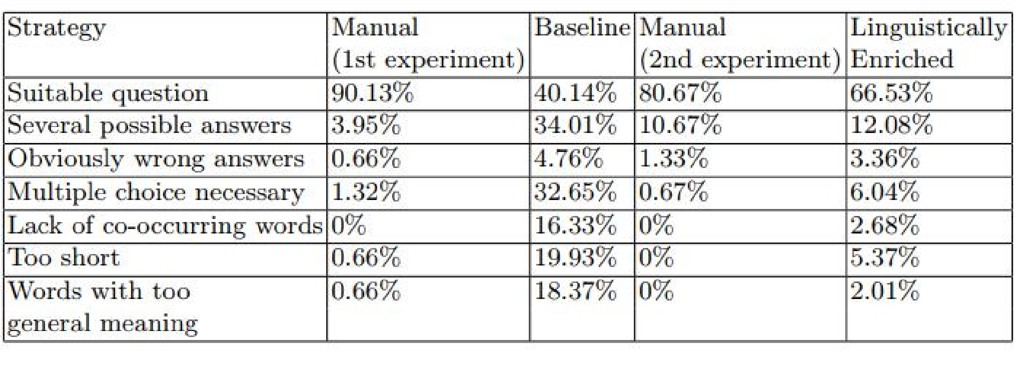
For open cloze questions, the teachers assessed the quality of the context on a one-to-four scale. They also indicated if the sentence was too long, too short or of the right length, and if the sentence was too simple, too difficult or at the appropriate level of English for upper-intermediate ESL students.
对于开放的完形填空问题，教师以一到四的比例评估背景的质量。 他们还表示，如果句子太长，太短或者长度不合适，并且句子对于中高级 ESL 学生而言过于简单，太难或适当的英语水平。

3.1 Cloze Questions
3.1 完形填空问题

The same teachers evaluated randomly ordered questions for both the baseline technique and the proposed technique. The target words for these questions were randomly sampled with replacement from a pool a word in both cases. Overall, the teachers judgments had a Fleiss' kappa6 agreement of 0.3. We therefore consider the agreement between the teachers to be “fair” [31]. We expect that agreement would have been higher if better training procedures had been in place, such as allowing teachers to practice on an independent set of items and then discuss their differences in order to better calibrate their judgments. It should be also noted that these teachers teach at a variety of levels, in France and the United States.
相同的教师评估了基线技术和提出的技术的随机排序问题。 这两个问题的目标词是随机抽样的，在两种情况下从池中替换一个词。 总体而言，教师判决的Fleiss'kappa6 协议为 0.3。 因此，我们认为教师之间的协议是“公平的”[31]。如果有更好的培训程序，我们希望协议会更高，例如允许教师练习一套独立的项目，然后讨论他们的差异，以便更好地校准他们的判断。 还应该指出的是，这些教师在法国和美国的各个层面进行教学。

Table 3 summarizes the results of the first two experiments about cloze questions. 40.14% of the questions generated by the baseline technique were judged acceptable, while our proposed strategy generated 66.53% of suitable questions.The difference was statistically significant (t(29) = 2.048, p < 0.025). This improved performance is still too low to plug the generated questions directly in a tutoring system; however, it appears to be high enough to allow us to build an authoring tool to improve authoring efficiency.
表 3 总结了关于完形填空问题的前两个实验的结果。 基线技术产生的问题中有 40.14%被认为是可接受的，而我们提出的策略产生了 66.53%的合适问题。差异具有统计学意义（t（29）= 2.048，p <0.025）。 这种改进的性能仍然太低，无法直接在辅导系统中插入生成的问题；但是，它似乎足以让我们构建一个创作工具来提高创作效率。

Table 3. Assessment of cloze questions



Strategy	Manual (1st experiment)	Baseline	Manual (2nd experiment)	Linguistically Enriched
Suitable question	90.13%	40.14%	80.67%	66.53%
Several possible answers	3.95%	34.01%	10.67%	12.08%
Obviously wrong answers	0.66%	4.76%	1.33%	3.36%
Multiple choice necessary	1.32%	32.65%	0.67%	6.04%
Lack of co-occurring words	0%	16.33%	0%	2.68%
Too short	0.66%	19.93%	0%	5.37%
Words with too general meaning	0.66%	18.37%	0%	2.01%

The most often cited reason for unacceptable questions was that several answers were possible. It was given for 34.01% of the baseline questions and 12.08% of the questions generated with the new strategy. In the baseline technique, there was no verification to determine if the distractors were semantically far enough from the target word. This flaw was corrected in the new strategy, but there is still room for improvement of the quality of the distractors. The second reason for unacceptable questions was that these questions, as open cloze questions, were impossible to answer, either because of a lack of co-occurring words, overly short sentences, or words with too general meaning. Again, the proposed strategy made a significant improvement over the baseline for all these aspects. However, the kappa values for inter-rater reliability were much lower for these points. The presence of obviously wrong answers as a reason for not acceptable questions was not often cited by the assessors either in the baseline strategy, or in the proposed strategy, probably because the distractors had the same part of speech as the correct answer and fit in the sentence at least grammatically.
不可接受的问题最常被引用的原因是有几个答案是可能的。它给出了 34.01％的基线问题和 12.08％的新策略产生的问题。在基线技术中，没有验证来确定干扰物在语义上是否远离目标词。这一缺陷在新战略中得到了纠正，但仍有改善干扰者素质的空间。不可接受的问题的第二个原因是，这些问题，如开放式完形填空问题，无法回答，或者是因为缺少共同发生的单词，过于短的句子，或者具有过于笼统意义的词语。同样，拟议的战略对所有这些方面的基线做出了重大改进。然而，这些点的评估者间可靠性的 kappa 值要低得多。评估人员经常在基线策略或拟议策略中引用明显错误答案作为不可接受问题的理由，可能是因为干扰者与正确答案具有相同的词性并且适合于句子至少是语法上的。

3.2 Open Cloze Questions
3.2 打开完形填空问题

Figures 4 and 5 show the distribution of the questions at levels 3 and 4 of context quality and for each set of question (manual, generated from dictionaries or from raw text). We consider that a sentence at level 3 and 4 of context quality is acceptable for an open cloze question. 61.82% of the automatically-generated questions for dictionaries are at levels 3 and 4 while 71.23% of the manuallygenerated questions were at levels 3 and 4.
图 4 和图 5 显示了上下文质量和每组问题（手册，从字典或原始文本生成）的第 3 和第 4 级问题的分布。 我们认为对于开放式完形填空问题，上下文质量级别 3 和 4 的句子是可接受的。 自动生成的词典问题的 61.82％处于 3 级和 4 级，而 71.23％的手动生成的问题处于 3 级和 4 级。

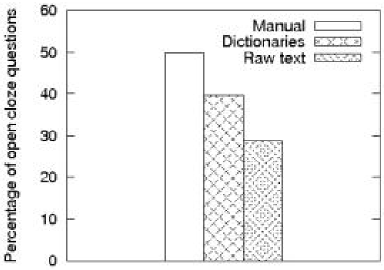


Fig. 4. Distribution of open cloze questions at context level 3 (well-defined context, two or three words can fit in the sentence)

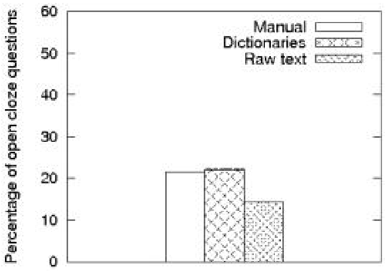
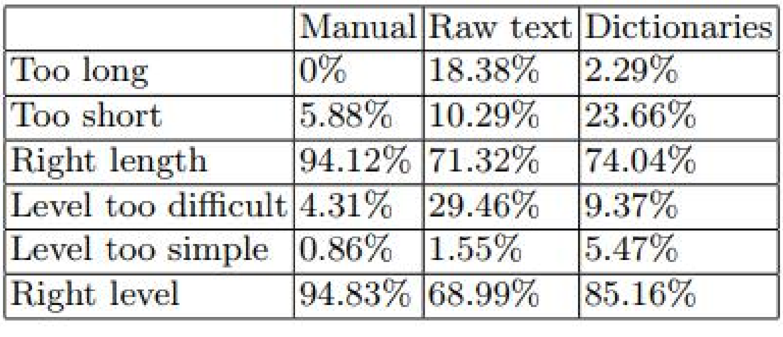


Fig. 5. Distribution of open cloze questions at context level 4 (very well-defined context, only one word can fit in the sentence)

When applied to dictionaries, our strategy is comparable to the human method for generating stems. The main difficulty encountered when generating cloze questions was the choice of the distractors. We can therefore focus on this task since the first one, namely generating stems, is relatively well mastered. However, the strategy did not perform as well on raw text. In raw text, the context is usually well-defined over several sentences but it is hard to find a single sentence that contains a rich context in itself. Analysis of length and level of difficulty shows that dictionary sentences tend to be too short but with the right level of difficulty and that raw text sentences tend to be too long and too difficult, as shown in Tab. 4.
当应用于词典时，我们的策略可与人类产生词干的方法相媲美。 生成完形填空问题时遇到的主要困难是选择分散注意力的人。 因此，我们可以专注于这项任务，因为第一个，即产生茎，相对较好掌握。 但是，该策略在原始文本上表现不佳。 在原始文本中，上下文通常在几个句子中被很好地定义，但是很难找到包含丰富上下文的单个句子。 对长度和难易程度的分析表明，字典句子往往太短但具有适当的难度，原始文本句子往往太长而且太难，如表格 4 所示。

Table 4. Length and difficulty level for open cloze questions



	Manual	Raw text	Dictionaries
Too long	0%	18.38%	2.29%
Too short	5.88%	10.29%	23.66%
Right length	94.12%	71.32%	74.04%
Level too difficult	4.31%	29.46%	9.37%
Level too simple	0.86%	1.55%	5.47%
Right level	94.83%	68.99%	85.16%

4 Conclusion and Future Work

4 结论和未来的工作

We have developed a strategy for selecting high quality cloze questions. This strategy was prompted by a baseline technique developed within the framework of the REAP system. It was developed by taking into account the weak points of the baseline technique shown by a first evaluation. The evaluation of our strategy showed that we are able to generate stems of good quality, and therefore open cloze questions of good quality. We believe that by generating distractors of better quality, for example using the technique described in [19], we will be able to improve the automatic generation of cloze questions. Our technique can be extended to other languages by using different parsers, dictionaries and corpora. It can also be used as a measure for the amount of context, or information, a sentence provides.

我们制定了一个选择高质量完形填空问题的策略。该策略是由 REAP 系统框架内开发的基线技术提出的。 它是通过考虑第一次评估所显示的基线技术的弱点而开发的。 对我们战略的评估表明，我们能够生成高质量的茎，因此打开质量好的完形填空问题。 我们相信通过产生更好质量的干扰物，例如使用[19]中描述的技术，我们将能够改进完形填空问题的自动生成。 我们的技术可以通过使用不同的解析器，词典和语料库扩展到其他语言。 它还可以用作句子提供的上下文量或信息量的度量。

We face two main challenges. First, distractors that fit in a sentence grammatically often also fit semantically. Choosing distractors with a large semantic distance from the correct answer does not always solve this problem. Similarly, open cloze questions rarely have only one possible answer. Second, corpus sentences inherently differ from dictionary sample sentences because the same amount of context is defined in several corpus sentences and in one dictionary sentence only.

我们面临两个主要挑战。 首先，在语法上适合句子的干扰者通常也在语义上适合。 选择与正确答案具有较大语义距离的干扰物并不总是能解决这个问题。 同样，开放式完形填空问题很少只有一个可能的答案。 其次，语料库句子本质上不同于字典样本句子，因为相同数量的上下文仅在几个语料库句子和一个字典句子中定义。

At a higher level, we have demonstrated the utility of linguistically motivated, statistical approaches for generating assessment and practice materials in the ill-defined domain of English vocabulary learning. Ill-defined domains often deal with processes, especially linguistic ones, that are probabilistic in nature or at least possess a degree of complexity which make them challenging to model with deterministic algorithms. As such, statistical methods, such as those applied in our approach to generating vocabulary assessments, can be particularly effective for

developing educational technology for ill-defined domains
在更高的层次上，我们已经证明了语言驱动的统计方法在英语词汇学习界限不明确的领域中产生评估和实践材料的效用。 不明确定义的域通常处理过程，特别是语言过程，这些过程本质上是概率性的，或者至少具有一定程度的复杂性，这使得它们难以用确定性算法建模。 因此，统计方法，例如我们用于生成词汇评估的方法中使用的统计方法，对于为不明确的领域开发教育技术可能特别有效。

5 Acknowledgments

We would like to thank Gregory Mizera, Stacy Ranson, Catherine Jutteau, Ren´ee Maufroid, Genevi`eve Jaslier and Annie Nottebaert for assessing the questions. We also would like to thank Catherine Jutteau and Sarah Carvalho for reviewing this paper and Le Zhao and Anagha Kulkarni for scientific input. This material is based on work supported by NSF grant SBE-0354420, the research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B040063 to Carnegie Mellon University. Any opinions, findings, conclusions or recommendations expressed in this material are the authors', and do not necessarily reflect those of the sponsor.

5 致谢

我们要感谢 Gregory Mizera，Stacy Ranson，Catherine Jutteau，Ren'ee Maufroid，Genevi`eve Jaslier 和 Annie Nottebaert 评估问题。 我们还要感谢 Catherine Jutteau 和 Sarah Carvalho 审阅本文以及 Le Zhao 和 Anagha Kulkarni 的科学意见。 该材料基于 NSF 资助 SBE-0354420 支持的工作，此处报告的研究部分由美国教育部教育科学研究所通过格兰特 R305B040063 向卡内基梅隆大学提供支持。 本材料中表达的任何意见，发现，结论或建议均为作者，并不一定反映赞助商的意见，发现，结论或建议。