ISCA Archive
http://www.isca-speech.org/archive

Speech and Language Technology in Education (SLaTE2009)
Wroxall Abbey Estate, Warwickshire, UK
September 3-5, 2009

Automated Generation of Example Contexts for Helping Children Learn Vocabulary

Liu Liu, Jack Mostow, and Gregory Aist

Project LISTEN, School of Computer Science
Carnegie Mellon University, Pittsburgh, Pennsylvania USA
{liuliu, mostow}@cs.cmu.edu, gregory.aist@alumni.cmu.edu

Abstract[1]

This paper addresses the problem of generating good example contexts to help children learn vocabulary. We construct candidate contexts from the Google N-gram corpus. We propose a set of constraints on good contexts, and use them to filter candidate example contexts. We evaluate the automatically generated contexts by comparison to example contexts from children's dictionaries and from children's stories.

1. Introduction

Vocabulary plays a critical role in reading comprehension. "A reader who can pronounce a word but does not know its meaning or crucial facts about it is at a disadvantage in comprehending the text in which it occurs" [1].

This paper focuses on one particular aspect of vocabulary learning – learning word meanings from example contexts. Word meaning includes both *denotation* (explicit definition) and *connotation* (implied meaning and associations) [2]. Readers must acquire both aspects, so effective vocabulary instruction combines explicit explanation with multiple encounters in varied contexts [3].

Contexts give clues to semantics but also convey many other different lexical aspects, such as part of speech, morphology, and pragmatics, which help enrich children's word knowledge base. However, not all contexts are equally useful; in fact, most natural contexts are insufficient to infer word meaning [4], especially for younger readers.

Accordingly, one key issue in vocabulary instruction is how to find or create good example contexts to help children learn a word. Context examples are usually created by teachers, lexicographers, or occasionally educational researchers [3, 5]. A human expert may generate excellent examples, but takes time, costs money, and may not be available when needed. Also, human-generated contexts are shaped by the cognitive retrieval and production processes of a person who knows the word, and may therefore overlook important uses. In contrast, computer-generated contexts can provide systematic, comprehensive coverage, and address specific learning goals.

[1] This work was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A080157 to Carnegie Mellon University. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute or the U.S. Department of Education. We thank Dr. Margaret McKeown.

This paper describes a system that generates example contexts to help children in grades 2-3 learn targeted vocabulary, by using language resources and multiple NLP technologies. The rest of this paper is organized as follows. Section 2 describes how our context generator uses Google N-gram data [6] to generate candidate contexts. Section 3 describes constraints on good contexts, and filters to operationalize them. Section 4 evaluates automatically generated example contexts against human-authored examples. Section 5 discusses limitations and future work. Section 6 concludes.

2. Context generation

Our contexts are sequences of overlapping Google n-grams.

2.1. Data set

The Google N-gram data set [6] contains billions of n-grams and their frequencies, based on over one trillion words of text extracted from public web pages and segmented into sentences. It has been used in many areas, including spelling correction and machine translation. The data set contains n-grams for *n* from one to five. We use five-grams in generating contexts, as five-grams are the longest so they provide more information about the target word. The data set contains all 1,176,470,663 five-grams that appeared at least 40 times, e.g.:

advantage in a competitive </S>	42
advantage in a competitive environment	66
advantage in a competitive job	69
advantage in a competitive market	219
advantage in a competitive world	94

Here *</S>* is the symbol for the end of a sentence, and the number after each five-gram is its frequency.

The entire data set is about 200 GB including indices, so we extract only the five-grams containing target vocabulary words to teach, and then save the five-grams for each target word in a separate database table to allow efficient access.

2.2. Generation method

Given a target vocabulary word, e.g. ***extinct***, the context generation process works as follows. First, choose a five-gram containing the target word as the initial context, e.g.:

*been **extinct** for millions of*

Then, repeatedly extend it one word to the left or right by choosing a five-gram (underlined here) that matches the first or last four words, e.g.:

*been **extinct** for millions of years*
*have been **extinct** for millions of years*

*Dinosaurs have been **extinct** for millions of years*

Continue until no further extension is possible. Our generator uses only five-grams containing the target word, so it generates sentences at most nine words long, with the target word in the middle and four words on each side of it.

This method is based on the consistency assumption that if one five-gram overlaps with another by four words, then both of them came from the same set of sentences in the original corpus. When this heuristic assumption holds true, the method reconstructs part or all of one of these sentences.

However, when it fails, the method can generate a novel word sequence. We call this phenomenon "crossover" because it combines five-grams from different sentences. The resulting sequence is still locally consistent because each successive five words constitute an authentic five-gram.

On the positive side, this ability to generate novel sentences can potentially produce example contexts that improve on the original sentences, for example by streamlining them to eliminate undesirable complexity. On the negative side, crossover can produce global inconsistencies, as we will see later. Fortunately, the 9-word limit restricts the opportunity for crossover.

2.3. Relation to prior work

Some related work has automated the generation of example sentences. Dowding et al. [7] used a grammar to generate example sentences containing specific words (e.g., ***pressure*** and ***commander*** in the sentence *Measure the **pressure** at the **commander**'s seat*) for targeted help in spoken dialogue systems. Our work involves a different population (children), purpose (vocabulary development), and method (generation using a corpus of n-grams).

Other related work [8-11] has automated the selection of example sentences for vocabulary learning and assessment. Some selection criteria [10, 11] resemble constraints we impose on the generation process. However, the selection methods *extract* complete sentences from an existing language corpus, but our method *generates* context sentences.

So far as we know, ours is the first study that uses Google five-grams to generate example sentences. N-grams aggregate information across sentences, so the frequency of n-grams reflects the typicality of contexts and usage. In contrast, the corpus frequency of most complete sentences is 1, which does not indicate whether or not their word usage is typical.

3. Context constraints

How can we ensure the generated contexts are good for vocabulary learning? We identified several constraints on good contexts, based partly on expert knowledge and partly on analyzing why some generated contexts were bad.

Sections 3.1-3.7 describe each constraint and operationalize it as one or more heuristic filters. These filters eliminate contexts that violate the constraint, or prefer contexts that satisfy it better or more probably. We compiled the filters into a heuristic search procedure using transformations described in [12], but space limitations preclude a detailed description of the resulting procedure.

3.1. Comprehensible to children

We want to generate contexts that assist the vocabulary development of children in primary school. For a context to be useful, the child must understand it. If the context contains many unfamiliar words besides the target word, the child will not understand the context well enough for it to help in learning the target word. For example, the context *It is time to **declare** victory and go home* is reasonably understandable, assuming the child knows the word *victory*. In contrast, any context containing *…penalties of perjury solemnly **declare**…* is useless for teaching ***declare*** to a child who does not know the words *penalties*, *perjury*, or *solemnly*.

One comprehensibility filter excludes examples containing more than two words rated above grade level 2 according to two leveled word lists [13, 14]. This threshold could easily be changed to fit students' reading level. Another filter removes examples containing relative pronouns (such as *who* or *that*), in order to limit sentence complexity.

3.2. Grammatically correct and complete

Good contexts should be complete, grammatical sentences. Some generated candidates are not grammatical, such as the list *Southpaw **Stout** Dem Blog The Scarlet*. Some candidates are incomplete sentences, such as *Jennifer is very **anxious** to know about the*.

To filter out incomplete or ungrammatical contexts, we use the Link Grammar Parser [15], a syntactic dependency parser, as a grammar checker. The parser rejects any context it fails to parse as syntactically valid English. A second filter requires that generated sentences must either start with <*S*> or a capitalized word, or end with </*S*> or punctuation. Another filter excludes sentences that end with modal or auxiliary verbs. The last two filters help favor complete sentences.

3.3. Sense-appropriate

A good context is consistent with target word meaning. A context that uses a different sense of the target word than the meaning to be taught is confusing, not helpful.

To filter out contexts where a word has a part of speech incompatible with its target meaning, the context generator checks the part of speech assigned by the Link Grammar Parser; if it does not match the target meaning, the filter excludes the context. For example, if the target meaning of ***stout*** is *sturdy*, this filter eliminates the context *Grant **Stout** added 16 points* because it uses ***stout*** as a (proper) noun.

A more sophisticated version of this filter would also exclude contexts with the right part of speech but a different sense of the target word. This capability would involve identifying the word sense used in the context and deciding if it is consistent with the target meaning.

3.4. Informative about word meaning

A highly informative context imposes strong semantic constraints on the target word. Experimental study confirmed that "the degree of semantic constraint for individual contexts played a substantial role in learning word meanings" [3].

The context generator operationalizes semantic constraint as multiple filters. One filter prefers longer sentences because they tend to provide richer information. Another filter prefers content words (such as nouns and verbs) because they tend to provide more meaning than function words. It eliminates sentences that contain fewer than three content words. A third filter specifically prefers words related to the

target word, i.e., that co-occur with the target word in many of the same five-grams in the corpus. It requires the initial five-gram to contain one or more related words.

Overall, these filters prefer sentences that contain more words overall, more content words, and more related words. For example, consider these two contexts:

*Find the strength and **courage** to take risks*

*We know it takes **courage** to do so*

Both contexts are 8 words long, but the filters prefer the first context because it contains more content words, including *strength* and *take*, which are related to ***courage***.

3.5. Ordinary prose

Good contexts use normal, classroom-appropriate English. However, we noticed that some of the generated contexts were very web specific, and likely unfamiliar to young children, e.g., *a **Merchant** ID and password*.

A filter to avoid web jargon eliminates contexts containing words much more common on the web than in print, such as *copyright*, *password*, and *download*, whose unigram frequencies in the Google corpus are disproportionally higher than in a conventional text corpus. Similar filters exclude sentences containing words from a list of taboo words, or special symbols such as @; sentences containing capitalized words other than the first word, the target word, or named entities; and sentences with more than four consecutive numerals or capitalized words.

3.6. Typical of usage and situation

Typicality is an important property of good contexts. They should show how words are commonly used, and in what situations. For example, ***celebrate*** is often used in situations like birthdays and anniversaries. We rely on five-gram frequency to quantify typicality.

Accordingly, one filter prefers high-frequency five-grams.

3.7. Varied and not redundant

Children need to see a word in several varied contexts to decontextualize their knowledge of the word's meaning and acquire enough retrieval cues to access it reliably and efficiently [3]. The Google corpus is large enough to generate diverse contexts for a target word, e.g.:

*Members are asked to **declare** that you are 18*

*He was forced to **declare** a state of emergency*

*It is time to **declare** victory and go home*

However, some generated contexts are very similar, e.g.:

*Just **declare** victory and go home*

*We should **declare** victory and go home*

*It's time to **declare** victory and go home*

A filter to eliminate such redundancy clusters the generated contexts and picks only one from each cluster.

4. Evaluation

How good are the generated contexts? Section 4.1 describes how we evaluated them. Section 4.2 presents the results.

4.1. Methodology

To evaluate our method, we selected 10 target words, generated contexts for them, and compared the generated contexts against human-authored contexts.

To choose the words, we started with the 789 words in Reading Tutor stories that our vocabulary expert Dr. Margaret McKeown had classified as "Tier 2" words [4], i.e., words used in many domains but unknown to most children, and thus important to teach. Of the 15 such words that occurred in exactly two stories, once in each story, we excluded 5 words with multiple parts of speech, and chose the other 10: ***anxious***, ***courage***, ***declare***, ***extinct***, ***merchant***, ***remarkable***, ***slender***, ***stout*** (because we didn't think of its noun sense), ***suspicious***, and ***tremendous***.

Of the contexts generated for each of these target words, we used the 6 rated highest by the context generator. For comparison we chose two types of human-authored contexts. As a sample of naturalistic contexts in which the child would encounter the word during normal reading, we used the two Reading Tutor story sentences containing the word. As a gold standard, we used all 1-3 example sentences from the WordSmyth children's dictionary (www.wordsmyth.net), crafted to illustrate the meaning of each word sense listed. The three source types totaled 98 contexts: 57 generated contexts, 20 story sentences, and 21 dictionary sentences.

Dr. McKeown scored all 98 sentences, blind to source type, on a five-point Likert scale (1=bad, 3=OK, 5=good), both in general quality, and on three specific aspects that influence it: (1) good use of words, i.e. correct or meaningful use in the intended target sense; (2) the degree to which the context is constraining, or reveals elements of the word meaning; (3) comprehensibility to children based on other words or concepts in the context, or syntactic complexity.

4.2. Results and discussions

Table 1 shows mean scores and standard errors for each type of context. ANOVA showed significant main effects for context source on all four measures. Pairwise comparison showed that dictionary contexts surpassed automatically generated examples in general score ($p<0.001$), in good use of words ($p<0.05$), in constraining context ($p<0.05$), and in comprehensibility to children ($p<0.05$). There was also a trend for the story sentences to be better than the generated contexts in general score ($p=0.051$). No other differences were significant.

Table 1: Expert scoring of contexts

Evaluation criteria	Mean (Standard Error)			
	Auto (all)	Auto (top half)	Story	Dictionary
General score	2.5 (0.21)	3.9 (0.15)	3.4 (0.28)	4.1 (0.21)
Good use	3.4 (0.22)	4.2 (0.21)	4.0 (0.28)	4.4 (0.20)
Constraining context	3.2 (0.21)	3.9 (0.19)	3.6 (0.21)	4.1 (0.19)
Comprehensibility	2.7 (0.23)	4.2 (0.18)	3.5 (0.33)	4.1 (0.25)

The top-scored half of the generated sentences compared favorably to story sentences, which suggests that refining the generator to filter out the bottom half would make its output as good as story sentences. Section 5 analyzes the bottom half to identify the main problems. However, predictions of the resulting performance are overly optimistic because we "tested on the training data" in that we designed some of the filters to eliminate bad contexts generated for the 10 test

words. This approach made sense as a first step; future work will test performance on unseen words.

5. Limitations and future work

We identified problems in automatically generated contexts on multiple levels, and possible approaches to some of them.

On the syntactic level, some generated sentences are incomplete or ungrammatical. To fix incomplete sentences, we plan to concatenate n-grams that start with <*S*> or end with </*S*> to lengthen and complete them. To filter out ungrammatical sentences more thoroughly, a better grammar checker would help. We also plan to explore the syntactic structure of generated sentences and restrict them to satisfy some syntactic constraints. For example, we could stick to sentences with simple parses such as [S [NP VP]] to improve comprehensibility, or with syntactic structures characteristic of more informative contexts. Such structures might be induced by analyzing a sufficiently large set of good sentences, such as a corpus of dictionary examples.

On the semantic and pragmatic levels, we found long-distance mismatches in some generated sentences, whose left half is not consistent with their right half, due to crossover. For example, *I will have a* ***tremendous*** *impact on my life* is semantically acceptable, but pragmatically problematic due to being self-evident. Another issue pertains to our particular application: non-kid-friendly sentences should not be used for vocabulary learning. Non-kid-friendly sentences include contexts unfamiliar to children, e.g. legal statements, and contexts that are inappropriate. Although we filter out sentences containing taboo words, some generated contexts are still inappropriate even though each word is fine, e.g. *She reaches her* ***slender*** *fingers towards my exploding*. Human judgment will likely remain necessary to detect such cases.

Another type of low-scored sentence is the spuriously high frequency context, e.g. *Please check the* ***merchant*** *store*. Such sentences are composed of five-grams with high frequency, and chosen by our context generator because it uses average five-gram frequency as a proxy for typicality of usage. However, their high frequency is not because they are actually very common in English but rather due to replication of documents on the Web. To combat this effect, we will start with high-frequency trigrams, extending them with five-grams. Because trigrams are shorter than five-grams, they occur much more often, so their frequencies are less distorted by replicated sentences, and hence reflect typicality better.

6. Conclusion

This paper makes three contributions to automated generation of good example contexts to help children learn vocabulary.

First, we introduce the problem of automatic context generation for learning vocabulary. Although the importance of context to learning vocabulary is well-known, context examples used in education have been created by hand and we know of no prior work to automate their generation.

Second, we show how to generate contexts by combining Google five-grams. We identify several constraints on good example contexts, and filters to operationalize them.

Third, we evaluate against dictionary examples and story sentences based on expert scoring. The top half of the generated contexts average as good as or better than the story sentences in which children would normally encounter them.

References

[1] K. Stanovich, R. West, and A. E. Cunningham, "Beyond phonological processes: Print exposure and orthographic processing," in *Phonological Processes in Literacy*, D. Shankweiler, Ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1992.

[2] D. Chandler, *Semiotics: The Basics*, 2 ed: Routledge, 2004.

[3] D. J. Bolger, M. Balass, E. Landen, and C. A. Perfetti, "Contextual variation and definitions in learning the meanings of words: An instance-based learning approach," *Discourse Processes*, vol. 45, pp. 122-159, 2008.

[4] I. L. Beck, M. G. McKeown, and L. Kucan, *Bringing Words to Life: Robust Vocabulary Instruction*. NY: Guilford, 2002.

[5] M. G. McKeown, "The acquisition of word meaning from context by children of high and low ability," *Reading Research Quarterly*, vol. 20, pp. 482-496, 1985.

[6] T. Brants and A. Franz, "Web 1T 5-gram Version 1," in *Linguistic Data Consortium, Philadelphia*, 2006.

[7] J. Dowding, G. Aist, B. A. Hockey, and E. O. Bratt, "Generating Canonical Example Sentences using Candidate Words," in AAAI Spring Symposium on Natural Language Generation in Spoken and Written Dialogue, Palo Alto, California, 2003, pp. 23-27.

[8] C.-L. Liu, C.-H. Wang, Z.-M. Gao, and S.-M. Huang, "Applications of lexical information for algorithmically composing multiple-choice cloze items," in Proceedings of the Second Workshop on Building Educational Applications Using NLP, Ann Arbor, Michigan, 2005, pp. 1-8.

[9] J. C. Brown, G. A. Frishkoff, and M. Eskenazi, "Automatic Question Generation for Vocabulary Assessment," in Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, B.C., Canada, 2005, pp. 819-826.

[10] J. Pino, M. Heilman, and M. Eskenazi, "A selection strategy to improve cloze question quality," in Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains. 9th International Conference on Intelligent Tutoring Systems, Montreal, Canada, 2008, pp. 22-32.

[11] T. K. Landauer, K. Kireyev, and C. Panaccione, "A New Yardstick and Tool for Personalized Vocabulary Building," in The 4th Workshop on Innovative Use of NLP for Building Educational Applications, Boulder, CO, USA, 2009, pp. 27-33.

[12] D. J. Mostow, "Machine transformation of advice into a heuristic search procedure," in *Machine Learning: An Artificial Intelligence Approach*, R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, Eds. Palo Alto, CA: Tioga: Springer, 1983, pp. 367-403.

[13] A. Biemiller, "Words worth teaching," *Columbus, OH: SRA/McGraw-Hill*, in press, 2008.

[14] E. B. Diane E. Paynter, Jane K. Doty, Nell K. Duke, "For the Love of Words: Vocabulary Instruction that Works, Grades K-6," pp. 127-202, 2005.

[15] D. Sleator and D. Temperley, "Parsing English with a link grammar," in Third International Workshop on Parsing Technologies, 1993, pp. 277-292.