A Challenge Set for Advancing Language Modeling

Geoffrey Zweig and Chris J.C. Burges
Microsoft Research
Redmond, WA 98052

Abstract

In this paper, we describe a new, publicly available corpus intended to stimulate research into language modeling techniques which are sensitive to overall sentence coherence. The task uses the Scholastic Aptitude Test's sentence completion format. The test set consists of 1040 sentences, each of which is missing a content word. The goal is to select the correct replacement from amongst five alternates. In general, all of the options are syntactically valid, and reasonable with respect to local N-gram statistics. The set was generated by using an N-gram language model to generate a long list of likely words, given the immediate context. These options were then hand-groomed, to identify four decoys which are globally incoherent, yet syntactically correct. To ensure the right to public distribution, all the data is derived from out-of-copyright materials from Project Gutenberg. The test sentences were derived from five of Conan Doyle's Sherlock Holmes novels, and we provide a large set of Nineteenth and early Twentieth Century texts as training material.

1 Introduction

Perhaps beginning with Claude Shannon's use of N-gram statistics to compute the perplexity of letter sequences (Shannon and Weaver, 1949), N-gram models have grown to be the most commonly used type of language model in human language technologies. At the word level, N-gram modeling techniques have been extensively refined, with state-of-the-art techniques based on smoothed N-gram counts (Kneser and Ney, 1995; Chen and Goodman, 1999), multi-layer perceptrons (Schwenk and Gauvain, 2002; Schwenk, 2007) and maximum-entropy models (Rosenfeld, 1997; Chen, 2009a; Chen, 2009b). Trained on large amounts of data, these methods have proven very effective in both speech recognition and machine translation applications.

Concurrent with the refinement of N-gram modeling techniques, there has been an important stream of research focused on the incorporation of syntactic and semantic information (Chelba and Jelinek, 1998; Chelba and Jelinek, 2000; Rosenfeld et al., 2001; Yamada and Knight, 2001; Khudanpur and Wu, 2000; Wu and Khudanpur, 1999). Since intuitively, language is about expressing *meaning* in a highly structured syntactic form, it has come as something of a surprise that the improvements from these methods have been modest, and the methods have yet to be widely adopted in non-research systems.

One explanation for this is that the tasks to which language modeling has been most extensively applied are largely soluble with local information. In the speech recognition application, there is a fundamental confluence of acoustic and linguistic information, and the language model can be thought of as resolving ambiguity only between acoustically confusable words (Printz and Olsen, 2002). Since words which are acoustically similar, e.g. "bill" and "spill" usually appear in very different textual contexts, the local information of an N-gram language model may be adequate to distinguish them. To a lesser degree, in a machine translation application,

1. One of the characters in Milton Murayama's novel is considered ______ because he deliberately defies an oppressive hierarchical society.
(A) rebellious (B) impulsive (C) artistic (D) industrious (E) tyrannical

2. Whether substances are medicines or poisons often depends on dosage, for substances that are _______ in small doses can be ______ in large.
(A) useless .. effective
(B) mild .. benign
(C) curative .. toxic
(D) harmful .. fatal
(E) beneficial .. miraculous

Figure 1: Sample sentence completion questions (Educational-Testing-Service, 2011).

the potential phrase translations may be similar in meaning and local information may again suffice to make a good selection.

In this paper, we present a language processing corpus which has been explicitly designed to be non-solvable using purely N-gram based methods, and which instead requires some level of semantic processing. To do this, we draw inspiration from the standardized testing paradigm, and propose a *sentence completion* task along the lines of that found in the widely used Scholastic Aptitude Test. In this type of question, one is given a sentence with one or two words removed, and asked to select from among a set of five possible insertions. Two examples of SAT test questions are shown in Figure 1.

As can be seen, the options available all make sense from the local N-gram point of view, and are all syntactically valid; only semantic considerations allow the correct answer to be distinguished. We believe this sort of question is useful for two key reasons: first, its full solution will require language modeling techniques which are qualitatively different than N-grams; and secondly, the basic task formulation has been externally determined and is a widely used method for assessing human abilities. Unfortunately, to date no publicly available corpus of such questions has been released.

The contribution of this work is to release a public corpus of sentence completion questions designed to stimulate research in language modeling technology which moves beyond N-grams to explicitly address global sentence coherence. The corpus is based purely on out-of-copyright data from Project Gutenberg, thus allowing us to distribute it. The test questions consist of sentences taken from five Sherlock Holmes novels. In each, a word has been removed, and the task is to choose from among five alternatives. One of the options is the original word, and the other four "decoys" have been generated from an N-gram language model using local context. Sampling from an N-gram model is done to generate alternates which make sense locally, but for which there is no other reason to expect them to make sense globally. To ensure that synonyms of the correct answer are not present, and that the options are syntactically reasonable, the decoys have been hand selected from among a large number of possibilities suggested by the N-gram model. The training data consists of approximately 500 out-of-copyright Nineteenth and early Twentieth century novels, also from Project Gutenberg.

We expect that the successful development of models of global coherence will be useful in a variety of tasks, including:

- the interactive generation of sentence completion questions for vocabulary tutoring applications;
- proof-reading;
- automated grading of essays and other student work; and
- sentence generation in free-form dialog applications.

The remainder of this paper is organized as follows. In Section 2, we describe the process by which we made the corpus. Section 3 provides guidance as to the proper use of the data. In Section 4, we present baseline results using several simple automated methods for answering the questions. Finally, in Section 5, we discuss related work.

2 The Question Generation Process

Question generation was done in two steps. First, a candidate sentence containing an infrequent word was selected, and alternates for that word were automatically determined by sampling with an N-gram language model. The N-gram model used the immediate history as context, thus resulting in words that may "look good" locally, but for which there is no a-priori reason to expect them to make sense globally. In the second step, we eliminated choices which are obviously incorrect because they constitute grammatical errors. Choices requiring semantic knowledge and logical inference were preferred, as described in the guidelines, which we give in Section 3. Note that an important *desideratum* guiding the data generation process was requiring that a researcher who knows exactly how the data was created, including knowing which data was used to train the language model, should nevertheless not be able to use that information to solve the problem. We now describe the data that was used, and then describe the two steps in more detail.

2.1 Data Used

Seed sentences were selected from five of Conan Doyle's Sherlock Holmes novels: *The Sign of Four (1890), The Hound of the Baskervilles (1892), The Adventures of Sherlock Holmes (1892), The Memoirs of Sherlock Holmes (1894),* and *The Valley of Fear (1915).* Once a *focus word* within the sentence was selected, alternates to that word were generated using an N-gram language model. This model was trained on approximately 540 texts from the Project Gutenberg collection, consisting mainly of 19th century novels. Of these 522 had adequate headers attesting to lack of copyright, and they are now available at the *Sentence Completion Challenge* website `http://research.microsoft.com/en-us/projects/scc/`.

2.2 Automatically Generating Alternates

Alternates were generated for every sentence containing an infrequent word. A state-of-the-art class-based maximum entropy N-gram model (Chen, 2009b) was used to generate the alternates. Ideally, these alternates would be generated according to $P(\text{alternate}|\text{remainder of sentence})$. This can be done by computing the probability of the completed sentence once for every possible vocabulary word, and then normalizing and sampling. However, the normalization over all words is computationally expensive, and we have used a procedure based on sampling based on the preceding two word history only, and then re-ordering based on a larger context. The following procedure was used:

1. Select a *focus word* with overall frequency less than 10^{-4}. For example, we might select "extraordinary" in "It is really the most extraordinary and inexplicable business."

2. Use the two-word history immediately preceding the selected focus word to predict alternates. We sampled 150 unique alternates at this stage, requiring that they all have frequency less than 10^{-4}. For example, "the most" predicts "handsome" and "luminous."

3. If the original (correct) sentence has a better score than any of these alternates, reject the sentence.

4. Else, score each option according to how well it and its immediate predecessor predict the next word. For example, the probability of "and" following "most handsome" might be 0.012.

5. Sort the predicted words according to this score, and retain the top 30 options.

In step 3, omitting questions for which the correct sentence is the best makes the set of options more difficult to solve with a language model alone. However, by allowing the correct sentence to potentially fall below the set of alternates retained, an opposite bias is created: the language model will tend to assign a lower score to the correct option than to the alternates (which were chosen by virtue of scoring well). We measured the bias by performing a test on the 1,040 test sentences using the language model, and choosing the *lowest* scoring candidate as the answer. This gave an accuracy of 26% (as opposed to 31%, found by taking the highest scoring candidate: recall that a random choice would give 20% in expectation). Thus although there is some remaining bias for the answer to be low scoring, it is small. When a language model other than the precise one used to generate the data is used, the score reversal test yielded 17% correct. The correct polarity gave 39%. If, however, just the single score used to do the sort in the last step is used (i.e. the probability of the immediate successor alone), then the lowest scoring alternate is correct about 38% of the time - almost as much as the language model itself. The use of the word score occurring two positions after the focus also achieves 38%, though a positive polarity is beneficial here. Combined, these scores achieve about 43%. Neither is anywhere close to human performance. We are currently evaluating a second round of test questions, in which we still sample options based on the preceding history, but re-order them according the the total sentence probability $P(w_1 \dots w_N)$.

The overall procedure has the effect of providing options which are both well-predicted by the immediate history, and predictive of the immediate future. Since in total the procedure uses just four consecutive words, it cannot be expected to provide globally coherent alternates. However, sometimes it does produce synonyms to the correct word, as well as syntactically invalid options, which must be weeded out. For this, we examine the alternates by hand.

2.3 Human Grooming

The human judges picked the best four choices of impostor sentences from the automatically generated list of thirty, and were given the following instructions:

1. All chosen sentences should be grammatically correct. For example: *He dances while he ate his pipe* would be illegal.

2. Each correct answer should be unambiguous. In other words, the correct answer should always be a significantly better fit for that sentence than each of the four impostors; it should be possible to write down an explanation as to why the correct answer is the correct answer, that would persuade most reasonable people.

3. Sentences that might cause offense or controversy should be avoided.

4. Ideally the alternatives will require some thought in order to determine the correct answer. For example:

   - *Was she his [client | musings | discomfiture | choice | opportunity] , his friend , or his mistress?*

   would constitute a good test sentence. In order to arrive at the correct answer, the student must notice that, while *"musings"* and *"discomfiture"* are both clearly wrong, the terms *friend* and *mistress* both describe people, which therefore makes *client* a more likely choice than *choice* or *opportunity*.

5. Alternatives that require understanding properties of entities that are mentioned in the sentence are desirable. For example:

   - *All red-headed men who are above the age of [800 | seven | twenty-one | 1,200 | 60,000] years , are eligible.*

   requires that the student realize that a *man* cannot be seven years old, or 800 or more. However, such examples are rare: most often, arriving at the answer will require thought, but not detailed entity knowledge, such as:

   - *That is his [generous | mother's | successful | favorite | main] fault , but on the whole he's a good worker.*

6. Dictionary use is encouraged, if necessary.

7. A given sentence from the set of five novels should only be used once. If more than one focus word has been identified for a sentence (i.e. different focuses have been identified, in different positions), choose the set of sentences that generates the best challenge, according to the above guidelines.

Note that the impostors sometimes constitute a perfectly fine completion, but that in those cases, the correct completion is still clearly identifiable as the most likely completion.

2.4 Sample Questions

Figure 2 shows ten examples of the Holmes derived questions. The full set is available at `http://research.microsoft.com/en-us/projects/scc/`.

3 Guidelines for Use

It is important for users of this data to realize the following: since the test data was taken from five 19th century novels, the test data itself is likely to occur in

1) I have seen it on him , and could _____ to it.
a) write b) migrate c) climb d) swear e) contribute

2) They seize him and use violence towards him in order to make him sign some papers to make over the girl's _____ of which he may be trustee to them.
a) appreciation b) activity c) suspicions d) administration e) fortune

3) My morning's work has not been _____ , since it has proved that he has the very strongest motives for standing in the way of anything of the sort.
a) invisible b) neglected c) overlooked d) wasted e) deliberate

4) It was furred outside by a thick layer of dust , and damp and worms had eaten through the wood , so that a crop of livid fungi was _____ on the inside of it.
a) sleeping b) running c) resounding d) beheaded e) growing

5) Presently he emerged , looking even more _____ than before.
a) instructive b) reassuring c) unprofitable d) flurried e) numerous

6) We took no _____ to hide it.
a) fault b) instructions c) permission d) pains e) fidelity

7) I stared at it _____ , not knowing what was about to issue from it.
a) afterwards b) rapidly c) forever d) horror-stricken e) lightly

8) The probability was , therefore , that she was _____ the truth , or , at least , a part of the truth.
a) addressing b) telling c) selling d) surveying e) undergoing

9) The furniture was scattered about in every direction , with dismantled shelves and open drawers , as if the lady had hurriedly _____ them before her flight.
a) warned b) rebuked c) assigned d) ransacked e) taught

10) The sun had set and _____ was settling over the moor.
a) dusk b) mischief c) success d) disappointment e) laughter

Figure 2: The first ten questions from the Holmes Corpus.

the index of most Web search engines, and in other large scale data-sets that were constructed from web data (for example, the Google N-gram project). For example, entering the string *That is his fault , but on the whole he's a good worker* (one of the sentence examples given above, but with the focus word removed) into the Bing search engine results in the correct (full) sentence at the top position. It is important to realize that researchers may inadvertently get better results than truly warranted because they have used data that is thus tainted by the test set. To help prevent any such criticism from being leveled at a particular publication, we recommend than in any set of published results, the exact data used for training and validation be specified. The training data provided on our website may also be considered "safe" and useful for making comparisons across sites.

4 Baseline Results

4.1 A Simple 4-gram model

As a sanity check we constructed a very simple N-gram model as follows: given a test sentence (with the position of the focus word known), the score for that sentence was initialized to zero, and then incremented by one for each bigram match, by two for each trigram match, and by three for each 4-gram match, where a match means that the N-gram in the test sentence containing the focus word occurs at least once in the background data. This simple method achieved 34% correct (compared to 20% by random choice) on the test set.

4.2 Smoothed N-gram model

As a somewhat more sophisticated baseline, we use the CMU language modeling toolkit [1] to build a 4-gram language model using Good-Turing smoothing. We kept all bigrams and trigrams occurring in the data, as well as four-grams occurring at least twice. We used a vocabulary of the 126k words that occurred five or more times, resulting in a total of 26M N-grams. Sentences were ordered according to their probability according to the language model: $P(w_1 \ldots w_N)$. This improved by 5% absolute on the simple baseline to achieve 39% correct.

4.3 Latent Semantic Analysis Similarity

As a final benchmark, we present scores for a novel method based on latent semantic analysis. In this approach, we treated each sentence in the training data as a "document" and performed latent semantic analysis (Deerwester et al., 1990) to obtain a 300 dimensional vector representation of each word in the vocabulary. Denoting two words by their vectors $\mathbf{x}, \mathbf{y}$, their similarity is defined as the cosine of the angle between them:

$$\mathtt{sim}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\| \mathbf{x} \| \| \mathbf{y} \|}.$$

To decide which option to select, we computed the average similarity to every other word in the sentence, and then output the word with the greatest overall similarity. This results in our best baseline performance, at 49% correct.

4.4 Benchmark Summary

Table 1 summarizes our benchmark study. First, for reference, we had an unaffiliated human answer a random subset of 100 questions. Ninety-one percent were answered correctly, showing that scores in the range of 90% are reasonable to expect. Secondly, we tested the performance of the same model (Model M) that was used to generate the data. Because this model output alternates that it assigns high-probability, there is a bias against it, and it scored 31%. Smoothed 3 and 4-gram models built with the CMU toolkit achieved 36 to 39 percent. Recall that the sampling process introduced some bias into the word scores at specific positions relative to the focus word. Exploiting the negative bias induced on the immediately following word, and combining it with the score of the word two positions in the future, we were able to obtain 43%. The simple 4-gram model described earlier did somewhat worse than the other N-gram language models, and the LSA similarity model did best with 49%. As a further check on this data, we have run the same tests on 108 sentence completion questions from a practice SAT exam (Princeton Review, *11 Practice Tests for the SAT & PSAT*, 2011 Edition). To train language models for the SAT question task, we used 1.2 billion words of Los Angeles Times data taken from the years 1985 through 2002. Results for the SAT data are similar, with N-gram language models scoring 42-44% depending on vocabulary size and smoothing, and LSA similarity attaining 46%.

Method	**% Correct** (N=1040)
Human	91
Generating Model	31
Smoothed 3-gram	36
Smoothed 4-gram	39
Positional combination	43
Simple 4-gram	34
Average LSA Similarity	49

Table 1: Summary of Benchmarks

These results indicate that the "Holmes" sentence completion set is indeed a challenging problem, and has a level of difficulty roughly comparable to that of SAT questions. Simple models based on N-gram statistics do quite poorly, and even a relatively sophisticated semantic-coherence model struggles to beat the 50% mark.

5 Related Work

The past work which is most similar to ours is derived from the lexical substitution track of SemEval-2007 (McCarthy and Navigli, 2007). In this task, the challenge is to find a replacement for a word or

[1] *http://www.speech.cs.cmu.edu/SLM/toolkit.html*

phrase removed from a sentence. In contrast to our SAT-inspired task, the original answer is indicated. For example, one might be asked to find replacements for *match* in "After the *match*, replace any remaining fluid deficit to prevent problems of chronic dehydration throughout the tournament." Scoring is done by comparing a system's results with those produced by a group of human annotators (not unlike the use of multiple translations in machine translation). Several forms of scoring are defined using formulae which make the results impossible to compare with correct/incorrect multiple choice scoring. Under the provided scoring metrics, two consistently high-performing systems in the SemEval 2007 evaluations are the KU (Yuret, 2007) and UNT (Hassan et al., 2007) systems. These operate in two phases: first they find a set of potential replacement words, and then they rank them. The KU system uses just an N-gram language model to do this ranking. The UNT system uses a large variety of information sources, each with a different weight. A language model is used, and this receives the highest weight. N-gram statistics were also very effective - according to one of the scoring paradigms - in (Giuliano et al., 2007); as a separate entry, this paper further explored the use of Latent Semantic Analysis to measure the degree of similarity between a potential replacement and its context, but the results were poorer than others. Since the original word provides a strong hint as to the possible meanings of the replacements, we hypothesize that N-gram statistics are largely able to resolve the remaining ambiguities, thus accounting for the good performance of these methods on this task. The Holmes data does not have this property and thus may be more challenging.

ESL synonym questions were studied by Turney (2001), and subsequently considered by numerous research groups including Terra and Clarke (2003) and Pado and Lapata (2007). These questions are easier than the SemEval task because in addition to the original word and the sentence context, the list of options is provided. For example, one might be asked to identify a replacement for "rusty" in "A [rusty] nail is not as strong as a clean, new one. *(corroded; black; dirty; painted)*." Jarmasz and Szpakowicz (2003) used a sophisticated thesaurus-based method and achieved state-of-the art performance on the ESL synonyms task, which is 82%. Again the Holmes data does not have the property that the intended meaning is signaled by providing the original word, thus adding extra challenge.

Although it was not developed for this task, we believe the recurrent language modeling work of Mikolov (2010; 2011b; 2011a) is also quite relevant. In this work, a recurrent neural net language model is used to achieve state-of-the-art performance in perplexity and speech recognition error rates. Critically, the recurrent neural net does not maintain a fixed N-gram context, and its hidden layer has the potential to model overall sentence meaning and long-span coherence. While theoretical results (Bengio et al., 1994) indicate that extremely long-range phenomena are hard to learn with a recurrent neural network, in practice the span of usual sentences may be manageable. Recursive neural networks (Socher et al., 2011) offer similar advantages, without the theoretical limitations. Both offer promising avenues of research.

6 Conclusion

In this paper we have described a new, publicly available, corpus of sentence-completion questions. Whereas for many traditional language modeling tasks, N-gram models provide state-of-the-art performance, and may even be fully adequate, this task is designed to be insoluble with local models. Because the task now allows us to measure progress in an area where N-gram models do poorly, we expect it to stimulate research in fundamentally new and more powerful language modeling methods.

References

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157 –166.

Ciprian Chelba and Frederick Jelinek. 1998. Exploiting syntactic structure for language modeling. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 225–231, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ciprian Chelba and Frederick Jelinek. 2000. Structured language modeling. *Computer Speech and Language*, 14(4):283 – 332.

Stanley F. Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 13(4):359 – 393.

S. Chen. 2009a. Performance prediction for exponential language models. In *NAACL-HLT*.

S. Chen. 2009b. Shrinking exponential language models. In *NAACL-HLT*.

S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(96).

Educational-Testing-Service. 2011. https://satonlinecourse.collegeboard.com/sr/ digital_assets/assessment/pdf/0833a611-0a43-10c2-0148-cc8c0087fb06-f.pdf.

Claudio Giuliano, Alfio Gliozzo, and Carlo Strapparava. 2007. Fbk-irst: Lexical substitution task exploiting domain and syntagmatic coherence. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 145–148, Stroudsburg, PA, USA. Association for Computational Linguistics.

Samer Hassan, Andras Csomai, Carmen Banea, Ravi Sinha, and Rada Mihalcea. 2007. Unt: Subfinder: Combining knowledge sources for automatic lexical substitution. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 410–413, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sanjeev Khudanpur and Jun Wu. 2000. Maximum entropy techniques for exploiting syntactic, semantic and collocational dependencies in language modeling. *Computer Speech and Language*, 14(4):355 – 372.

R. Kneser and H. Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of ICASSP*.

Jarmasz M. and Szpakowicz S. 2003. Roget's thesaurus and semantic similarity. In *Recent Advances in Natural Language Processing (RANLP)*.

Diana McCarthy and Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53.

Tomas Mikolov, Martin Karafiat, Jan Cernocky, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proceedings of Interspeech 2010*.

Tomas Mikolov, Anoop Deoras, Stefan Kombrink, Lukas Burget, and Jan Cernocky. 2011a. Empirical evaluation and combination of advanced language modeling techniques. In *Proceedings of Interspeech 2011*.

Tomas Mikolov, Stefan Kombrink, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. 2011b. Extensions of recurrent neural network based language model. In *Proceedings of ICASSP 2011*.

Sebastian Pado and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics, 33 (2)*, pages 161–199.

Harry Printz and Peder A. Olsen. 2002. Theory and practice of acoustic confusability. *Computer Speech and Language*, 16(1):131 – 164.

Ronald Rosenfeld, Stanley F. Chen, and Xiaojin Zhu. 2001. Whole-sentence exponential language models: a vehicle for linguistic-statistical integration. *Computer Speech and Language*, 15(1):55 – 73.

R. Rosenfeld. 1997. A whole sentence maximum entropy language model. In *Proceedings ASRU*.

Holger Schwenk and Jean-Luc Gauvain. 2002. Connectionist language modeling for large vocabulary continuous speech recognition. In *Proceedings of ICASSP*.

Holger Schwenk. 2007. Continuous space language models. *Computer Speech and Language*, 21(3):492 – 518.

Claude E. Shannon and Warren Weaver. 1949. *The Mathematical Theory of Communication*. University of Illinois Press.

Richard Socher, Cliff Chiung-Yu Lin, Andrew Y. Ng, and Christopher D. Manning. 2011. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 2011 International Conference on Machine Learning (ICML-2011)*.

E. Terra and C. Clarke. 2003. Frequency estimates for statistical word similarity measures. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Peter D. Turney. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *European Conference on Machine Learning (ECML)*.

Jun Wu and Sanjeev Khudanpur. 1999. Combining nonlocal, syntactic and n-gram dependencies in language modeling. In *Proceedings of Eurospeech*.

Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pages 523–530, Stroudsburg, PA, USA. Association for Computational Linguistics.

Deniz Yuret. 2007. Ku: word sense disambiguation by substitution. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 207–213, Stroudsburg, PA, USA. Association for Computational Linguistics.