Exploiting Linguistic Features for Sentence Completion

Aubrie M. Woods
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213, USA
amwoods@cmu.edu

Abstract

This paper presents a novel approach to automated sentence completion based on pointwise mutual information (PMI). Feature sets are created by fusing the various types of input provided to other classes of language models, ultimately allowing multiple sources of both local and distant information to be considered. Furthermore, it is shown that additional precision gains may be achieved by incorporating feature sets of higher-order n-grams. Experimental results demonstrate that the PMI model outperforms all prior models and establishes a new state-of-the-art result on the Microsoft Research Sentence Completion Challenge.

1 Introduction

Skilled reading is a complex cognitive process that requires constant interpretation and evaluation of written content. To develop a coherent picture, one must reason from the material encountered to construct a mental representation of meaning. As new information becomes available, this representation is continually refined to produce a globally consistent understanding. Sentence completion questions, such as those previously featured on the Scholastic Aptitude Test (SAT), were designed to assess this type of verbal reasoning ability. Specifically, given a sentence containing 1-2 blanks, the test taker was asked to select the correct answer choice(s) from the provided list of options (College Board, 2014). A sample sentence completion question is illustrated in Figure 1.

Certain clear patterns in the metamorphosis of a butterfly indicate that the process is ——.
(A) systematic
(B) voluntary
(C) spontaneous
(D) experimental
(E) clinical

Figure 1: An example sentence completion question (The Princeton Review, 2007).

To date, relatively few publications have focused on automatic methods for solving sentence completion questions. This scarcity is likely attributable to the difficult nature of the task, which occasionally involves logical reasoning in addition to both general and semantic knowledge (Zweig et al., 2012b). Fundamentally, text completion is a challenging semantic modeling problem, and solutions require models that can evaluate the global coherence of sentences (Gubbins and Vlachos, 2013). Thus, in many ways, text completion epitomizes the goals of natural language understanding, as superficial encodings of meaning will be insufficient to determine which responses are accurate.

In this paper, a model based on pointwise mutual information (PMI) is proposed to measure the degree of association between answer options and other sentence tokens. The PMI model considers multiple sources of information present in a sentence prior to selecting the most likely alternative.

The remainder of this report is organized as follows. Section 2 describes the high-level characteristics of existing models designed to perform automated sentence completion. This prior work provides direct motivation for the PMI model, introduced in Section 3. In Section 4, the model's performance on the Microsoft Research (MSR) Sentence Completion Challenge and a data set comprised of SAT questions are juxtaposed. Finally, Section 5 offers concluding remarks on this topic.

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 438–442,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

2 Background

Previous research expounds on various architectures and techniques applied to sentence completion. Below, models are roughly categorized on the basis of complexity and type of input analyzed.

2.1 N-gram Models

Advantages of n-gram models include their ability to estimate the likelihood of particular token sequences and automatically encode word ordering. While relatively simple and efficient to train on large, unlabeled text corpora, n-gram models are nonetheless limited by their dependence on local context. In fact, such models are likely to overvalue sentences that are locally coherent, yet improbable due to distant semantic dependencies.

2.2 Dependency Models

Dependency models circumvent the sequentiality limitation of n-gram models by representing each word as a node in a multi-child dependency tree. Unlabeled dependency language models assume that each word is (1) conditionally independent of the words outside its ancestor sequence, and (2) generated independently from the grammatical relations. To account for valuable information ignored by this model, e.g., two sentences that differ only in a reordering between a verb and its arguments, the labeled dependency language model instead treats each word as conditionally independent of the words and labels outside its ancestor path (Gubbins and Vlachos, 2013).

In addition to offering performance superior to n-gram models, advantages of this representation include relative ease of training and estimation, as well as the ability to leverage standard smoothing methods. However, the models' reliance on output from automatic dependency extraction methods and vulnerability to data sparsity detract from their real-world practicality.

2.3 Continuous Space Models

Neural networks mitigate issues with data sparsity by learning distributed representations of words, which have been shown to excel at preserving linear regularities among tokens. Despite drawbacks that include functional opacity, propensity toward overfitting, and elevated computational demands, neural language models are capable of outperforming n-gram and dependency models (Gubbins and Vlachos, 2013; Mikolov et al., 2013; Mnih and Kavukcuoglu, 2013).

Log-linear model architectures have been proposed to address the computational cost associated with neural networks (Mikolov et al., 2013; Mnih and Kavukcuoglu, 2013). The continuous bag-of-words model attempts to predict the current word using *n* future and *n* historical words as context. In contrast, the continuous skip-gram model uses the current word as input to predict surrounding words. Utilizing an ensemble architecture comprised of the skip-gram model and recurrent neural networks, Mikolov et al. (2013) achieved prior state-of-the-art performance on the MSR Sentence Completion Challenge.

3 PMI Model

This section describes an approach to sentence completion based on pointwise mutual information. The PMI model was designed to account for both local and distant sources of information when evaluating overall sentence coherence.

Pointwise mutual information is an information-theoretic measure used to discover collocations (Church and Hanks, 1990; Turney and Pantel, 2010). Informally, PMI represents the association between two words, *i* and *j*, by comparing the probability of observing them in the same context with the probabilities of observing each independently.

The first step toward applying PMI to the sentence completion task involved constructing a word-context frequency matrix from the training corpus. The context was specified to include all words appearing in a single sentence, which is consistent with the hypothesis that it is necessary to examine word co-occurrences at the sentence level to achieve appropriate granularity. During development/test set processing, all words were converted to lowercase and stop words were removed based on their part-of-speech tags (Toutanova et al., 2003). To determine whether a particular part-of-speech tag type did, in fact, signal the presence of uninformative words, tokens assigned a hypothetically irrelevant tag were removed if their omission positively affected performance on the development portion of the MSR data set. This non-traditional approach, selected to increase specificity and eliminate dependence on a non-universal stop word list, led to the removal of determiners, coordinating conjunctions,

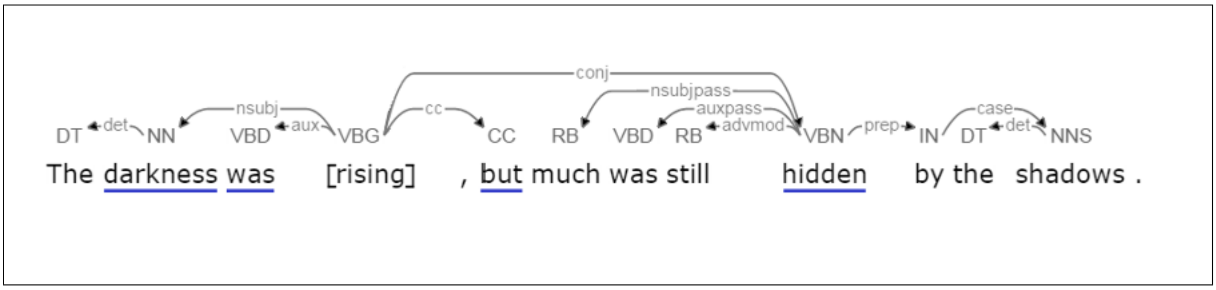
conj
nsubjpass
auxpass
advmod
nsubj
cc
det
aux
prep
case
det
DT NN VBD VBG CC RB VBD RB VBN IN DT NNS
The darkness was [rising] , but much was still hidden by the shadows .

Figure 2: The dependency parse tree for Question 17 in the MSR data set. Words that share a grammatical relationship with the missing word *rising* are underscored. Following stop word removal, the feature set for this question is [*darkness*, *was*, *hidden*].

pronouns, and proper nouns.[1] Next, feature sets were defined to capture the various sources of information available in a sentence. While feature set number and type is configurable, composition varies, as sets are dynamically generated for each sentence at run time. Enumerated below are the three feature sets utilized by the PMI model.

1. **Reduced Context**. This feature set consists of words that remain following the pre-processing steps described above.

2. **Dependencies.** Sentence words that share a semantic dependency with the candidate word(s) are included in this set (Chen and Manning, 2014). Absent from the set of dependencies are words removed during the pre-processing phase. Figure 2 depicts an example dependency parse tree along with features provided to the PMI model.

3. **Keywords**. Providing the model with a collection of salient tokens effectively increases the tokens' associated weights. An analogous approach to the one described for stop word identification was applied to discover that common nouns consistently hold greater significance than other words assigned hypothetically informative part-of-speech tags.

Let *X* represent a word-context matrix with *n* rows and *m* columns. Row $x_{i:}$ corresponds to word *i* and column $x_{:j}$ refers to context *j*. The term *x*(*i*,*j*) indicates how many times word *i* occurs in context *j*. Applying PMI to *X* results in the *n* x *m* matrix *Y*, where term *y*(*i*,*j*) is defined by (1). To avoid overly penalizing words that are unrelated to the context, the positive variant of PMI is considered, in which negative scores are replaced with zero (4).

$$P(i,j) = \frac{x(i,j)}{\sum_{i=1}^{n}\sum_{j=1}^{m} x(i,j)} \tag{1}$$

$$P(i*) = \frac{\sum_{j=1}^{m} x(i,j)}{\sum_{i=1}^{n}\sum_{j=1}^{m} x(i,j)} \tag{2}$$

$$P(*j) = \frac{\sum_{i=1}^{n} x(i,j)}{\sum_{i=1}^{n}\sum_{j=1}^{m} x(i,j)} \tag{3}$$

$$pmi(i,j) = max\left\{0, log\left(\frac{P(i,j)}{P(i*)P(*j)}\right)\right\} \tag{4}$$

In addition, the discounting factor described by Pantel and Lin (2002) is applied to reduce bias toward infrequent words (7).

$$mincontext = min(\sum_{k=1}^{n} x(k,j), \sum_{k=1}^{m} x(i,k)) \tag{5}$$

$$\delta(i,j) = \frac{x(i,j)}{x(i,j)+1} \cdot \frac{mincontext}{mincontext+1} \tag{6}$$

$$dpmi(i,j) = pmi(i,j) \cdot \delta(i,j) \tag{7}$$

$$similarity(i,S) = \sum_{j \in S} dpmi(i,j) \cdot \gamma \tag{8}$$

The PMI model evaluates each possible response to a sentence completion question by substituting each candidate answer, *i*, in place of the blank and scoring the option according to (8). This equation measures the semantic similarity between each candidate answer and all other words in the sentence, *S*. Prior to being summed, individual PMI values associated with a particular word *i*

[1] Perhaps counterintuitively, most proper nouns are uninformative for sentence completion, since they refer to specific named entities (e.g. people, locations, organizations, etc.).

and context word *j* are multiplied by γ, which reflects the number of feature sets containing *j*. Ultimately, the candidate option with the highest similarity score is selected as the most likely answer.

Using the procedure described above, additional feature sets of bigrams and trigrams were created and subsequently incorporated into the semantic similarity assessment. This extended model accounts for both word- and phrase-level information by considering windowed co-occurrence statistics.

4 Experimental Evaluation

4.1 Data Sets

Since its introduction, the Microsoft Research Sentence Completion Challenge (Zweig and Burges, 2012a) has become a commonly used benchmark for evaluating semantic models. The data is comprised of material from nineteenth-century novels featured on Project Gutenberg. Each of the 1,040 test sentences contains a single blank that must be filled with one of five candidate words. Associated candidates consist of the correct word and decoys with similar distributional statistics.

To further validate the proposed method, 285 sentence completion problems were collected from SAT practice examinations given from 2000-2014 (College Board, 2014). While the MSR data set includes a list of specified training texts, there is no comparable material for SAT questions. Therefore, the requisite word-context matrices were constructed by computing token co-occurrence frequencies from the New York Times portion of the English Gigaword corpus (Parker et al., 2009).

4.2 Results

The overall accuracy achieved on the MSR and SAT data sets reveals that the PMI model is able to outperform prior models applied to sentence completion. Table 1 provides a comparison of the accuracy values attained by various architectures, while Table 2 summarizes the PMI model's performance given feature sets of context words, dependencies, and keywords. Recall that the n-gram variant reflects how features are partitioned.

It appears that while introducing phrase-level information obtained from higher-order n-grams leads to gains in precision on the MSR data set, the same cannot be stated for the set of SAT questions. The most probable explanation for this is twofold. First, informative context words are much less likely to occur within 2-3 tokens of the target word. Second, missing words, which are selected to test knowledge of vocabulary, are rarely found in the training corpus. Bigrams and trigrams containing these infrequent terms are extremely uncommon. Regardless of sentence structure, the sparsity associated with higher-order n-grams guarantees diminishing returns for larger values of *n*. When deciding whether or not to incorporate this information, it is also important to consider the significant trade-off with respect to information storage requirements.

Language Model	MSR
Random chance	20.00
N-gram [Zweig (2012b)]	39.00
Skip-gram [Mikolov (2013)]	48.00
LSA [Zweig (2012b)]	49.00
Labeled Dependency [Gubbins (2013)]	50.00
Dependency RNN [Mirowski (2015)]	53.50
RNNs [Mikolov (2013)]	55.40
Log-bilinear [Mnih (2013)]	55.50
Skip-gram + RNNs [Mikolov (2013)]	58.90
PMI	**61.44**

Table 1: Best performance of various models on the MSR Sentence Completion Challenge. Values reflect overall accuracy (%).

Features	MSR	SAT
Unigrams	58.46	**58.95**
Unigrams + Bigrams	60.87	58.95
Unigrams + Bigrams + Trigrams	**61.44**	58.95

Table 2: PMI model performance improvements (% accurate) from incorporating feature sets of higher-order n-grams.

5 Conclusion

This paper described a novel approach to answering sentence completion questions based on pointwise mutual information. To capture unique information stemming from multiple sources, several features sets were defined to encode both local and distant sentence tokens. It was shown that while precision gains can be achieved by augmenting these feature sets with higher-order n-grams, a significant cost is incurred as a result of the increased data storage requirements. Finally, the superiority of the PMI model is demonstrated by its performance on the Microsoft Research Sentence Completion Challenge, during which a new state-of-the-art result was established.

References

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 740–750.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

The College Board. 2014. Sat reading practice questions: Sentence completion. Retrieved from https://sat.collegeboard.org/practice/sat-practice-questions-reading/sentence-completion.

Joseph Gubbins and Andreas Vlachos. 2013. Dependency language models for sentence completion. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1405–1410.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Workshop Proceedings of the International Conference on Learning Representations*.

Piotr Mirowski and Andreas Vlachos. 2015. Dependency recurrent neural language models for sentence completion. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 511–517. Association for Computational Linguistics.

Andriy Mnih and Koray Kavukcuoglu. 2013. Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in Neural Information Processing Systems 26*, pages 2265–2273. Curran Associates, Inc.

Patrick Pantel and Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 613–619. Association for Computing Machinery.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2009. English gigaword fourth edition ldc2009t13.

The Princeton Review. 2007. *11 Practice Tests for the SAT and PSAT, 2008 Edition*. Random House, Inc., New York City, NY.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, volume 1, pages 252–259. Association for Computational Linguistics.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

Geoffrey Zweig and Christopher J.C. Burges. 2012a. A challenge set for advancing language modeling. In *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, pages 29–36. Association for Computational Linguistics.

Geoffrey Zweig, John C. Platt, Christopher Meek, Christopher J.C. Burges, Ainur Yessenalina, and Qiang Liu. 2012b. Computational approaches to sentence completion. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 601–610. Association for Computational Linguistics.