英語穴埋め問題の自動解法

Automatic Solver of English Incomplete Sentence Questions

白石　裕次郎　　佐々木　裕
Yujiro Shiraishi　Yutaka Sasaki

豊田工業大学
Toyota Technological Institute

1. はじめに

本研究では，英語穴埋め問題の解答方法を検討する．対象として TOEIC※(Test of English for International Communication)[1]の空所穴埋め問題を用いることとした．コンピューターが自動で問題文と４つの選択肢を読み取り，問題文の空所に入る最も適切な選択肢を選ぶ解答システムの構築を目標とする．

まず我々が通常英語穴埋め問題を解答するためには，単語の意味を理解する語彙力，文脈を理解するために必要な文法構造の理解，頻出するフレーズに対する知識など，様々な要素の言語知識を使って解答を選択する[2]．そのため，この穴埋め問題は，大学入試や英検・TOEIC※などの英語学習者の言語理解度を測るテストでよく用いられる．このタイプの問題の利点としては，１つの設問に必ず１つの明確な正解があるため，解答者の言語理解度をスコアとしてはっきりと表すことができるという特徴がある．そのような問題に対して，回答者は自身の言語知識を基に，空所に入る最も適切な１つの解答を選ぶが，そこで高い正解率を得ることができれば，回答者は英語の言語を理解しているとされる．

そこで我々は，コンピューターに対しても同様のことが言えると考えた．英語穴埋め問題を自動で解答するシステムの構築し，そのシステムが正しく正解を選択できれば，そのシステムの言語処理技術の高さを示せるのではないかと考えた．当然，解答を選択するためには何らかの英文コーパスや文法知識を必要とするが，本研究の提案手法では Google n-gram[3]のデータを用い，そこから得た bigram～5-gram の共起頻度を総合的に判断し，１つの選択肢を決定する解答システムを構築した．

評価実験では，Webanhvan[4]から取得した英語穴埋め問題を使用し，そのシステムの言語処理能力の評価を行った．本研究で TOEIC の模擬問題を利用した背景としては，TOEIC は近年多くの日本人英語学習者が受験しているテストであるため，作成した解答システムと TOEIC 受験者の比較ができるのではないかと考えた．

```
<DOCID>TOEIC-519</DOCID>
<COMMENT> http://www.webanhvan.com
</COMMENT>
<TEXT>
<QUESTION>A lot of …… are parked at the street.
</QUESTION>
<SELECT>
<OPTION VALUE="A">books</OPTION>
<OPTION VALUE="B">people</OPTION>
<OPTION VALUE="C">teachers</OPTION>
<OPTION VALUE="D">cars</OPTION>
</SELECT>
<ANS>D</ANS>
</TEXT>
```

図１　英語穴埋め問題の xml データ例

2. 前提条件

本研究での解答システムが問題を解くにあたって幾つかの条件を設定した．

i. 問題の解法に焦点を当てるため，通常人間が試験を受ける際に読む必要がある“指示文”の理解は割愛した．解答システムには問題文，４つの選択肢，また答え合わせの際に必要な正解解答が書かれた xml 形式のファイルを与えることとする．図１に実際に使用したファイルの一部示す．

ii. 問題の種類は特定しない．TOEIC の空欄穴埋め問題では，語彙・品詞・接続詞・前置詞などを選択する様々なジャンルの問題があるが，ある特定のジャンルに焦点を当てるのではなく，これら全ての問題を解くことにチャレンジした．しかし，実際の TOEIC の試験は公式には公開されていないため，公式の問題は入手できない．よって本研究では，web で公開されていた類似の練習問題[2]を使って評価実験を行った．

※TOEIC is a registered trademark of Educational Testing Service

produce the engineers	53
produce the exaggerated	78
produce the express	78
produce the imaginative	63
produce the imbalance	64
produce the important	518

図２　google trigram 例

3. 提案手法

本研究の提案手法では Google n-gram データを使用した.Google n-gram とは米 Google 社が作成した，一般の web ページにある約１兆もの英単語を英文コーパスとして bigram～5-gram で共起頻度をカウントしたデータで，その内容は 2～5 単語の“キーワード”とその“共起頻度”で構成されている．図２に 3gram の例を示す．このデータは，web 上にある一般的な文章から作成された膨大な量の n-gram データを扱えるという利点がある．一方で，扱うデータ量の観点から，本研究ではこの Google n-gram の txt データを GDBM ファイルに変換して使用することとした.

3.1 システム概要

まずコンピューターが問題文と４つの選択肢を読み取る．続いて，問題文と選択肢を組み合わせてキーワードを作り，それらのキーワードを Google n-gram を参照してそこから共起頻度を得る.我々の提案する解答システムでは，この共起頻度が“存在する”または“存在しない”か，また存在する場合はそれらの数値の差を考慮することで，正解を選択することができるのではないかと考えた．このシステムの構成図を図３に示す．例題１として“How (　) you been ?”と言う問題文があり，選択肢として(A) are, (B) have, (C) will, (D) is というような語句が並んでいたとすると，問題と選択肢をそれぞれ組み合わすことで４つの 5-gram キーワードが作成できる．それらのキーワードに対して Google n-gram を参照して共起頻度を得ると表１のような結果になった.

表１　例題１に対するキーワードとその共起頻度

選択肢	5-gram キーワード	共起頻度
(A)	How **are** you been ?	0
(B)	How **have** you been ?	3948
(C)	How **will** you been ?	0
(D)	How **is** you been ?	0

この表を参照すると，問題文と正解選択肢(B) have を組み合わせた場合のキーワードでは高い共起頻度を示しているが，その他のキーワードでは共起頻度は“存在しない”という結果になっていることが分かる．このように，不正解選択肢から作られたキーワードには文法的な間違いがあるか，または前後の語句の繋がりを考えた時にどこか不自然な部分が生まれ，この共起が存在しないかまたは頻度が小さくなる傾向にあると考えた．提案手法では，この特性を利用して正解を選ぶシステムとした．特に先ほどの例では，不正解選択肢からなるキーワードでは文法的な間違いが発生しているため，共起頻度は“存在しない”となり，正解と不正解の違いがはっきりと示された．次にもう１つ違う例題２を考えてみる．問題文が“A lot of (　) are ….”となっていて，選択肢には(A)books (B)people (C)teachers (D)cars という語句が並んでいたとする．４つの選択肢は全て名詞なので，空欄に適切な名刺を入れる語彙問題となっている．先程と同様にキーワードを作成し，Google n-gram を参照して共起頻度を得ると表２のような結果になった.

図３　提案手法のシステム構成図

表２　例題２に対するキーワードとその共起頻度

選択肢	5-gram キーワード	共起頻度
(A)	A lot of **books** are	78
(B)	A lot of **people** are	10546
(C)	A lot of **teachers** are	109
(D)	A lot of **cars** are	50

この例では，例題１とは違い全てのキーワードから共起頻度が得られた．この場合選択肢(B)people は“A lot of people are”という組み合わせで web 上の英文コーパス中に 10546 回出現したことになり，その他の共起頻度と比較してみても最も高い数字を示しているため，この問題の正解選択肢であることが高いのではないかと考えられる．しかし，この問題文には続きがあり，完全な文は“A lot of (　) are parked at the street.”であったとする．そうなると，実際この問題からは他にもいくつもキーワードが作成できることになる．例えば，“lot of

( ) are parked”や“( ) are parked at the”などが挙げられる．そこで，後者の単語の組み合わせに選択肢を組み合わすことで，４つの新たなキーワードを作成する．それらに対して同様に共起頻度を得ること，結果は表３のようになった．

表３　例題２に対するキーワードとその共起頻度

選択肢	5-gram キーワード	共起頻度
(A)	**books** are parked at the	0
(B)	**people** are parked at the	0
(C)	**teachers** are parked at the	0
(D)	**cars** are parked at the	73

先ほどの結果では選択肢(B)を正解候補として示していたのに対し，今回の結果では選択肢(D)が高い共起頻度を示し，この問題の正解候補となる．この様に１つの問題に対していくつもキーワードが作成でき，またそれぞれの共起頻度のスコアから違う考察ができるため，本研究の提案手法では Google n-gram より取得したそれぞれの共起頻度のスコアを総合的に判断し，１つの解答選ぶこととした．

3.2 キーワード作成

提案手法では，問題文より空欄を含む連続する２～5 単語の全通りの組み合わせを作成し，それに対して選択肢を含めてキーワードを作成した．図４に 5-gram キーワード作成の様子を示す．この例からは合計で 16 個のキーワードを作成することができる．5-gram のキーワードは，それが正解選択肢から作られたものだとしても，Google n-gram から共起頻度を得ることができないことがある．そうなる解答選択の手がかりがなくなり，１つの解答を選ぶことができなくなる．そのため，共起頻度を得ることができる可能性の高い bigram，trigram，4-gram についても同様の方法でキーワードを作成した．

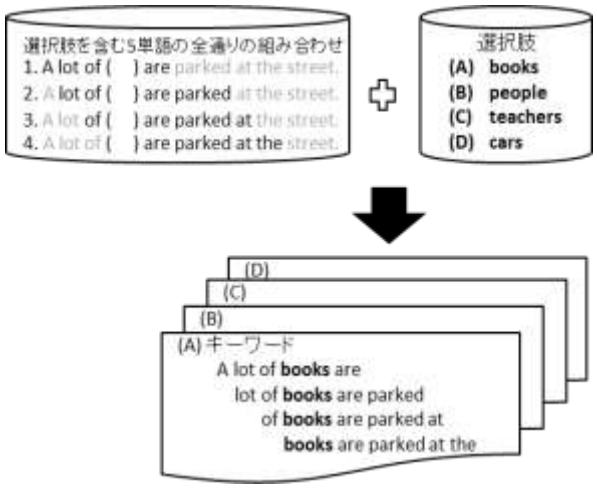

図４　5 単語の組み合わせ作成の様子

3.3 解答選択

キーワードから得た共起頻度のスコアを比較し，１つの解答を決定する工程に移る．3.1 節で示した用に正解選択肢からなるキーワードが必ずしも高いスコアを示すとは限らない．そこで，本研究の提案手法では，まず共起頻度が“存在する”または“存在しない”この違いに注目した．キーワードが英文法の間違った文章で構成されている場合や【表１，(A)(C)(D)】，通常使われることのない主語と動詞の組み合わせで構成されている場合など【表３，(A)(B)(C)】，Google n-gram 内に共起頻度が存在しない．また動詞と目的語に対しても同様のことが言え，例えば“eat the cake”と“eat the bike”いう２つのキーワードを比較してみると，前者のスコアは 136 であったのに対し，後者は“eat”が目的語として“bike”をとることは珍しく，このキーワードを Google n-gram で参照しても共起頻度は存在しない．そこで我々の提案する解答システムでは，１つの問題に対して作成したすべてのキーワードに対して共起頻度を参照し，そのスコアの有無を計算することで正解選択肢を１つ選べるのではないかと考えた．例題２では選択肢毎に 13 のキーワードが作成でき，それぞれ共起頻度の有無を調べた結果 表４の様な結果となった．正解の選択肢(D)が単

表４　13 のキーワードに対して共起頻度の得られた数

選択肢	共起頻度の得られた数
(A)	7
(B)	7
(C)	7
(D)	11

独で高い数字を表しているので，この問題についてはこの手法で正解選択肢を選ぶことができる．しかし，全ての問題でこの数値に差が出ると言う訳ではない．表５のように，幾つかの選択肢でこの数値が対立してしまった場合，正解選択肢が１つに絞れなくなる．そのような時には表２で示したように，共起頻度のスコアの差を解答選択の考察に取り入れた．表５を例にとって考えると，対立した選択肢(B)，(D)に対して共起頻度の差をキーワード毎に

表５　選択肢でスコアが対立した例

選択肢	共起頻度の得られた数
(A)	5
(B)	7
(C)	4
(D)	7

表６　対立した選択肢を共起頻度で比較

		各キーワードに対する共起頻度						
		bigram		trigram		4-gram 5-gram		
		1	2	3	4	…	13	14
選択肢	(B)	1000	1100	50	60	…	0	0
	(D)	800	1500	100	70	…	0	0
優位性		(B)	(D)	(D)	(D)	…	×	×

計算し，どちらの選択肢が優位であるかをそれぞれ決定する．表６ではこの優位性を比較した様子を示したが，これ見ると選択肢(D)が(B)よりも優っている様子が分かる．この過程において，4-gram や 5-gram のようなキーワードからは共起頻度が得られていないことも多く，優位性を評価できないことも珍しくない．一方で，我々の提案手法では，そのようなキーワードで高い共起頻度を示している場合は，その選択肢が正解である可能性が高いと考え，優位性の結果に重み付けを行い，その最終的な結果を評価し解答を選択するシステムとした．

４．評価実験

4.1 練習問題を使ったシステムの評価

作成した解答システムに対して webanhvan[4]より取得した練習問題615問を利用して実験を行った(表７)．また，前章で述べた重み付けを行ったシステムが，行わなかった方に比べ 4.1%正解率が高く，この方法の有効性が示された．解答に要した時間は 5 分で，1 問当たり約 0．5 秒で解答しているという結果になった．

表７　解答システムの正解率

	問題数(問)	正解数(問)	正解率(%)
解答システム	615	450	73.1

4.2 学生の能力と解答システムの比較

この解答システムと人間の解答能力を比較した．事前に 20 名の学生に協力を依頼し，実際にサンプル問題 40 問を解いてもらった．解答システムとの比較を表８に示す．このサンプル問題での受験者の得点と，彼らの現在の TOEIC スコアの分布を以下の図５に示す．このグラフに今回の解答システムの正解数を★でプロットした．このグラフをより解答システムは TOEIC で 700 点台の人と同等の正解率であると言える．

表８　サンプル問題での正解率

	問題数(問)	正解数(問)	正解率(%)
解答システム	40	31	77.5
学生の平均	40	26	65.0

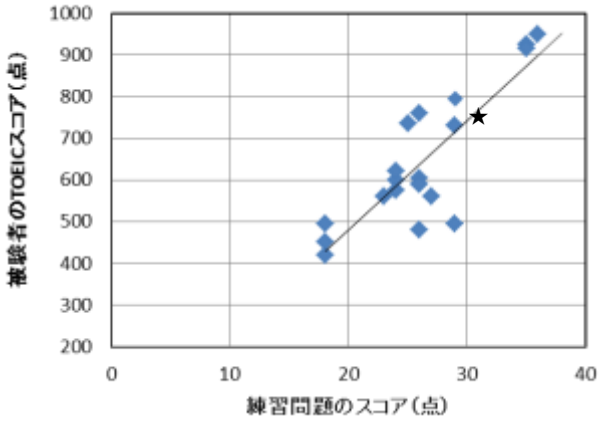

図５　解答システムと学生のスコアの比較

５．おわりに

本稿では，Google n-gram データを用いた英語穴埋め問題の解法を提案した．本手法では，問題文と選択肢によって作成されたキーワードを Google n-gram データを参照し，そこでの共起頻度の有無やスコアの差を総合的に評価して１つの解答を選択する．実際の練習問題を使用した実験においては，73.1%の確率で正しい正解を選択できていることが確認できた．

今後の課題として，構文解析システム Enju の利用や，意味的類似度を示す BLEU の評価値などを参考に，空欄前後の語句だけを評価するのではなく，文全体の構文理解や意味的理解を行い，より精度の高い解答システムへと改善する必要がある[5][6]．また，キーワード内に地名・人名などといった固有名詞が含まれることで，共起頻度が得られなくなり，正確なスコアの比較ができずに正解を選べなかったこと実も確認された．より正確な共起頻度が得られるよう，このような固有名詞への対策も必要であると考えられる．

参考文献

[1]. TOEIC, http://www.ets.org/toeic

[2]. 西村芳康，新 TOEIC®テストの特徴，電気通信大学紀要 19 巻 1･2 合併号，pp. 101-116, 2006

[3]. Google n-gram data, http://www.ldc.upenn.edu/Catalog/

[4]. Webanhvan http://www. webanhvan. com/

[5]. 乾健太郎，浅原正幸 ： 自然言語処理の再挑戦～統計的言語処理を超えて～,知能と情報(日本知能情報ファジィ学会誌)vol.18, No.5, pp.669-681 ， 2006.

[6]. 村瀬隆久ほか ： 依存関係を用いた斬新的構文解析の効率化，言語処理学会 第 6 回年次大会発表論文集，2000.