G11-18

英文穴埋め問題の N-gram データを用いた解法

◎泉野 和弘　岩村 雅一　黄瀬 浩一
大阪府立大学大学院工学研究科
izukazu@m.cs.osakafu-u.ac.jp {masa|kise}@cs.osakafu-u.ac.jp

1 はじめに

本研究では，英文穴埋め問題の解答方法を検討する．今回は前置詞を対象とした問題である．穴埋め問題とは，英文中の空欄箇所に入る単語を複数の正解候補の中から選ぶ問題である．この穴埋め問題を解く手段として，身近に使えるものに，検索エンジンがある [1]．例えばフレーズ “launched 〜 the orbit” の〜に入る単語を推定するには，〜に候補となる単語を入れフレーズごとの検索件数を比較すれば良い．これは，よく出現するフレーズが一般的なフレーズであるという考えに基づいている．

検索エンジンを使う手法の問題点は，問題文中のフレーズがデータに登録されていない場合，件数を比較できない点である．しかし，同様の処理が可能な N-gram データを用いれば，データが手元にあるので用途に応じてデータの改良が出来る．本稿では，この利点を生かし，問題文中の単語と N-gram データ中の単語を予め分類することによって上述の問題の解決方法を検討する．フレーズの長さが 1 単語から 5 単語までの出現頻度が登録されている Google N-gram data [2] を使用する．実験では，単語の分類を施した N-gram データと分類していないデータを使ったときの穴埋め問題の正解率の違いについて検討を行う．また，問題文から予め不要と考えられる単語を削除することの正解率への影響について検討を行う．

2 N-gram データを使った解答方法とデータの改良

本稿では，N-gram データを使った解答方法を二つ提案する．提案手法を説明する前に，N-gram データを用いて頻度の高いフレーズを決定する方法について例文 “. . . which was launched 〜 the orbit 30 years. . . ” を使って説明する．3-gram データを使う場合，まず，3 単語のフレーズ “was launched 〜”，“launched 〜 the”，“〜 the orbit” を問題文から切り出す．次に，切り出したフレーズに正解候補を入れて，3-gram データでフレーズの頻度を調べる．頻度が最大となったフレーズが最も尤もらしいとして出力する．

上述の手法に基づいて N-gram データを用いた提案手法を述べる．一つ目の手法は，上述の方法を 5-gram で行い，解答できない問題を 4-gram，3-gram，2-gram まで多段階処理を行う．これは，5-gram では解答できない問題 (フレーズがデータにないため，頻度の比較が出来ない問題) が多いが正解率は高く，2-gram では解答できない問題は少ないが正解率が低かったためである．二つ目の手法は，上述の方法で頻度を正規化する方法である．例えば，“was launched 〜” の頻度を問題箇所以外の “was launched” の頻度で割って正規化する．これによって，検索するフレーズに含まれる高頻度の単語が解答に悪影響を及ぼす問題を解決できると考えられる．

次に前述の手法の改良について二種類の方法を検討する．一つ目として，上述の例題の “30” のような数値の単語は他にも “10”，“20” のように数限りなく存在するため，N-gram データに未登録の場合が多くなり，正解率の低下が懸念される．そこで，予め数値を表す単語をクラス分けする．このクラス分けを N-gram データと問題文の両方に施す．二つ目として，副詞は文法上の位置を指定されていないので，予め削除する．

表 1: 問題文と N-gram を改良して解答した結果

		数値データのグループ化	
		なし	あり
問題文中からの副詞の削除	なし	頻度 62.67 % 正規化 66.07 %	頻度 62.34 % 正規化 65.93 %
	あり	頻度 61.35 % 正規化 66.39 %	頻度 61.09 % 正規化 66.26 %

3 実験

実験では，前節で提案した頻度をそのまま用いる手法と頻度を正規化する手法の有効性を確認する．また，数値と副詞に対処した手法についても検証する．

穴埋め問題は AP 通信の新聞記事の一部から 9270 問作成した．空欄箇所に入る正解候補として，前置詞を 63 種類用意した．表 1 に実験結果を示す．頻度情報を用いた多段階処理よりも頻度の正規化を行った多段階処理のほうが正解率が高かった．これは，フレーズの頻度を問題箇所以外の単語列の頻度で正規化することの有効性を示している．数値のグループ化の有無について見ると，頻度を用いた場合と正規化を用いた場合のどちらもグループ化によって正解率が僅かに低下した．この問題に対する検討は今後の課題とする．また，副詞の削除の有無では，削除によって頻度を使った場合は正解率が下がったが，正規化の場合は正解率が上がった．これは，問題文から副詞を削除すると未登録の単語や正解に関係する単語が検索するフレーズに入るためと考えられる．

4 考察とまとめ

本稿では，N-gram データを用いて穴埋め問題を解答する方法と，予めデータや問題文を改良したものを使うことによってどのような効果があるのかを検討した．また，N-gram データの頻度情報を用いた場合と頻度情報を正規化したものとを比較した．その結果，データの使い方については，頻度情報を正規化したほうが正解率が良い事が分かった．数値以外の単語についてもグループ化を検討することが今後の課題である．

参考文献

[1] 大鹿広憲:”検索エンジンを使った英作文支援システムの構築”, Waseda University, 2004, http://hdl.handle.net/2065/779.

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