E-008

構文情報を考慮した検索英文集合に対する汎化手法

An Approach to Simplify Retrieved English Sentences in Consideration of Their Structures

天野翼† 渡部孝幸‡ 田中省作* 宮崎佳典**
Tsubasa Amano Takayuki Watabe Shosaku Tanaka Yoshinori Miyazaki

1. はじめに

英作文をする際，学習者が適切な語や構文を選択するために例文を参考にするという方法がしばしば見受けられる．しかし，参考にする例文が学習者にとって過度に複雑である場合，学習者が例文のどこに注目して良いのか判断するのは一般的に容易ではなく，結果として折角の例文が活用されない可能性がある．そこで我々は，例文を簡略化して提示する手法として汎化を提案してきた[1][2]．例文を汎化することで，学習者の英作文作業に貢献する可能性が高くはないと考えられる語が品詞に置き換えられ，学習者は参考になる語を把握しやすくなると考えられる．

本発表では，従来提案してきた汎化手法を改良し，「学習者が英作文を行う上で参考になると思われる語の提示」，「参考になる語と強い関連を持つ語の提示」，「構文構造に基づく語数の削減」という 3 点を満たす新たな汎化手法と，その手法を導入したシステムについて述べる．

学習者が例文を参照する状況として，学習者が何らかの方法で大量の英文から複数の例文を取り出し，それら例文を閲覧するという場合を考える．例文の抽出方法には「特定の語を含む例文を取り出す」，「入力した英文と類似する文を取り出す[3]」といったケースが想定されるだろう．学習者が抽出した例文中に共通して含まれる表現は学習者が英作文を行う上で参考となる表現である可能性が高いと思われる．そこで，学習者が取り出した例文の中で低頻度・低文数である語から順に品詞化する．

品詞化する度合（回数）は学習者によって決められる．そのため，“cache memory”のような共起関係にある語の片方のみが品詞化された状態で提示されてしまうことがある．しかし，“cache”と“memory”は二語で 1 つのまとまった意味を有しているため，一方のみが残されても，残された語を英作文に活用できないと推測される．そこで，例文中に共起関係にある語が存在するのであれば，それらの語を同時に品詞化するのが賢明である．

例文の複雑さは単に語を品詞化するだけでは減少させることができず，可読性向上のため，必要に応じて語数を減らす必要があると考えられる．そこで，構文構造に基づき，例文中の句内の各単語が，品詞化されている あるいは 機能語である，という条件を満たす場合，その句全体を，句を表す記号に置き換えることによって語数を削減する．

†静岡大学情報学部 Faculty of Informatics, Shizuoka University
‡静岡大学自然科学系教育部 Graduate School of Science and Technology, Shizuoka University
*立命館大学文学部 College of Letters, Ritsumeikan University
**静岡大学大学院総合科学技術研究科情報学領域 Department of Informatics, Graduate School of Integrated Science and Technology, Shizuoka University

2. 汎化

2.1 システムの構成

本システムにおける例文汎化処理の概念図を図 1 に示す．

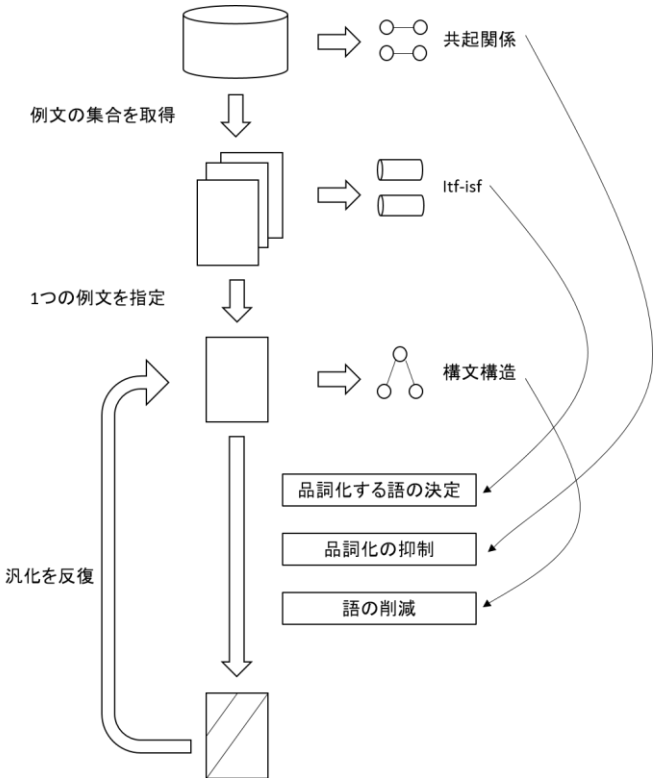

図 1 例文汎化処理の概念図

まず，学習者は大量の英文の集合から複数の例文を取得する．取り出された例文の集合をS，個々の例文をs_iとする．また，例文$s = w_1 w_2 \ldots w_n$とし，sの構文木をT_sとする．

あらかじめ，システムでは文の集合から共起関係の情報を取得している．そして，学習者が取得した例文中の各語から itf-isf（itf-isf については次節で説明する）を算出する．学習者が取得した例文のうち 1 つを選択すると，選択された例文の構文解析が行われる．

その後，学習者は選択した文に汎化を繰り返し行う．汎化は「品詞化する語の決定」，「品詞化の抑制」，「語の削減」の 3 つの要素で構成され，あらかじめ取得していた共起関係，itf-isf，構文構造の情報を用いて行われる．個々の要素については本節後に具体的に述べる．

2.2 品詞化する語の決定

品詞化する語は，学習者にとって参考となる程度が低いと推測されるべきである．そこでSに含まれる全ての語w（語はレンマ化されているものとする）について，次の値を計算する．

$$\text{itf–isf}(w;S) = \frac{1}{f(w)} \log\frac{|S|}{sf(w)} \quad (1)$$

$f(w)$はSに含まれるwの数，$sf(w)$はSのうちwを含む文の数，$|S|$は取り出された例文の総数を表す．

この itf-isf は低頻度・低文数であるほど大きな値をとる指標である．itf-isf の値が大きくなる語は，学習者が取得した例文の集合Sにおいて共通して出現する語ではなく，文もしくはそれが含まれた文書のトピックに強く依存したものであると考えられる．すなわち，itf-isf の値が高い語は，英作文の参考にならないと考えられる．そこで，itf-isf の値が最大となる語（itf-isf が等しくなるような語が複数存在する場合は，それらの語すべて）を品詞に置き換える．このような品詞化を繰り返すことで，例文中の個別的表現がマスクされる．そのため，学習者に英作文の参考になると思われる語を提示することができると考えられる．

例として，学習者が次のような例文の集合Sを取得したと仮定する．

s_1: We propose the new method in this paper.
s_2: We propose the new method in the paper.
s_3: We propose the new solution in the paper.
s_4: We propose the solution in the paper.
s_5: We propose the new solution.

上記の例文の集合Sにおいて語の itf-isf を算出すると，this，method，solution の順に値が大きくなる．ここで一度品詞化することによって，this が品詞に置き換わる．さらにもう二度品詞化すると method と solution も品詞に置き換わる．それぞれの品詞は<DT>（限定詞），<NN>（名詞），<NN>（名詞）であるので文集合Sは次のようになる．

s_1: We propose the new <NN> in <DT> paper.
s_2: We propose the new <NN>in the paper.
s_3: We propose the new <NN>in the paper.
s_4: We propose the <NN>in the paper.
s_5: We propose the new <NN>.

2.3 品詞化の抑制

語のみではなくその語を中心とする表現に着目し品詞化を行うことで，学習者に有用な情報を与えることができると考えられる．例えば，cache と memory が共起関係にあるとみなされ，その表現が例文中に含まれているという状況を考える．memory の itf-isf の値が cache より高かった場合，memory は品詞化され，先ほどの表現は cache <NN>という形で表示される．cache と memory は 2 語で 1 つのまとまりを成しており，他方だけ（ここでいう cache）が残されたとしても，ユーザに資する情報となるとは考えにくい．そのため，共起関係にある語は併せて品詞化する必要があると考えられる．

故に，品詞化されていない語w_jと強い共起性が認められた語w_iについては，たとえ itf-isf が高くとも品詞化せず，w_jを品詞化すると同時にw_iも品詞化することとする．具体的には，memory の itf-isf の大きさが m 番目の場合であっても，itf-isf の大きさが n（＞m）番目である cache と共起関係にあるので memory を品詞化しない．その後，itf-isf の大きさが n 番目である cache が品詞化されるのであれば，それと同時に memory も品詞化する．

このことから，適当な共起指標 cooc の下で，維持すべき（品詞化しない）s中の語w_iは次のいずれかを満たすものとし，また，これを$m(w_i;s,S)$あるいは$m(w_i)$とする．

i. $\text{itf–isf}(w_i;S) < \alpha$
あるいは (2)
ii. $\exists w_{j(1\le j\ne i\le n)}[\text{cooc}(w_i,w_j) > \beta \wedge \text{itf–isf}(w_j;S) < \alpha]$

なお，ここにα，βはユーザまたはシステムが定める適当な定数とする．

2.4 語の削減

例文を品詞化しても語数は変わらず，品詞化を行うたびに品詞は多くなる．このことは，学習者が例文を読みづらくなるといった事態を招く可能性がある．そこで，品詞化後の文の語数を減らし，可読性を高めることを目指す．

例として，ある例文を何度か品詞化した結果，次の文が得られたとする．

<PP> propose a <JJ> solution for the <NN> of <DT> <NN>.

この例文の末尾は句内が品詞と機能語の羅列（下線部）であり，英作文をする上で参考になるような情報を有していないと考えられ，冗長である．これらの語を句を表す記号に置き換えることによって，文全体の語数を削減する．

2.4.1 構文木の非冗長性

語数を減らすために，最初に例文を構文解析し構文木Tを得る．構文木Tにおける語や中間ラベルは，Tの走査順序などで区別されるものとする[1]．また，中間ラベルXから直接導出される語や中間ラベルの集合を$X\downarrow_T$と表す．ただし，Tが明らかな場合は，$X\downarrow$と省略することとする．T内で葉であるラベルは，何も導出しないので$X\downarrow_T = \emptyset$となる．

構文木で，$m(w)$を満たさないような語は不要であると考えられるので，その品詞もしくは中間ラベルに置き換えられることになる．Xを頂点とする部分木で，品詞あるいは中間ラベルのみを導出するものは結局Xの文法構造であり，内容や語に依存しない文法的な情報を示唆するだけで，冗長である．したがって，Xを頂点とする冗長な部分木であること *R(X;T)* は，次のように再帰的に定式化される．

i. $W(X) \wedge [\neg m(X) \vee I_N(X)]$
あるいは (3)
ii. $\neg W(X) \wedge \forall Y \in X\downarrow [R(Y;T)]$

ここに$W(X)$はXが語なら 1，そうでないなら 0 の値を取り，$I_N(X)$はXが特別にマスクする語である場合にのみ 1 の値を取るものである．

(3)から，Xを頂点とする部分木が非冗長であること$U(Y;T)$は，次のように定式化される．

[1] たとえば“the”が複数回T内に現れていても，走査順序等で区別され，異なる“the”として考える．

i. $W(X) \rightarrow (m(X) \wedge \neg I_N(X))$
かつ (4)
ii. $\neg W(X) \rightarrow \exists Y \in X \downarrow [U(Y;T)]$

2.4.2 木構造の非冗長化と文のマスク

与えられた文の木構造の非冗長化は，(3)が成り立つ中間ラベル下の部分木を全て削除すれば，(4)を満たす構文木となる．実装方法は以下のとおりである．

1. 構文木中で，次のような語や中間ラベルを算出する．

$$\Gamma = \{X | \left[W(X) \wedge \left(\neg m(X) \vee I_N(X)\right)\right] \vee [\neg W(X) \wedge X \downarrow = \emptyset]\}$$

2. それぞれの$X \in \Gamma$に対して，$(X\uparrow)\downarrow \subset \Gamma$ならば，$\Gamma$より$(X\uparrow)\downarrow$を取り除き，$X\uparrow$から導出される部分木を$X\uparrow$に置換する．
3. 構文木が変化しなかった場合は構文木を出力し，終了．そうでなければ 1 へ．

2.5 汎化の例

例として以下の状況を設定する．

・ “We propose a new solution for the equation of the language model.” を，汎化対象とする例文とする．
・“propose, a, solution, the, for, of”は itf-isf の値が小さく，汎化されない語とする．
・“propose”と“solution”，“solution”と“for”が共起する．

例文を構文解析することで得られる構文木を図 2 に示す．

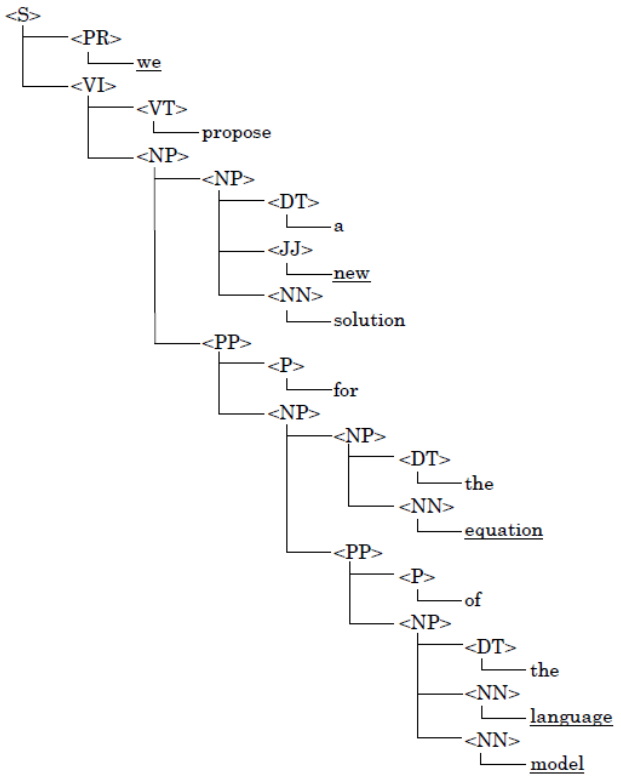

図 2　例文から得られる構文木

この例文に汎化を繰り返し行うことによって，itf-isf の値が大きい語から汎化されてゆき，図 2 で下線が引かれた語が汎化される．このため，下線部の葉がすべて削除され，品詞に置き換えられる．すると，構文木は以下の図 3 のようになる．

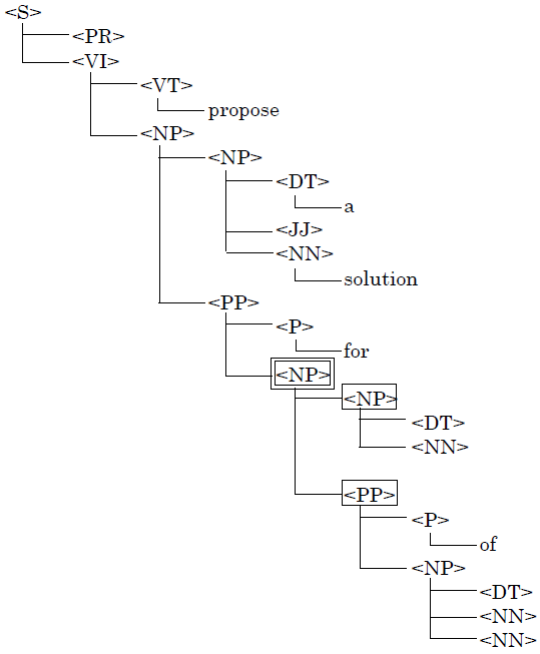

図 3　語が汎化された構文木

図 3 において，四角形で囲まれている節<NP>，<PP>はすべての子が「品詞」あるいは「品詞化されていない語と共起関係を持たない機能語」を表す．このため，この箇所は品詞に置き換えられる．同様の処理を行うことで，二重線で囲まれた節<NP>の子が削除され，品詞に置き換えられる．すると，構文木は図 4 のようになる．

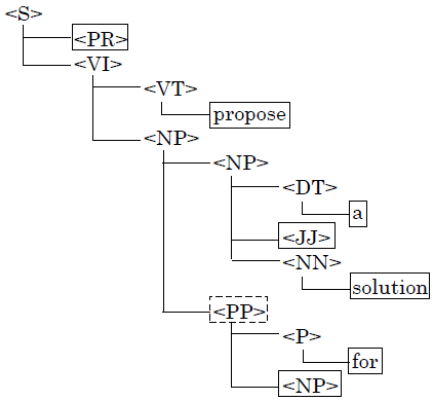

図 4　語数の削減を行った構文木

for と solution は共起関係にあるとみなされているので，破線で囲まれた<PP>には置き換えは適用されない．以上の手順によって，例文を汎化した場合は以下のようになる．

<PR> propose a <JJ> solution for <NP>.

このように例文の構文構造を考慮することで，the <NN> of <DT> <NN> <NN>という箇所を，句として置換することができた．それにより語数が減り，学習者にとって見やすい例文になると期待される．

2.6 学習者による汎化回数の指定

英語の習熟度は学習者ごとに異なるため，汎化する回数は学習者が指定できるようにする．学習者は自身にとって例文が複雑であると感じるのであれば汎化する回数を多くし，そうでなければ汎化を行わなくても良い．

2.7 最低一語の品詞化

学習者が 2.2 節の例で述べた例文の集合 S を取得し，例文を汎化する場合，1 つの例文に注目していると考えられる．品詞化される語は，例文の集合 S から決定されるため，学習者が注目している例文中に存在しているとは限らない．そのため一度汎化をしても，注目している例文の語が品詞化されない場合がある．例えば，学習者が以下の例文に注目していたとする．

s_2: We propose the new method in the paper.

一回目の汎化で品詞化される語は this であり，学習者が注目している例文中には存在しない．そのため，学習者が一度汎化しても，注目している例文の語は品詞化されない．この挙動は学習者を混乱させてしまう．そこで，文内汎化という手法を用いる．文内汎化とは，学習者が汎化を選択した際に，例文の集合 S 中の一語を品詞化するのではなく，個々の例文s_i中の一語を品詞化することである．例えば，先の例であれば学習者が注目している例文中の語で method の itf-isf の値が最も大きいので一度汎化を行うと，method が品詞に置き換わる．

3. システムの実装

我々は，提案手法のプロトタイプを web アプリケーションとして実装した．そのインタフェースを図 5 に示す．

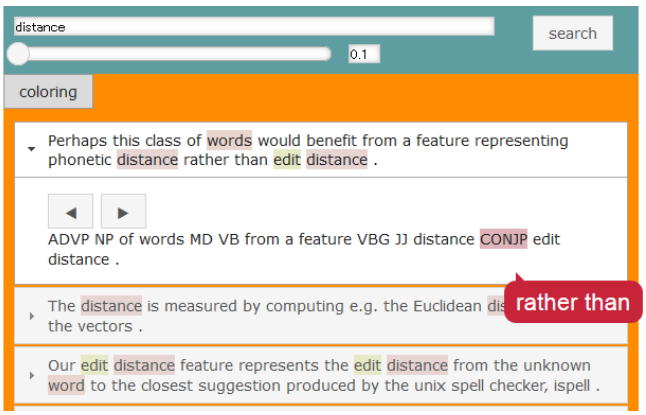

図 5　web アプリケーションのインタフェース

例文は，入力した 1 つ以上の語をもとに構成される特徴ベクトルと，例文から構成される特徴ベクトルとの間のコサイン類似度が高いものから順に表示される．特徴ベクトルは，語の頻度を特徴量とするものである．語を入力する部分の直下にあるスライダーは閾値を指定するためのものであり，スライダーで設定された閾値以上の類似度を持つ例文が取得される．

例文をクリックすると，汎化ビューが表示される．汎化ビューの ▶ ボタンをクリックすると文内汎化が 1 度行われ，◀ ボタンをクリックすると文内汎化が 1 段階前に戻る．汎化された箇所にカーソルを合わせると，汎化前の語の列が表示される．

coloring ボタンをクリックすると，例文集合において頻度の高い語をハイライト表示する．図 5 は，coloring ボタンをクリックした状態である．

4. まとめと今後の展望

本発表では，英作文支援として学習者が参考にする例文の汎化を提案した．itf-isf を算出し，学習者にとって参考にならないと思われる語を品詞化することで，「学習者が英作文を行う上で重要であると思われる語を提示」することができると考えられる．加えて，共起関係と構文構造を考慮することによって「重要度の高い語と強い関連を持つ語の提示」と「構文構造に基づく語数の削減」を可能にした．

英文の集合から抽出される例文は学習者が意図しているものであるという前提で汎化手法について述べたが，コサイン類似度を用いて例文抽出を行ったとしても学習者が意図していない例文が多く含まれてしまう可能性がある．学習者が意図するような例文を抽出するための研究は行われており，これらの研究を本手法に組み込むことで，より学習者の意図を汲んだアプリケーションに昇華してゆくと期待する．

今後は，本手法が実際に英作文に効果を発揮することを検証する予定である．特に，技術文書のような複雑な英作文を対象として検証を行うことを計画している．例えば既出の[3]は，技術文献コーパスを用いた例文提示型英文書作成支援ツールを開発するものであり，提示される例文が往々にして長文ゆえ，常に学習者の英作文に資する形で提供できているとは限らない．本研究はこの問題に対する 1 つの有力な解法になると考えられる．

参考文献

[1] 田中省作, 宮崎佳典, 池本孝徳, 小山由紀江, 大規模 n-gram データベースを活用した英作文支援のための文集合の汎化に関する一考察, 統計数理研究所共同研究リポート 254, pp. 1-19 (2011).

[2] 宮崎佳典, 田中省作, 小山由紀江, コーパスを用いた英語技術文書作成補助ツールの評価 -データ分析を通じて-, 統計数理研究所共同研究リポート 276, pp. 1-21 (2012).

[3] 戸沢信晴, 宮崎佳典, 田中省作, 技術文献コーパスを用いた例文提示型英文書作成支援ツールの開発, 電子情報通信学会技術研究報告信学技報, 114(82), pp. 69-72 (2014).