Vol.2017-NL-233 No.16
2017/10/25

カーネル埋め込みを用いた英語学習者向けの用例検索

塩田 健人[1,a)]　小町 守[1,b)]　池谷 瑠絵[†1,c)]　持橋 大地[†2,d)]

概要：
我々は英作文支援のアプローチの一つである用例検索に取り組む．一般的なキーワード検索においてユーザーが言語学習者である場合，情報要求に即した適切なクエリをユーザーが入力できない問題がある．そこで本研究では，クエリの背景にある潜在的な情報要求を考慮するために，カーネル埋め込みを用いた用例検索モデルを提案する．カーネル埋め込みと内積に基づく単純なクエリ-文間の類似度計算手法では，クエリと関係の弱い単語がノイズとなるが，N-gram 窓の導入によって我々はこの問題を解決した．英語学習者によって収集されたクエリ-適合文のデータセットによる実験の結果，提案手法は文間類似度タスクの先行研究における教師なし手法より高い適合率を達成した．

1. はじめに

近年，多くの英作文支援ツールが研究されている．英作文の支援をすることは，英語学習者にとって有意義である．しかし，熟練した英語学習者であっても特定のドメインにおいて適切な表現や様式に沿って英文を書くことは難しい．従って，英語学習者が英文を書く際に書きたい文に関係するキーワードを用いて特定ドメインのコーパスに基づき、英文を検索して表示するシステムは有益である．英語学習者が書きたい文に関係する英文を検索する場合，Google や Yahoo!などの既存の検索エンジンを用いてキーワードに関連する英文を検索することがあると考えられる．しかしながら，既存の検索エンジンでは英語学習者が英文を書く際に用例検索をすることに最適化されていないため，英語学習者が期待するような検索結果を得られることは難しいと考えられる．

また，既存の英作文検索ツールは，学習者がクエリに入力した単語の表層を利用して用例文を検索するものが多い．そのようなツールにおいては，英語学習者の書きたい英文を表すクエリ，つまり，学習者が持つ情報要求に即したクエリを考えて入力することが前提とされている．しかし，学習者にとっては英文を表現する適切なクエリを考えることは困難であると考えられる．そこで，我々はクエリの背景にある潜在的な情報要求を満たす新しい文検索手法を提案する．

我々が提案する手法は，クエリと検索対象の文に潜在的な確率分布が存在すると仮定することにより，各分布間の距離が近いものを潜在的な意味を考慮したクエリと文の組み合わせとして扱うことを可能にする．さらに，潜在的な確率分布を考慮することにより，文検索においてクエリに表現力を追加することができる．潜在的な確率分布と分布間の距離を表現するために，分布のカーネル埋め込みの枠組みを用いてこの問題に取り組む．

加えて，クエリと検索対象文の分布間の距離は内積によって計算されるが，この方法ではクエリの単語と全く関係がない文中の単語まで計算に考慮されてしまい，クエリの潜在的な意味が十分に反映されない問題がある．そこで，我々は N-gram 窓を用いることにより，クエリと関係度が高い文中の単語をピンポイントで考慮することを可能にすることを示し，文検索において高い適合率を達成した．

本研究の貢献は以下の 3 つである．

- 分布のカーネル埋め込みと N-gram 窓を用いた新しい文検索の類似度計算法を提案した．
- 大学広報に関するコーパスを作成し，2 語のクエリに関連する英語学習者のための例文をアノテーションした．
- 我々が作成したコーパスを用いた実験で，提案手法が先行研究である教師なし文間類似度計算法に対して高い適合率を達成した．

[1] 首都大学東京
Tokyo Metropalitain University
[†1] 情報・システム研究機構
Research Organization of Information and Systems
[†2] 統計数理研究所
The Institute of Statistical Mathematics
[a)] shioda-kent@ed.tmu.ac.jp
[b)] komachi@tmu.ac.jp
[c)] ikeya@rois.ac.jp
[d)] daichi@ism.ac.jp

2. 分布のカーネル埋め込みによるクエリ-文間類似度計算法

我々は単語が潜在的な確率分布を持つと仮定することにより，分布のカーネル埋め込み [1] と呼ばれる手法を利用してクエリと文の潜在的な確率分布を比較できる新しい文検索手法を提案する．分布のカーネル埋め込みとは，確率分布をカーネル k によって定められる高次元空間上にマップすることである．この手法により，クエリが持つ潜在的な意図を考慮することが可能になる．

さらに，通常高次元空間上で分布間の類似度を計算する際には内積が用いられる．内積の計算には，文中に含まれる全ての単語を考慮するため，文の長さによって検索結果に悪影響が出てしまう知見が予備実験を通して得られた．そこで，我々は計算する際に N-gram で文を区切ることにより，文中のクエリとの関係性が高い部分のみを考慮することを可能にし，この問題を解決した．

従って，我々の手法は 2 単語を入力とし，出力として文中の N-gram に基づいてクエリと関係のある文を検索する．以下のサブセクションでは，分布のカーネル埋め込みをどのように文検索タスクに適応させ，適合率を上げるためどのように N-gram 窓を取り入れたのかを説明する．

2.1 分布のカーネル埋め込み

Yoshikawa ら [2] は異なるドメイン間のインスタンスの類似度を計算する手法を提案した．Yoshikawa らの手法は，Smola ら [1] が提案した分布のカーネル埋め込みの枠組みを用いて，各ドメインの全てのインスタンスの素性を潜在的共有空間に埋め込むことによって類似度計算を可能にしている．分布のカーネル埋め込みとは，任意の空間 $\mathcal{X}$ 上の確率分布 $\mathbb{P}$ をカーネル k で定義される再生核ヒルベルト空間（RKHS）$\mathcal{H}_k$ に埋め込む際に使用される．ここで，マップされた確率分布 $\mathbb{P}$ は RKHS 上のインスタンスとして表現される．

我々は Yoshikawa らの手法を拡張し，文検索タスクに適応させた．我々の手法は，クエリや文を単語の集合とみなし，さらに各単語には高次元空間である RKHS 上にマップされる潜在的な確率分布が存在すると仮定する．以上の仮定より，クエリと文は RKHS 上のインスタンスとして表現され，マップされたインスタンス間の類似度を計測することによりクエリと文との類似度を比較することができる．本研究では、潜在的な確率分布を表現するために word2vec によって学習された単語分散表現を使用する．クエリ q と文 s に含まれる単語の分散表現 $\vec{q_i}$ と $\vec{s_j}$ は，カーネル k で決定される RKHS $\mathcal{H}_k$ 上のインスタンス $\mu_{\mathbb{P}_q}$，$\mu_{\mathbb{P}_s}$ として表される．ここで，本研究で扱う単語分散表現は独立同分布なサンプルとする．以下に RKHS 上に表現されるクエリのインスタンスを示す．文のインスタンスも同様に決定される．

$$\mu_{\mathbb{P}_q} = \frac{1}{|q|}\sum_{l=1}^{|q|} k(\cdot, \vec{q_l}) \in \mathcal{H}_k \tag{1}$$

次に，RKHS 上のインスタンス間の類似度計算手法を示す．2 つの集合が独立同分布であると仮定すると，同じ空間上の集合 $X = \{\mathrm{x}_l\}_{l=1}^{n}$，$Y = \{\mathrm{y}_{l'}\}_{l'=1}^{n'}$ は分布のカーネル埋め込みによって RKHS 上で $\mu_{\mathbb{P}_X}, \mu_{\mathbb{P}_Y}$ と表現される．これら 2 つのインスタンス間の距離 $D(X,Y)$ は以下の式によって計算される．

$$\begin{aligned} D(X,Y) &= ||\mu_{\mathbb{P}_X} - \mu_{\mathbb{P}_Y}||^2_{\mathcal{H}_k} \\ &= \langle \mu_{\mathbb{P}_X}, \mu_{\mathbb{P}_X} \rangle_{\mathcal{H}_k} + \langle \mu_{\mathbb{P}_Y}, \mu_{\mathbb{P}_Y} \rangle_{\mathcal{H}_k} - 2\langle \mu_{\mathbb{P}_X}, \mu_{\mathbb{P}_Y} \rangle_{\mathcal{H}_k} \end{aligned} \tag{2}$$

従って，式 2 の第 3 項は集合 X と Y に基づいた RKHS 上の両インスタンス $\mu_{\mathbb{P}_X}$，$\mu_{\mathbb{P}_Y}$ が依存する項である．よってクエリ q と文 s の距離 $D(q,s)$ を式 2 から導出し，我々は以下に示すようにクエリ-文間類似度 sim_{ke} を定義する．

$$\begin{aligned} sim_{ke}(q,s) &= \langle \mu_{\mathbb{P}_q}, \mu_{\mathbb{P}_s} \rangle_{\mathcal{H}_k} \\ &= \Big\langle \frac{1}{|q|}\sum_{i=1}^{|q|} k(\cdot, q_i), \frac{1}{|s|}\sum_{j=1}^{|s|} k(\cdot, s_j) \Big\rangle_{\mathcal{H}_k} \\ &= \frac{1}{|q||s|}\sum_{i=1}^{|q|}\sum_{j=1}^{|s|} k(\vec{q_i}, \vec{s_j}) \end{aligned} \tag{3}$$

2.2 N-gram 窓

分布のカーネル埋め込みを用いた手法は，キーワードベースの文検索タスクにおいて再現率を上げることに関して強みがあると考えられる．一方で類似度を計算する際，単純に内積を使う手法だと文中に含まれる全ての単語を考慮するため，クエリと全く関連がないとされる単語まで考慮され，精度が下がってしまう可能性がある．そこで，類似度を計算する際に文を N-gram で区切ることによりこの問題を解決する．

我々が用いたアルゴリズムを Algorithm 1 に示す．はじめに検索対象の文を N-gram に区切り，クエリと各 N-gram の類似度を計算する．そして，クエリと全ての N-gram の中から類似度が最大になるものをクエリと文の類似度とみなす．

3. 英語学習者向けの用例検索実験

3.1 実験設定

我々は，Google News dataset を用いて word2vec で学習済みの公開されている単語分散表現 [*1] を使用した．また，文

[*1] https://code.google.com/archive/p/word2vec/

Algorithm 1 文間類似度計算

```
Input: sentence, query, Output: similarity
max_SIM ← 0
for each N-gram ∈ sentence do
    SIM ← sim_ke(query, N-gram)
    if SIM > max_SIM then
        max_SIM ← SIM
    end if
end for
return max_SIM
```

をトークナイズする際に Stanford Core NLP tokenizer (Ver. 3.6.0)[*2] を使用した．計算処理をする際，トークナイズされた単語は全て小文字化した．我々は式 3 中のカーネル k としてコサイン類似度と RBF カーネルを実験に使用した．それぞれのカーネル k を以下に示す．

$$k_{cos}(q_i, s_j) = \frac{\langle q_i, s_j \rangle}{|q_i||s_j|} \tag{4}$$

$$\begin{aligned} k_{\mathrm{RBF}}(q_i, s_j) &= \exp\left(-\frac{||q_i - s_j||^2}{2\sigma^2}\right) \\ &= \exp\left(-\gamma||q_i - s_j||^2\right) \end{aligned} \tag{5}$$

本研究では RBF カーネルのハイパーパラメータである γ を $\gamma \in \{10^{-1}, 10^0, 10^1, 10^2\}$ の範囲で用いて予備実験を行い，その結果を踏まえ $\gamma = 10^1$ に設定した．

3.2 データ

本研究では，大学のプレスリリースドメインの記事に対して実験を行う．我々は英語学習者に向けた文検索に使用するデータセットを以下に示すように構築した．

はじめに，“.edu” をドメインネームの末尾に含むウェブページから本文を 579,867 文抽出し，コーパスを作成した．2 語からなる 30 組のクエリをアノテーター 1 名によって作成し，それぞれのクエリに含まれる 2 単語を完全一致で含む文を抽出した．さらに，アノテーターは抽出した文がクエリの検索結果として適切か否かを評価した．アノテーターによって評価されたデータを実験での評価データとした．

次に，テストデータに関して説明する．我々はアノテーターによって適切と判断された文が最低 10 文あるクエリを 10 組選択し，各クエリにつき 10 文を正解文とした．不正解文として，各クエリにおいてアノテーターに検索結果として適切でないと判断された文を 90 文選択した．適切でないと判断された文が 90 文存在しない場合，評価データから不正解文が 90 文になるようにランダムにサンプリングした．アノテーターによってクエリの検索結果として適切であると判断された文の全てにクエリを構成している 2 単語が含まれている．また，実験で使用したテストデータの平均文長は 30 単語であった．

[*2] http://nlp.stanford.edu/software/stanford-corenlp-full-2015-12-09.zip

3.3 評価

Precision@k（以下，p@k）で検索結果を評価し，以下に示すベースラインと比較した．

3.3.1 単語ベクトルの平均による類似度

シンプルなベースラインとして，クエリと文に含まれている単語のベクトルの平均類似度 sim_{ave} を使用した．単語ベクトルとして word2vec の単語分散表現を使用した．ここで sim_{ave} を式 6 に示す．クエリのベクトルはクエリ q に含まれる各単語のベクトルの平均を取ったものとし，文ベクトルも同様に文 s に含まれる単語のベクトルの平均を使用した．我々は式 6 中のカーネル k にコサイン類似度と RBF カーネルを用いた．

$$sim_{ave}(q, s) = k(\vec{q}, \vec{s}) \tag{6}$$

$$\vec{q} = \frac{1}{|q|}\sum_{i=1}^{|q|}\vec{q}_i, \qquad \vec{s} = \frac{1}{|s|}\sum_{j=1}^{|s|}\vec{s}_j$$

3.3.2 アライメントベースの類似度

Song and Roth [3] によって提案された文間類似度計測手法の一つをベースラインとして使用する．彼らの手法は，Semantic Textual Similarity (STS) タスクにおいて当時の最高精度を達成した教師なし文間類似度計算法である．我々はその中で, 以下に示す分散表現のアライメント (Maximum Alignment) に基づいた手法をベースラインとして用いた．

$$sim_{max}(q, s) = \frac{1}{|q|}\sum_{i=1}^{|q|}\max_j k(\vec{q}_i, \vec{s}_j) \tag{7}$$

この手法は，クエリ q の単語分散表現 $\vec{q}_i$ と文 s に含まれる単語分散表現 $\vec{s}_j$ 間の類似度の最大値をクエリの単語数 $|q|$ で割ったものをクエリと文の類似度とするものである．また，本研究では全ての単語間の類似度を使用した．ここで，我々は式 7 中のカーネル k として提案法並びに，平均ベクトルによる類似度と同じくコサイン類似度と RBF カーネルを使用した．

3.4 実験結果

図 1 と図 2 に実験結果を示す．本研究では，1-gram から 40-gram までの N で実験をし，1-gram から 5-gram に加えて 10-gram ずつプロットした．

図 1 ではカーネル k にコサイン類似度を使用したものを示した．3.3.1 に示したベースラインのコサイン類似度を使

表 1
19-gram を用いた際に出力された正解文（RBF カーネル）と不正解文（3.3.1 で示したコサイン類似度）

kernel	label	input query: partnership support
RBF	✓	The advisers work in *partnership* with the college staff and other university offices to provide information and *support* for all students and to offer programs on community issues as well as small-scale social activities.
Cosine	×	The Robert Mehrabian CIC is a *partnership* between Carnegie Mellon, the Carnegie Museums, and local economic development organizations and is funded with $8 million in Commonwealth of Pennsylvania tax *support*.

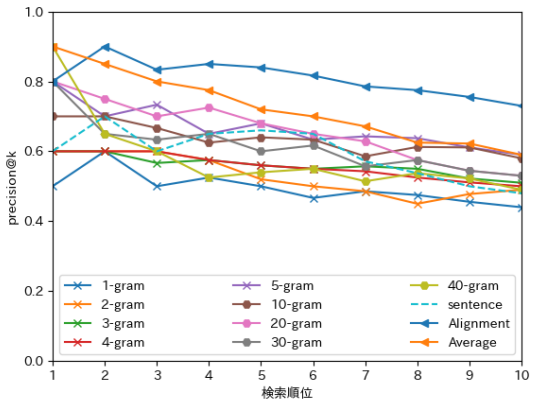

図 1 コサイン類似度の p@k

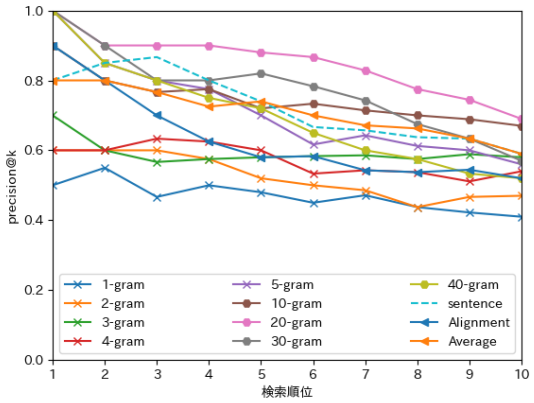

図 2 RBF カーネルの p@k

用したものは，分布のカーネル埋め込みを用いた手法と比較して低い適合率となった．また，上位 1 位を除き，3.3.2 のアライメントベースの手法が最も高い適合率を示した．

図 2 にはカーネル k に RBF カーネルを使用したものを示した．RBF カーネルを使用した場合，N-gram 窓を使用した方が高い適合率が得られる結果となった．加えて，上位 5 位において RBF カーネルと大きな N-gram 窓を組み合わせたモデルが最も良い結果を得られた．図 2 より，N-gram の窓幅は 20-gram が最も効果的であることが見受けられる．

しかしながら，1-gram から 3-gram の窓枠と RBF カーネルの組み合わせは低い適合率となることが観察された．次の節では，なぜ短い窓枠が低い適合率となってしまったのかを議論する．

3.5 考察

我々は追加実験の結果から一番高い適合率を示した RBF カーネルと 19-gram の組み合わせについてエラー分析を行った．表 1 はクエリ：partnership，support に対して検索結果の上位 10 件に出力されたカーネル k に RBF カーネルを用いた場合の正解文と 3.3.1 に示したベースラインとしてコサイン類似度を使用した際の不正解文の一例を示した．RBF カーネルによって出力された文は，partnership と support が文中で並列で使用されている．従ってこれらの単語は文中で比較的重要な役割をしており，このことから英語学習者が partnership と support をキーワードとして検索してきた際に，検索結果として参考になると判断していると考えられる．一方で，コサイン類似度を使用して出力された例の場合，キーワードの 2 語はそれぞれ文中で関連のない使われ方をしていることがわかる．このことは，潜在的なクエリの意図を考慮することができないため，例に示したようなクエリと関連がないとアノテーターによって判断されてしまった文が出力されてしまう．

次に，我々は検索対象の文中でクエリに含まれる 2 単語と完全一致する単語が何単語離れているかを計測した．結果として，2 語間は平均して 11.8 単語離れていた．また，正解文の 72%においてキーワードとされるクエリの単語が同じ節内にあった．短い窓幅の場合において低い適合率になった結果とこれらの事実から，このタスクにおいて英語学習者に有益とされる英文は文の中でキーワードとなる単語同士が近すぎず，かつ同じ節内に 2 つの単語が存在することであると考えられる．

最後に，我々はアライメントベースの手法と分布のカーネル埋め込みを用いた手法を比較する．アライメントベースの手法は，クエリと文中に含まれる最も類似度の高い単語のみを考慮している．一方で，分布のカーネル埋め込みによる類似度は文中に現れる単語を包括的に計算に組み込むことができる．さらに，N-gram 窓と組み合わせることにより，クエリと類似度の高い単語の周辺の単語を集中的に考慮することが可能になる．これらのことが N-gram 窓とカーネル埋め込みによる我々の提案法が，アライメントベースの手法より高い適合率を示した理由であると考えられる．

4. 関連研究

近年，多くの英作文支援システムが開発されている．その中の 1 つとして，松原らが作成した英文検索システム ESCORT [4] が挙げられる．このシステムは学術論文や調査報告などを書くときに使用されることを想定され，ユーザーが英文を作成する際に用いる単語の使い方の用例を見せることを目的としている．入力となるキーワード間に構文的な関係が存在する場合，それらキーワードを構文解析し，同じ構文構造をしている文を英語論文から取り出されてきた大量の文から構成されるコーパスから検索して出力するシステムである．しかし，このシステムは入力のキーワードからコーパス中の文を検索する際に語幹の完全一致で構文解析を行う．そのため，単語の分散表現で得られるような周辺文脈は同じであるが表記上は違う単語については検索対象から外れてしまう問題がある．また，ユーザーが思いつくキーワードに必ずしも構文構造があるとは限らない．この問題はキーワード間に構文構造がある前提で設計されているシステムにおいては致命的な問題である．また，完全一致で検索しているため，松原らの手法ではクエリの潜在的な意図をモデル化できていない．

一方，Chen ら [5] は英語学習者に向けて英作文支援ツールを開発した．この FLOW と呼ばれるシステムは，非ネイティヴの語彙力を補うことができる．英語学習者が英語を語彙力不足で書くことができない場合でも，FLOW を使えば彼らは第一言語で文の途中から書き進めることが可能である．FLOW は既に書かれている英文から文脈を認識することができ，文脈を考慮して書き手の第一言語を英語へと翻訳する．このシステムは書き手の潜在的な意図を第一言語で書くことを許容することにより考慮できていると言える．しかし，我々の手法では分布のカーネル埋め込みを使用することにより，文のモデル化を改善できる．

加えて，Hayashibe ら [6] は書き手が書くのと同時に英文を書く支援をするツールを開発した．彼らのシステムは，英語に加えてローマ字で書かれた日本語を入力として受け入れており，Chen ら [5] のように書き手の第一言語を考慮可能である．このツールはクエリに入力された情報に基づいて文脈に合った句を書き手に提示する．一方で，我々は 2 語だけを入力として要求している．また，検索する際に彼らの手法は入力の完全一致を使用しているため，検索結果における再現率に悪影響を与えている可能性がある．

5. おわりに

本研究では，英作文を支援するために単語に潜在的な確率分布が存在すると仮定し，分布のカーネル埋め込みを利用した新しい文検索手法を提案した．我々の RBF カーネルと N-gram 窓を組み合わせた分布のカーネル埋め込みによる手法は，単純なコサイン類似度とアライメントベースの手法と比較した際に高い適合率を示した．今後はクエリの入力単語数を 3 単語以上に増やすことによってもこの手法が有効かどうかを立証する．

参考文献

[1] Smola, A., Gretton, A., Song, L. and Schölkopf, B.: A Hilbert space embedding for distributions, *Proceedings of International Conference on Algorithmic Learning Theory*, pp. 13–31 (2007).
[2] Yoshikawa, Y., Iwata, T., Sawada, H. and Yamada, T.: Cross-Domain Matching for Bag-of-Words Data via Kernel Embeddings of Latent Distributions, *Advances in Neural Information Processing Systems 28*, pp. 1405–1413 (2015).
[3] Song, Y. and Roth, D.: Unsupervised Sparse Vector Densification for Short Text Similarity, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1275–1280 (2015).
[4] Matsubara, S., Kato, Y. and Egawa, S.: ESCORT: example sentence retrieval system as support tool for English writing, *Journal of Information Processing and Management*, Vol. 51, No. 4, pp. 251–259 (2008).
[5] Chen, M.-H., Huang, S.-T., Hsieh, H.-T., Kao, T.-H. and Chang, J. S.: FLOW: a first-language-oriented writing assistant system, *Proceedings of the ACL 2012 System Demonstrations*, pp. 157–162 (2012).
[6] Hayashibe, Y., Hagiwara, M. and Sekine, S.: phloat : Integrated Writing Environment for ESL learners, *Proceedings of the Second Workshop on Advances in Text Input Methods*, pp. 57–72 (2012).