16890523

修士論文

カーネル埋め込みを用いた英語学習者向け用例検索

塩田 健人

2018 年 2 月 23 日

首都大学東京大学院
システムデザイン研究科 情報通信システム学域専攻

本論文は首都大学東京大学院システムデザイン研究科に
修士（工学）授与の要件として提出した修士論文である.

塩田 健人

審査委員：

小町 守 准教授　（主指導教員）
石川 博 教授　（副指導教員）
片山 薫 准教授　（副指導教員）

カーネル埋め込みを用いた英語学習者向け用例検索*

塩田 健人

内容梗概

近年，英語学習者向けの学習支援に関する研究が多くされている．学習支援システムの中には，学習者が書いた英文の誤りを発見するものやその誤りを訂正するもの，また英文を書く際に補助となるような英作文支援システムが存在している．英作文の支援をすることは，英語学習者の英作文時の負荷を減らすと同時に間違いを減らすことに貢献するので有意義である．

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．そこで，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．

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，学習者が検索をするために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，学習者にとっては自身の情報欲求を再現するような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，我々は学習者が考えた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それ

*首都大学東京大学院 システムデザイン研究科情報通信システム学域専攻 修士論文, 16890523, 2018 年 2 月 23 日.

ぞれを表すような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と考えられる．我々は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んだ．分布のカーネル埋め込みとは，カーネルで表現される高次元空間上に分布をマップすることである．この手法を使うことにより，分布間の類似度を比較的容易に計算することが可能となる．加えて，クエリと検索対象文の分布間の距離はそれぞれのインスタンス間の内積によって計算されるが，単純なこの方法でのクエリ-文間の類似度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た．

英語学習者によってアノテーションされたクエリ-適合文のデータセットを使った Precision@k による評価実験の結果，我々の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．

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評価実験で，先行研究である教師なし文間類似度計算法に対して提案手法が高い適合率を達成した．

キーワード

修士論文，首都大学東京大学院，自然言語処理，学習支援，用例検索

Suggesting Sentences for English as a Second Language using Kernel Embeddings*

Kent Shioda

Abstract

In recent years, there are many studies on learning support for English as a Second Language (ESL) learners. There are some writing assistant tools that find and correct errors in English sentences written by learners. Assisting English writing is meaningful as it contributes to reducing the burden of ESL English composition at the same time as reducing the mistakes.

However, even for advanced ESL learners, it is difficult to write sentences conforming to the styles and expressions in a specific domain. In existing search engines, systems are not optimized for retrieving example sentences when writing English sentences, so it is considered difficult to obtain query intent expected by English learners. Also, existing tools search for example sentences using the surface layer of the words that the learner has entered in the query because the query is assumed to represent an English sentence that the ESL learner wishes to write. However, it is considered difficult for ESL learner to think of an appropriate query that expresses English sentences that reproduce his or her information need. Therefore, it is beneficial for non-native speakers to search for sentences using keywords that the writer aims to use. So, we tackle example sentence search with latent intent to support English composition for ESL learners.

ESL learners are familiar with web search engines, but generic web search results may not be adequate for composing documents in a specific domain. However, if we build our own search system specialized to a domain, it may be

*Master's Thesis, Department of Information and Communication System, Graduate School of System Design, Tokyo Metropolitan University, 16890523, February 23, 2018.

subject to the data sparseness problem.

Recently proposed word2vec partially addresses the data sparseness problem, but fails to extract sentences relevant to queries owing to the modeling of the latent intent of the query. We address this problem by using a kernel embeddings framework. Kernel embeddings make it possible to add expressiveness to the query in sentence retrieval by using latent probability distribution. In addition, our method of taking N-gram windows boosts the precision of sentence retrieval by considering words that are highly related to the query. This method implicitly models latent intent of query and sentences, and alleviates the problem of noisy alignment. Our results show that our method achieved higher precision in sentence retrieval for ESL in the domain of a university press release corpus, as compared to a previous unsupervised method used for a semantic textual similarity task.

The main contributions of this study are as follows:

- We propose a novel sentence similarity metric based on kernel embeddings and N-gram windows.
- We build a corpus of university press releases and annotated example sentences for ESL, given a query of two words.
- We show that our proposed method outperforms unsupervised baselines on our dataset.

Keywords:

Master's Thesis, TMU, Natural Language Processing, Learning Support, Example Sentences Retrieval

目次

図目次

第 1 章　はじめに

1.1　背景

近年，英語学習者向けの学習支援に関する研究が多くされている．学習支援システムの中には，学習者が書いた英文の誤りを発見するものやその誤りを訂正するもの，また英文を書く際に補助となるような英作文支援システムが存在している．英作文の支援をすることは，英語学習者の英作文時の負荷を減らすと同時に間違いを減らすことに貢献するので有意義である．

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．そこで，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．

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，学習者が検索をするために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，学習者にとっては自身の情報欲求を再現するような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，我々は学習者が考えた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それぞれを表すような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と考えられる．我々は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

題に取り組んだ．分布のカーネル埋め込みとは，カーネルで表現される高次元空間上に分布をマップすることである．この手法を使うことにより，分布間の類似度を比較的容易に計算することが可能となる．加えて，クエリと検索対象文の分布間の距離はそれぞれのインスタンス間の内積によって計算されるが，単純なこの方法でのクエリ-文間の類似度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た．

英語学習者によってアノテーションされたクエリ-適合文のデータセットを使った Precision@k による評価実験の結果，我々の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．

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評価実験で，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．

本論文の構成は以下のようになっている．第 1 章では本研究全体の概要，貢献を述べる．第 2 章では英語学習者向けの英作文支援の先行研究について述べる．第 3 章では分布のカーネル埋め込みによるクエリ-文間類似度計算法について詳しく述べる．第 4 章では英語学習者向けの用例検索実験ついて述べる．第 5 章では第 4 章での実験結果をもとに考察を述べる．最後に第 6 章で本研究のまとめ，今後の展望について述べる．

第 2 章　関連研究

2.1　英作文支援ツール

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 [1] 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．図 2.1 と図 2.2 に例を示す．入力となるキーワード間に図 2.1 のような構文関係が存在する場合，それらキーワードを構文解析し，図 2.2 のような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．

図 2.1　“We developed a system” の係り受け例

図 2.2　図 2.1 を入力とした出力例

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，単語の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 [2] 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 [3] 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 [2] 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．

2.2 カーネル埋め込み

自然言語処理を行うにあたって，単語や文などの意味を考慮した類似度を計算することは重要なタスクである．類似度を計算する際に，Glove [4] や word2vec [5] に代表される単語の分散表現を使用する手法が自然言語処理の様々なタスクにおいて成果を出している．単語の分散表現とは，単語をベクトル化する際に 1-of-K に代表されるような疎なベクトルではなく，密なベクトルとしてベクトル化を行う．この手法により， 1-of-K のベクトル表現では十分に考慮することができなかった単語の意味を考慮することが可能になり，単語の意味の演算が可能になった．しかし，単語ではなく文やフレーズ単位での意味表現をベクトル空間上にどのように落とし込むのか，ベクトル表現にした後にどのように類似度を計算するのかに関しては様々な研究が行われている．

また，カーネル埋め込みと呼ばれる手法は，確率分布が持つ高次元情報をカーネ

ルによって定義される再成核ヒルベルト空間上の点として表現する [6] [7]．この手法は異種データ間での分類，または画像や自然言語処理の分類タスクにおいて研究されており比較されている従来手法と比較していづれも高い成果を収めている [8] [9]．カーネル埋め込みを用いて，単語ベクトルを潜在変数として扱うことにより，低次元では汲み取ることの出来なかったより高次な情報を扱うことができる．

本研究では，カーネル埋め込みの枠組みを用いて用例検索タスクにおける新しいクエリと文の類似度計算法を提案する．この手法により，従来のコサイン類似度などの類似度計算手法だけでは考慮することが出来なかったクエリの意味，つまりクエリを入力するユーザの情報要求をより考慮することが可能であると証明した．

第 3 章　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 [6] 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．

3.1　分布のカーネル埋め込み

Yoshikawa ら [10] 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 [6] 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．

$$\mu_{\mathbb{P}_q} = \frac{1}{|q|}\sum_{l=1}^{|q|} k(\cdot, \vec{q_l}) \in \mathcal{H}_k \tag{3.1.1}$$

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．

$$\begin{aligned} D(X,Y) &= ||\mu_{\mathbb{P}_X} - \mu_{\mathbb{P}_Y}||^2_{\mathcal{H}_k} \\ &= \langle \mu_{\mathbb{P}_X}, \mu_{\mathbb{P}_X} \rangle_{\mathcal{H}_k} + \langle \mu_{\mathbb{P}_Y}, \mu_{\mathbb{P}_Y} \rangle_{\mathcal{H}_k} - 2\langle \mu_{\mathbb{P}_X}, \mu_{\mathbb{P}_Y} \rangle_{\mathcal{H}_k} \end{aligned} \tag{3.1.2}$$

従って，式 3.1.2 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 3.1.2 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．

$$\begin{aligned} sim_{ke}(q,s) &= \langle \mu_{\mathbb{P}_q}, \mu_{\mathbb{P}_s} \rangle_{\mathcal{H}_k} \\ &= \Big\langle \frac{1}{|q|}\sum_{i=1}^{|q|} k(\cdot, q_i), \frac{1}{|s|}\sum_{j=1}^{|s|} k(\cdot, s_j) \Big\rangle_{\mathcal{H}_k} \\ &= \frac{1}{|q||s|}\sum_{i=1}^{|q|}\sum_{j=1}^{|s|} k(\vec{q_i}, \vec{s_j}) \end{aligned} \tag{3.1.3}$$

3.2 文間類似度の数学的解釈

一方，Berger and Lafferty ら [11] は情報検索において，情報要求をクエリとして変換するプロセスを確立モデルとして示している．彼らの考え方を用例検索に適用すると，ユーザの情報要求に基づく文 s はベイズの定理より以下のように定式化できる．また，q はクエリ， w はクエリの言い換えを表す．

$$\begin{aligned} p(s|q) &= \frac{p(q|s)p(s)}{p(q)} \\ &\propto p(q|s)p(s) \\ &\simeq \frac{1}{|q||w|}\sum_{q}\sum_{w} p(q|w)p(w|s)p(s) \\ &= \frac{1}{|q||w|}\sum_{q}\sum_{w} \frac{p(w,s)p(q,w)p(s)}{p(w)p(s)} \end{aligned} \tag{3.2.1}$$

ここで，式 3.2.1 中の q と w の同時確率 $p(q,w)$ を用例検索において，検索対象の文 s 中の単語 w_j と クエリ q_i の類似度とすると，以下のように式変形される．また，k はカーネル埋め込みの際に使用するカーネル k によって定義される．

$$\begin{aligned} p(s|q) &= \frac{1}{|q||w|}\sum_{q}\sum_{w} \frac{p(w,s)p(q,w)p(s)}{p(w)p(s)} \\ &\approx \frac{1}{|q||w|}\sum_{q}\sum_{w} \frac{p(w,s)k(q,w)p(s)}{p(w)p(s)} \end{aligned} \tag{3.2.2}$$

また，式 3.2.2 中の $\frac{p(w,s)}{p(w)p(s)}$ は自己相互情報量 PMI 逆対数である．したがって，式 3.2.3 のように変形できる．

$$\begin{aligned} p(s|q) &= \frac{1}{|q||w|}\sum_{q}\sum_{w} \frac{p(w,s)k(q,w)p(s)}{p(w)p(s)} \\ &= \frac{1}{|q||w|}\sum_{q}\sum_{w} k(q,w)\exp\left(\mathrm{PMI(w,s)}\right)p(s) \end{aligned} \tag{3.2.3}$$

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

以上より，式 3.2.3 を式 3.1.3 と比較すると，式 3.1.3 に PMI の逆対数，および検索対象の文の言語モデルを掛けた値となることが示される．ここで，PMI を用いたクエリ-文間類似度 sim_{pmi} を式 3.2.3 のように定義する．

3.3 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．

第 4 章　英語学習者向けの用例検索実験

4.1　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[a]を使用した．また，文をトークナイズする際に Stanford Core NLP tokenizer (Ver. 3.6.0) [12] [b]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 3.1.3 および，式 3.2.3 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．

$$k_{cos}(q_i, s_j) = \frac{\langle q_i, s_j \rangle}{|q_i||s_j|} \tag{4.1.1}$$

$$\begin{aligned} k_{\mathrm{RBF}}(q_i, s_j) &= \exp\left(-\frac{||q_i - s_j||^2}{2\sigma^2}\right) \\ &= \exp\left(-\gamma||q_i - s_j||^2\right) \end{aligned} \tag{4.1.2}$$

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．

4.2　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 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

[a] https://code.google.com/archive/p/word2vec/
[b] http://nlp.stanford.edu/software/stanford-corenlp-full-2015-12-09.zip

た．さらに，アノテータ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．

4.3 用例検索実験

以下に提案法と比較したベースラインを示す．また，評価方法に関しては情報検索の評価指標として用いられる Precision@k [13] （以下，p@k）で検索結果を評価した．

4.3.1 単語ベクトルの平均による文間類似度

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 4.3.1 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 4.3.1 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．

$$sim_{ave}(q, s) = k(\vec{q}, \vec{s}) \tag{4.3.1}$$

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

4.3.2 アライメントベースの文間類似度

Song and Roth [14] 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．

$$sim_{max}(q, s) = \frac{1}{|q|} \sum_{i=1}^{|q|} \max_{j} k(\vec{q}_i, \vec{s}_j) \tag{4.3.2}$$

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 4.3.2 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．

4.3.3 実験結果

図 4.1 と図 4.2 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 4.1 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 4.2 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 4.2 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

表 4.1

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（4.1 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*.

適合率となることが観察された.

4.4 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 4.1 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

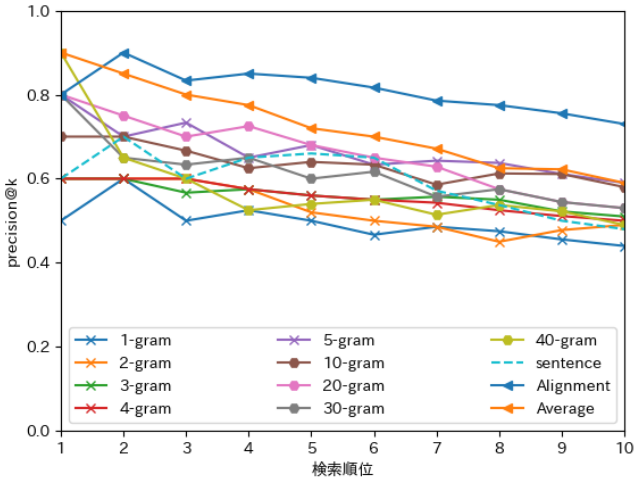


図 4.1　コサイン類似度の p@k

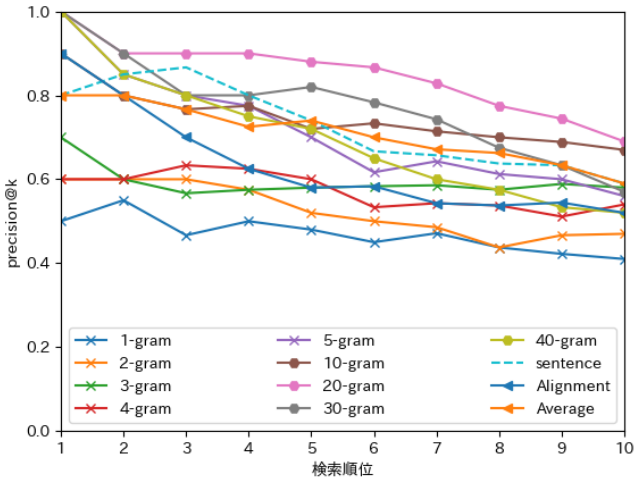


図 4.2　RBF カーネルの p@k

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．

第 5 章　英語学習者向けの作文支援実験

我々が提案した手法が英作文支援に対して有効であるかどうか評価するため，日本語文を英訳するタスクに英作文支援システムを使用して主観および客観評価する実験を行った．

5.1　実験設定

本実験では，システムの評価者はブラウザを使用し，WEB 上で我々が作成したシステムを使用して日英翻訳実験を行うことによりシステムを評価した．我々が作成したシステムのクエリの入力例，英文の出力例のスクリーンショットを図 5.1 に示す．システムの評価者が英作文実験を行う際に，検索エンジン，翻訳エンジンは使うことは禁止した．また，辞書の取り扱いについては，単語が分からずに英作文ができない場合は使用を許可した．その際に使用する辞書は，例文ができるだけ表示されていない辞書サイト[a]を指定し，検索は単語についてのみ許可し，フレーズに関しての検索は禁止にした．また，各文につき必ずシステムを 1 回以上使うよう協力者に指示をした．

ベースラインとして式 4.3.1 で示した単語ベクトルの平均による文間類似度を使用し，提案法として式 3.2.3 で示した PMI の逆対数を考慮した文間類似度を比較した．これら 2 つのシステムの出力が，ベースラインと提案法のどちらであるか協力者に分からないように見せ、どちらが英作文をする際により参考になったかを記録してもらった．さらに，1 文を訳す都度システムをどのように使用したかを自由記述形式でアンケートを取った．また，各文ごとにどのようなクエリで検索をしたのかが分かるようにログを採取した．協力者が持つ語学能力がどの程度であるのか把握するため，協力者が過去に受験経験のある英語能力試験の種類，受験時期，点数を自己申告制のアンケートとして提出してもらった．そのアンケートに関して，図 5.1 に示した．

本実験では日本語を母語とする英語学習者 4 名により実験を行った．

[a] goo 和英辞書 https://dictionary.goo.ne.jp/en/

表 5.1　実験協力者の英語能力

No.	試験名	スコア
協力者 1	TOEIC	650 / 990
協力者 2	TOEIC	730 / 990
協力者 3	英検	2 級
協力者 4	未受験	–

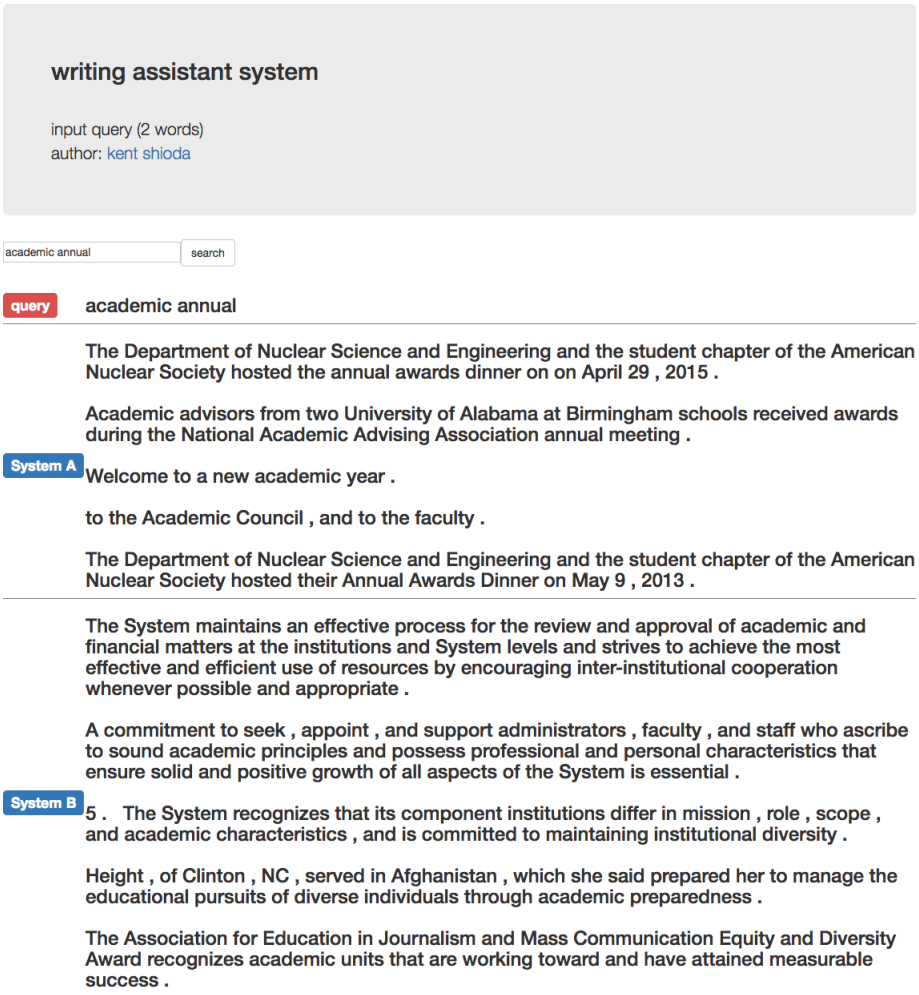


図 5.1　システムの出力画面

表 5.2　協力者による英作文の際に参考になったシステムの主観評価

ベースライン	提案法
2	17

5.2　データ

テストデータは，サンプリングした英文を日本語を母語とするアノテータ 2 名が日本語に訳したものを使用する．我々が作成した英作文支援システムは，4.2 節で示したようにアメリカの大学のウェブサイトに基づいて作成したコーパスから文を検索する．したがって，本実験の評価データはテストデータに含まれないカナダの大学[bcde]の “News” コンテンツに含まれる文からランダムに 12 文サンプリングした．

5.3　評価方法

原文と協力者が作成した作文の結果を自動で評価するために N-gram の一致率を用いた機械翻訳の評価尺度として使用される BLEU [15] を使用する．

また，協力者はシステムに対しての主観評価をアンケート形式で回答した．

5.4　実験結果

今回の翻訳タスクにおいて，協力者は一文を訳すにあたり平均 5.9 回例文検索システムを使用していた．協力者によるアンケート結果を 5.2 に示した．また，前節で定義した自動評価尺度による評価結果を 5.3 に，表 5.4 に協力者が訳した英文とその正解となる英文，正解文の日本語訳の一例を示した．

[b]University of Waterloo `https://uwaterloo.ca/`
[c]University of Alberta `https://www.ualberta.ca/`
[d]University of McGill `https://www.mcgill.ca/`
[e]University of British Columbia `https://www.ubc.ca/`

表 5.3　BLEU を用いた英訳文自動評価

評価尺度	スコア
BLEU	59.2

表 5.4

日英翻訳タスクの一例

原文	Seniors need twice as long as young adults to realize they are falling, a delay that puts them at increased risk for serious injury, according to a new study from the University of Waterloo.
日本語訳	Waterloo 大学の新しい研究によると、高齢者は、彼らが落ちていることを認識するために若者の 2 倍の時間を要し、その遅れは彼らが重傷を負うリスクを高める。
英訳結果	According to a new research of Waterloo university, the elder take twice time to recognize falling themselves as the youth and the delay raise risk to injure them seriously.
検索クエリ	(university, research), (raise, risk), (heavy, damage), (injury, heavy), (elderly, risk), (old, risk), (take, time), (twice, time), (need, time), (twice, than), (double, time), (twice, as), (falling, be)

5.5　考察

協力者によるシステムの主観評価によると，ベースラインより我々の提案法が英作文の際に参考になったという結果が得られた．また，協力者の英語力に依存するが，BLEU による自動評価結果からは極めて高いスコアとなる英文を協力者が作文したことを示している．

協力者の自由記述形式のアンケートによると，提案法に関する回答として “充電という意味の charge を調べたい時に値段の方が出力されたため、役に立たなかった。” との回答が得られた．このことから，用例検索システムにおいて入力単語数が少ないとクエリに与えられる表現力には限界があると考えられる．我々のシステ

ムにおける語義曖昧性解消に対して考えられる手法としては，クエリへ入力する単語数を増やしてクエリにより表現力を追加すること．また，検索対象の文の文脈を抽出して文が表しているトピックに関する上位概念を検索する際に考慮に入れることなどを行う必要がある．

第 6 章　おわりに

本研究では，英語学習者のための学習支援である，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んだ．英作文支援の中で，特に特定ドメインに関しての英作文を支援するためのシステムを構築した．

我々は，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．また，分布のカーネル埋め込みによるクエリ-文間類似度と確立モデルを利用したクエリ-文間類似度との間に PMI と言語モデルの関係性があることを明らかにした．さらに，大学の広報に関してのコーパスを作成し，英語学習者のための例文をアノテーションした．

用例検索実験において，我々が提案した 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．

また，英語学習者による英作文支援実験を行った．単純なコサイン類似度の平均を用いた手法と比較した際に，我々が提案したカーネル埋め込みと PMI の逆対数，並びに言語モデルを考慮した手法が実験協力者の主観評価が高い結果が得られた．

付録

本研究において，作文支援実験を行うにあたりに使用した実験への協力依頼書と指示書を図 6.1 と図 6.2 に示す.

実験への協力のお願い

１．研究課題名
英語学習者向け英作文支援システムの評価実験

２．研究目的
英語学習者にとって特定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タスクです。そこで、我々は特定ドメインにおける英作文支援システム作成することにより、英語学習者の英作文支援を行います。
今回の実験で行ってもらうタスクは、英語学習者向け英作文支援の評価実験です。本実験では、日本語文を英訳するタスクを行っていただきます。そのタスクを解く際に我々が作成した英作文支援システムを使用することにより、システムの性能を測定します。

３．実験内容
１）研究協力者：情報・システム研究機構の職員および首都大学東京の学生約10名
２）作業内容：英作文支援システムを用いて、日本語文15文の英訳作業

４．健康へのリスク
特になし

５．倫理的配慮
個人が特定できない手法（IDをテキストデータと紐付けない方法）でデータを収集します。自由英作文ではないため、書かれた英作文には個人情報は含まれません。

６．研究者と連絡先
１）研究代表者
小町守（首都大学東京システムデザイン学部）
住所：191-0065　日野市旭が丘6-6　首都大学東京システムデザイン学部
電話番号：042-585-8606
E-mail：komachi@tmu.ac.jp

２）共同研究者
塩田健人（首都大学東京システムデザイン研究科）
池谷瑠絵（情報・システム研究機構）
持橋大地（統計数理研究所）

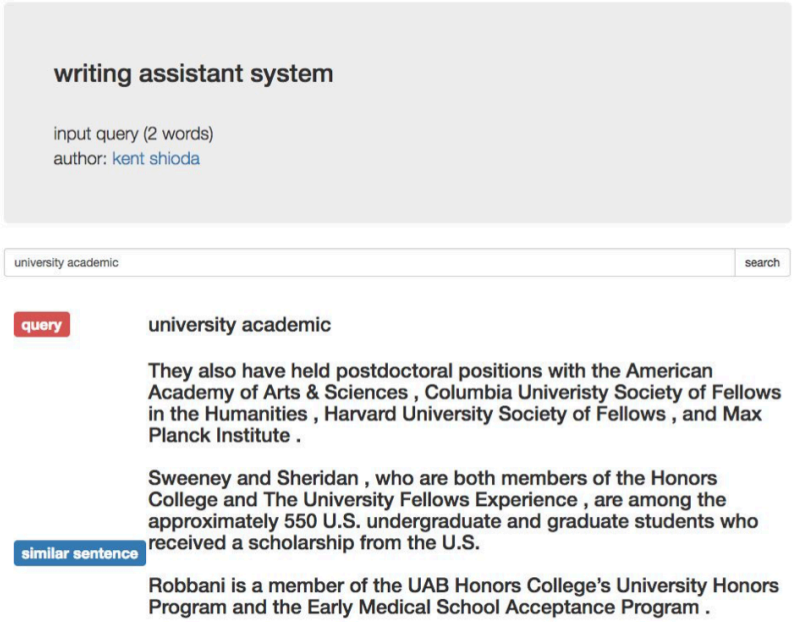
図 6.1　実験への協力依頼書

実験指示書

英作文支援システムの評価を行っていただきます。

12文の日本語を英訳してください。
その際に、2つのシステムを使っていただきます。

クエリに単語を入れると、その単語を考慮した英文が5文出力されます。
（クエリは2単語、半角スペース区切りで入力してください。）



英作文をする際に辞書を使うことは禁止します。
ただし、英作文をする際に単語が思い浮かばず、英文を書き進めることができない場合のみ和英辞書を使っていただいて結構です。
その際使用していただく辞書はこちらになります。

- goo和英辞書

ただし、辞書で検索するのは単語のみとしフレーズを検索、また辞書内に表示されている例文を見ることは禁止します。
・各設問にある日本語文を英訳していただいたのちに、アンケートにお答えください。
アンケートには作文をする際にどのようにシステムを使ったかを記入してください。

図 6.2　実験指示書

謝辞

本論文の執筆にあたり，指導教員である小町守准教授に心より感謝いたします．自然言語処理をはじめ，情報工学に関する知識が無い状態からここまで来ることができたのは，学部 4 年生の時に小町研究室に配属されてからの 3 年間，教育熱心な小町先生にご指導賜りましたおかげです．

また，本論文の執筆にあたり終始非常に有益なアドバイスをして頂いた統計数理研究所の持橋大地准教授，情報・システム研究機構の池谷瑠絵さんに深く感謝いたします．

研究室に配属されてすぐに研究の仕方の基礎を叩き込んで頂き，その後も数々のアドバイスにより助けて頂いた梶原智之さんには非常に感謝しております．今日の自分がいるのは梶原さんのおかげです．

また，プログラムを書く際に非常に丁寧に指導して頂いた研究室の先輩である宮崎亮輔さん，佐藤貴之さん，堺澤勇也さん，研究を進める上で有意義な議論をさせて頂いた研究室の同期である金子正弘さん，小平知範さん，関沢祐樹さんありがとうございました．この場をお借りしてお礼を申し上げたいと思います．

最後に，博士前期課程修了まで私を応援してくださった友人，家族には大変感謝しております．

充実した 3 年間を本当にありがとうございました．

発表文献一覧

国内会議・研究会論文（査読なし・ポスター発表）

- 塩田健人, 梶原智之, 小町守. 使用者数による語彙制限を用いた日本語学習者のための文章読解支援. 情報処理学会第 224 回自然言語処理研究会, Vol.2015-NL-224, No.6, pp.1-6. December 2015.

国内会議・研究会論文（査読なし・口頭発表）

- 塩田健人, 小町守, 瀬戸口光宏, 市橋立. 法律相談 SNS におけるユーザー投稿文書を用いた著者役割推定. 情報処理学会第 232 回自然言語処理研究会, Vol.2017-NL-232, No.1, pp.1-7. July 2017.
- 塩田健人, 小町守, 池谷瑠絵, 持橋大地. カーネル埋め込みを用いた英語学習者向けの用例検索. 情報処理学会第 233 回自然言語処理研究会, Vol.2017-NL-233, No.16, pp.1-5. October 2017.

国際会議（査読あり・ポスター発表）

- Kent Shioda, Mamoru Komachi, Rue Ikeya, Daichi Mochihashi. Suggesting Sentences for ESL using Kernel Embeddings. In Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications, pp.64-68. Taipei, Taiwan, December 2017.

参考文献

[1] S. Matsubara, Y. Kato, and S. Egawa, “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.

[2] M.-H. Chen, S.-T. Huang, H.-T. Hsieh, T.-H. Kao, and J.S. Chang, “Flow: a first-language-oriented writing assistant system,” Proceedings of the ACL 2012 System Demonstrations, pp.157–162, 2012.

[3] Y. Hayashibe, M. Hagiwara, and S. Sekine, “phloat : Integrated writing environment for ESL learners,” Proceedings of the Second Workshop on Advances in Text Input Methods, pp.57–72, Dec. 2012.

[4] J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” Proceedings of the 2014 conference on empirical methods in natural language processing, pp.1532–1543, 2014.

[5] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” Advances in neural information processing systems, pp.3111–3119, 2013.

[6] A. Smola, A. Gretton, L. Song, and B. Schölkopf, “A Hilbert space embedding for distributions,” Proceedings of International Conference on Algorithmic Learning Theory, pp.13–31, 2007.

[7] M. Kanagawa, Y. Nishiyama, A. Gretton, and K. Fukumizu, “Monte carlo filtering using kernel embedding of distributions.,” Association for the Advancement of Artificial Intelligence, pp.1897–1903, 2014.

[8] Y. Yoshikawa, T. Iwata, and H. Sawada, “Latent support measure machines for bag-of-words data classification,” Advances in Neural Information Processing Systems 27, eds. by Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, pp.1961–1969, 2014.

[9] Y. Yoshikawa, T. Iwata, and H. Sawada, “Non-linear regression for bag-of-words data via gaussian process latent variable set model.,” Association for the Advancement of Artificial Intelligence, pp.3129–3135, 2015.

[10] Y. Yoshikawa, T. Iwata, H. Sawada, and T. Yamada, "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.

[11] A. Berger and J. Lafferty, "Information retrieval as statistical translation," Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.222–229, SIGIR '99, 1999.

[12] C.D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S.J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," Association for Computational Linguistics System Demonstrations, pp.55–60, 2014.

[13] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of ir techniques," ACM Transactions on Information Systems, vol.20, no.4, pp.422–446, 2002.

[14] Y. Song and D. Roth, "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, May–June 2015.

[15] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp.311–318, ACL '02, 2002.