

# 単語分散表現を用いた超球面空間における確率的言語モデル: 情報検索におけるクエリ尤度モデリングへの適用

増村 亮 浅見 太一 政瀧 浩和 貞光 九月 西田 京介 東中 竜一郎

日本電信電話株式会社 NTTメディアインテリジェンス研究所

{masumura.ryo, asami.taichi, masataki.hirokazu, sadamitsu.kugatsu,  
nishida.kyosuke, higashinaka.ryuichiro}@lab.ntt.co.jp

## 1 はじめに

分類対象のクラス数が明確である文書分類などの識別タスクにおいては、深層学習技術等の教師あり学習が、強力な手段となり得る。一方で、分類対象や検索対象のクラス数が容易に変わり得る情報検索などのタスクにおいては、識別境界をあらかじめ教師あり学習しておくことは難しい場合が多い。本稿では、教師あり学習が困難である情報検索タスクにおいて、近年注目を集めている教師なし学習により獲得可能な単語分散表現 [1] を活用する方法を検討する。

識別境界を陽にモデル化することなく分類や検索を実現する方法として、本稿では確率的言語モデルに基づく方法に焦点を当てる。確率的言語モデルを用いる方法では、識別境界を陽にモデル化することは行わず、識別対象や検索対象となるカテゴリごとに構築した言語モデルについての生成確率基準で識別を行うため、新たに分類対象や検索対象を増やす場合においても容易に拡張可能である。確率的言語モデルを用いる方法は、分類においては単純ベイズ分類器 [2,3]、検索においてはクエリ尤度モデリング [4,5] が代表的な方法として知られている。従来、これらのモデル化には単語を確率変数とした離散空間におけるモデル化が適用されてきた。しかしながら、離散空間におけるモデル化では、単語がそれぞれ異なる次元で表現されるため、単語間の関係性を直接捉えることができず、語彙のミスマッチに弱いことが知られている。そのため、離散空間におけるモデル化の範囲内で単語間の関係性を考慮するために、これまでではクエリ拡張や潜在変数モデルの導入が検討されてきた [6]。

これに対して本研究では、確率的言語モデルに基づく分類や検索方法の改善のために、離散空間におけるモデル化を用いないアプローチを試みる。そのための我々のアイデアは、言語モデルの確率変数に、単語ではなく単語分散表現を用いることである。単語分散表現は、単語の意味を連続値ベクトルとして表現するための方法として近年注目されており、Skip-gram モデルや Continuous Bag-of-Words (CBOW) モデルなどの手法を用いることで、大量の言語資源から教師なし学習により獲得することができる [1]。そして、この単語分散表現を用いることで、単語間の文脈的類似性をコサイン類似度によりうまく捉えられることが知られている。したがって、確率変数として単語分散表現を用いるアプローチをとることで、言語モデル自体に複雑な工夫を入れることなく、単語間の関係性まで考慮した情報検索の実現が期待できる。

そこで本稿では、単語分散表現自体の生成確率を扱う確率的言語モデルを提案する。提案法では、単語分散表現間のコサイン距離が意味を持つという知見を考慮したモデル化を行うために、超球面空間上のモデル化として利用可能な混合 Von Mises-Fisher 分布 [7-9]

を言語モデルのために導入する。つまり提案法では、ユークリッドノルムが1となるように正規化した単語分散表現を確率変数として扱う。本モデル化により、超球面空間上で単語の生成確率を考えることができ、従来の離散空間上のモデル化と同様の考え方で、確率的言語モデルに基づく情報検索を実現できる。本稿の貢献は以下の3点にまとめられる。

- 単語分散表現を用いた超球面空間における確率的言語モデルを一般化し、混合 Von Mises-Fisher 分布を導入した場合のモデル化方法を定式化する (3 節)。
- 超球面空間における確率的言語モデルについて、情報検索におけるクエリ尤度モデリングとしての解釈を与える。また、従来の離散空間におけるモデル化との関係を明かにする (4 節)。
- コンタクトセンタ音声書き起こしに関する類似ドキュメント検索タスクにおいて、代表的な検索モデル、および分散表現を用いたドキュメント間の距離尺度を用いる方法と比較して、超球面空間における確率的言語モデルを用いた情報検索手法の優位性を明かにする (5 節)。

## 2 関連研究

本研究は、単語分散表現を用いて情報検索手法を改善する研究に関連がある。代表的な検討は、翻訳言語モデル [10] の拡張の枠組みで単語分散表現を導入する方法であり、単語の言い換わり確率を算出する際に、単語分散表現間のコサイン距離を応用している [11-15]。また、Paragraph Vector [16] を用いて、クエリ尤度モデルを拡張する方法も主流な検討である [17,18]。これらの研究は、単語を確率変数とした離散空間におけるモデル化の範囲内での検討であるが、本研究は、単語分散表現自体を確率変数とした超球面空間におけるモデル化を新たに導入している点が大きく異なる。

また本研究では、関連ドキュメント検索タスクでの評価を行うが、分散表現を用いたドキュメント間の距離尺度指標にも関連がある。典型的な方法は、ドキュメントに出現する単語の分散表現について、平均化や重み付けによる足し合わせによりドキュメントベクトルを構成し、コサイン距離を使ってドキュメント間の距離を算出する方法である [19,20]。また、ドキュメント自体を Paragraph Vector として直接分散表現に変換する方法も取ることができる [16]。さらに、Word Mover's Distance (WMD) は、分散表現を用いてドキュメント間の距離を頑健に測る方法として有用であることが報告されている [21]。本稿では、関連ドキュメント検索タスクにおいて、これらの手法との比較も実施する。

なお、単語分散表現についての生成モデル化は、トピックモデリングでも検討が始められており、有用であることが示されている [22,23]。これに対して本研究は、分類や検索を目的とした確率的言語モデルについての検討であり、これまで単語を確率変数として扱ってきたナイーブベイズ分類器やクエリ尤度モデルに、超球面空間上の確率変数を扱う生成モデルを導入した検討は、我々の知る限り他にない。

### 3 超球面空間における言語モデル

#### 3.1 定式化

本研究では、単語分散表現を超球面空間上に配置して、超球面空間上の確率変数を扱う確率的言語モデルを新たにモデル化する。本モデル化では、Skip-gram モデルや CBOW モデルなどに基づき、単語を固定長の連続値ベクトルとして表現できることを前提とする [1]。以下では、あらかじめ学習された単語分散表現に基づき、ある単語  $w$  を長さ 1 に正規化された単語分散表現  $\mathbf{w}$  に変換できるものとする。

確率的言語モデルは、通常単語列に対する生成確率をモデル化するものである。我々は、単語列  $W = w_1, \dots, w_T$  の生成確率ではなく、長さ 1 に正規化された単語分散表現系列  $\mathbf{W} = \mathbf{w}_1, \dots, \mathbf{w}_T$  の生成モデルとして、(1) 式の通りモデル化する。

$$p(\mathbf{W}|\Theta) = \prod_{t=1}^T p(\mathbf{w}_t|\Theta) \quad (1)$$

ここで、 $p(\mathbf{w}_t|\Theta)$  は、単語分散表現  $\mathbf{w}_t$  に対する確率密度分布、 $\Theta$  は言語モデルのモデルパラメータを表す。

単語分散表現の確率密度分布  $P(\mathbf{w}_t|\Theta)$  は、混合 Von Mises-Fisher 分布によってモデル化を行う [7-9]。Von Mises-Fisher 分布は、超球面空間上のガウス分布であり、角度情報により単語間の類似性を捉えることが可能な単語分散表現のモデル化を行うために適切と考えられる。確率密度  $p(\mathbf{w}_t|\Theta)$  は、(2-4) 式の通りモデル化される。

$$p(\mathbf{w}_t|\Theta) = \sum_{m=1}^M \alpha_m f(\mathbf{w}_t|\theta^m) \quad (2)$$

$$f(\mathbf{w}_t|\theta^m) = C_d(\kappa) \exp(\kappa \boldsymbol{\mu}_m^\top \mathbf{w}_t) \quad (3)$$

$$C_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)} \quad (4)$$

ここで、 $M$  は混合数、 $\alpha_m$ 、 $\boldsymbol{\mu}_m$  がモデルパラメータである。 $\kappa$  も通常はパラメータであるが、学習により頑健に推定することが困難であるため、本稿では人手で与えることとする。なお、 $d$  は単語分散表現の次元数、 $I$  は第 1 種変形 Bessel 関数を表す。

#### 3.2 モデルパラメータ推定

次に、混合 Von Mises-Fisher 分布に基づく超球面空間における言語モデルのモデルパラメータ推定について述べる。

学習データの単語列  $W = w_1, \dots, w_T$  についての単語分散表現系列  $\mathbf{W} = \mathbf{w}_1, \dots, \mathbf{w}_T$  が与えられた場合、各パラメータの Maximum likelihood (ML) 推定は Expectation Maximization (EM) アルゴリズムに基づ

く、EM アルゴリズムによる  $M$  ステップのパラメータ更新は (5-7) 式に従う。

$$\alpha_m = \frac{1}{T} \sum_{t=1}^T q(m|\mathbf{w}_t, \Theta) \quad (5)$$

$$\mathbf{r}_m = \sum_{t=1}^T \mathbf{w}_t q(m|\mathbf{w}_t, \Theta) \quad (6)$$

$$\boldsymbol{\mu}_m = \frac{\mathbf{r}_m}{\|\mathbf{r}_m\|} \quad (7)$$

ここで、 $q(m|\mathbf{w}_t, \Theta)$  は、E ステップにおいて計算される  $m$  番目の Von Mises-Fisher 分布に対する  $\mathbf{w}_t$  の負荷率であり、ソフト方式とハード方式の 2 種類の計算方法がある。ソフト方式、ハード方式は、それぞれ (8) 式、(9) 式に従う。

$$q(m|\mathbf{w}_t, \Theta) = \frac{\alpha_m f(\mathbf{w}_t|\theta^m)}{\sum_{l=1}^M \alpha_l f(\mathbf{w}_t|\theta^l)} \quad (8)$$

$$q(m|\mathbf{w}_t, \Theta) = \begin{cases} 1 & m = \underset{l}{\operatorname{argmax}} \alpha_l f(\mathbf{w}_t|\theta^l) \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

ソフト方式が EM アルゴリズムの厳密解であるが、ハード方式によるモデル化は高速に動作するため、本研究ではハード方式を用いている。

次に、モデル適応について述べる。ここでは、学習済みのパラメータ  $\Theta = \{\alpha_1, \dots, \alpha_M, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_M\}$  を持つ超球面空間における言語モデルに対する Maximum a Posteriori (MAP) 適応を扱う。

適応データの単語列  $W = \bar{w}_1, \dots, \bar{w}_T$  についての単語分散表現系列  $\bar{\mathbf{W}} = \bar{\mathbf{w}}_1, \dots, \bar{\mathbf{w}}_T$  が与えられた場合、混合重み、および平均値の各 MAP 推定値  $\bar{\alpha}_m$ 、 $\bar{\boldsymbol{\mu}}_m$  は (10-14) 式に従う。

$$\beta_m = \frac{1}{T} \sum_{t=1}^T q(m|\bar{\mathbf{w}}_t, \Theta) \quad (10)$$

$$\mathbf{e}_m = \sum_{t=1}^T \bar{\mathbf{w}}_t q(m|\bar{\mathbf{w}}_t, \Theta) \quad (11)$$

$$\bar{\alpha}_m = \frac{\tau}{T + \tau} \alpha_m + \frac{T}{T + \tau} \beta_m \quad (12)$$

$$\bar{\mathbf{r}}_m = \frac{\tau}{T + \tau} \mathbf{r}_m + \frac{T}{T + \tau} \mathbf{e}_m \quad (13)$$

$$\bar{\boldsymbol{\mu}}_m = \frac{\bar{\mathbf{r}}_m}{\|\bar{\mathbf{r}}_m\|} \quad (14)$$

ここで、 $\tau$  は適応化を調整するためのハイパーパラメータである。適応化されたモデルは、各 MAP 推定値をパラメータとしたモデルとして表される。

## 4 超球面空間における言語モデルの情報検索への適用

### 4.1 クエリ尤度モデル

確率的言語モデルに基づく情報検索では、文書  $D$  が検索クエリ  $Q$  に適合する確率に基づき情報検索を実施する [4]。この場合、文書  $D$  に関する言語モデルが

検索クエリ  $Q$  を生成する確率  $P(Q|\Theta_D)$  をモデル化することで実現する。

$$P(D|Q) \propto P(Q|\Theta_D) \quad (15)$$

ここで  $\Theta_D$  は、文書  $D$  についての確率的言語モデルのモデルパラメータであり、 $D$  についてのクエリ尤度モデル (Query Likelihood Models; QLMs) と呼ばれる。

このモデル化は一般的にカテゴリカル分布に基づく離散空間上のモデル化が一般的であったが、本稿では、超球面空間における言語モデルをクエリ尤度モデルとして適用することを試みる。

## 4.2 超球面空間クエリ尤度モデルと離散空間上のモデル化との関係

本稿では、従来の離散空間上のモデル化の拡張としての解釈を与えることが可能な超球面クエリ尤度モデル (Hyperspherical QLMs; HQLMs) を提案する。

超球面クエリ尤度モデルのモデル化は、語彙空間の各単語ごとに Von Mises-Fisher 分布を割り当てることで実現できる。つまり、語彙空間  $\mathcal{V}$  と定義した場合、混合 Von Mises-Fisher 分布の混合数  $M$  を語彙サイズ  $|\mathcal{V}|$  とすることで、そのようなモデル化が可能となる。なお、語彙空間の各単語  $v_m \in \mathcal{V}$  について、長さ 1 に正規化された単語分散表現を  $v_m$  と記述する。また、学習時に必要な負荷率の計算の際は、ハード方式の計算を適用することとする。

文書  $D$  から ML 推定を行う場合、文書  $D$  についての超球面空間クエリ尤度モデルは (16–17) 式で表すことができる。

$$p(w_t|\Theta_D^{\text{ML}}) = \sum_{m=1}^{|\mathcal{V}|} \frac{c(v_m, D)}{|D|} f(w_t|\theta_D^m) \quad (16)$$

$$f(w_t|\theta_D^m) = C_d(\kappa) \exp(\kappa v_m^\top w_t) \quad (17)$$

ここで、 $c(v_m, D)$  は、 $D$  に含まれる  $v_m$  の頻度、 $|D|$  は  $D$  の総単語数である。(16) 式には、カテゴリカル分布を用いる際の ML 推定値が含まれている。つまり、超球面空間クエリ尤度モデルは、従来の単語の出現頻度を考慮したモデル化と、単語間関係を捉えたモデル化を両立させた方法となっている。なお、本モデル化と従来の離散空間におけるモデル化との関係は、(18) 式で表現できる。

$$\lim_{\kappa \rightarrow \infty} p(w_t|\Theta_D^{\text{ML}}) = \frac{c(w_t, D)}{|D|} \quad (18)$$

つまり、離散空間におけるモデル化は本モデル化の一形態とみなすことができる。

さらに、確率的言語モデルを用いた情報検索では、文書集合から学習したパラメータを用いてスムージングを行うことで、高い性能を実現できることが知られている [5]。ここでは、一度構築した超球面クエリ尤度モデルに対してモデル適応を行う場合について述べる。文書集合  $C$  により学習したパラメータに対して、文書  $D$  を用いて MAP 適応を行う場合、文書  $D$  についての超球面空間クエリ尤度モデルは (19) 式で表すことができる。

$$p(w_t|\Theta_D^{\text{MAP}}) = \sum_{m=1}^{|\mathcal{V}|} \frac{c(v_m, D) + \tau \frac{c(v_m, C)}{|C|}}{|D| + \tau} f(w_t|\theta_D^m) \quad (19)$$

ここで、 $c(v_m, C)$  は、 $C$  に含まれる  $v_m$  の頻度、 $|C|$  は  $C$  の総単語数である。また、 $f(w_t|\theta_D^m)$  は (17) 式に一致する。このモデル化において、カテゴリカル分布を用いる際のディリクレスムージングによる推定値 [5] が含まれており、スムージングを行う場合に置いても、離散空間におけるモデル化は本モデル化の一形態とみなすことができる。なお、以下の評価実験では、MAP 適応を行ったモデルを用いることとする。

## 5 評価実験

### 5.1 実験条件

超球面空間における確率的言語モデルの有効性を明らかにするために、コンタクトセンタ音声書き起こしに関する類似ドキュメント検索タスクにおいて、評価実験を行った。各ドキュメントはオペレータとユーザの書き起こしのペアとなっており、ドキュメントごとに、コンタクトセンタの業種ラベルが付与されている。本評価実験における類似ドキュメント検索では、ドキュメント集合に対して、入力クエリのドキュメントと同一の業種ラベルを持つドキュメントが検索上位となるかを評価した。検索対象となるドキュメント集合のドキュメント数は 391 である。また、クエリはドキュメント集合には含まないものであり、ドキュメント数は 308 である。なお、業種の種類数は 6 であり、検索対象となるドキュメント集合、クエリともに、各業種のドキュメント数の偏りはない。各ドキュメントは平均 1200 単語である。

さらに、約 50 億単語 (異なり語彙数は約 500 万) の日本語 Web ドキュメントをクロールにより準備した。この日本語 Web ドキュメントは、様々な分野についてクロールを行ったものであり、単語分散表現のモデル化、比較対象として利用する Paragraph Vector のモデル化、単語の Inverse Document Frequency (IDF) 値の算出、クエリ尤度モデル学習における MAP 適用によるスムージングに利用する。単語分散表現のモデル化には CBOW モデルを利用し、前後 5 単語をコンテキストとして考慮するように Negative Sampling を用いて学習した。学習した分散表現は 200 次元である。また、Paragraph Vector のモデル化には、Distributed Bag-of-Words (PV-DBOW) を利用し、分散表現が 200 次元となるよう学習した。

以下、本評価実験で用いるドキュメント検索手法を述べる。

- VSM: 単語 TF-IDF 値による文書ベクトルを用いて、コサイン距離基準で検索。
- CWV: 単語分散表現の平均ベクトルを用いて、コサイン距離基準で検索。
- WCWV: 単語分散表現を IDF 値で重み付けして足し合わせたベクトルを用いて、コサイン距離基準で検索。
- PV: Paragraph Vector を用いて、コサイン距離基準で検索。
- WMD: 単語分散表現を用いて、Word Mover's Distance 基準で検索。
- QLM: カテゴリカル分布のクエリ尤度モデルを MAP 適応によりモデル化 ( $\tau = 10$ )。クエリ尤度基準で検索。

表 1: 関連ドキュメント検索の実験結果

	MAP	Precision@10
VSM	0.305	0.470
CWV	0.339	0.485
WCWV	0.429	0.644
PV	0.390	0.583
WMD	0.383	0.592
QLM	0.512	0.801
HQLM	<b>0.543</b>	<b>0.826</b>

- **HQLM**: 超球面空間におけるクエリ尤度モデルを MAP 適応によりモデル化 ( $\tau = 10, \kappa = 20$ ) . クエリ尤度基準で検索 .

なお, 各手法について, ストップワードフィルタやSTEMMINGなどの処理は行わないこととした .

## 5.2 実験結果

各ドキュメント検索手法について, 各クエリで検索対象のドキュメント集合に対して検索を行った場合の結果を評価した . 評価には, Mean Average Precision (MAP) と, Precision@10 を用いた . 評価結果を表 1 に示す .

まず, クエリ尤度モデルを用いた手法 (QLM, HQLM) は, 一般的な単語文書ベクトルを用いた方法 (VSM) や, 分散表現に基づくドキュメント距離を用いる手法 (CWV, WCWV, PV, WMD) と比較して高い性能を示すことを見て取れる . このことから, クエリ尤度モデルに基づく情報検索が有望な手法であると言える . 分散表現の考え方は連続値ベクトルに情報を埋め込むことであるが, 教師なし学習により求めた分散表現を用いてドキュメントを連続値ベクトルとして表現し, それに基づき情報検索を行うのみでは不十分であることが分かった . 次に, 通常のカテゴリカル分布を用いる QLM と, 提案法である HQLM を比較すると, HQLM の方が高い性能を示すことが見て取れる . この性能改善は, 超球面空間上で単語を扱うことにより, 単語間の関係性まで考慮できることに起因したものであると考えられる . 本実験では, HQLM のパラメータを  $\kappa = 20$  としたが, この値により単語間の関係性についての考慮の度合いが変化するため,  $\kappa$  の調整が重要と考えられる . なお,  $\kappa$  を十分大きくした場合 ( $\kappa = 500$ ), 通常の QLM とほぼ同等の性能となり, (18) 式の関係性をモデル化できていることも確認できた .

## 6 まとめ

本稿では, 単語分散表現を用いた超球面空間における確率的言語モデルを提案し, 情報検索におけるクエリ尤度モデリングへの適用を試みた . 提案モデルは, あらかじめ学習済みの単語分散表現を用いて, 混合 Von Mises-Fisher 分布によりモデル化できる . また, 提案モデルを情報検索タスクで用いる場合に利用可能な超球面クエリ尤度モデルは, カテゴリカル分布を用いたモデル化の拡張として表現でき, 従来のスムージ

ング理論をそのまま適用することが可能である . コンタクトセンタ音声書き起こしに関する類似ドキュメント検索タスクにおいて, 代表的な検索モデル, および分散表現を用いたドキュメント間の距離尺度を用いる方法と比較して, 超球面空間における確率的言語モデルを用いた情報検索手法が有効であることを示した . 今後は, さらに大規模な検索タスクでの評価を行うとともに, 超球面空間における確率的言語モデルのモデル化についてさらなる分析を試みる予定である .

## 参考文献

- [1] T. Mikolov *et al.*, “Distributed representation of words and phrases and their compositionality,” *In Proc. NIPS*, pp. 3111–3119, 2013.
- [2] A. McCallum and K. Nigam, “A comparison of event models for naive bayes text classification,” *In AAAI-98 Workshop on Learning for Text Categorization*, pp. 41–48, 1998.
- [3] F. Peng *et al.*, “Language and task independent text categorization,” *In Proc. HALT-NACCL*, pp. 110–117, 2003.
- [4] J. M. Ponte and W. B. Croft, “A language modeling approach to information retrieval,” *In Proc. SIGIR*, pp. 275–281, 1998.
- [5] C. Zhai and J. Lafferty, “a study of smoothing methods for language models applied to adhoc information retrieval,” *In Proc. SIGIR*, pp. 334–342, 2001.
- [6] X. Wei and W. B. Croft, “Lda-based document models for ad-hoc retrieval,” *In Proc. SIGIR*, pp. 178–185, 2006.
- [7] A. Banerjee *et al.*, “Clustering on the unit hypersphere using von mises-fisher distribution,” *Journal of Machine Learning Research*, vol. 6, pp. 1345–1382, 2005.
- [8] S. Gopal and T. Yang, “Von Mises-Fisher clustering models,” *In Proc. ICML*, 2014.
- [9] S. Sra, “Directional statistics in machine learning: a brief review,” *arXiv preprint arXiv:1605.00316*, 2016.
- [10] A. Berger and J. Lafferty, “Information retrieval as statistical translation,” *In Proc. SIGIR*, pp. 222–229, 1999.
- [11] G. Zuccon *et al.*, “Integrating and evaluating neural word embeddings in information retrieval,” *In Proc. ADCS*, pp. 12:1–12:8, 2015.
- [12] D. Ganguly *et al.*, “A word embedding based generalized language model for information retrieval,” *In Proc. SIGIR*, pp. 795–798, 2015.
- [13] H. Zamani and W. B. Croft, “Estimating embedding vectors for queries,” *In Proc. ICTIR*, pp. 123–132, 2016.
- [14] H. Zamani and W. B. Croft, “Embedding-based query language models,” *In Proc. ICTIR*, pp. 147–156, 2016.
- [15] S. Kuzi *et al.*, “Query expansion using word embeddings,” *In Proc. CIKM*, pp. 1929–1932, 2016.
- [16] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” *In Proc. ICML*, pp. 1188–1196, 2014.
- [17] Q. Ai *et al.*, “Improving language estimation with the paragraph vector model for ad-hoc retrieval,” *In Proc. SIGIR 2016*, pp. 869–872, 2016.
- [18] Q. Ai *et al.*, “Analysis of the paragraph vector model for information retrieval,” *In Proc. ICTIR*, pp. 133–142, 2016.
- [19] A. Kosmopoulos *et al.*, “Biomedical semantic indexing using dense word vectors in bioasr,” *Journal Of Biomedical Semantics*, 2015.
- [20] G.-I. Brokos *et al.*, “Using centroids of word embeddings and word movers distance for biomedical document retrieval in question answering,” *In Proc. BioNLP*, pp. 114–118, 2016.
- [21] M. J. Kusner *et al.*, “From word embeddings to document distances,” *In Proc. ICML*, 2015.
- [22] R. Das *et al.*, “Gaussian LDA for topic models with word embeddings,” *In Proc. ACL*, pp. 795–804, 2015.
- [23] K. Batmanghelich *et al.*, “Nonparametric spherical topic modeling with word embeddings,” *arXiv preprint arXiv:1604.00126*, 2016.