DEIM2020 P2-24(day2 p97)

語義と分散表現を用いたランキング学習

欅　悼志† 杉山　一成††

† 株式会社デンソーアイティーラボラトリ
†† 京都大学 情報学研究科
E-mail: †akeyaki@d-itlab.co.jp, ††kaz.sugiyama@i.kyoto-u.ac.jp

あらまし　本研究では，大域的な意味情報として分散表現を，局所的な意味情報として語義を用いたランキング学習手法を提案する．それぞれの意味情報を駆使することで，情報検索において主要な課題の一つである文書とクエリの語彙の乖離を軽減することを目指す．語義を用いた情報検索に関する既存の研究では，主に文書に対して正確に語義の付与することに注力してきたが，本研究ではクエリに対する語義の付与方法を提案する．また，単語レベルの局所的意味情報として語義を用いた文書とクエリの類似度計測手法 (語義スコア) を提案する．対して，文書 · レベルの大域的意味情報として分散表現を用いた文書とクエリの類似度計測手法 (分散表現スコア) の提案を行う．その際，両スコアはランキング学習に適用可能な構造に設計する．評価実験の結果，提案手法を用いることで検索結果上位の検索結果の精度を統計的有意に改善することができた．その際，語義スコアと分散表現を組み合わせた手法が最も高精度を達成した．

キーワード　ランキング学習，語義曖昧性解消，分散表現

1 はじめに

文書とクエリの語彙の乖離は情報検索における主要な課題の一つである．TF-IDF [36]，BM25 [35]，クエリ尤度モデル [33] といった，広く利用されている単語の重み付けに基づくスコリング手法には，例えクエリと同一の概念を含んでいたとしても，クエリ語を含まない適合文書を発見することができないという短所が存在する．例えば，クエリ “情報検索システム” は “サーチエンジン” を含む文書と照合できない．このとき，“システム” と “エンジン” は必ずも常に同義語関係にあるとは限らず，両単語はあるときは可換であり，またあるときは非可換であるという点において問題の難易度を増している．これは自然言語の持つ曖昧性に由来する現象であり，多くの単語は多義性を持ち，その意味は単語が出現するコンテキストに依存して決定される．なお，コンテキストとは，例えば，周辺の語やその品詞，それらの並び順などである．従って，クエリと文書が同一の概念を共有しているのかどうかを判定する上で，コンテキスト情報を考慮する必要がある．

クエリと文書の語彙の乖離を解消する方法の一つとして分散表現 [9] が着目されている．分散表現を用いることで，単語の表層的な情報だけではなく，単語間の意味的距離を考慮したクエリ語と文書の照合を行うことが可能である．実際，さまざまな情報検索研究において分散表現を利用されている [21], [47], [48]．分散表現は，文書やクエリレベルの意味情報を扱うため，本稿では分散表現は大局的な意味情報を表すとみなす．

その一方で，近年，語義のような単語レベルの意味情報はそれほど着目されていない．語義を用いた情報検索は存在するものの [5], [6], [8], [13], [37], [40], [43], [50]，語義の曖昧性解消が常に検索精度を向上させることができるかどうかについては結論が導き出されていない．なお，本研究では語義は局所的な意味情報として捉える．

前述の語義の曖昧性解消を用いた情報検索は主に文書への正確な語義の付与に着目しており，クエリに対する自動かつ正確な語義の付与については十分に取り組まれていない．これはクエリの持つ特徴である，高々数語から構成され，自然言語の文法にも基づかないクエリからは限られたコンテキスト情報のみしか取得することができないということに起因する．その結果，自然言語を対象として提案された語義の曖昧性解消手法をクエリに適用したとしても正確な付与を行うことができない．従って，本研究では，利用可能なコンテキスト情報を利用したクエリに対する語義の曖昧性解消手法を提案する．更に，我々は局所的意味情報として，語義に基づくクエリと文書の類似度計測手法の定式化を行う．同様に，大域的意味情報として，分散表現に基づくクエリと文書の類似度計測手法の提案も行う．我々は，大域的と局所的な意味情報の両者を用いることで検索精度の向上が実現可能であると考える．具体的には，大域的 · 局所的意味情報によってクエリ語をほとんど含まない適合文書を発見することを目指す．

ランキング学習は多くの情報検索研究にて用いられるフレームワークであり，適合文書をより上位に再ランキングする用途で用いられる．我々の知る限り，本研究は初の語義を用いたランキング学習の研究である．従って，上記の類似度計測手法の設計においてランキング学習に適した定式化を行う必要がある．これらを踏まえ，本研究では，語義と分散表現を用いたランキング学習を提案する．

2 関連研究

2.1 単語の照合と分散表現

TF-IDF [36]，BM25 [35]，クエリ尤度モデル [33] といった，広く利用されている単語の重み付けに基づくスコリング手法で

は，類似度を計測する手順の中でクエリと文書の比較が行われる．その際，クエリと同一の概念を含んでいたとしても，クエリ語を含まない適合文書を発見することができないという短所が存在する．この問題を解決するため，クエリ語の関連語をクエリに追加するクエリ拡張や，ユーザのフィードバックを用いたクエリ修正などが提案された．近年の主流は word2vec [24] や GloVe [31]，BERT [7] といった分散表現を用いるアプローチであり，各単語は数百次元のベクトルとして表現される．クエリや文書も同様に数百次元のベクトルで表現され，それぞれ doc2vec [15] と query2vec [21] と呼ばれる．doc2vec と query2vec は，文書とクエリの比較においてベクトルレベルで評価される．

なお，BERT は多くの end-to-end の NLP タスクで特筆すべき成果を挙げた最先端のモデルであるものの，本研究では分散表現として word2vec を用いる [1] 従って，本研究における分散表現として BERT を利用可能かどうかの検証は今後の課題である．

2.2 語義の曖昧性解消による情報検索

語義の曖昧性解消は代表的な自然言語処理分野のタスクの一つであり，周辺の単語やその品詞を用いて対象語の語義を付与する．語義の曖昧性解消は (1) コンテキストベースの手法と (2) 知識ベースの手法が存在する．主に教師あり学習は (1) コンテキストベースの手法に分類され，ルールベースや教師なし学習の手法は (2) 知識ベースの手法に分類される．本研究では，一般的により高精度であると報告されている (1) コンテキストベースの手法を採用する．

Yarowsky [45] は語義の曖昧性解消において，二つの仮説を提案した．すなわち，(i) 語義を判定する上で対象語の周辺語は重要なヒントとなる，(ii) 一連の議論の中ではある語の語義は一貫性を持つ，である．これらの仮説のもとに，語義付きコーパスから判別器を学習する研究が多数取り組まれた [16], [28], [30]．その際に多用されるコーパスは SemCor [25]，Senseval-3 [38]，One Million Sense-Tagged Instances (OMSTI) [3] などであり，語義候補は WordNet で定義される語義リストが用いられる．WordNet には固有名詞はほとんど含まれておらず，固有名詞の曖昧性解消は entity linking [34] と呼ばれるタスクにて取り組まれている．

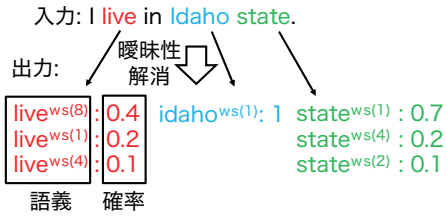

図 1　IMS の実行例

"It Makes Sense" (IMS) [49] [2] は大規模コーパスである OMSTI から構築された語義曖昧性解消ツールであり，(1) コンテキストベースの手法である．IMS は入力として文を受け取ると，出力として単語ごとに語義候補とその確率を返す．図 1 に例を示す．"I live in Idaho state." が入力されたとすると，"live" の候補語義は $live^{ws(8)}$，$live^{ws(1)}$，$live^{ws(4)}$ ($t^{ws(i)}$ は 単語 t の i 番目の語義)，$live^{ws(8)}$ の確率は 0.4 である．なお，IMS は文レベルに対して適用されることを想定されているため，自然言語文法に基づかないクエリに対しては正確な語義の付与ができない．

語義曖昧性解消を用いた情報検索に関する既存の研究 [8], [13], [37], [40], [43] では，語義によって真に検索精度が改善するのかどうか決断がなされていない．また，精度が改善するかどうかはクエリによって異なるということや，語義の曖昧性解消の性能が検索精度に対しても強く影響を及ぼすということが報告されている．なお，これらの研究で採用された語義曖昧性解消手法は，人手で語義を付与しているものや，実用レベルには達していない性能の語義曖昧性解消手法であった．これらの多くの語義曖昧性解消手法は (2) 知識ベースの手法に分類される Lesk アルゴリズム [17] の拡張であり，WordNet の定義文や例を用いる．

また，情報検索分野における語義曖昧性解消としては，辞書ベースの言語横断情報検索が存在する [32]．中でも，ソース言語とターゲット言語間の語の共起に基づく語義曖昧性解消手法が存在する [1], [4]．

Zhong and Ng [50] は語義曖昧性解消手法として IMS を用いた．クエリ語は語義ベクトルとして表現され，それぞれの要素は各語義の確率が格納される．その確率は，クエリから得られる上位 k 件の文書を用いて計算される．なお，各単語は個別に曖昧性解消が行われるため，クエリの持つコンテキスト情報のうち主要な情報である単語の共起情報を用いていない．

より近年の研究 [5], [6] では，同じ語義を持つ文書やクエリが同じクラスタに割り当てられるようにクラスタリングを行う．つまり，クエリが含まれるクラスタに属する文書はクエリと同じ語義を持つ．このアプリーチと比較し，IMS は，複数の語義候補を確率的に持つことができるという点において優れている．また，IMS は WordNet の語義集合を用いるため，例えば同義語などの関連語を取得するなどの拡張が容易である．実際，文献 [14], [22] では WordNet 上の関連語を用いた検索を行ってお

1：理由は下記の通りである．

（1） BERT の token size は 512 であるため，文書のような長文を扱う方法についての結論は出ていない．例えば，512 単語以上の文字列は除外する，文書を分割するなどである．後者に関して Yilmaz ら [46] はパッセージ単位に分割している．ただし，これらのうちいずれの方式が適切であるかについては結論が出ていない．また，ほとんどの研究者にとって，512 単語よりも大きな token size のモデルを学習することは計算機の性能の制約で困難である．

（2） BERT で成果を挙げているタスクは主にテキストシーケンスを入力とするタスクである．例えば BERT の代表的なタスクである機械読解タスクにおける質問は自然文で記述されているおり，Web クエリのようにキーワード集合に対しても適切にコンテキストを考慮した埋め込みを行うことができるかどうかについては検証されていない．

（3） word2vec は既にさまざまな情報検索検索研究で利用され成果を挙げている．また，word2vec はアナロジータスクを通じて，Web クエリの分散表現構築で必須である単語の合成性 (例えば $\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} = \overrightarrow{queen}$) についての有用性が評価されている．

2：https://www.comp.nus.edu.sg/~nlp/software.html

り，言語横断情報検索においても同様である [32].

2.3 ランキング学習

ランキング学習はランキングタスクにて広く利用される (半) 教師あり学習の一種である．ランキングアルゴリズムや一般的な特徴量など，ランキング学習の基本事項については 文献 [19] に纏められている．また，大規模検索エンジン企業の持つクリックログや閲覧履歴などのユーザの振る舞いデータも精度改善に有向である．ランキング学習のアルゴリズムにおいて，LambdaMART [2] は多くのランキング学習のデータセットにおいて高いパフォーマンスを示す手法として多用されている．その一方で，Coordinate Ascent (CA) [23] は教師データのラベルが不均衡な状況においても頑健なアルゴリズムであると報告されている．本研究で用いる Web 検索のデータセットである TREC Web Track[3] ではクエリごとに適合データの個数と非適合データの個数が大きく異なるため，本研究では CA を採用する．

ランキング学習はさまざまな研究によって拡張されている．分散表現を組み合わせた研究も多数存在する [21], [47], [48]. Sousa et al. [39] はリスクを加味した特徴量選択手法を提案した．Ifada and Nayak [11] や Ustinovskiy [42] はラベルの再重み付け問題に取り組んだ．Su et al. [41] は，従来のクエリ依存モデルと対比して，クエリ依存の距離学習に適応するために理想的な候補文書集合の概念を導入した．Lucchese et al. [20] は，モデル構築時に有用なデータのみ学習に用いる Selective Gradient Boosting (SelGB) を提案した．Wang et al. [44] はオンラインランキング学習における勾配の探索にて，位置バイアスを用いることで分散を小さくする手法を提案した．

2.4 ニューラルランキング

深層学習ベースのランキングモデルはニューラルランキングと呼ばれ，人手による特徴量選択が不要となる [10], [26], [27]. Yilmaz ら [46] は BERT を用いた文書検索モデルを提案している．さまざまな研究において目覚ましい成果が報告されている一方で，ニューラルランキングモデルの性能はチューニングされたベースライン手法と同等であるという報告 [18] も存在する．つまり，ニューラルランキングモデルが真に検索精度向上に貢献するかどうかは結論が出ていない．

従って，本研究では局所的 · 大域的意味情報が検索精度向上に寄与するかどうか，人手によって設計された特徴量を用いて分析を行う．

3 提 案 手 法

本研究において，分散表現である大局的意味情報と語義である局所的意味情報の二つを扱う．それぞれの意味情報をランキング学習に組み込む．

提案するランキング学習システムでは，語義に基づく特徴量と分散表現に基づく特徴量を追加の特徴量として用いる．提案手法の概要を図 2 に示す．オフライン処理として，文書集合，

3 : https://lemurproject.org/clueweb09.php/

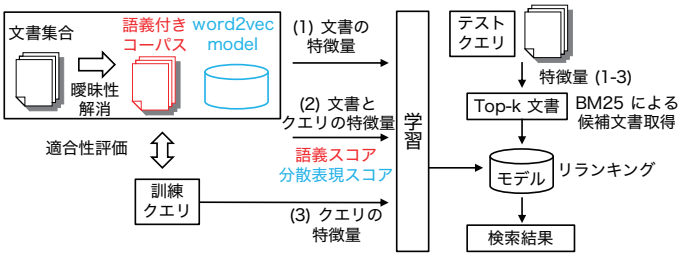

図 2 ランキング学習システムの概要図

訓練クエリ，適合性評価を用いてランキングモデルを構築する．その際，クエリや文書から抽出されたさまざまな特徴量と適合性評価のラベルを用いて，各クエリ-文書ペアに対してベクトルを作成する．これらのベクトルを用いてランキングモデルを構築する．オンラインの処理では，クエリが入力されると，ランキングモデルは検索結果候補に対して適合性スコアを算出し，適合度の降順に並んだ検索結果を提示する．

3.1 ベースライン手法

予備実験の結果，高い性能を示したため，最も単純なベースライン手法として BM25 を採用する．一般的なランキング学習システムにおいて利用される特徴量 [19] から構築されたモデルを一般特徴量モデルと呼ぶ．それらの特徴量は三種類に分類され，その一部を下記に示す．

(1) 文書特徴量: PageRank [29]，文書長，
(2) 文書-クエリ特徴量: 語の重み付け手法 (TF-IDF [36], BM25 [35]，クエリ尤度モデル [33] など)，
(3) クエリ特徴量: クエリ語，クエリ語の品詞，クエリ長．

なお，ランキング学習システムは一般的に，計算コスト削減のため，単純なスコアリング手法で得られた上位 k 件の文書のリランキングを行う．

3.2 語義スコア

語義スコアは，新たに追加される文書-クエリ特徴量の一つであり，クエリの語義と文書の語義の一致度を表す．

まずはじめに，文書とクエリに対して語義曖昧性解消を行う．

3.2.1 文書の曖昧性解消

文書中の語の曖昧性解消に IMS [49] を用いる．例を図 1 に掲載する．IMS が文 “I live in Idaho state.” を受け取ったとすると，“live” と “Idaho”，“state” が曖昧性解消され，“live” と “state” が多義語，“Idaho” は単義語である．$state^{ws(1)}$ (WordNet における state%1:15:01::) の定義は “the territory occupied by one of the constituent administrative districts of a nation” であり，コンテキストに適切な語義が付与されている．

3.2.2 クエリの曖昧性解消

IMS は自然言語文法に基づくデータから訓練されている一方で，クエリは自然言語文法に基づいていない．従って，クエリに対して直接 IMS を適用した場合には問題が生じる．そこで，我々は二種類のクエリの曖昧性解消手法の提案を行う．

意図曖昧性解消手法 本研究で扱うクエリは，一般的に Web

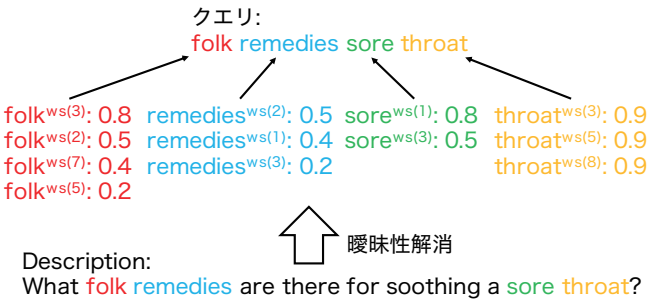

図 3　意図曖昧性解消手法の実行例

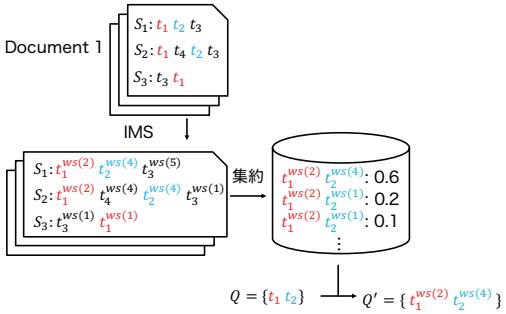

図 4　共起語曖昧性解消手法の手順

クエリと同様に，語の羅列で構成される．各クエリには *description* と呼ばれる情報が付与されており，これはクエリの検索意図が自然文で記述され，その中には大部分のクエリ語が含まれる．例えば，クエリ “folk remedies sore throat” (ID 261) の description は “What folk remedies are there for soothing a sore throat?” であり，全てのクエリキーワードを含む．description 中のクエリ語とクエリ中のクエリ語の語義は一致していると考えるのが自然である．IMS は文単位の語義曖昧性解消ツールであり，description は自然文法に基づく文である．従って，description に対して IMS を適用した結果は高精度であることを見込まれる．従って，本研究では，文献 [8] と同様に，クエリの語義曖昧性解消に description を用いる．図 3 に各クエリ語に対する語義の曖昧性解消の結果の例を示す．

共起語曖昧性解消手法　一般的に，Web 検索においてクエリの検索意図を知ることは困難であるため，意図を用いずに正確な語義曖昧性解消が望ましい．また，クエリにおいてクエリ語の並び順を考慮することはかえって語義曖昧性解消において悪影響を及ぼす可能性がある．この観点に基づいて，我々は，自然言語文法，特に語の並び順に依存しない語義曖昧性解消手法の提案を行う．その際，文献 [1] のアイデアである “*the correct translations of query terms should co-occur in target language documents and incorrect translations should tend not to co-occur.*” を踏襲する．ただし，本研究におけるソース言語はタグなしコーパス，ターゲット言語はタグありコーパスとする．つまり，我々は語の共起に基づく語義曖昧性解消手法を提案する．その処理手順を図 4 に示す．なお，図 4 では単純化のためにクエリは 2 から構成されているものの，3 語以上への拡張は同様にして行われる [4]．事前に大規模 Web コーパスに対して IMS を適用する．例では，document 1 は 3 個の文 S_1, S_2, S_3 を含む．また，S_1 は 3 単語 t_1, t_2, t_3 から構成される．このときに，IMS は t_1 の語義を語義集合の二つ目の語義であると判別したことを $t_1^{ws(2)}$ と表す．コーパス全体に対して IMS が適用されれば，その結果を語義データベースに集約する．語義データベースには，任意の共起語とその語義の組合せが確率とともに格納されている．なお，確率は，共起語と語義の組合せの出現回数を共起回数で割って求められる．t_1 と t_2 の共起に対して 3 件のエントリーが存在し，$t_1^{ws(2)}$ と $t_2^{ws(4)}$ の確率値として 0.6 が付与される．クエリ $Q = \{t_1\, t_2\}$ は，最も確率の高いエントリーである $Q' = \{t_1^{ws(2)}\, t_2^{ws(4)}\}$ として曖昧性解消される．

3.2.3　語義スコアの定式化

文書とクエリ双方に対して語義曖昧性解消が完了すれば，語義スコアの算出を行う．我々は二種類の語義スコア算出方法を提案する．なお，クエリ Q に含まれるクエリ語 q の候補語義 ($i = 1, \cdots, N$ (ただし N は単語によって異なる))，クエリ Q ($|Q|$ はクエリ長)，文書 D に含まれる単語 t の語義を ($t^{ws(i)}$ は t の i 番目の語義) と表す．

全語義スコア IMS の出力結果として各単語は確率的に複数の語義を付与されているため，本手法では全ての候補語義候補を用いる．この方針は，IMS が適切な語義に最も高く確率を付与できていない場合において有効である．全語義スコア SC_{All} を下記の通り定式化する:

$$SC_{All}(Q, D) = \frac{1}{|Q|} \sum_{q \in Q} prob_{q,t} \tag{1}$$

$$prob_{q,t} = \begin{cases} \sum_{i=1}^{N} p(q^{ws(i)}) \cdot p(t^{ws(i)}) & (t \in D \wedge t = q) \\ 0 & (t \notin D \vee t \neq q) \end{cases} \tag{2}$$

ただし $p(q^{ws(i)})$ と $p(t^{ws(i)})$ は q と t の i 番目の候補語義の確率とする．同一のクエリ語と文書中の語に対して，語義ごとに $prob_{q,t}$ の算出を行う．その際，クエリと文書両者の語義確率が高いときにスコアが高くなる．更に，全語義スコアは単語が複数の語義を持つ場合にも対応可能な手法である．

最大確率語義スコア全語義スコアはロングテール部分の語義の悪影響が受ける可能性があるため，最大確率語義スコアでは最も高い確率が付与された語義に決め打ちしてスコア計算を行う．最大確率語義スコアは下記の通り定式化される:

$$SC_{Best}(Q, D) = \frac{1}{|Q|} \sum_{q \in Q} prob_{q,t} \tag{3}$$

$$prob_{q,t} = \begin{cases} p(q^{ws(best)}) \cdot p(t^{ws(best)}) & (t \in D \wedge t = q) \\ 0 & (t \notin D \vee t \neq q) \end{cases} \tag{4}$$

ただし，$q^{ws(best)}$ と $p(q^{ws(best)})$ は q の最も高い確率を持つ語義とその確率，$t^{ws(best)}$ と $p(t^{ws(best)})$ は $q^{ws(best)}$ と同じ語義

4：一語クエリについては単に最頻語義を付与する．

を持つ t とその確率とする．最大確率語義スコアの設計方針は，最も高い確率を持つ語義のみに着目するという点を除くと全語義スコアと同様である．

3.3 分散表現スコア

近年，多くの研究において word2vec [24] や GloVe [31], BERT [7] といった分散表現が着目されている．分散表現では各単語が数百次元の密なベクトルとして表される．同様に，文書とクエリは doc2vec [15] と query2vec [21] は数百次元のベクトルで表現され，それらの比較は単語レベルではなくベクトルレベルで行われる．なお，本研究では，同一の空間上での比較を行うため，クエリと文書の分散表現はそれぞれに含まれる単語の word2vec の平均ベクトルと (それぞれクエリ平均 wor2vec 及び 文書平均 word2vec) する．

分散表現は，クエリと文書の語彙の乖離を埋めるため，ランキング学習において多用される．それらの多くは doc2vec をそのまま特徴として用いているものの，これは訓練クエリの中にテストクエリと同一もしくは非常に似たクエリが含まれているときに有効である．つまり，テストクエリの適合文書においてどのような doc2vec が有効な特徴を持つのかが明らかな場合に doc2vec 特徴量が有効ということである．しかし，本研究では完全に未知のクエリが問い合わせられる状況を前提としている．従って，本研究では文書平均 word2vec をそのまま特徴として使うのではなく，分散表現に基づきクエリと文書の (非) 類似度を計測し，分散表現スコアを算出する．なお，分散表現スコアも文書-クエリ特徴量の一種である．これらの議論を踏まえて，下記の通り分散表現スコアを定式化する:

差分分散表現スコアクエリと文書が完全に同一の概念を共有している場合には両者の差ベクトルはゼロベクトルとなる．従って，クエリ平均 word2vec と 文書平均 word2vec の差ベクトルを差分分散表現スコアとして特徴量に追加する．なお，特徴量の次元数はクエリ平均 word2vec や文書平均 word2vec の次元数と同次元である．

距離分散表現スコア距離分散表現スコアでは，類似度尺度として，クエリ平均 word2vec と文書平均 word2vec の差ではなくユークリッド距離を用いる．

コサイン類似度分散表現スコアコサイン類似度分散表現スコアでは，多くの情報検索研究において類似度尺度として用いられるコサイン類似度に基づきクエリ平均 word2vec と文書平均 word2vec の類似度を算出する．なお，次元数は 1 次元に縮約される．

4 評価実験

4.1 データセット

評価実験に用いたデータセットは，5,000 万件の英文 Web ページが含まれる ClueWeb-09 Category B document corpus[5] である．加えて，2009 年から 2012 年までの TREC Web Track Topics[6] 合計 200 件を用いた．文書とクエリに対する適合性評価も利用可能である．

ランキング学習のデータセットとしては MSLR-WEB10K や MSLR-WEB30K[7] が利用されることが多いものの，これらのデータセットでは文書集合が提供されておらず，3.2 節で提案した語義スコアを計算することができないため，本研究では ClueWeb-09 を用いることとする．

4.2 実験準備

4 分割交差検定による評価実験を行う．それぞれの分割データセットでは，3 年間分のクエリ 150 個を訓練クエリとして用いて残りの 1 年分のクエリ 50 個をテストクエリとして用いる．

評価尺度としては，検索結果上位の性能を評価するために，上位 20 文書に対する normalized discounted cumulative gain [12] (NDCG)，つまり，nDCG@20 を用いる．同様に，より多くの検索結果及び検索結果全体の性能を評価するため，上位 50 文書の精度 (P@50) と mean average precision (MAP) も計測する．

評価実験では，提案した二種類の語義スコア (全語義スコアと最大確率語義スコア) と三種類の分散表現スコア (差分分散表現スコア，距離分散表現スコア，コサイン類似度分散表現スコア) を評価する．モデルの構築と評価には RankLib[8] を用い，分散表現スコア計算に用いる word2vec は Google's pre-trained model[9] を用いた．同様に，古典的にスコアリング手法の BM25 と，一般的に特徴量 (3.1 節 参照) から構築したベースラインのランキング学習手法 (一般特徴量) の評価も行った．また，全てのランキング学習モデルは，BM25 で得られた上位 100 文書に対してリランキングを行った．

4.3 予備実験

4.3.1 クエリへの語義曖昧性解消手法の評価

語義スコアの計算に必要となる，クエリの語義曖昧性解消に用いる手法 (3.2 節の意図曖昧解消手法と共起語曖昧性解消手法参照) の評価を行う．表 1 に結果を示す．

表 1 クエリの語義曖昧性解消手法の評価

	nDCG@20	MAP
意図曖昧解消手法	.341	.397
共起語曖昧性解消手法	.289	.363

意図曖昧解消手法は共起語曖昧性解消手法よりも高精度を示した．このことから，クエリの語義曖昧性解消は挑戦的な取り組みであり，並び順の持つ文脈情報は語義曖昧性解消において重要であるということが示唆された．以降の実験では，語義曖昧性解消手法として意図曖昧解消手法を用いる．

5 : https://lemurproject.org/clueweb09.php/

6 : https://trec.nist.gov/data/webmain.html

7 : https://www.microsoft.com/en-us/research/project/mslr/

8 : https://sourceforge.net/p/lemur/wiki/RankLib/

9 : https://code.google.com/archive/p/word2vec/ 1000 億語を含む Google News dataset から学習.

クエリ ID	順位	文書 ID	適合性	語義スコア
1	1	10	0	0.9
1	2	6	1	0.7
1	3	5	2	0.5
1	4	2	0	0.4
1	5	7	0	0.3

適合文書:

$NAR_{rel} = (2+3) \cdot \frac{1}{2} \cdot \frac{1}{5} = 0.5$

非適合文書:

$NAR_{irr} = (2+3+5) \cdot \frac{1}{3} \cdot \frac{1}{5} = 0.67$

図 5　適合文書と非適合文書の正規化平均順位

4.3.2　語義スコアと分散表現スコアの予備実験

語義スコアは単体のスコアリング手法として用いられるように設計されていないため，nDCG のような評価指標を用いて直接的に性能評価を行うことは困難である．そのため，語義スコアの値の大きさと適合性判定の相関関係を調査することで，間接的に語義スコアの有用性を評価する．直感的には，適合文書と非適合文書の語義スコアの平均順位を比較する．

まず，図 5 の通り，クエリごとに語義スコアの降順に文書をソートする．クエリ 1 の適合文書に着目すると，文書 5, 6 の順位はそれぞれ 2 位と 3 位である．これらの合計である 5 を適合文書数である 2 で割ると，適合文書の平均順位が求められる．ただし，クエリごとに文書数が異なるため，平均順位の期待値の値は異なり，平均順位のままでは他のクエリとの比較を行うことができない．従って，平均準備は文書数である 5 で割られる．これにより，平均順位は 0 から 1 の範囲に正規化され，クエリごとのバイアスは消えてクエリ間の比較が可能となる．文書は語義スコアの降順でソートされているため，正規化平均順位の値が小さいほど，語義スコアが高い場合に適合文書を高い順位に配置できていることを表す．

式 5 と式 6 はそれぞれ適合文書の正規化平均順位 NAR_{rel} と非適合文書の正規化平均順位 NAR_{irr} の (クエリ語 q についての) 定式化である．

$$NAR_{rel} = \sum_{q \in Q} \left(\frac{1}{|D_{rel}|} \frac{1}{|D|} \sum_{d_{rel} \in D_{rel}} rank_{d_{rel}} \right) \quad (5)$$

$$NAR_{irr} = \sum_{q \in Q} \left(\frac{1}{|D_{irr}|} \frac{1}{|D|} \sum_{d_{irr} \in D_{irr}} rank_{d_{irr}} \right) \quad (6)$$

ただし，Q はクエリ集合，D_{rel} は適合文書集合，D_{irr} は非適合文書集合，$rank_d$ は文書 d の順位，$|D|$ $(= |D_{rel}| \cup |D_{irr}|)$ は合計文書数とする．全てのクエリを用いた平均である MNAR は下記の取りである．

$$MNAR = \sum_{q \in Q} NAR(q) \quad (7)$$

また，DR スコアの正規化平均順位は同様に式 5 と式 6，式 7 で算出され，文書が分散表現スコアでソートされるという点のみ異なる．

表 2 に，語義スコア (最大確率語義スコア) と分散表現スコア (コサイン類似度分散表現スコア) の適合文書の正規化平均順位 $MNAR_{rel}$ と非適合文書の正規化平均順位 $MNAR_{irr}$ を示す．語義スコアの $MNAR_{rel}$ は $MNAR_{irr}$ よりも小さいため，高い語義スコアを持つ文書は適合文書である傾向を持つ．同様に，分散表現スコアにおいても，$MNAR_{rel}$ は $MNAR_{irr}$ よりも小さいため，高い分散表現スコアを持つ文書は適合文書である傾向を持つ．これらの観測から，適切な語義スコアと分散表現スコアが算出されている．

表 2　適合文書と非適合文書の正規化平均順位

	適合文書	非適合文書
語義スコア (最大確率語義スコア)	.382	.546
分散表現スコア (コサイン類似度分散表現スコア)	.450	.510

表 3　提案ランキング学習モデルの評価実験結果．“*” と “**” は，下線が引かれた一般特徴量と，それぞれ有意水準 $p < 0.05$ と $p < 0.001$ において提案手法の有意差が統計的に有意に差があることを表す．

	nDCG@20	P@50	MAP
[ベースライン]			
BM25 [35]	..267	324	.282
一般特徴量	.333	.318	.398
[語義スコア]			
最大確率語義スコア	.341*	.319	.397
全語義スコア	.331	.321	.388
[分散表現スコア]			
差分分散表現スコア	.338	.308	.391
距離分散表現スコア	.339	.312	.386
コサイン類似度分散表現スコア	.341	.312	.392
最大+コサイン	.349**	.312	.397

4.4　実験結果

4.4.1　手法間の分析

表 3 に実験結果を示す．語義スコア (3.2 節参照) に関して，最大確率語義スコアは，特に nDCG@20 において全語義スコアよりも高い結果を示した．このことから，意図曖昧性解消手法によってクエリに適切な語義に最も高い確率が付与され，適合文書を高いランクに配置できていることが示唆される．分散表現スコア (3.3 節参照) に関して，コサイン類似度分散表現が最も高精度であり，差分分散表現と距離分散表現も一般特徴量の精度を上回った．最終的に，最大確率語義スコアとコサイン類似度分散表現スコアの組合せ (最大+コサイン) が最も高い精度を示した．符号検定の結果，BM25 及び一般特徴量と比較して，最大+コサインは，統計的に有意に精度が向上した (p<0.001)．

最大+コサインは 105 個のクエリ (53%) にて一般特徴量よりも高精度を示し，62 個のクエリ (31%) において劣っている．最大+コサイン及び一般特徴量において，残りの 33 個のクエリ (16%) から適合文書を発見できなかった (以降，null query と表現)．分析の結果，適合文書が少ないクエリほど null query となる傾向が観測された．この結果より，null query はランキング学習の性質というよりもクエリの難易度によるものであると示唆される．

P@50 と MAP において，語義スコア，分散表現スコア，最大+コサインは一般特徴量と同等の精度を示した．この結果から，提案手法はより多くの検索結果や検索結果全体の観点でも検索精度を低下させていないということを示した．

続いて，提案手法が効果的な状況を分析した．まず，提案手法はクエリ長が長いほどクエリのコンテキスト情報を利用できて高精度であることが期待されるため，クエリ長に着目する．表 4 は，クエリ長ごとの nDCG@20 における最大+コサイン

表 4　クエリ長と精度差

クエリ長	クエリ数	一般特徴量	最大+コサイン	差分
1	56	.298	.301	.003
2	65	.393	.405	.012
3	62	.312	.337	.025
4	14	.301	.335	.034
5	2	.335	.384	**.049**

表 5　適合文書数と精度差

適合文書数	クエリ数	一般特徴量	最大+コサイン	差分
1-9	22	.117	.153	**.036**
10-99	104	.250	.265	.015
100-	68	.558	.568	.010

と一般特徴量の精度の差分を示す．この結果から，クエリ長が大きくなるほど精度の差分が大きくなることから，提案手法は適切にクエリのコンテキスト情報を利用できていることが示唆された．

次に，適合文書数と精度の関係を分析する．特に，適合文書数が少ない “高難易度クエリ” に着目する．適合文書数に応じてクエリを 3 種類に分類する．すなわち，適合文書数が 10 個未満，10 個から 99 個，100 個以上である．表 5 にクエリのカテゴリごとの nDCG@20 における最大+コサインと一般特徴量の精度の差分を示す．適合文書数が少ないときほど差分が大きく，提案手法は高難易度クエリに対してより効果的であることが判明した．

4.4.2　語義スコアの分析

以降，個別のクエリに着目して語義スコアの影響について議論する．“state” の持つ語義には “the territory occupied by one of the constituent administrative districts of a nation” や “the way something is with respect to its main attributes” が存在する．クエリ “idaho state flower” (ID 197) の “state” の語義は前者であるものの，BM25 は後者の語義を持つ文書に高いスコアを付与しているのに対して，語義スコアは正しい語義を持つ文書に高いスコアを付与している．その結果，語義スコアは適合文書を発見することに成功した．

“news” の語義として “information about recent and important events” や “information reported in a newspaper or news magazine” が存在する．クエリ “rocky mountain news” (ID 139) の “news” は両方の語義が当てはまると考えられるものの，最大確率語義スコアは前者のみを正しいと判断した結果，後者の語義を持つ適合文書のいくつかを発見できなかった．このような問題を解決するうえで，類似語義の集約を行うことは有効であると考えられる．その他の解決方針としては，最大確率語義スコアと全語義スコアをクエリごとに切り替えて用いる方針が考えられる．最大確率語義スコアは 200 個中 99 個のクエリで全語義スコアより高い精度を示し，68 個のクエリにて劣っている．より高精度を示した手法をクエリごとに切り替えて用いた場合，nDCG@20 は .349 となり，それぞれの手法を単体で用いた場合よりも高い精度を達成することができた．従って，いずれの手法がより効果的な適切に判別することができれば，より高精度を目指すことができると示唆される．

表 6　最大+コサインが一般特徴量より高精度のクエリ上位 10 件

id	クエリ	適合	差分	理由
102	**fickle creek farm**	**1**	.247	正確な語義
79	voyager	69	.241	正確な語義
122	**culpeper national cemetery**	**2**	.175	正確な語義，適切な分散表現
121	sit and reach test	19	.169	正確な語義，適切な分散表現
118	poem in your pocket day	11	.161	正確な語義，適切な分散表現
85	milwaukee journal sentinel	52	.157	適切な分散表現
1	obama family tree	246	.145	正確な語義
29	ps 2 games	245	.144	正確な語義
93	raffles	16	.132	正確な語義，適切な分散表現
140	**east ridge high school**	**1**	.123	適切な分散表現

表 7　一般特徴量より最大+コサインがより低精度のクエリ下位 10 件

id	クエリ	クエリ長	差分	理由
81	**afghanistan**	**1**	-0.190	不適切な分散表現
200	ontario california airport	3	-0.146	不適切な分散表現
24	**diversity**	**1**	-0.129	誤った語義
144	trombone for sale	2	-0.105	語義付与失敗
139	rocky mountain news	3	-0.104	語義付与失敗
168	**lipoma**	**1**	-0.090	不適切な分散表現
108	ralph owen brewster	3	-0.085	語義付与失敗
10	cheap internet	2	-0.073	誤った語義
159	**porterville**	**1**	-0.070	不適切な分散表現
171	ron howard	2	-0.057	語義付与失敗

また，“volvo” (ID 21), “starbucks” (ID 27), “atari” (ID 31) といった単語は語義集合の中に含まれていなかったため，これらのクエリに対しては語義曖昧性解消が失敗し，語義スコアが適切に機能しなかった．これらの問題は entity linking [34] によって解決可能であると考えられる．

以上の結果から，先行研究 [8], [13], [37], [40], [43] と同様の結論として，語義スコアは語義曖昧性解消の性能に大きく依存するという結果が得られた．また，定量的 · 定性的分析を通して，本研究の語義曖昧性解消手法は実用的なレベルであり，クエリ及び文書に対して正確に語義を付与できていると判明した．

4.4.3　分散表現スコアの分析

“milwaukee journal sentinel” (ID 85), “the secret garden” (ID 43), “east ridge high school” (ID 140) といったクエリでは分散表現スコアを適用することで検索精度が向上した．その一方で，“lipoma” (ID 168), “ontario california airport” (ID 200), “afghanistan” (ID 81) といったクエリではかえって検索精度が低下した．

特に，一般特徴量では適合文書である Wikipedia 記事を発見できているのに対して，最大+コサインでは “afghanistan” (ID 81) に対して適合文書を全く発見することができなかった．Wikipedia 記事のように多様なトピックについて記述されている文書ではクエリ平均 word2vec と 文書平均 word2vec の類似度が低くなる．その結果，分散表現スコアは低くなり，最大+コサインでは適合文書を発見できなかったと考えられる．このことから，文書から作成される分散表現よりも，例えばフレーズ，パッセージ，センテンスといったより小さな粒度の分散表現とクエリの分散表現と比較することが有用であることを示唆する．

4.4.4　成功 · 失敗したクエリ

表 6 と表 7 に，最大+コサインと一般特徴量の精度の差が大きい 10 個のクエリを掲載する．表 6 は最大+コサインの精度が高い 10 個のクエリであり，表 **??** は一般特徴量の精度が高い

10 個のクエリである．

表 7 に関して，最大+コサインは “fickle creek farm” (ID 102) と “voyager” (ID 79) にておいて適合文書を検索結果上位に提示したのに対して，一般特徴量では適合文書が発見できなかった．一般特徴量は多数のクエリキーワードを含む非適合文書を上位にランキングしている．それに対して，最大+コサインでは，語義スコアと分散表現スコアを持ちることで，クエリキーワードをほとんど含まない適合文書を上位にリランキングすることに成功した．理由からも，語義スコアと分散表現スコア双方が効果的であり，10 個のクエリのうち語義スコアが 8 個，分散表現スコアが 6 個のクエリの精度向上に役立った．

その一方で，表 7 の通り，最大+コサインは “afghanistan” (ID 81) に対して適合文書を全く発見できなかった．クエリ長の小さいクエリではクエリのコンテキスト情報が十分に活用できず語義スコアの向上は困難であるため，分散表現の高性能化によって適切なスコアを付与することを目指す．更に，未定義語による検索精度の悪化が確認された．

5 おわりに

本稿では，局所的意味情報として語義スコアを，大域的意味情報として分散表現スコアを加味したランキング学習手法を提案した．実験の結果，提案手法は一般的な特徴量を用いた構築したモデルと比較して，統計的有意に検索精度を向上させることができた．

今後は類似語義の統合，異なる分散表現手法や，分散表現粒度の検討などが今後の課題である．

謝　辞

本研究の一部は，JSPS 科研費 (JP15K20990，JP17K12684)，JST ACT-I の助成を受けたものである．ここに記して謝意を表す．また，本研究を遂行するうえで有益なコメントを頂いたシンガポール国立大学の Min-Yen Kan 准教授に謝意を表す．

文　献

[1] L. A. Ballesteros and W. B. Croft. Resolving Ambiguity for Cross-Language Retrieval. In *Proc. of SIGIR*, pages 64–71, 1998.
[2] C. J. C. Burges. From RankNet to LambdaRank to LambdaMART : An Overview. Technical Report MSR-TR-2010-82, Microsoft Research Technical Report, 2010.
[3] Y. S. Chan and H. T. Ng. Scaling Up Word Sense Disambiguation via Parallel Texts. In *Proc. of AAAI*, pages 1037–1042, 2005.
[4] H.-H. Chen, G.-W. Bian, and W.-C. Lin. Resolving Translation Ambiguity and Target Polysemy in Cross-Language Information Retrieval. In *Proc. of ACL*, pages 215–222, 1999.
[5] A. G. Chifu, F. Hristea, J. Mothe, and M. Popescu. Word Sense Discrimination in Information Retrieval: A Spectral Clustering-based Approach. *Information Processing and Management*, 51(2):16–31, 2015.
[6] A. G. Chifu and R. T. Ionescu. Word Sense Disambiguation to Improve Precision for Ambiguous Queries. *Central European Journal of Computer Science*, 2(4):398–411, 2012.
[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*, pages 4171–4186, 2019.
[8] J. Guyot, G. Falquet, S. Radhouani, and K. Benzineb. Analysis of Word Sense Disambiguation-Based Information Retrieval. In *Proc. of CLEF*, pages 146–154, 2008.
[9] Z. S. Harris. Distributional Structure. *WORD*, 1954.
[10] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck. Learning Deep Structured Semantic Models for Web Search using Clickthrough Data. In *Proc. of CIKM*, pages 2333–2338, 2013.
[11] N. Ifada and R. Nayak. How Relevant is the Irrelevant Data: Leveraging the Tagging Data for a Learning-to-Rank Model. In *Proc. of WSDM 2016*, pages 23–32, 2016.
[12] K. Järvelin and J. Kekäläinen. IR Evaluation Methods for Retrieving Highly Relevant Documents. In *Proc. of SIGIR*, pages 41–48, 2000.
[13] R. J. Krovetz and W. B. Croft. Lexical Ambiguity and Information Retrieval. *ACM Transactions on Information Systems (TOIS)*, 10(2):115–141, 1992.
[14] M. Lapata and F. Keller. An Information Retrieval Approach to Sense Ranking. In *Proc. of HLT/NAACL*, pages 348–355, 2007.
[15] Q. V. Le and T. Mikolov. Distributed Representations of Sentences and Documents. In *Proc. of ICML*, pages 1188–1196, 2014.
[16] Y. K. Lee and H. T. Ng. An Empirical Evaluation of Knowledge Sources and Learning Algorithms for Word Sense Disambiguation. In *Proc. of EMNLP*, pages 41–48, 2002.
[17] M. Lesk. Automatic Sense Disambiguation Using Machine Readable dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *in Proc. of SIGDOC*, pages 24–26, 1986.
[18] J. Lin. The Neural Hype and Comparisons Against Weak Baselines. *ACM SIGIR Forum*, 52.
[19] T.-Y. Liu. Learning to Rank for Information Retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.
[20] C. Lucchese, F. M. Nardini, R. Perego, S. Orlando, and S. Trani. Selective Gradient Boosting for Effective Learning to Rank. In *Proc. of SIGIR*, pages 155–164, 2018.
[21] C. Luo, Y. Liu, M. Zhang, and S. Ma. Query Ambiguity Identification Based on User Behavior Information. In *Proc. of AIRS*, pages 36–47, 2014.
[22] R. Mandala, T. Tokunaga, and H. Tanaka. The Use of WordNet in Information Retrieval. In *Workshop on WordNet@ACL/COLING 1998*, 1998.
[23] D. Metzler and W. B. Croft. Linear Feature-based Models for Information Retrieval. *Information Retrieval*, 10(3):257–274, 2007.
[24] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. In *Proc. of ICLR*, pages 1–12, 2013.
[25] G. A. Miller, C. Leacock, R. Tengi, and R. T. Bunker. A Semantic Concordance. In *Proc. of HLT*, pages 303–308, 1993.
[26] B. Mitra and N. Craswell. An Introduction to Neural Information Retrieval. *Foundations and Trends in Information Retrieval*, 31.
[27] B. Mitra, F. Diaz, and N. Craswell. Learning to Match using Local and Distributed Representations of Text for Web Search. In *Proc. of WWW*, pages 1291–1299, 2017.
[28] H. T. Ng. Exemplar-Based Word Sense Disambiguation: Some Recent Improvements. In *Proc. of EMNLP*, pages 41–48, 1997.
[29] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report SIDL-WP-1999-0120, Stanford Digital Library Technologies Project, 1998.
[30] T. Pedersen. A Simple Approach to Building Ensembles of Naive Bayesian Classifiers for Word Sense Disambiguation. In *Proc. of NAACL*, pages 63–69, 2000.
[31] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proc. of EMNLP*, pages 1532–1543, 2014.
[32] A. Pirkola, T. H. aind Heikki Keskustalo, and K. Jäärvelin. Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings. *Foundations and Trends in Information Retrieval*, 4(3–4):209–230, 2001.
[33] J. M. Ponte and W. B. Croft. A language Modeling Approach to Information Retrieval. In *Proc. of SIGIR*, pages 275–281, 1998.
[34] J. Raiman and O. Raiman. DeepType: Multilingual Entity Linking by Neural Type System Evolution. In *Proc. of AAAI*, 2018.
[35] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proc. of TREC-3*, pages 109–126, 1995.
[36] G. Salton and C. Buckley. Term-Weighting Approaches in Automatic Text Retrieval. *Journal of Information Processing and Management*, 24(5):513–523, 1988.
[37] M. Sanderson. Word Sense Disambiguation and Information Retrieval. In *Proc. of SIGIR*, pages 142–150, 1994.
[38] B. Snyder and M. Palmer. The English All-Words Task. In *Proc. of Senseval-3 Workshop*, 2004.
[39] D. X. D. Sousa, S. D. Canuto, T. C. Rosa, W. S. Martins, and M. A. Gonçalves. Incorporating Risk-Sensitiveness into Feature Selection for Learning to Rank. In *Proc. of CIKM*, pages 257–266, 2016.
[40] C. Stokoe, M. Oakes, and J. Tait. Word Sense Disambiguation in Information Retrieval Revisited. In *Proc. of SIGIR*, pages 159–166, 2003.
[41] Y. Su, I. King, and M. Lyu. Learning to Rank Using Localized Geometric Mean Metrics. In *Proc. of SIGIR*, pages 45–54, 2017.
[42] Y. Ustinovskiy, V. Fedorova, G. Gusev, and P. Serdyukov. An Optimization Framework for Remapping and Reweighting Noisy Relevance Labels. In *Proc. of SIGIR*, pages 105–114, 2016.
[43] E. M. Voorhees. Using WordNet to Disambiguate Word Senses for Text Retrieval. In *Proc. of SIGIR*, pages 171–180, 1993.
[44] H. Wang, S. Kim, E. McCord-Snook, Q. Wu, and H. Wang. Variance Reduction in Gradient Exploration for Online Learning to Rank. In *Proc. of SIGIR*, pages 835–844, 2019.
[45] D. Yarowsky. Unsupervised Word Sense Disambiguation Rivaling Supervised MethodS. In *Proc. of ACL*, pages 189–196, 1995.
[46] Z. A. Yilmaz, W. Yang, H. Zhang, and J. Lin. Cross-domain modeling of sentence-level evidence for document retrieval. In *Proc. of EMNLP-IJCNLP*, pages 3488–3494, 2019.
[47] H. Zamani and W. B. Croft. Embedding-based Query Language Models. In *Proc. of ICTIR*, pages 147–156, 2016.
[48] G. Zheng and J. Callan. Learning to Reweight Terms with Distributed Representations. In *Proc. of SIGIR*, pages 575–584, 2015.
[49] Z. Zhong and H. T. Ng. It Makes Sense: A Wide-Coverage Word Sense Disambiguation System for Free Text. In *Proc. of ACL*, pages 78–83, 2010.
[50] Z. Zhong and H. T. Ng. Word Sense Disambiguation Improves Information Retrieval. In *Proc. of ACL*, pages 273–282, 2012.