意味的逆引き辞書『真言(まこと)』におけるスコア付け

村脇 有吾 †　粟飯原 俊介 ‡　原田 泰佑 ††　長尾 真 ⋆　田中 久美子 †
† 九州大学大学院システム情報科学研究院　‡ アクセンチュア株式会社
†† 九州大学工学部　⋆ 京都大学名誉教授
{murawaki, tai_harada, kumiko}@cl.ait.kyushu-u.ac.jp
shunsuke.aihara@accenture.com　maknag@fm2.seikyou.ne.jp

1　はじめに

本稿では、意味的逆引き辞書『真言』の精度向上に向けた取り組みを報告する。ここで、意味的逆引き (以下逆引きとよぶ) とは、意味を説明する自然文を入力したとき、その意味を表す言葉を得ることである。例えば、クエリ「すぐれた洞察力を持つこと」から「炯眼」を得る。逆引きは、度忘れした単語を思い出したり、適切な表現を探すのに役立つことが期待される。

前回の報告 [6] では、既存の人間用辞書を用いた実現方法を提案した。人間用辞書を用いることで、逆引きを自然文クエリと辞書中の説明文とのマッチング問題として定式化し、情報検索的解法を示した。すなわち、辞書中の各項目を文書とみなし、各項目の説明文から転置インデックスを作成した。そして、各項目を自然文クエリによってスコア付けし、スコア順に利用者に提示した (図 1)。

逆引きの難しい点として、クエリと説明文との間の表現のずれが挙げられる。例えば、「地質や地形が悪い」と「軟弱な地層から成る」は似た意味を表すが、内容語をまったく共有しない。しかも、通常の文書検索とくらべて文書が極端に短く、説明文が数単語に過ぎないことも少なくない。そのため、クエリ中の単語が文書中の他の箇所で出現することも期待できない。

本稿では、このような課題を考慮して、スコア付けの改善を試みる。本稿で試す様々なスコア付け手法は、着目する手がかりをもとに、次の 3 種類に整理できる。

単語一致　クエリと説明文との間の単語の一致は、表現のミスマッチのおそれがあるとはいえ、やはり重要な手がかりとなる。

トピック一致　クエリと説明文を潜在意味空間に写像し、その空間上で類似度を測る。これにより表現のミスマッチの影響を緩和する。

範疇一致　クエリと説明文から対象となる言葉の範疇 (品詞) を推定し、その一致度を測る。

図 1: 意味的逆引き辞書『真言』の実行例

これらの手法は、単独で用いることも可能だが、組み合わせることで互いの欠点を補うことが期待できる。

実験では、これらの手法を複数の辞書で検証した。その結果、単独では単語一致がもっとも有効だが、これにトピック一致と範疇一致を手がかりとして加えると精度が向上することが確認できた。

2　関連研究

逆引きは含意関係認識や言い換えと関連する。ただし、同じ語の同じ語義の説明文であっても、必ずしも含意や言い換えの関係にあるとは限らない。より緩やかな関係を捉える必要がある。

逆引きは、文書が短いという点で、一般的な文書検索よりも、むしろ Twitter に代表されるマイクロブログの処理に近い。[5] は BTM (Biterm Topic Model) という LDA (Latent Dirichlet Allocation) の拡張を提案している。このトピックモデルは、文書中の単語の共起を明示的に組み込むことにより、短いテキストから得られる限られた情報を有効に活用する。

[3] は重み付き行列因子分解 (WMF) を用いたトピックモデル (WTMF) を提案している。彼らは文のペアの類似度判定に取り組んでおり、辞書のペアを用いた自動評価実験を行っているという点で本稿と共通する。

3 問題設定

逆引きの実現のために既存の人間用辞書を用いる。辞書には**見出し語**が並ぶ。見出し語は 1 個以上の語義を持ち、各語義について短い**説明文**が与えられる。見出し語、説明文の組を**項目**とよぶ。その他に、品詞やその他の付加情報が記述されている場合がある。範疇一致を手がかりとする場合は、あわせて品詞も用いる。

逆引きを検索エンジンとして実装する。各項目を自然文クエリによってスコア付けし、スコア順に提示する。したがって、適切な見出し語に高い順位を与えるスコア付け手法の設計が目標となる。

4 提案手法

スコア付け手法として、単語一致、トピック一致、範疇一致の 3 種類を試す。いずれの手法も $f_*(q,d)$ の形で表され、自然文クエリ q と各項目の説明文 d の組に対してスコアを返す。また、それらのスコアの線形和 $\sum_i w_i f_i(q,d)$ も新たなスコアとする。ここで、$w_i > 0$ はスコアの重みである。

4.1 単語一致

クエリと説明文との単語の共有は、表現のミスマッチのおそれがあるとはいえ、やはり重要な手がかりとなる。そこで、クエリと説明文が共有する単語にスコアを与え、単語スコアの総和を各項目のスコアとする。

前回の報告 [6] では、単語一致の手法として TF-IDF を用いた。本稿では、TF-IDF の拡張であり、予備実験でより良い精度が得られた Okapi BM25[4] を採用する。

4.2 トピック一致

単語一致では、「地質や地形が悪い」と「軟弱な地層から成る」のように、似た意味を表す別の表現がマッチしないという問題がある。この問題に対処するためにトピックモデルを用いる。クエリ q と説明文 d をそれぞれ K 次元の潜在意味ベクトルに写像し、両者のコサイン類似度をスコアとする。いま、トピック k が地質に関するトピックを表すとすると、上記 2 例に対応する潜在意味ベクトルは k 番目の要素に大きな値を持つ。これにより、単語を共有しない場合にも、大きなスコアを与えることが可能となる。

前回の報告 [6] では、LSI (Latent Semantic Indexing) を用いた。しかし、予備実験で、LDA でより高い精度が得られたことから、LDA を採用する。LDA の推論には Gibbs sampling を用いる [2]。

潜在意味ベクトルとして、文書のトピック分布を用いるのが自然である。しかし、[3] が指摘するように、わずか 1、2 個のトピックにほとんどの確率質量が割り当てられ、類似度尺度としては良い性能が得られない。そこで、[3] に従い、単語のごとにトピック確率 $P(z|w)$ を求め、その総和をクエリ・説明文の潜在意味ベクトルとする。

LDA に加えて、短いテキストを考慮したモデルである BTM [5] も試す。[5] は推論に Gibbs sampling を用いている。しかし、実行に非常に長い時間を要したため、本稿ではより収束の速い変分法を採用する [1]。潜在ベクトルとしては、LDA と同様に、単語のトピック分布の和を用いる。

4.3 範疇一致

上記 2 手法は単語の並びを無視する。また、日本語の「こと」、英語の “to” などの機能語はスコア付けに役に立たないと仮定している。しかし、文書の短さを考えると、これらが提供する手がかりは無視できない。

単語の並びや機能語が提供する手がかりとして、範疇に着目する。例えば、日本語の場合、「〜すること。」というクエリ・説明文から、対象が名詞であると推測できる。英語の場合も同様に、‘to make” のように “to” で始まるなら動詞、“having” のような分詞で始まるなら形容詞と推測できる。

クエリ・説明文間の範疇一致を測るために、辞書に記載された品詞を教師として、多値分類器を学習する。多値分類器を用いて、以下の手順で説明文・クエリをベクトルに変換する。まず、説明文・クエリを多値分類器に入力し、品詞スコア $\boldsymbol{s} = (s_1, \cdots, s_C)$ を得る。ここで、C は品詞の異なり数。次に、$\boldsymbol{s}$ に softmax 関数を適用し、ベクトル $\boldsymbol{v} = (v_1, \cdots, v_C)$ を得る。

$$v_i = \frac{\exp(v_i/\tau)}{\sum_{j=1}^{C} \exp(v_j/\tau)}$$

ここで温度係数 τ は分配の度合いを制御する。このベクトルのコサイン類似度を最終的なスコアとする。

多値分類器としては線形の Passive-Aggressive アルゴリズム [1] を採用する。分類に用いる特徴量は、日本語の場合を図 2 に、英語の場合を図 3 に示す。特徴量はいずれも 2 値である。日本語は主辞後置型であり、文末の情報だけでほぼ分類可能である。英語の場合は、逆に、文頭を見ればわかる場合が多い。ただし、英語の係り受けは双方向であり、“cruel, violent, or unfair treatment” のように長い前方修飾句を伴うことがある。文頭の数語だけを見る現在の特徴量には限界がある。係り受けの利用が有効と思われるが、今後の課題としたい。

[1] ただし、小規模な予備実験では、LDA、BTM ともに Gibbs sampling の方が変分法よりも良い性能を示した。

	日本語			英語		
名称	広辞苑	J_WORD	JAWiktionary	Cambridge	E_WORD	ENWiktionary
用途	辞書構築	評価	評価	辞書構築	評価	評価
見出し語数	216,463	267,692	7,176	46,345	172,404	501,171
共通見出し語数	—	90,662	2,500	—	31,048	35,573
説明文の単語数 平均 (中央値)	19.2 (12)	6.9 (6)	10.4 (7)	13.5 (13)	6.6 (5)	7.5 (4)

表 1: 辞書の諸元

⟨W0⟩	⟨W-1⟩	⟨W-2⟩
⟨PF0⟩	⟨PF-1⟩	⟨PF-2⟩
⟨PF0, PS0⟩	⟨PF-1, PS-1⟩	⟨PF-2, PS-2⟩
⟨W-1, W0⟩	⟨PF-1, PF0⟩	⟨W-2, W-1, W0⟩
⟨PF-1, PS-1, PF0, PS0⟩		
⟨PF-2, PF-1, PF0⟩		
⟨PF-2, PS-2, PF-1, PS-1, PF0, PS0⟩		

図 2: 日本語品詞推定の特徴量テンプレート。W は表層形、PF は品詞 1 段目、PS は品詞 2 段目、それらに続く数字は単語の位置。0 は最初の文の句点の直前の単語、-1 はその 1 個前、-2 は 2 個前の単語。

⟨W0⟩	⟨W1⟩	⟨W2⟩
⟨W3⟩	⟨W4⟩	⟨P0⟩
⟨P1⟩	⟨P2⟩	⟨P3⟩
⟨P4⟩	⟨W0, W1⟩	⟨W1, W2⟩
⟨W3, W4⟩	⟨P0, P1⟩	⟨P1, P2⟩
⟨P3, P4⟩	⟨P0, P1, P2⟩	⟨P1, P2, P3⟩
⟨P2, P3, P4⟩	⟨P0, P1, P2, P3⟩	⟨P1, P2, P3, P4⟩
⟨P0, P1, P2, P3, P4⟩		

図 3: 英語品詞推定の特徴量テンプレート。W は表層形、P は品詞、それらに続く数字は単語の位置。0 は文の先頭の単語、1 はその次の単語。

5 実験

5.1 実験手順

手法の有効性検証のために、実際の自然文クエリとそれに対応する正解の組を大量に確保するのは困難である。そこで、近似的だが大規模な自動評価を行う。

逆引き辞書構築に用いた辞書 A とは別に辞書 B を用意し、A、B に共通して収録されている見出し語に着目する。各共通見出し語について、B 中の説明文をクエリとして与え、該当する見出し語の順位によって性能を評価する [2]。ただし、見出し語には一般に複数の語義があり、語義同士の対応は分からない。B の見出し語に複数の語義が与えられている場合は、第 1 語義のみを評価に用いる。A の複数の語義については、最上位の候補を機械的に正解とみなす。

評価尺度として Mean Reciprocal Rank (MRR) を用いる。MRR は順位の逆数の平均である。ただし、上位 1,000 件に正解が入らなければ、正解なしとみなす。

[2]該当見出し語より上位の項目が必ずしも不適切とは限らないため、見かけ上の精度よりも良い結果を返しているかもしれない。

5.2 データ

日本語と英語の辞書を用いた (表 1)。日本語では、逆引き辞書構築に**広辞苑**第 6 版を、評価用に EDR 辞書の日本語単語辞書 (**J_WORD**)、および日本語版 Wiktionary の 2013 年 7 月 5 日付けのダンプ (**JAWiktioanry**) を用いた。

英語では、逆引き辞書構築に Cambridge Advanced Learner's Dictionary 第 4 版 (**Cambridge**) を、評価用に EDR 辞書の英語単語辞書 (**E_WORD**)、および 2013 年 5 月 27 日付けの英語版 Wiktionary (**ENWiktionary**) を用いた。

各辞書から、見出し語、説明文の組を抽出した。また、広辞苑と Cambridge からはあわせて品詞も抽出した。ただし、広辞苑の品詞はそのまま用いるのではなく、活用形と動詞の自他の区別を廃した。例えば、「自五」、「他上一」などは「動詞」に集約した。

5.3 各手法の設定

単語一致手法 (**BM25**) のための単語抽出は以下の手順で行った。日本語は MeCab (辞書は ipadic) によって形態素解析を行い、形態素列から品詞に基づく規則によって内容語を抽出した。また、活用語は原形に戻した。英語は stopword を除外し、Porter stemmer によってステミングを行った。

トピックモデルでは単語一致の場合と同じ単語集合を用いた。予備実験では次元が大きいほど良い精度が得られたが、速度の面から 500 次元を採用した。**LDA** の Gibbs sampling は burn-in 100 反復のあと、10 個の連続するサンプルを採取し平均した。Dirichlet 分布のハイパーパラメータは最尤推定により求めた。**BTM** の変分推論は 20 回の反復を行った。

範疇一致 (**POS**) の特徴量抽出のために、日本語では MeCab による形態素解析、英語では NLTK 付属の最大エントロピー法による品詞タグ付けを行った。訓練には品詞付きの項目すべてを用いた。多値分類器の学習には PA-I を採用し、パラメータ $C = 0.1$、訓練の反復回数は 10、温度係数 $\tau = 0.5$ とした。

手法の組み合わせにおける重み付けでは、発見的に様々な値を試した。開発セットを用いたパラメータチューニングは今後の課題とする。

	日本語		英語	
	広辞苑 – J_W.	広辞苑 – JAWikt.	Camb. – E_W.	Camb. – ENWikt.
BM25	.284	.193	.130	.181
LDA	.170	.124	.088	.123
BTM	.132	.091	.058	.083
POS	.038	.020	.010	.010
BM25 + 10 × LDA	.286	.195	.131	.183
BM25 + 10 × POS	.291	.194	.145	.200
BM25 + 10 × LDA + 10 × POS	**.292**	**.195**	**.146**	**.201**

表 2: 各手法の精度 (MRR)。手法横の数字はスコアの重み。

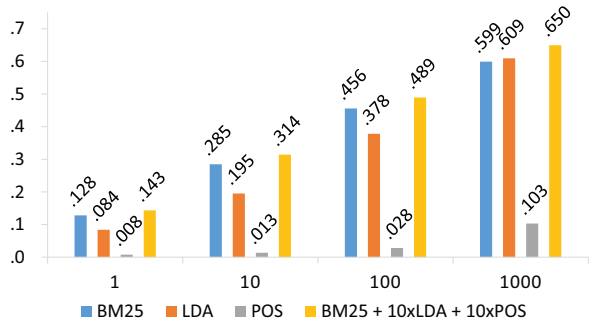

図 4: Recall@N (Cambridge – ENWiktionary)

5.4 結果

各手法の精度を表 2 に示す。スコアを単独で用いた場合、単語一致の BM25 がもっとも高精度で、トピック一致の LDA、BTM がそれに続いた。BTM は短いテキストを考慮したモデルであり、LDA を上回る精度を期待したが、一貫して下回った。予想される通り、範疇一致が一番精度が低かった。

手法の組み合わせでは、BM25+LDA、BM25+POS の両者で、BM25 単独とくらべて精度が向上した。単独では低精度の範疇一致が、特に英語において精度向上に大きく貢献しているのは興味深い。範疇一致は、例えば、クエリ「思いがけなくめぐり合うこと。」(正解は「邂逅」) について、BM25 単独では上位に来る「対戦相手としてめぐり合う。 」(「当たる・中る」) のように品詞が一致しない候補の順位を引き下げた。さらに 3 手法を組み合わせた場合、すべての辞書ペアで最高精度を達成した。

評価用辞書としては、EDR 辞書よりも Wiktionary の方が難しく、また実際の利用者の入力に近いと事前に予想した。前者は作成者が専門家という点で逆引き辞書構築用の辞書と共通するのに対して、後者は一般利用者が作成しているからである。結果として、日本語版では J_WORD が Wiktionary よりも高い精度を出したが、英語版では、予想に反して、Wiktionary が E_WORD よりも精度が高くなった。また、広辞苑は Cambridge よりも見出し語数が多いにもかかわらず、より高精度となる傾向が見られた。

上位 N 件中に正解見出し語が存在する割合 (Recall@N) を図 4 に示す。3 手法を組み合わせると、約 3 割の正解が上位 10 件に入った。一方、4 割近くが上位 1,000 件にも入らず、実用面で課題が残る。

6 おわりに

本稿では、意味的逆引き辞書『真言』の精度向上を目的に、単語一致、トピック一致、範疇一致という 3 種類の手がかりでスコア付けを行った。これらを組み合わせることにより精度が向上することを示した。

各手がかりを利用する手法は、本稿で検証したものに限らない。例えば、トピック一致に用い得るトピックモデルとしては、WTMF[3]、連続空間トピックモデル [7] を含む多くの手法が提案されている。今後はそれらの有効性も検証したい。

既存の人間用の辞書 (のみ) を用いて逆引きを実現することの限界も実感している。手法を工夫したところで、1、2 語の短い説明文はいかんともしがたい。今後の方向性として、人間用の辞書以外の資源の利用を考えている。

謝辞

岩波書店には『広辞苑』第 6 版の利用を許可いただきました。Dictionary and thesaurus content Copyright © Cambridge University Press, courtesy of Cambridge Dictionaries Online: `http://dictionary.cambridge.org`

参考文献

[1] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, Vol. 7, pp. 551–585, 2006.

[2] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *PNAS*, Vol. 101, pp. 5228–5235, 2004.

[3] Weiwei Guo and Mona Diab. Modeling sentences in the latent space. In *Proc. of ACL*, pp. 864–872, 2012.

[4] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline M. Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-3. In *Proc. of TREC*, pp. 109–126, 1995.

[5] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A biterm topic model for short texts. In *Proc. of WWW*, pp. 1445–1456, 2013.

[6] 粟飯原俊介, 長尾真, 田中久美子. 意味的逆引き辞書『真言』. 言語処理学会第 19 回年次大会 発表論文集, pp. 406–409, 2013.

[7] 持橋大地, 吉井和佳, 後藤真孝. ガウス過程に基づく連続空間トピックモデル. NL 研, Vol. 2013, No. 11, pp. 1–8, 2013.