文章埋め込み表現を用いた文章検索におけるコサイン類似度の補正

濱﨑 利之
株式会社 NDKCOM
hamasaki.toshiyuki@ndkcom.co.jp

要旨

文章埋め込み表現を用いた文章検索におけるコサイン類似度の補正について提案する．提案手法では，検索語句と検索語句を多く含む検索対象とがなす角を，適切に補正することで，そのコサイン類似度を大きくするように構成している．提案手法は，検索語句としてスペースで区切られた断片的なキーワードを入力した場合に有用である．

1. はじめに

文章を高次元空間に埋め込み，ベクトルとして扱う文章埋め込み表現は，２つの文章の意味的な類似性を定量的に扱う手法であり，検索語句として入力した文章に対して意味的な類似文章の検索が可能となる．したがって，文章埋め込み表現はオペレータ業務のように検索を行い情報を提供する業務の作業効率化に貢献すると期待される．

しかしながら，文章埋め込み表現を用いた検索では，入力した検索語句が文章として十分な意味を成さない場合には文章の類似を適切に判断できない．オペレーター業務では，聞き取った内容を断片的なキーワードで入力することもあり，この場合は検索者の期待とかけ離れた検索結果が得られることがある．文章の類似性を定量的に扱う利点を生かしながら，キーワード検索の利点を持つ検索ができれば，より良い検索結果を提供できると考える．

本稿では，文章埋め込み表現を使用した検索にて，検索語句が文章として十分な意味を成さない断片的なキーワードを入力した場合であっても，検索者が期待する結果を得るための，類似度計算の補正方法について提案する．

2. 文章埋め込み表現

文章埋め込み表現とは，文章を高次元空間に埋め込んだ意味的な情報を持つベクトル (特徴ベクトル) のことである．文章埋め込み表現を生成する手法には，Universal Sentence Encoder (USE) [1] および BERT [2] があるが，本稿では USE を使用した．USE は Google の研究者によって 2018 年に Tensorflow Hub で公開された文書埋め込み表現生成手法である．USE は多言語に対応しており，異なる言語間でも定量的に類似した文章の判断が可能である．USE により生成された文書埋め込み表現は，512 次元空間上の点として表現され，また規格化されているため，512 次元球面上の点と考えることができる．

3. コサイン類似度

2 つの文章埋め込み表現がどれだけ似ているかを定量的に判断する尺度としてはコサイン類似度が一般的である．2 つの文章埋め込み表現を $\boldsymbol{v}$, $\boldsymbol{w}$ とすると，コサイン類似度 S は

$$S = \frac{\boldsymbol{v} \cdot \boldsymbol{w}}{||\boldsymbol{v}||\,||\boldsymbol{w}||} = \cos\theta \tag{1}$$

で与えられる．ここで，最初の等式の $\cdot$ はベクトルの内積を表し，$||\cdots||$ はベクトルの大きさ (L2 ノルム) を表す．また θ は 2 つの文章埋め込み表現のなす角である．なお，USE の文章埋め込み表現は規格化されているため ($||\boldsymbol{v}||=||\boldsymbol{w}||=1$)，USE の文章埋め込み表現を用いる場合，式 (1) の分母を計算する必要は無い．コサイン類似度の値の範囲は $-1 \leq S \leq 1$ であり，2 つの文章埋め込み表現のなす角 θ が小さいほど類似度が高い (ベクトルが同じ方向であるほど類似度が高い)．

4. 提案手法

2 つの文章埋め込み表現の類似度を定量的に評価する方法として式 (1) のコサイン類似度を使用した場合，検索語句としてスペースで区切られた断片的なキーワードを入力すると，入力した検索語句を含まない文章の類似度を高く評価してしまう場合がある．例えば，「ソフトウエアシンポジウム 仙台」を検索語句とした場合，「ソフトウエアシンポジウム」や「仙台」を含まないものが高い類似度となることがある．これは，検索語句として入力したものが文章として十分な意味を成さないことにより，適切な埋め込みができないためと考えられる．そこで，式 (1) の類似度に対して下記手順で補正を行う:

1. 検索語句をスペースで分解する．分解して得られる個々の検索語句をキーワードと呼ぶ．
2. 検索語句に含まれるキーワードの数を N とする．
3. 検索対象に含まれるキーワードの数をカウントする．その数を x ($\leq N$．つまり重複は排除) とする．
4. N, x を使って次の式でコサイン類似度を補正する．

$$\alpha = 1 - 0.5 \times \frac{x}{N} \tag{2}$$

$$S' = \cos(\alpha \times \theta) \tag{3}$$

ここで，S' は補正後のコサイン類似度，θ は補正前のコサイン類似度のなす角である．

コサイン類似度では，2 つのベクトルのなす角が小さいほど類似度が大きくなり，なす角が大きいほど類似度は小さくなる．そこで，キーワードを多く含むものは，なす角を最大で 2 分の 1 にすることで，類似度を大きくするように補正している．

5. 結果

独自に収集したヘルプデスクへの問い合わせ約 4000 件に対して，検索語句に「変更契約 金額」を入力した場合の補正の有無における検索結果上位 5 件を，表 1 および 表 2 に示す．表 1 が補正無しの場合の検索結果であり，表 2 が補正有りの場合の検索結果である．

両者を比較すると，順位の 1 および 2 番目については差異は無いが，順位の 3 番目以降について異なる検索結果が得られている．補正無しの場合の順位 3 番目の内容は，「変更契約を入れたいが，入れれない.」であり，「金額」というキーワードを含んでない．対して，補正有りの場合の順位 3 番目の内容は，「変更契約画面で，今回請負金額の金額が間違っている．どうしたらいいか.」であり，検索語句のキーワードを全て含むものとなっている．このように，補正後の検索結果では，検索語句に入力したキーワードを含むものが高い類似度を示すように補正されており，検査語句を含むものを検索結果に期待する検索者にとっては，望んだ結果に近づくと考えられる．

表 1. 補正無における「変更契約 金額」での検索結果

順位	類似度	内容
1	0.7219	変更契約を伴わない金額変更についてどのように入れたらいいか.
2	0.7052	変更契約画面で金額の変更はないはずなのに，今回請負金額の箇所に違う金額が入っている.
3	0.6220	変更契約を入れたいが，入れれない.
4	0.6082	支払条件の変更は，変更契約が終わっていても行うことができるか.
5	0.6055	業者を介さない形で当初の契約額を変更したいどうしたらいいか.

表 2. 補正有における「変更契約 金額」での検索結果

順位	類似度	内容
1	0.9278	変更契約を伴わない金額変更についてどのように入れたらいいか.
2	0.9233	変更契約画面で金額の変更はないはずなのに，今回請負金額の箇所に違う金額が入っている.
3	0.8920	変更契約画面で，今回請負金額の金額が間違っている．どうしたらいいか.
4	0.8878	変更契約画面で，登録をしようとした際に，「契約額の合計と請負金額が等しくありません.」というメッセージが表示される.
5	0.8872	変更契約画面で，再見積もりにより金額が減額になったため，その金額を登録をしようとした際に，「契約額の合計と請負金額が等しくありません.」というメッセージが表示される.

6. 課題

(検索結果の妥当性検証)

本稿で示した結果では，検索者がどのような検索結果

を期待して，検索語句を選んだのかを調査していない．つまり，「変更契約 金額」を入力した検索者にとって，補正有りの検索結果がどの程度妥当であったかは示していない．今後の課題として，実際の業務に即した形で本提案の補正を施した検索を行い，検索結果がどの程度妥当であるかを検証する必要がある．

(補正の計算式)

本稿で示した式 (2) の補正では，キーワードの数に対して直線的になす角を小さくする補正を行っている．なす角を小さくする方法については，直線的な補正ではなく，例えば，tanh ようなシグモイド型の非線形関数を使用することも考えられる．また，なす角の補正は，式 (2) の通り，最大で 2 分の 1 まで小さくするよう構成しているが，この 2 分の 1 という値についても，検討の余地がある．さらに，本稿では類似度の計算にコサイン類似度を使用したが，コサイン類似度以外の類似度計算方法 [3] について同様の補正が可能であるか，およびコサイン類似度と比較して結果にどのような変化があるかは一考する価値がある．より効果的な補正の計算式については今後の課題である．

7. おわりに

本稿では，文章埋め込み表現を用いた文章検索におけるコサイン類似度の補正について提案した．提案した補正方法では，検索語句が文章として十分な意味を成さないようなスペースで区切られた断片的なキーワードを入力した場合に，入力したキーワードを含む文章のコサイン類似度を大きくするように構成した．

提案した補正方法は，類似度を用いた検索の本来の仕組みを変更するものではない．つまり，検索語句に文章を入力した場合は，当該文章と類似したものを検索でき，類似度を用いた検索の機能を損なうことは無い．提案した補正方法は，類似度を用いた検索の仕組みの中でキーワード検索のような検索結果を得ることを可能にしており，これは，キーワード検索の検索結果を期待する場合に，より良い検索結果を提供できると考えられる．

類似度を用いた検索の利点は，細かい表現の差異や同義語に依らず，意味が同じと考えらるものが検索できることである．類似度を用いた検索では，検索対象に対する理解が浅い人 (例えば業務の初心者) の知識の不足や記憶の曖昧さを補うことが可能となる．一方，キーワード検索の利点は，検索語句が明確に定まっている場合に，当該検索語句を含むものを検索できることである．キーワード検索では，検索対象に対する理解が深い人 (例えば業務の熟練者) が使用すると，求める検索結果を少ない労力で取得することが可能である．本稿の提案は，このような業務の初心者にも，業務の熟練者にも，心地よい検索体験を継ぎ目なく提供する試みでもあり，課題に記載した検索結果の妥当性について，業務の初心者および熟練者両者による検証は特に重要である．

参考文献

[1] Daniel Cer, et al.: Universal Sentence Encoder *arXiv preprint*, arXiv:1803.11175.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding *NAACL-2019*, pp. 4171–4186 (2019).

[3] 吉成未菜里, 岡本圭史: 要求文のベクトル化による要求文間の包含関係検出法の提案, 情報処理学会第 78 回全国大会講演論文集 2016 (1), pp. 427–428 (2016).