4B4-2

文構造を考慮した発話理解に基づく自然文検索

Natural Language Information Retrieval
using Query Understanding with Sentence Structure

大塚 淳史 Atsushi OTSUKA　別所 克人 Katsuji BESSHO　平野 徹 Toru HIRANO　東中 竜一郎 Ryuichiro HIGASHINAKA　浅野 久子 Hisako ASANO
松尾 義博 Yoshihiro MATSUO

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
NTT Media Intelligence Laboratories, Nippon Telegraph and Telephone Corporation

Natural language retrieval is one of the most important techniques for AI agents. In this paper, we present a novel phrase based natural language information retrieval method using sentence structures and neural networks. We create concept vectors from sentence dependency structures by Recursive Neural Networks, and compare the similarity between a user query and documents. Experimental results show the proposed method outperformed previous methods in a FAQ search task.

1. はじめに

自然文でシステムに問い合わせることで所望の情報を入手する自然文検索は，スマートフォンによる音声検索や，音声対話エージェントなど様々な場面で活用されている．近年では，コンタクトセンタにおいて，電話での問い合わせ内容（クエリ）と “よくある質問集（FAQ）” を照合し，最も関連する QA の回答を提示する FAQ 検索型のオペレータ支援技術としても注目を集めている [菊地 16]．

自然文検索では，自然文中の重要なキーワードを抽出し，その抽出したキーワードに基づいて検索を行う．しかしながら，自然文からキーワードを抽出した後は，クエリを単なるキーワードの集合（Bag-of-words）と見なして検索を行うため，クエリが本来持っていた語順や構文といった情報が欠落してしまうという問題がある．特に，FAQ 検索では “メールが送れない” と “送れないメールがある” のように使用されているキーワードは同じでありながら言い回しでニュアンスが異なる QA が混在している場合もあり，自然文で入力されるクエリの意味をより正確に理解した検索が求められる．

自然文のクエリを正確に理解した情報検索では，クエリの質問タイプに基づく検索 [永田 06] や，述語項構造解析の適用 [山田 11]，係り受け表現を用いた手法 [新里 09] など自然言語処理技術を応用した技術がこれまで多く提案されている．また，自然文であるという特徴から，文節や係り受け表現など単語よりもより広範囲のパターンを照合するフレーズ検索も提案されている [Eric 01]．近年では DeepLearning 技術を適用した手法 [Minwei 15] も提案されている．DeepLearning では単語や文の意味をベクトルによる分散表現として扱っているという特徴がある．

本論文では，自然文で入力されたクエリの意味を語順や係り受けといった文構造も含めて理解することによる自然文検索手法を提案する．クエリと検索対象文書中の単語や文節，係り受け表現の意味を全て分散表現に変換し，ベクトル間の距離に基づく類似度により検索を行う．このとき，クエリと検索対象文書の単語同士の類似度，文節同士の類似度，係り受け表現同士の類似度のように，同一粒度の表現同士で類似度を計算する．そして，粒度の異なる類似度の差により，最終的なクエリと検索対象文書との類似度を算出する．実験では FAQ 検索において，提案手法が Bag-of-words および Bag-of-phrase よりも高精度に検索を実現できることを明らかにする．

本論文の構成は以下のとおりである．まず 2. 節で関連研究について述べる．次に 3. 節で提案手法について説明し，4. 節にて評価実験を行う．最後に 5. 節でまとめを述べる．

連絡先: 大塚淳史，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，〒 239-0847，神奈川県横須賀市光の丘 1-1，otsuka.atsushi@lab.ntt.co.jp

2. 関連研究

自然文で入力されるユーザの発話や質問を理解する技術は対話システムや質問応答システムにおいて自然言語理解技術として多く研究されている．Yazdani ら [Yazdani 15] は，ごく僅かな事例から学習を行う Zero-shot learning を自然言語理解に導入し，ユーザ発話と知識ベースの照合を行う手法を提案している．杉山ら [杉山 15] は，雑談のように任意の話題に関するユーザ発話について，発話中の係り受け関係にある文節ペアを話題として抽出する手法を提案している．関根ら [関根 05] は，百科辞典を活用した質問応答技術を実現するため，ユーザの質問発話を固有表現抽出から取得する “項目” と，長さ，大きさと行った質問内容を表す “属性” のパターンとして抽出する技術を提案している．

近年，ニューラルネットワークを用いた DeepLearning 技術は，自然言語処理や情報検索分野においても様々な適用が報告されている．Zeng ら [Zeng 14] は，文中に出現する単語の意味の関係性を Deep Learning を用いて推定している．Huang ら [Huang 13] は，クリックスルーログとニューラルネットワークを用いて Web 検索のための意味ベクトルモデルを作成し，LSA などの従来の意味モデルと比較することで，DeepLearning モデルが Web 検索においても効果があることを報告している．

本論文は，ユーザの質問発話（クエリ）をニューラルネットワークによって分散表現に変換することで，クエリと文書の類似度比較による自然文検索を実現する．従来の検索での分散表現の活用とは異なり，単語だけでなく文節や係り受け表現など文の構造も分散表現のベクトル中に内包させ利用することに特徴がある．

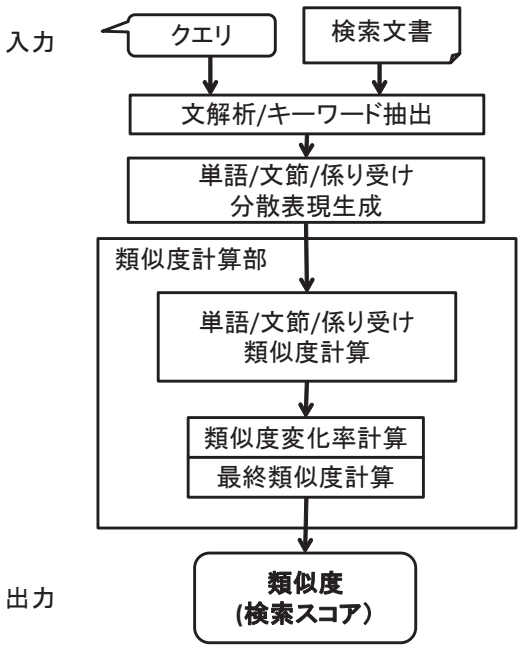

図 1: 提案手法の処理の流れ

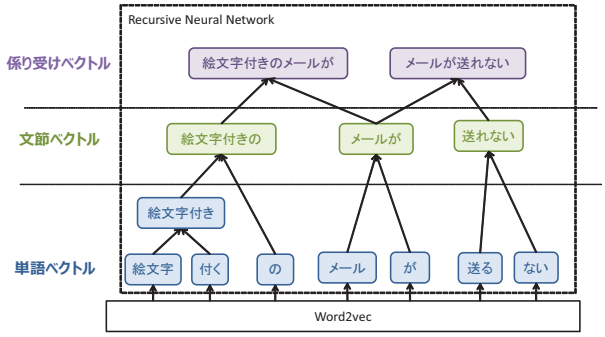

図 2: ニューラルネットワークによる，構文を考慮したベクトル合成

3. 提案手法

自然文検索のための類似度計算手法の流れを図 1 に示す．自然文のクエリと検索対象文書に対して係り受け解析を行い，単語，文節，係り受け表現を抽出する．また，単語の中で内容語や自立語といった検索のキーワードとして使用できるものを抽出する．文解析の後は，各単語，文節，係り受け表現に分散表現であるベクトルを付与し，類似度計算を行う．

以降は 3.1 節で，分散表現の生成と付与について説明し，3.2 節でベクトルの距離に基づく類似度計算手法について詳述する．

3.1 ニューラルネットワークによる分散表現の付与

ニューラルネットワーク言語モデルを用いて単語の意味をベクトルとして表現する分散表現学習の代表例として Mikolov ら [Mikolov 13] の Word2vec[*1] がある．Word2vec ではニューラルネットワークによって単語の周辺文脈を予測するモデルにより分散表現を学習する．一方で，単語だけでなく文節や係り受け表現などの単語よりの粒度の大きいフレーズ表現に関しても分散表現を付与する研究も多く取り組まれている．Socher ら [Socher 11] は，単語の意味ベクトルをボトムアップ式に合成を繰り返すことによってフレーズの分散表現を生成する Recursive Neural Network を提案している．

本論文では，単語の分散表現モデルである Word2vec と，フレーズベクトル生成技術である Recursive Neural Network を組み合わせて，単語，文節，係り受け表現の意味ベクトルを生成する．ベクトル生成モデルを図 2 に示す．まず，Word2vec により全ての単語にベクトルを付与する．Word2vec のモデルは検索対象文書に関連するドメインのコーパスから事前に学習しておく．次に Recursive Neural Network により，2 つの単語ベクトルの合成を繰り返すことにより，文節，係り受け表現のベクトルを生成する．合成ベクトル p_1 は，符号化重み行列 W_e とバイアス項 b_e，活性化関数 f からなるニューラルネットワークにより以下の通り計算できる．

$$p_1 = f(W_e[c_1; c_2] + b_e) \tag{1}$$

ここで，$[c_1; c_2]$ は合成元となるベクトル c_1 と c_2 を連結させたベクトルである．文節ベクトルは，文節内の単語を先頭から順に合成していくことで生成する．係り受け表現ベクトルは，係り受け解析によって得られた全ての係り受け関係に対して，文節ベクトル同士を合成することによって生成する．Recursive Neural Network では，文構造に基きベクトルを合成していく．ベクトルを合成して文ベクトルを作成する方法には LSTM[Sutskever 14] や Sentence2vec[Le 14] があるが，LSTM は文の先頭から順に合成を繰り返し，Sentence2vec は Word2vec モデルを生成する過程で同時に文ベクトルを学習してくという違いがあり，文中の構文に沿った部分表現のベクトルを抽出することは難しい．

3.2 ベクトル間の距離に基づく類似度計算

ニューラルネットワークにより，クエリと検索対象文書それぞれで単語，文節，係り受け表現の意味ベクトルを生成した．ベクトル間の類似度はコサインによって算出する．

類似度計算の流れを図 3 に示す．クエリ中の各キーワードに対して，単語，文節，係り受け表現の単位で類似度を計算していき，最後に単語と文節，文節と係り受け表現という表現の間の類似度の差（類似度変化率）を計算することでクエリと検索対象文書間の類似度を計算する．

提案手法の類似度計算の特徴は，クエリと検索対象文書の単語ベクトル同士，文節ベクトル同士，係り受け表現ベクトル同士で類似度計算を行う点にある．最初に，クエリ中のキーワード w_q と検索対象文書の全キーワードで類似度計算を行い，w_q と最も類似度が高いキーワード w_r を抽出する．次に，クエリからキーワード w_q を含む文節 s_q と，検索対象文書中のキーワード w_r を含む文節 s_r を抽出し，s_q と s_r の類似度を計算する．最後に，文節 s_q を含む係り受け表現 d_q をクエリから，s_r を含む係り受け表現 d_r を検索対象文書から抽出し係り受け表現同士の類似度を計算する．

単語同士，文節同士，係り受け表現同士の類似度を算出した後，類似度変化率を計算する．類似度変化率とは，単語から文節，文節から係り受け表現のように，比較する意味の粒度を変えたとき，類似度がどの程度変化するかということを示す尺度である．単語同士での類似度が高いキーワードであっても，文節同士の類似度が低い場合は文節単位まで見るとクエリと検索対象文書はあまり似ておらず，検索における重要度は低くなる．一方で，単語単位から係り受け表現まで一貫して，類似度が高いキーワードに関しては，検索時により重要なキーワード

*1 https://code.google.com/archive/p/word2vec/

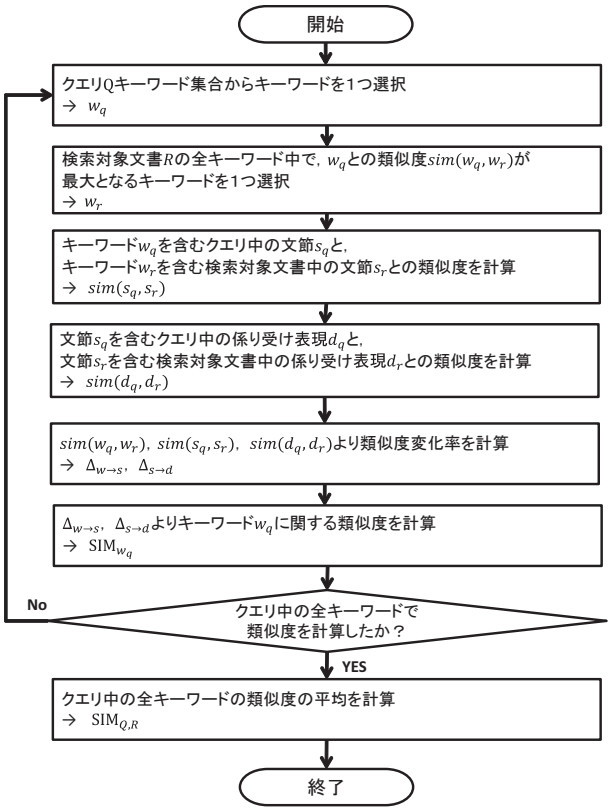

図 3: 類似度計算の流れ

として扱うことができる．類似度変化率は単語と文節に関する類似度変化率 $\Delta_{w\to s}$，文節と係り受け表現に関する類似度変化率 $\Delta_{s\to d}$ を計算する．類似度変化率は次式で計算する．

$$\underset{w\to s}{\Delta} = sim(s_q,s_r) - sim(w_q,w_r) \quad (2)$$

$$\underset{s\to d}{\Delta} = sim(d_q,d_r) - sim(s_q,s_r) \quad (3)$$

ここで，$sim(w_q,w_r)$ はクエリ Q と検索対象文書 R の単語の類似度，$sim(s_q,s_r)$ は文節同士の類似度，$sim(d_q,d_r)$ は係り受け表現同士の類似度である．

類似度変化率により，最終的な類似度を計算する．ここで，類似度変化率によって計算される類似度は，クエリ Q 中の 1 つキーワードについての類似度である．クエリと検索対象文書との類似度は，クエリの全キーワードの類似度の平均をとったものとなる．クエリ Q 中のキーワード w_q と検索対象文書 R との類似度 $Sim(w,R)$，そしてクエリ Q と検索対象文書 R の類似度 $SIM(Q,R)$ は以下のとおりで計算できる．

$$Sim(w_q,R)=\frac{sim(w_q,w_r)}{3}\{1+(1-|\underset{w\to s}{\Delta}|)+(1-|\underset{w\to s}{\Delta}|)(1-|\underset{s\to d}{\Delta}|)\} \quad (4)$$

$$SIM(Q,R) = \frac{1}{|Q|}\sum_{w_q\in Q} Sim(w_q,R) \quad (5)$$

ここで，$|Q|$ はクエリ Q 中のキーワード数である．検索の際には，クエリ Q について，全検索対象文書に対しての類似度を計算し，類似度を降順に並べた結果を検索結果として出力する．

表 1: FAQ 検索タスクでの精度評価

	Bow	Bop	提案手法
MAP	0.509	0.502	**0.526**
MRR	0.643	0.680	**0.682**
precision@1	0.580	**0.620**	**0.620**
precision@2	0.390	0.390	**0.415**
precision@3	0.297	0.307	**0.320**
precision@4	0.245	0.255	**0.270**
precision@5	0.212	0.214	**0.228**

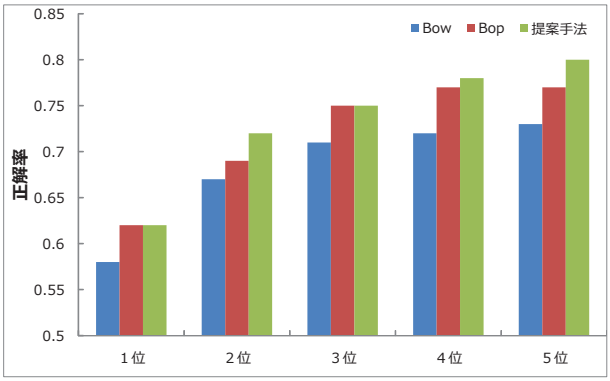

図 4: FAQ 検索タスクでの正解率比較

4. 評価実験

本節では，提案手法の有効性を示すための評価実験について述べる．まず，4.1 節で，実験設定について説明し，4.2 節で実験結果と考察について詳述する．

4.1 実験設定

提案手法の有効性を検証するため，自然文による FAQ 検索タスクでの評価実験を行った．検索対象の FAQ 文書は，Web 上のインターネットのサポートに関する FAQ から取得した 578 の QA セットを用いる．クエリについては人手で作成した 100 の自然文質問を用いる．各クエリには 1 つ以上の正解 QA が付与されている．

提案手法では，単語ベクトルを作成するための Word2vec モデル，そして，Recursive Neural Network のためのパラメータを学習しておく必要がある．Word2vec のモデルに関しては，Web の質問回答サービス（CQA）のインターネットカテゴリから収集した 116,763 の QA セットにより学習した．Recursive Neural Network のパラメータは約 10 万の自然文発話データから学習した．

実験の比較手法として，単語の出現頻度のベクトルを用いた Bag-of-words（Bow）．単語に加えて，文節や係り受け表現の出現頻度ベクトルを用いた Bag-of-phrase(Bop) による FAQ 検索を行う．Bow，Bop での検索では単語，文節，係り受け表現には BM25 によって計算された重みスコアを用いる．

検索精度の評価尺度には，検索結果上位 5 件での平均適合率（MAP），最初の正解が出現したときの平均逆順位（MRR）．precision@1〜5，そして上位 N 件に正解が 1 つでも含まれている質問の割合である正解率を用いる．

4.2 実験結果・考察

実験結果の MAP，MRR，precision@1〜5 を表 1 に示す．MAP では提案手法が比較手法である Bow，Bop を上回ってい

ることがわかる．このことより，提案手法は比較手法よりも多くの正解を高精度で検索できているといえる．一方で，MRRにおいては，提案手法と Bop との差は 0.002 ポイントとなっている．Bop においてはクエリと FAQ 文書の文節，係り受け表現が一致するのかを評価することにより検索される．クエリとの文節や係り受け表現が完全に一致している QA はクエリに関係している可能性が高いということになる．実際，precision@1 の Bop のスコアは提案手法と同値となっている．しかしながら，Bop の場合はクエリと QA の文節，係り受け関係が文字列として完全一致している必要がある．そのため助詞など僅かな違いに対してでも，クエリと検索対象文書が一致しなくなるという問題がある．そのため，Bop は precision@1 は高いスコアとなっているが，precision@2〜5 に関しては Bow とほぼ同じスコアとなってしまっている．一方で提案手法は，precision@2〜5 においても，比較手法よりも高いスコアとなっている．これは，文節や係り受け表現の意味をベクトルとして表現しているため，文字列の完全一致ではなく意味の比較が行えているからであると考えられる．

1 位から 5 位までの正解率の結果を図 4 に示す．正解率とは，上位 N 位までに 1 つ以上正解 QA が含まれていたクエリの割合である．1 位，3 位の正解率は Bop と提案手法は同値であるが，上位 5 位では提案手法が最も高い正解率となった．5 位までの検索結果に対して正解，不正解の McNemar 検定を実施した．その結果，Bow と Bop に関しは $p <= 0.523$ という結果になり，有意差は存在しなかった．一方，Bow と提案手法では $p <= 0.0156$ となり，5% の有意水準において，有意差が存在しているといえる．

評価実験により，本論文で提案した検索手法は自然文検索において，従来よりも高精度に関連する文書を検索できることが明らかとなった．特に，クエリに対して，文字列として見れば異なるような文書に対しても，ベクトルによる意味の比較を行うことで，上位の関連文書として検索することが可能になったといえる．提案手法で不正解となった例としては，“動画が見づらい” というクエリに対して，“回線速度が遅い時の対処法” という文書を検索するようなタスクがある．これは “動画が見づらい” ことが “回線速度が遅い” から発生するという関連付けを必要とする検索である．このような検索に対しては提案手法であるベクトルの類似度による検索手法では対応は難しく今後の課題と考えている．

5. おわりに

本論文では，文節や係り受け表現など，文の構造を含めてクエリの意味を理解することによる情報検索手法を提案した．ニューラルネットワークによる単語やフレーズの意味ベクトル作成技術を用いてた類似度計算手法により自然文クエリに対しても，クエリの意味を的確に捉えて，関連する文書を検索することができる．実験では，自然文 FAQ 検索タスクにおいて，単語やフレーズの一致による従来の検索手法よりも高精度に検索できることを明らかにした．

今後の課題は，より多様な自然文クエリに対する検索精度の向上が挙げられうる．クエリと文書間の因果関係など一種の推論を必要とするような関連文書に対しても，クエリの意味を正しく解釈することで，検索可能になることを目指したいと考えている．

参考文献

[Eric 01] Eric, B., Jimmy, L., Michele, B., Dumais, S., and Ng, A.: Data-Intensive Question Answering, *Proc of Text REtrieval Conference (TREC 2001)* (2001)

[Huang 13] Huang, P.-S., He, X., Gao, J., Deng, L., Acero, A., and Heck, L.: Learning Deep Structured Semantic Models for Web Search Using Clickthrough Data, *Proc of the 22nd ACM International Conference on Information & Knowledge Management (CIKM2013)* (2013)

[Le 14] Le, Q. and Mikolov, T.: Distributed Representations of Sentences and Documents, *Proc of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1188–1196 (2014)

[Mikolov 13] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient Estimation of Word Representations in Vector Space, *CoRR*, Vol. abs/1301.3781, (2013)

[Minwei 15] Minwei, F., Bing, X., Michael, G., Lidan, W., and Bowen, Z.: Applying Deep Learning to Answer Selection: A Study and An Open Task, *Proc of The 2015 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015)* (2015)

[Socher 11] Socher, R., Huang, E. H., Pennin, J., Manning, C. D., and Ng, A. Y.: Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection, *Proc of Advances in Neural Information Processing Systems (NIPS 2011)*, pp. 801–809 (2011)

[Sutskever 14] Sutskever, I., Vinyals, O., and Le, Q. V.: Sequence to Sequence Learning with Neural Networks., *CoRR*, Vol. abs/1409.3215, (2014)

[Yazdani 15] Yazdani, M. and Henderson, J.: A Model of Zero-Shot Learning of Spoken Language Understanding, in *Proc of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP2015)*, pp. 244–249 (2015)

[Zeng 14] Zeng, D., Liu, K., Lai, S., Zhou, G., and Zhao, J.: Relation Classification via Convolutional Deep Neural Network, *Proc of the 25th International Conference on Computational Linguistics(COLING 2014)*, pp. 2335–2344 (2014)

[永田 06] 永田 昌明, 齋藤 邦子, 松尾 義博：日本語自然文検索システム Web Answers, 言語処理学会第 12 回全国大会論文集, pp. 320–323 (2006)

[関根 05] 関根 聡, 須藤 清, 安藤 まや：属性値の自動抽出と質問文パターンを使った百科事典質問応答システム, 言語処理学会第 11 回全国大会論文集 (2005)

[菊地 16] 菊地 勝由（編）: BUSINESS COMMUNICATION 2016 年 2 月号, ビジネスコミュニケーション社 (2016)

[山田 11] 山田 浩之, ジェプカ ラファウ, 荒木 健治：述語項構造解析による検索対象細分化を用いた自然文検索, 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション, pp. 59–64 (2011)

[新里 09] 新里 圭司, 黒橋 禎夫：クエリの語句の重要度と係り受けを考慮した自然文検索, 情報処理学会研究報告情報学基礎（FI）, pp. 113–120 (2009)

[杉山 15] 杉山 弘晃, 目黒 豊美, 東中 竜一郎, 南 泰浩：任意の話題を持つユーザ発話に対する係り受けと用例を利用した応答文の生成, 人工知能学会論文誌, pp. 183–194 (2015)