分散ベクトルに基づく文書のアライメント
—AKB48の歌詞の類似性解析—

竹中要一 [1,2]
[1] 関西大学　総合情報学部　　[2] 大阪大学 大学院医学系研究科
takenaka@kansai-u.ac.jp

1　はじめに

与えられた２つの文の類似性を評価する多くの手法が存在する。文字や単語を集合とみなし、Jaccard 係数や Dice 係数のような集合の類似度で評価する手法、あるいは文字や単語を単位とする編集距離や最長共通部分文字列、最長共通部分列で定義する方法である。これらの方法は文字や単語の一致、不一致に基づき評価するため、単語の有する意味を考慮することができない。

単語の意味を表現する方法としては、word2vec に代表される分散表現が挙げられる [1]。この分散表現を用いて２文の類似性を評価する方法としては、Average Alignment, Maximum Alignment, Hungarian Alignment, Word Mover's Distance などがある [2]。また、近年では文を一つのベクトルとして表現し、ベクトルの類似度で表現する方法も提案されている [3]. しかし、これらの方法では文の類似性を評価する事はできるが、類似している部位を同定する事はできない。たとえば、ある一文節だけが類似しており他が異なる２文において、類似する一文節を抜き出すという用途に用いることはできない。

本研究では、単語分散表現と最長共通部分列に基づき単語分散表現を用いる事により、２文間に類似する部分を抽出する方法を提示する。その有効性を歌詞の類似性解析より明らかにする。

2　文書の局所アライメント

与えられた２文の類似した領域を特定できるように並べる事をシーケンス・アライメント（以下アライメント）と呼ぶ。このアライメントは生物学において、塩基を単位とする文である遺伝子、アミノ酸を単位とする文であるタンパク質の解析に適用され、進化や機能の解析に用いる基礎技術となっている。この手法は文全体の類似性を評価する大域アライメントと、文のうち特に類似した領域を特定する事に特化した局所アライメントに大別される。両アライメントは最長共通部分列に基づいたアルゴリズムで計算される。大域アライメント、すなわち遺伝子の核酸配列やタンパク質のアミノ酸配列への最長共通部分列の応用は 1970 年に S.B. Needleman と C.D. Wunsch によって提案されている [4]。そして類似部分を抽出する局所アライメントは 1981 年、T. Smith と M. Waterman が提案している [5]。

アライメントが提案されて以降、長い年月を経ても自然言語へと応用されていなかったのは、単語間の類似性評価が困難であったからである。核酸配列やアミノ酸配列の文字数はそれぞれ 4 個、20 個であり、化学・物理学・生物学的な特性に基づく類似性評価尺度を総当たり的に決める事が可能である。一方、自然言語の場合は類似性の評価単位は単語となるため、10 万を超えるような単語間の類似性評価を網羅的に行う事が従来困難であった。しかし、分散表現の出現が網羅的な類似性評価を可能とした。本研究は分散表現で単語の類似性を評価する事で局所アライメントが可能になる事を示す。

表１に自然言語のアライメント例を示す。文１と文２の単語が文頭から文末まで一対一で整列している。また各単語は分散表現を有するため、相関係数のような類似性評価が可能である。もし類似性評価を行わない場合、「柴犬」と「猫」は異なるという評価しかされない。「柴犬」からみれば「猫」も「広場」も「いる」も異なるという同じ評価になり、意味を斟酌することができない。

２文が冒頭から末尾まで一対一で整列する場合は稀であり、表層的には文の長短、細かくは修飾節の有無などといった差異が存在する。表２は、表１の例に修飾句が加わった文のアライメント例である。相互の文に対応する修飾句がないため空欄となっている。この

表 1: 2 文のアライメント（ギャップなし）

文 1	文 2	同一性	相関係数
柴犬	猫	×	0.716
が	が	○	1
草原	広場	×	0.437
を	を	○	1
走っ	歩い	×	0.786
て	て	○	1
いく	いる	×	0.627
評価		○ 3 個	平均 0.795

空欄をギャップと呼ぶ。

3 アルゴリズム

2 つの文書を A, B とする。A,B はそれぞれ単語列 $A = (a_1, a_2, \cdots, a_n),\ B = (b_1, b_2, \cdots, b_m)$ で表される。ただし、$a_i, b_j (1 \leq i \leq n), (2 \leq j \leq m)$ は、全て分散表現を有する単語とする。単語 a, b 間の類似度を $\phi(a, b)$ と表す。

アライメントで対応する単語がない事をギャップと呼び、記号 " $-$ " で表す。アライメントにおいてギャップが発生した時のペナルティ値を gap とする。

この時、アライメントを計算する動的計画法のアルゴリズムは次のようになる。

1. 表の作成 大きさ $n{+}1 \times m{+}1$ の表 T, D を作成する。添字の範囲 $T[i,j], D[i,j]$ は $0 \leq i \leq n, 0 \leq j \leq m$ とする。

2. 表の初期化 $T[0,j], T[i,0], D[0,j], D[i,0]$ を 0 で初期化する。ただし、$0 \leq i \leq n, 0 \leq j \leq m$ とする。

3. ます目の計算 次式に従い表 T, D の左上 [1,1] より順番にます目 $[i,j]$ の値を計算する。

$$T[i,j] \leftarrow \max \begin{cases} 0 & \text{停止} \\ T[i-1,j-1] + \phi(a_i, b_j) & \text{対角} \\ T[i,j-1] - gap & \text{水平} \\ T[i-1,j] - gap & \text{鉛直} \end{cases}$$

$D[i,j] \leftarrow \{$ 停止, 対角, 水平, 鉛直 $\}$ のうち、上式で選ばれた項右側の文字列

表 2: 対応しない文節がある 2 文のアライメント

文 1	文 2	相関係数
小さく		
て		
可愛い		
柴犬	猫	0.716
が	が	1
	町	
	の	
	大きな	
草原	広場	0.488
を	を	1
走っ	歩い	0.786
て	て	1
いく	いる	0.627

4. 最大値の検出 表 T の最大値のます目 $[i_{max}, j_{max}]$ を探す。

5. 足跡をたどる $D[i,j]$ の値に従い、次表の動作を繰り返し実行する。

なお初期値は $[i,j] \leftarrow [i_{max}, j_{max}]$ とし、$Align$ はアライメントを格納するリスト、演算子 $+$ はリストの前方に要素を追加する演算とする。

$D[i,j]$	動作
停止	終了
対角	$Align \leftarrow (a_i, b_j) + Align$ $(i,j) \leftarrow (i-1, j-1)$
水平	$Align \leftarrow (-, b_j) + Align$ $(i,j) \leftarrow (i, j-1)$
鉛直	$Align \leftarrow (a_i, -) + Align$ $(i,j) \leftarrow (i-1, j)$

4 実験

単語の分散表現に基づく局所アライメントが類似文書検索に有効であり、かつ単語を単位とした詳細な解析に有用である事を明らかにするため、日本歌謡曲の歌詞に対して適用し、その結果を提示する。

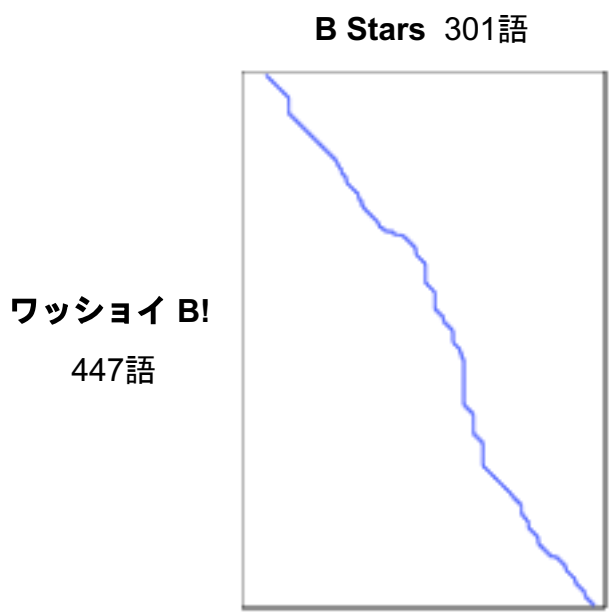

図 1: スコア最大の組合せ楽曲のアライメント足跡

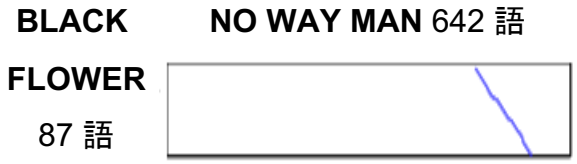

図 2: 局所スコア – 大域スコア最大の組合せ楽曲のアライメント足跡

4.1 実験条件

本研究では、歌手が”AKB48”として登録されている 371 曲を解析対象とした。単語の分散表現には国立国語研究所の形態論情報付きの表 [6] を用いた。これは延べ 1,738,455 単語に 200 次元の分散表現を付与している。形態素解析には MeCab [7] を用いた。形態素解析に用いる辞書は、Unidic-mecab の version 2.1.2 を用いた。最新の辞書ではなく古い Version 2.1.2 を用いた理由は、国立国語研究所が配布する分散表現を計算する際に用いた辞書と同じ条件で形態素解析を行うためである。単語間の距離はコサイン類似度を用いた。

371 曲の全組み合わせ $371 \times 370/2 = 68{,}635$ 組に対してアライメントを行った。また、アフィンギャップペナルティ $g(L) = 0.5 + 0.1(L-1)$ を用いた。ここで、L はギャップ長である。

4.2 実験結果

局所アライメントスコアの上位５組を表３に、上位５０組をの図３に記す。図３は曲が頂点、曲組を辺で表している。図より３つのクラスタの存在、多くの楽曲と類似する曲の存在（小池）が観測される。

全楽曲の組合せ中、スコアが２番目に高かった組合せのアライメント・マトリックスを図１に示す。なお、最上位は同一歌詞の楽曲である。図２は、大域アライメントとのスコア差が最大であった組合せである。図中の線は、歌詞中における単語の対応関係を表す。

図１は対応関係が楽曲全体に及ぶ事を、図２より曲名 Back Flower の対応関係が No Way Man の一部に集中していることを示している。図２のように類似性が一部分にとどまる場合、文の類似性をスカラーで表現する手法で類似性解析を行う事は困難である。

5 まとめ

本研究では、生物学の基礎的な解析技術である局所アライメントを自然言語処理に拡張する事を提案し、その有効性を明らかにした。自然言語処理の局所アライメント解析は、従来手法でも行われてきたクラスタ分析に加え、文の類似性部分の抽出が可能である。類似部分を抽出可能である点を活用する事により、本研究の手法が情報検索、記述式問題やレポートの採点支援等にも効果的であると考えている。

参考文献

[1] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, Vol. abs/1301.3781, , 2013.

[2] Tomoyuki Kajiwara and Mamoru Komachi. Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 1147–1158, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

表 3: 局所アライメントスコア上位 5 組

歌 1	歌 2	局所	大域	局所 – 大域
桜の花びらたち	桜の花びらたち 2008	306.6	306.6	0.0
B Stars	ワッショイ B!	117.3	113.8	3.5
アイドルの夜明け	B Stars	111.2	105.1	6.0
アイドルの夜明け	ワッショイ B!	98.6	94.1	4.5
小池	16 人姉妹の歌	83.9	82.9	1.0

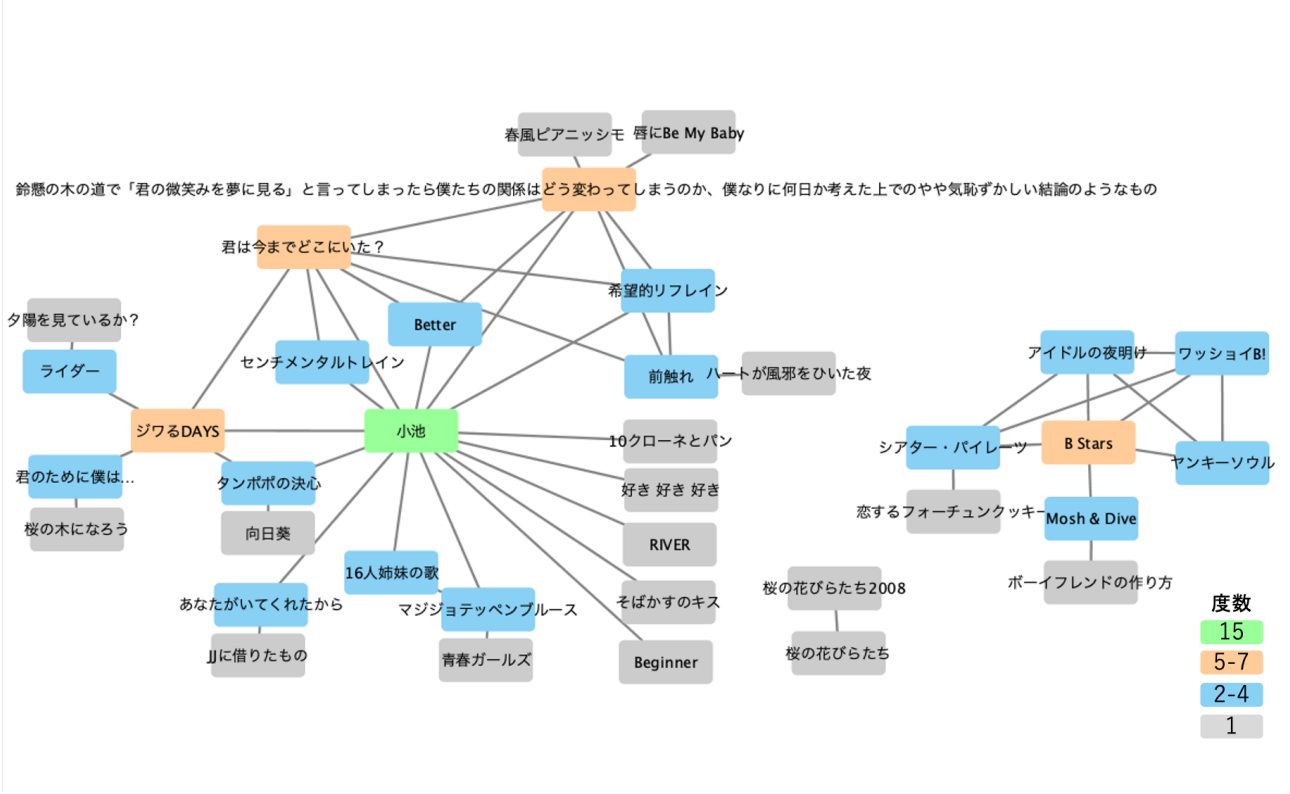

図 3: 局所アライメントのスコア上位 50 位の曲の組合せ

[3] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. *CoRR*, Vol. abs/1405.4053, , 2014.

[4] S.B. Needleman and C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, Vol. 48, pp. 443–453, 1970.

[5] T. Smith and M. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, Vol. 147, pp. 195–197, 1981.

[6] Masayuki Asahara. NWJC2Vec: Word embedding dataset from 'NINJAL Web Japanese Corpus'. *Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication*, Vol. 24, No. 2, pp. 7–25, Feb. 2018.

[7] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Vol. (EMNLP-2004), pp. 230–237, 2004.