単語分散表現に基づく単一言語内フレーズアライメント手法

吉仲 真人　　梶原 智之　　荒瀬 由紀
大阪大学
{yoshinaka.masato, arase}@ist.osaka-u.ac.jp, kajiwara@ids.osaka-u.ac.jp

1 はじめに

単一言語内フレーズアライメントは自然言語理解における基礎タスクである．本タスクは，同一言語の文対に含まれる意味的に対応するフレーズについて，それらのアライメントを取ることを目的とする．単一言語内フレーズアライメントの応用は多岐に渡るが，特に関連の深いタスクは言い換え認識および含意関係認識や意味的類似度推定である．

先行研究では，大規模な辞書あるいは高品質な構文解析器やチャンカーが利用できることが前提となっており，英語以外の言語への適用に関して大きな制約が存在する．既存手法である Jacana-phrase [1] や SemAligner [2] は WordNet [3] や大規模な言い換え知識である PPDB [4] を使用して素性を抽出している．また，Arase and Tsujii [5] や Ouyang and McKeown [6] は高品質な構文解析器あるいはチャンカーを用いてフレーズの構造を獲得している．

一方，統計的機械翻訳の分野で研究されてきた対訳フレーズアライメント手法の多くは，パラレルコーパスにのみ依存している．対訳フレーズアライメントの一般的なアプローチでは，最初に単語アライメントを獲得し，ヒューリスティクスに基づいてそれらをフレーズ対へ拡張する．ただし，対訳フレーズアライメントの目的は大量の対訳フレーズ対を獲得することであり，ある文対において尤もらしいフレーズアライメントを決定することではない．また，大規模な単一言語のパラレル（言い換え）コーパスが充実している言語は限定されているため，単一言語内フレーズアライメントへの適用は難しい．

本研究では，対訳フレーズアライメント手法の利点を活用したシンプルな単一言語内フレーズアライメント手法を提案する．提案手法では，まず訓練済みの単語分散表現を用いて単語アライメントを獲得し，対訳フレーズアライメント手法に基づき単語アライメントをフレーズ対へ拡張する．獲得したフレーズアライメントの候補の中から，入力文対において尤もらしいフレーズアライメントを探索する．

英語のフレーズアライメントのベンチマーク [7] を用いた実験により，提案手法は既存手法 [1, 6] を大幅に上回る最高性能を達成しており，高精度なフレーズアライメント手法を実現した．既存手法とは異なり，提案手法が言語資源として必要とするのは単語分散表現を訓練するための生コーパスのみである．そのためあらゆる言語に容易に適用可能である．

提案手法は単一言語内フレーズアライメントツール SAPPHIRE として一般公開している．[1]

2 提案手法

本節では，最初にフレーズアライメント問題を定義した後，提案手法の詳細について説明する．

2.1 問題定義

入力される文対をそれぞれ $|X|$，$|Y|$ 個の単語からなる文 $X = x_0, \ldots, x_{|X|}$，$Y = y_0, \ldots, y_{|Y|}$ とする．また，文 X 中の p 番目から q 番目までの単語からなるフレーズを $x_p^q = x_p, \ldots, x_q$，文 Y 中の r 番目から s 番目までの単語からなるフレーズを $y_r^s = y_r, \ldots, y_s$ とする．このとき，矛盾のないフレーズアライメントの集合を $A = \{a_k = (x_p^q, y_r^s) | x_p^q \in X, y_r^s \in Y\}$ とする．ただし，アライメントを持たないフレーズも存在し得るため，x_p^q，y_r^s は空 ($\emptyset$) となり得る．

本研究では，フレーズアライメント間において単語の重複が存在しない一対一のフレーズアライメントの組を矛盾のないアライメント集合と定義する．つまり A は $\exists a_k = (x_p^q, y_r^s) \in A$ および $\exists a_l = (x_f^g, y_m^n) \in A$ $(k \neq l)$ について，文 X 中で p から q までのスパンと f から g までのスパンは重ならず，かつ文 Y 中でも r

[1] https://github.com/mybon13/sapphire

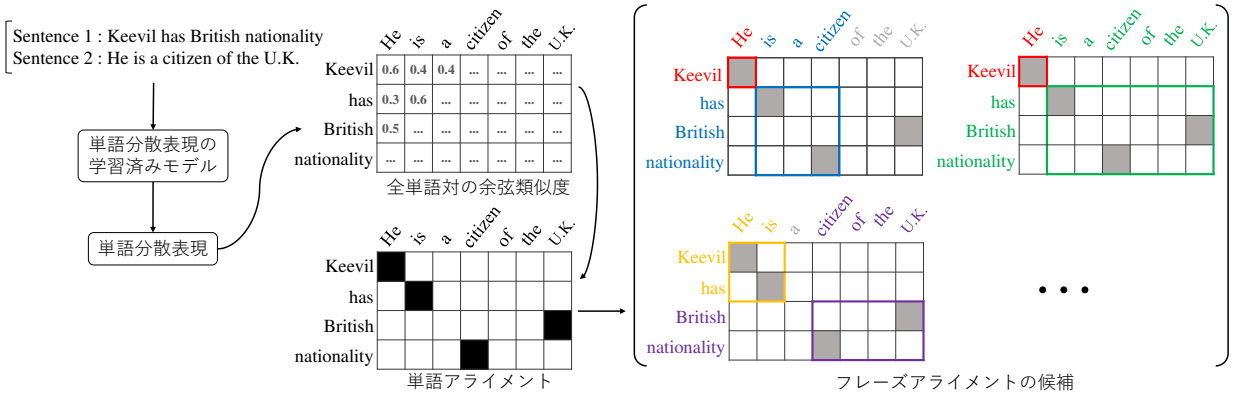

図 1: 提案手法におけるフレーズアライメント候補の抽出

から s までのスパンと m から n までのスパンは重ならない，という制約を満たす．

1 つの文対の中には矛盾のないフレーズアライメントの集合が複数存在し得る．その数は文長について指数関数的に大きくなるため，矛盾のない最適なフレーズアライメントの集合を決定するのは計算コストが非常に高い．そこで，本手法ではパラフレーズが持つ特徴を考慮した探索方法を用いて，最適なフレーズアライメントの近似解を求める．

2.2 提案手法の概要

提案手法では，まず図 1に示す通り，入力文対に含まれる全ての単語分散表現間の余弦類似度に基づいて単語アライメントを得る．そして対訳フレーズアライメント手法で用いられるヒューリスティクスで単語アライメントをフレーズ対に拡張する．最後に，獲得したフレーズアライメントの候補の中から矛盾のないフレーズアライメントを探索する．

2.3 単語アライメント

本手法では，統計的機械翻訳の分野で培われてきたヒューリスティクスである grow-diag-final[2] を用いて，単語分散表現間の余弦類似度から単語アライメントの候補を得る．獲得した候補の中から一定以上の余弦類似度を持つ候補のみを最終的な単語アライメントとして採用する．

具体的には，まず単語 x_i をそれと最も高い余弦類似度を持つ単語 y_j に対応付ける．

$$(x_i, y_j) = \arg\max_k \cos(\mathbf{e}_{x_i}, \mathbf{e}_{y_k}), \tag{1}$$

ここで，$\mathbf{e}_{x_i}$，$\mathbf{e}_{y_j}$ はそれぞれ単語 x_i，y_j の分散表現を表す．同様にして y_j を x_l と対応付ける．

$$(y_j, x_l) = \arg\max_k \cos(\mathbf{e}_{y_j}, \mathbf{e}_{x_k}). \tag{2}$$

両方向からのアライメントの積集合を単語アライメントの最初の集合とする．次に，grow-diag-final のヒューリスティクスにより，アライメントを行列とみなして以下の条件を満たすアライメントを両方向のアライメントの和集合から追加する．

- 縦・横・斜めのいずれかの方向に積集合のアライメントが存在する
- 積集合のアライメントにおいてアラインされていない単語を持つ

得られた単語アライメントについて，余弦類似度が λ 以上のものを最終的な単語アライメントとする．

2.4 フレーズ対の抽出

対訳フレーズアライメントのヒューリスティクスに基づいて単語アライメントをフレーズ対へ拡張する．最初に単語アライメントの任意の対を含むようなフレーズ対を作り，それに隣接する単語アライメントが存在する場合，そのアライメントを含むように拡大させていく．ここで，フレーズとは 1 単語から文全体までの単語 n-gram とする．[3]

[2] ハンガリアン法を用いた実験も実施したが，grow-diag-final 法の方が高い性能を示した．

[3] ただし，この段階でフレーズに文法的な制約を課すことで，文法的なフレーズアライメントとなるように容易に変更できる．

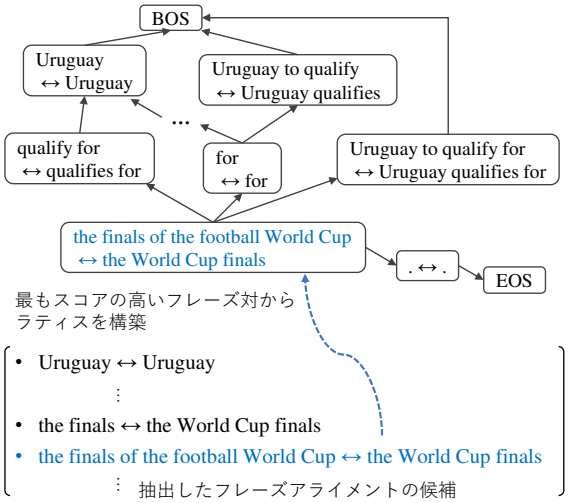

図 2: フレーズアライメント候補からのラティス構築

次に，獲得したフレーズ対のアライメントの尤度を表すスコアを求める．スコアはフレーズの分散表現に基づいて計算し，閾値 δ 以上のスコアを持つフレーズ対のみをアライメントの候補として保持する．本研究では最も簡単なフレーズの分散表現として，フレーズ内の単語分散表現の平均を用い，スコアはフレーズの分散表現間の余弦類似度を用いて計算するものとする．フレーズ長が長くなるほどスコアが低くなってしまう問題を緩和するために，余弦類似度に加えてフレーズ長を考慮するように，フレーズ対 (x, y) のスコアを以下のように設計する．

$$\mathrm{score}(x, y) = \cos(\mathbf{f}_x, \mathbf{f}_y) - \alpha \cdot \frac{1}{|x| + |y|}, \qquad (3)$$

ここで，$\mathbf{f}_x$，$\mathbf{f}_y$ はそれぞれ x，y 中の単語の分散表現を平均したフレーズの分散表現，$|\cdot|$ はフレーズ長を計算する関数，α はフレーズ長のバイアスの重みを制御するハイパーパラメータである．

2.5 矛盾のないアライメント集合の探索

獲得したフレーズアライメント候補の組み合わせの中から矛盾のないフレーズアライメントを探索する．2.1節で議論したように，最適解の探索は計算量の観点から現実的ではないため，矛盾のないフレーズアライメント集合のうち高いアライメント尤度をもつ集合を効率的に探索し，近似解を求める．

フレーズアライメントにおいては，意味的に等価なフレーズが文対内で出現する位置は自由度が高く，また文頭にアライメント尤度の高いフレーズ対が出現するとも限らない．そのため，フレーズアライメントは文の先頭から順に決定するのが最適とは考えにくい．

そこで本研究では，2.4節で得たアライメントスコアが最大のフレーズアライメントから探索をスタートし，矛盾がなくかつ最もスコアの平均が大きいアライメント集合を決定するように探索アルゴリズムを設計する．図 2に示すように，スコアが最大のフレーズ対をラティス構築の開始点として，矛盾のないフレーズアライメント候補を前後両方向に追加していく．その後，最もスコアの平均の高い経路を最終的なフレーズアライメントとして出力する．

予備実験により，この方法は一般的に使われる文を左から右へ捜査する方法よりも高い性能を示すことを確認している．

3 実験

3.1 データセットと評価指標

評価には MSR RTE corpus [7] を用いる．MSR RTE corpus は含意関係認識の評価のために構築されたもので，2 つの入力文について人手による単語アライメントが付与されている．単語アライメントはアノテータの確信度に応じて，Sure または Possible ラベルが付与されている．開発セットとテストセットのそれぞれ 800 文対から構成され，単語アライメントをフレーズアライメントに変換することでフレーズアライメントの評価に用いられてきた [1, 6].

Yao ら [1] は Sure ラベルが付与された単語アライメントのみを用いて，連続する単語アライメントから疑似的にフレーズアライメントを構成している．具体的には，データセットを OpenNLP chunker[4] を用いてチャンキングし，任意のチャンク対のうち含まれる単語全てが一対一にアラインされているものをフレーズアライメントとする．Yao らの正解セットに含まれるフレーズアライメントの割合は約 23% である．

より長いフレーズに対する評価を行うために，Ouyang and McKeown [6] は Possible アライメントを含めてフレーズアライメントを構成した．Possible アライメントを含めると，1 単語が複数単語とアラインされるため，チャンキングを行わずともフレーズアライメントを得ることができる．この評価設定では，少なくとも 1 つ以上の Possible アライメントを含む文対

[4] https://opennlp.apache.org

手法	P%	R%	F_1%	E%
Jacana-token [9]	**92.9**	66.1	77.2	13.5
Jacana-phrase [1]	83.5	77.0	**80.1**	14.3
提案手法	50.5	71.7	59.3	**33.6**

表 1: Yao ら [1] の正解セットでの実験結果

のみに評価対象を限定し，開発セットとテストセットからそれぞれ 487，441 文対のみを用いて評価を行う．

これら先行研究では，正解のフレーズアライメント中のすべての単語間にアライメントが存在するものとし，単語アライメントの適合率（P），再現率（R），F 値（F_1）を評価指標としている．さらに，フレーズアライメントの直接的な評価として，フレーズアライメントの完全一致の割合（E）も評価する．[5]

提案手法の実装においては，単語分散表現として訓練済みの fastText [8] [6] を用いる．またハイパーパラメータの値は開発セットでの F 値を最大化するようにグリッドサーチにより決定する．

3.2 結果

表 1および表 2にテストセットでの評価の結果を示す．比較手法の評価値は [1, 6] で報告されているものを使用している．

表 1では，提案手法はフレーズアライメントの完全一致率で既存手法の 2 倍以上の値を達成している．Yao ら [1] の正解セットでは約 77%が単語アライメントであるため，手法がフレーズアライメントを出力すればする程，適合率は下がってしまう．比較手法では適合率が高く，フレーズの完全一致率が低いことを考慮すると，出力の多くが単語アライメントとなっていると考えられる．一方，提案手法では，出力の約 40%がフレーズアライメントであった．表 2では，より長いフレーズアライメントに対する評価が可能となっており，提案手法は F 値で最高性能を達成している．

以上の結果より，提案手法は既存手法よりも高いフレーズアライメント性能を持つことが示された．

[5]ただし，Ouyang and McKeown [6] はアラインされるフレーズ長の問題から完全一致の評価を報告していないため，本論文でも省略する．

[6]wiki-news-300d-1M-subword: `https://fasttext.cc/docs/en/english-vectors`

手法	P%	R%	F_1%
Jacana-phrase [1]	5.2	6.7	5.8
pointer-aligner [6]	23.4	**47.7**	31.4
提案手法	**31.6**	40.6	**35.5**

表 2: Ouyang and McKeown [6] の正解セットでの実験結果

4 おわりに

本研究では，生コーパスのみに依存する単一言語内フレーズアライメント手法を提案した．実装はツール SAPPHIRE[1]として公開している．MSR RTE corpus を用いた評価実験により，提案手法はフレーズアライメントにおいて既存性能を大幅に上回る最高性能を達成することを確認した．

今後は本手法を様々なドメインや言語へ適用し，その性能を評価する予定である．

謝辞 本研究は，JST，ACT-I，JPMJPR16U2 の支援を受けたものである．

参考文献

[1] Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. Semi-Markov Phrase-Based Monolingual Alignment. In *Proc. of EMNLP*, pp. 590–600, 2013.

[2] Nabin Maharjan, Rajendra Banjade, Nobal Bikram Niraula, and Vasile Rus. SemAligner: A Method and Tool for Aligning Chunks with Semantic Relation Types and Semantic Similarity Scores. In *Proc. of LREC*, pp. 1207–1211, 2016.

[3] George A. Miller. WordNet: A Lexical Database for English. *Commun. ACM*, Vol. 38, No. 11, pp. 39–41, 1995.

[4] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. PPDB: The Paraphrase Database. In *Proc. of NAACL-HLT*, pp. 758–764, 2013.

[5] Yuki Arase and Junichi Tsujii. Monolingual Phrase Alignment on Parse Forests. In *Proc. of EMNLP*, pp. 1–11, 2017.

[6] Jessica Ouyang and Kathy McKeown. Neural Network Alignment for Sentential Paraphrases. In *Proc. of ACL*, pp. 4724–4735, 2019.

[7] Chris Brockett. Aligning the RTE 2006 Corpus. Technical report, Microsoft Research, 2007.

[8] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. *TACL*, Vol. 5, pp. 135–146, 2017.

[9] Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. A Lightweight and High Performance Monolingual Word Aligner. In *Proc. of ACL*, pp. 702–707, 2013.