C3-3

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

https://github.com/m-yoshinaka/sapphire

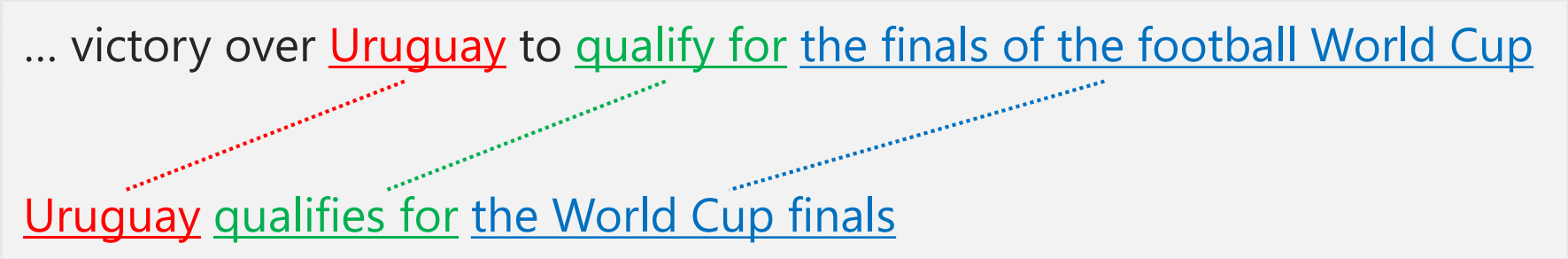
吉仲真人　梶原智之　荒瀬由紀

大阪大学

研究背景

単一言語内フレーズアライメント

- 単一言語の文対の中で意味的に対応するフレーズを対応付ける



- 言い換え/含意関係認識タスクなどへ応用
- 自然言語理解の重要な基礎技術のひとつ

既存手法の概観

既存手法	フレーズの単位	依存する 主な言語資源
MANLI [1] Jacana-phrase [2]	単語n-gram	• WordNet • PPDB
Arase and Tsujii [3] Pointer-aligner [4]	文法に基づく フレーズ	• 構文解析器 • チャンカー

[1] MacCartney et al. (2008). A Phrase-Based Alignment Model for Natural Language Inference. In Proc. of EMNLP, pp. 802-811.
[2] Yao et al. (2013). Semi-Markov Phrase-Based Monolingual Alignment. In Proc. of EMNLP, pp. 590-600.
[3] Arase and Tsujii. (2017). Monolingual Phrase Alignment on Parse Forests. In Proc. of EMNLP, pp. 1–11.
[4] Ouyang and McKeown. (2019). Neural Network Alignment for Sentential Paraphrases. In Proc. of ACL, pp. 4724–4735.

既存手法の課題

- 既存手法：**辞書**や**構文解析器**に依存

- 辞書や構文解析器などの言語資源：
  - ✓ 高精度なアライメントを得るための強力な資源
  - × 英語以外の多くの言語では充実していない
  - → 英語以外の言語への適用が困難

提案手法のアプローチ

- 学習済みの**単語分散表現のみを用いる**
  - ✓ 辞書や構文解析器が不要*
  - ✓ 豊富に存在する生コーパスのみに依存

- 任意の単語n-gramフレーズをアライン
  : 各アライメントの単語が重複しないような
  一対一のフレーズアライメントを求める

* 単語分割器は利用可能であると仮定

提案手法

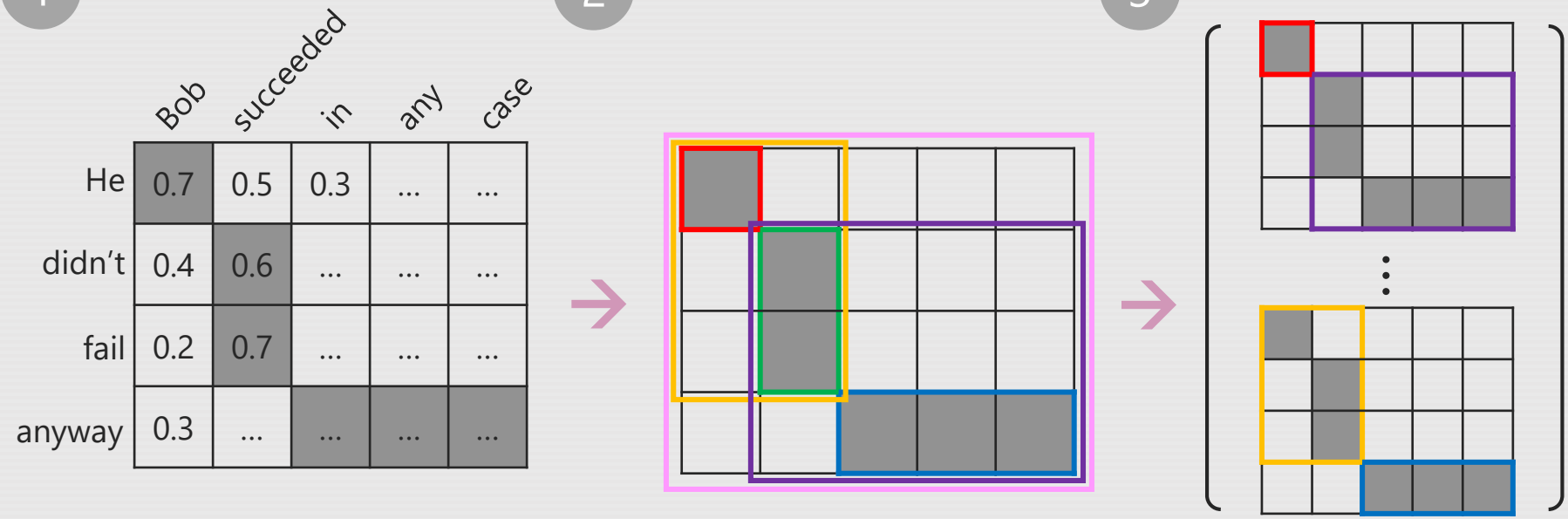
提案手法の概要

3つのステップでフレーズアライメントを獲得

1 単語分散表現に基づく単語アライメント

2 単語アライメントからフレーズ対への拡張

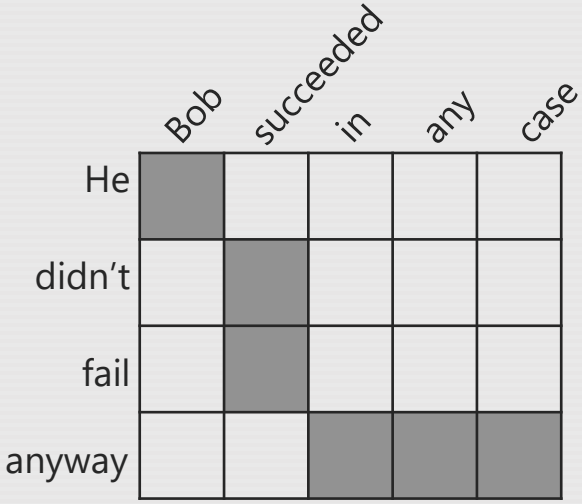
3 単語の重複がない フレーズ対の組の探索



単語アライメント

- 全単語対の分散表現間の余弦類似度を行列で表し **grow-diag-final** [5] を用いて求める

	Bob	succeeded	in	any	case
He	0.7	0.5	0.3	…	…
didn't	0.4	0.6	…	…	…
fail	0.2	0.7	…	…	…
anyway	0.3	…	…	…	…

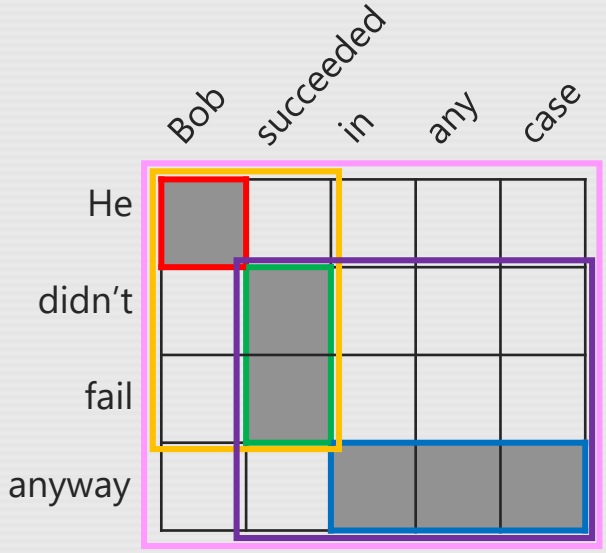


- 余弦類似度で単語アライメントを限定 (閾値：λ)

[5] Koehn et al. (2003). Statistical Phrase-Based Translation. In Proc. of NAACL-HLT, pp. 127–133.

フレーズ対への拡張

- SMTのヒューリスティクスで単語アライメントを
  フレーズ対へ拡張 → フレーズアライメント候補に



=

- Bob ↔ He
- Bob succeeded ↔ He didn't fail
  ⋮
- succeeded ↔ didn't fail
- succeeded in any case
  ↔ didn't fail anyway
  ⋮

- 後述のスコアで候補を限定 (閾値：δ)

フレーズ対のスコア

- フレーズ対が**どの程度対応し得るか**を表すスコア
- フレーズ長を考慮してスコアリング

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

フレーズ対 x, yのスコア

フレーズ対 x, yの類似度

フレーズ長のバイアス

$\mathbf{f}_x, \mathbf{f}_y$ ：フレーズの分散表現 (＝単語分散表現の平均)

$|x|, |y|$ ：フレーズ長 (＝単語数)

α ：フレーズ長のバイアスの重み

3 フレーズアライメントの探索

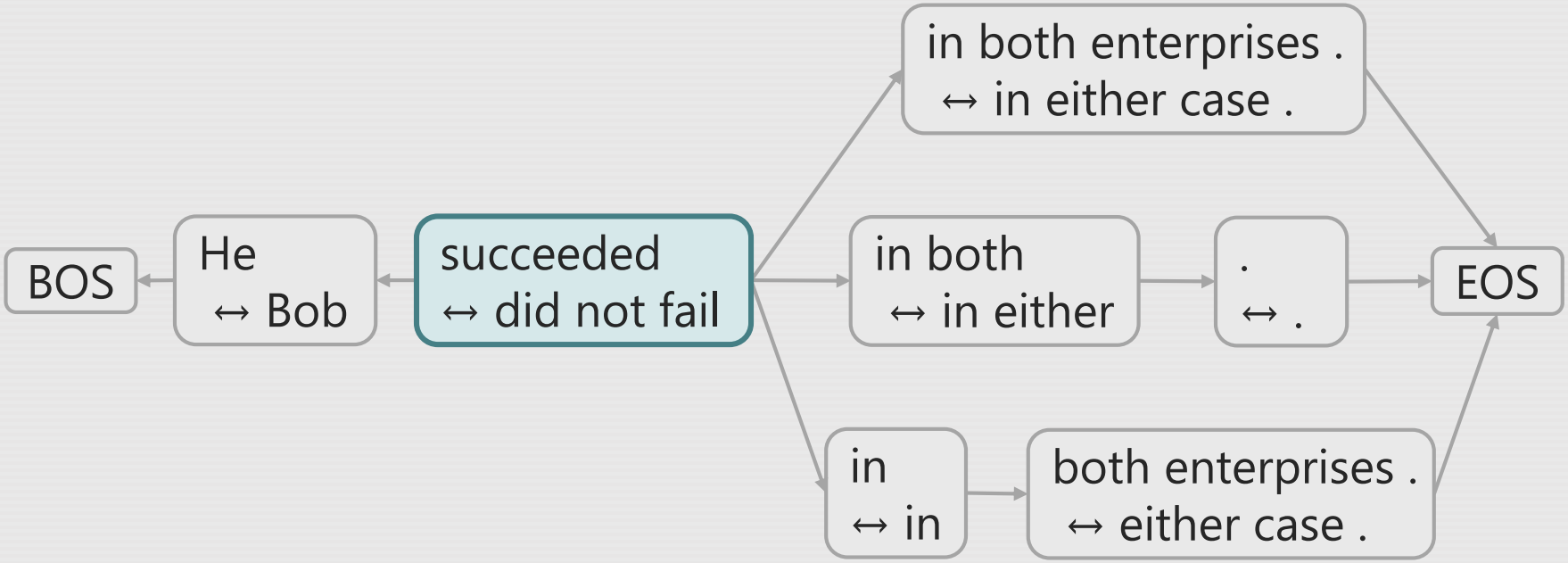
- 単語の重複がないフレーズ対の組み合せを探索
  → 全探索は計算量の観点から困難
  → 探索対象を限定

- フレーズアライメントは**文頭から順に決定できるとは限らない**
  → 最もスコアの高いフレーズ対から探索

3

ラティスを用いた探索

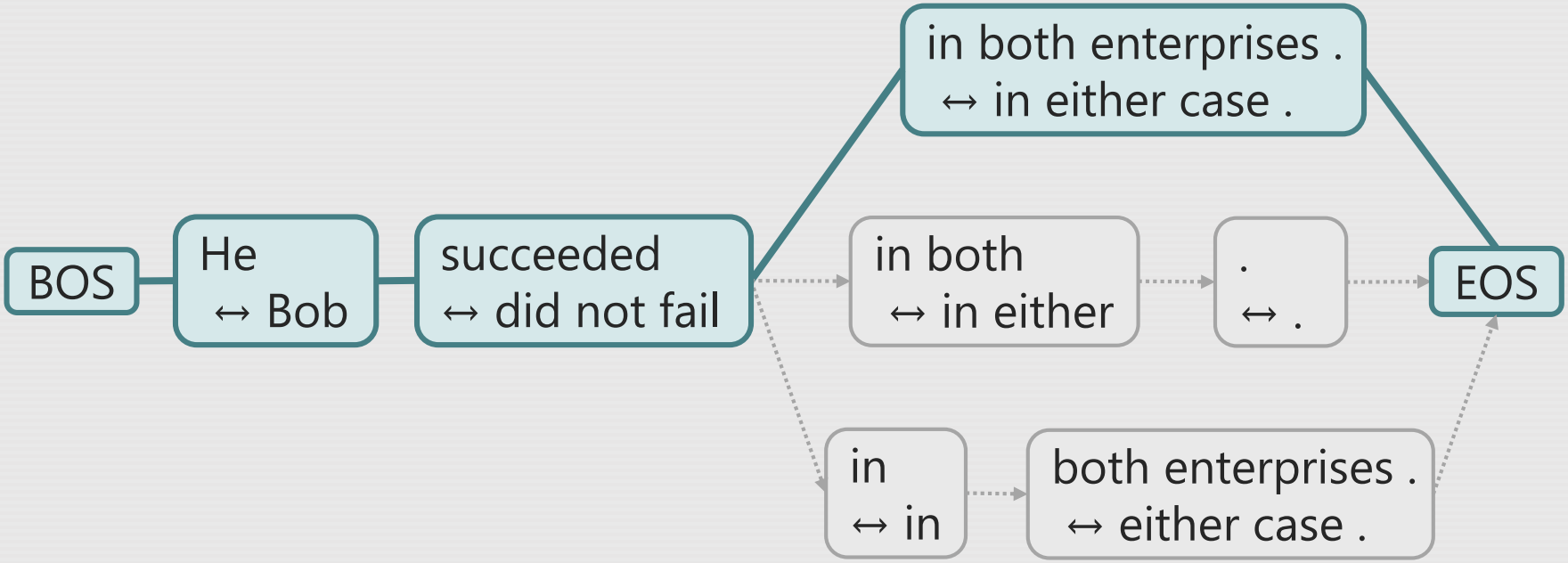
- **最もスコアの高いフレーズ対**の前後両方向へ
  単語が重複しないフレーズ対を追加して動的に探索



3

フレーズアライメントの決定

- 文頭から文末までの**最も平均スコアの高い経路**を適当なフレーズアライメントとみなして出力



英語での評価実験

データセット

MSR RTE corpus [6]

- dev/testセット：各800文対
- アノテータの確信度に応じて単語アライメントにSure/Possibleラベルが付与 (Possibleの97.5%は多対多)
- フレーズアライメントの評価は単語アライメントから擬似的に構成された正解セットで行う [2, 4]

[6] Brockett. (2007). Aligning the RTE 2006 Corpus. Technical report, Microsoft Research.

フレーズアライメントの正解セットの構成方法

方法	用いるラベル	文対数 dev / test	評価指標
単語が1対1に対応したチャンク対で構成 [2] (OpenNLP chunker)	Sure	800 / 800	完全一致率 (E)
多対多の単語アライメントを増やすことで構成 [4]	Sure Possible	487 / 441*	単語レベル** の適合率，再現率，F値 (P) (R) (F_1)

* 1つ以上のPossibleアライメントを含む文対に限定

** 出力中の任意の単語対を単語アライメントとみなして計算

提案手法の実装

単語分散表現

- fastText [7] の学習済みモデル*

ハイパーパラメータ $(\lambda, \delta, \alpha)$

- MSR RTE corpusのdevセットでF値が
  最大となるようにグリッドサーチして決定

[7] Bojanowski et al. (2017). Enriching Word Vectors with Subword Information. TACL, 5:135–146.
* wiki-news-300d-1M-subword: `https://fasttext.cc/docs/en/english-vectors`

実験結果

評価方法 1

方法	用いる ラベル	文対数 dev / test	評価指標
単語が1対1に対応した チャンク対で構成 [2] (OpenNLP chunker)	Sure	800 / 800	完全一致率 (E)
多対多の単語アライメント を増やすことで構成 [4]	Sure Possible	487 / 441*	単語レベル** の 適合率，再現率，F値 (P) (R) (F_1)

* 1つ以上のPossibleアライメントを含む文対に限定

** フレーズアライメント中の任意の単語を単語アライメントとみなして計算

結果

手法	E %
Jacana-token [8] (単語アライナー)	13.5
Jacana-phrase [2]	14.3
提案手法	**33.6**

既存手法の限定的な性能に対して 19.3pt の性能向上

[8] Yao et al. (2013). A Lightweight and High Performance Monolingual Word Aligner. In Proc. of ACL. pp. 702-707.

評価方法 2

方法	用いる ラベル	文対数 dev / test	評価指標
単語が1対1に対応した チャンク対で構成 [2] (OpenNLP chunker)	Sure	800 / 800	完全一致率 (E)
多対多の単語アライメント を増やすことで構成 [4]	Sure Possible	487 / 441*	単語レベル** の 適合率，再現率，F値 (P) (R) (F_1)

* 1つ以上のPossibleアライメントを含む文対に限定

** フレーズアライメント中の任意の単語を単語アライメントとみなして計算

結果

手法	P %	R %	F_1 %
Jacana-phrase [2]	5.2	6.7	5.8
Pointer-aligner [4]	23.4	**47.7**	31.4
提案手法	**31.6**	40.6	**35.5**

SoTAの既存手法の 4.1pt 高いF値を達成

出力例

- All that changed in 1922 , when **Tutankhamun 's tomb was discovered by** Egyptologist **Howard Carter** on behalf of his patron **Lord Carnarvon .**
- **Tutankhamun 's Tomb was unearthed by Howard Carter** and **Lord Carnarvon .**

MSR RTE corpus の **Sure** + Possible アライメント

- All that changed in 1922 , when Tutankhamun 's tomb was discovered by Egyptologist Howard Carter on behalf of his patron Lord Carnarvon .
- Tutankhamun 's Tomb was unearthed by Howard Carter and Lord Carnarvon .

提案手法の出力 (同じ色のフレーズ間にアライメントが存在)

日本語への適用

データセット・実装

データセット

- 首都大言い換えコーパス (TMUP) [9]
- 全655件の日本語言い換え文対

提案手法の実装

- 単語分散表現 : fastTextの日本語の学習済みモデル*
- 英語の実験 (評価方法２) でのハイパーパラメータ

[9] Suzuki et al. (2017). Building a Non-Trivial Paraphrase Corpus using Multiple Machine Translation Systems. In Proc. of ACL-SRW, pp. 36-42.

* cc.ja.300: `https://fasttext.cc/docs/en/crawl-vectors`

出力例

- これ は 、 一般 的 な 薬 として 利用 可能 です 。
- ジェネリック 医薬品 として 入手 でき ます 。

- 彼 は イースト・アングリア 大学 の 卒業生 です 。
- 彼 は 東 アングリア 大学 を 卒業 し て い ます 。

パラメータ調整無しで概ね適切なアライメントを出力

まとめ

背景

- 単一言語内フレーズアライメントの既存手法は辞書や構文解析器に依存

提案手法 https://github.com/m-yoshinaka/sapphire

- 学習済みの単語分散表現のみに依存

英語の評価実験

- ２つの実験設定で比較手法を上回る性能を達成