

アテンション機構を組み込んだモデルを用いた
StackOverflow投稿の関連性予測

Attention-based model for predicting question relatedness on Stack Overflow

Jiayan Pei, Yimin Wu, Zishan Qin, Yao Cong, Jingtao Guan
South China University of Technology, Guangzhou, MSR2021

背景

- 情報産業の継続的な進歩によりソフトウェア従事者が増加
 - Stack Overflow等のプログラミングに関するQAが人気
- 意味的に関連性のある質問 (duplicate) が多数存在する
 - コミュニティ内の評価の高いプログラマが重複した質問を手動でラベル付けしている
- 手動で行うには限界があり，非効率である
 - 質問間の意味的関連性を予測することに役立つ自動化されたアプローチを提案する

関連研究

○ StackOverflowが提案する関連質問

- 単語の重複に基づいているため信用性に欠ける[10]

○ 特徴量やヒューリスティックに基づいた研究 [2] ~ [4]

- 質問間の意味的関連性を自動的に計算できるが、質問の意味的情報を利用していない [5]
- 共通する単語が少ない場合もあるため、精度が低い

○ 深層学習を用いた研究 [13] ~ [16]

- 上記の研究よりは精度が高い
- 質問間の意味的関連性をベクトル距離に基づいて計算しているだけ
→ 2つの質問間の相互作用情報を十分に考慮することができない

アテンション機構を質問の関連性予測に用いることで
モデルは質問間の相互作用情報をより効果的に抽出できる

先行研究と提案手法

○ Shiraniら [12]

- 深層学習とSVM分類器に基づく2つのモデルをそれぞれ構築
 - DOT-BILSTM (最先端) とSOFTSVM
- 学習データ
 - Stanford[22]が公開している事前学習済みのGloVe単語埋め込みを使用
(ソフトウェア工学とは無関係なデータ = 曖昧な結果になる可能性がある[18])

○ 提案手法

- マルチクラス分類タスクとみなし, ASIM (Attention-based Sentence pair Interaction Model) を提案
- 学習データ
 - ソフトウェア工学のコーパスを構築(15GB)を構築し,
ドメイン固有の単語埋め込みを学習

データセット

○ Shiraniら [12]

- 347,372組の知識ユニット(質問とその回答)を構築
- 知識ユニットの関連性を重複, 直接的, 間接的, 無関係の4つで分類

重複	2つの知識ユニットの質問は重複した質問である
直接的	1つの知識ユニットの情報は, 別の知識ユニットの質問を直接解決することができる
間接的	1つの知識ユニットの情報は, 別の知識ユニットの質問の解決に 役立つが, その情報だけでは質問を直接解決できない
無関係	2つの知識ユニットの間には, 意味的な関連性はない

目的

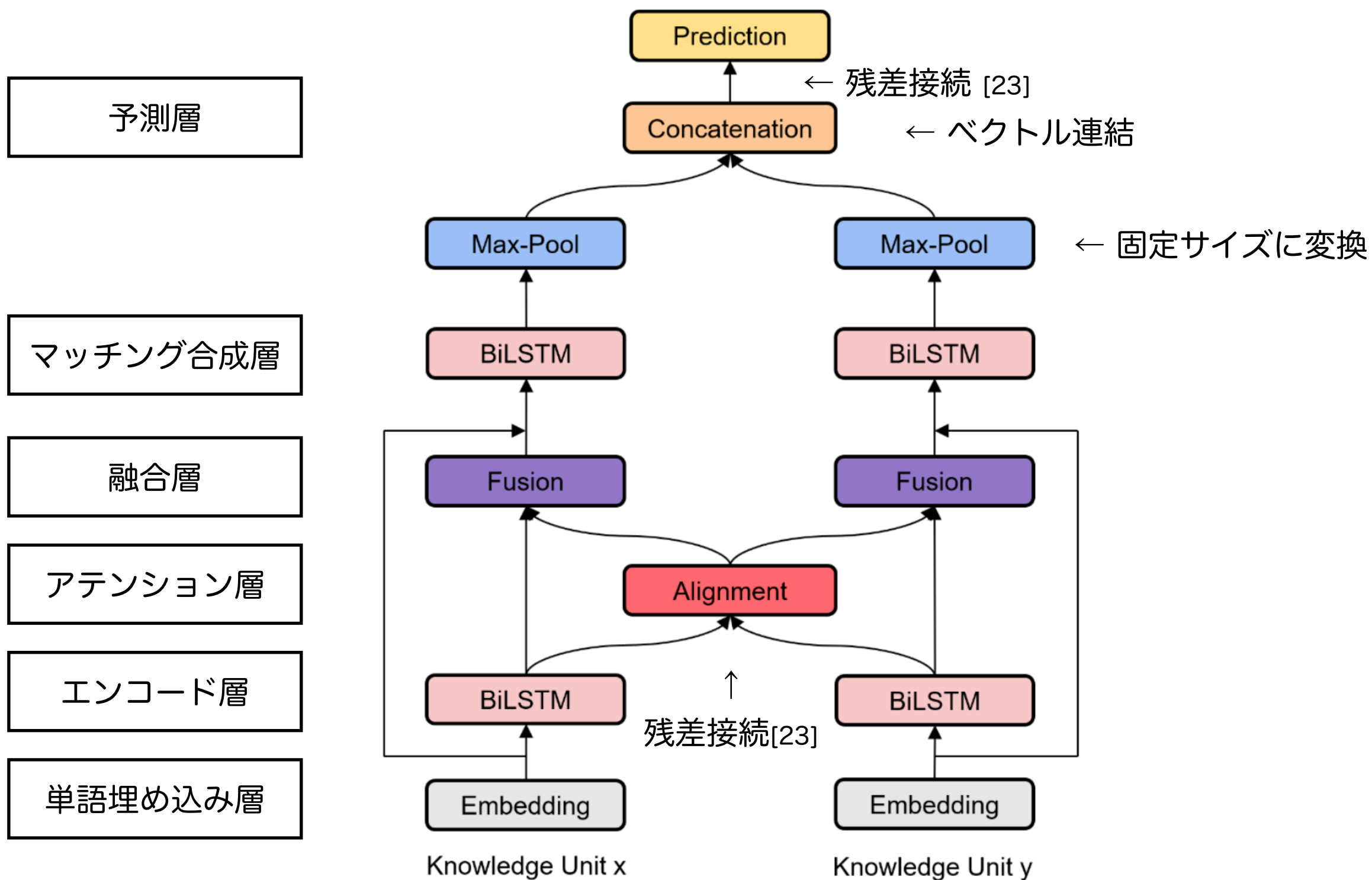
質問間の意味的関連性をより正確に予測する

→ ドメイン固有の学習データを用い、
アテンション機構を適用したマルチタスク分類モデルを構築

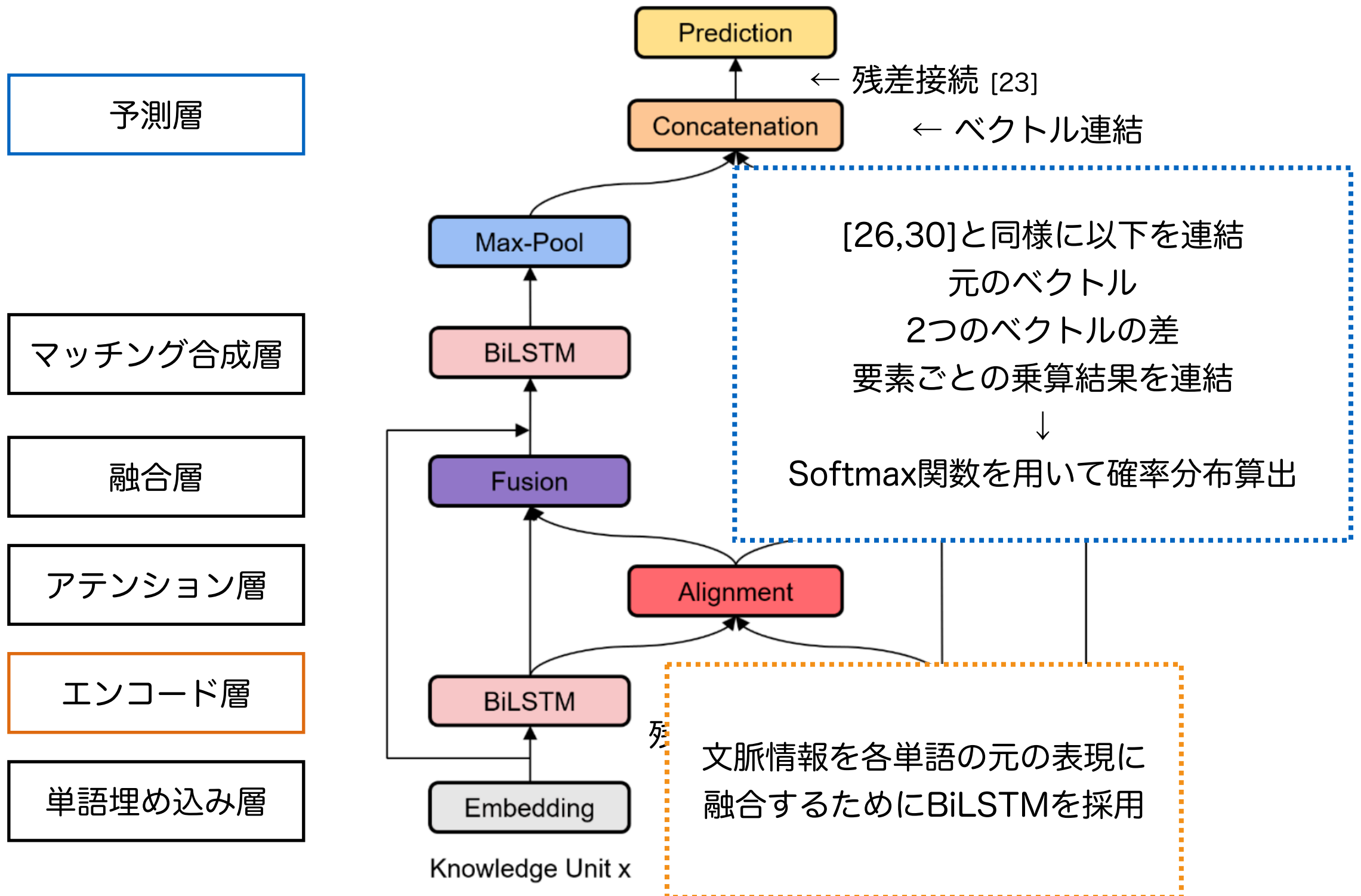
○ 貢献

- 重複質問を管理するコストの低減
- 構築したコーパスは他のソフトウェアエンジニアタスクの
モデル設計にも役に立つ

提案手法 | ASIMの概要



提案手法 | ASIMの概要



- RQ1 : (おそらく)同一クラスの知識ユニットペアを予測した際の精度の比較
- RQ2 : 異なる関連性クラスの知識ユニットペアを予測した際の精度の比較
- RQ3 : アテンション機構はASIMの改善にどの程度の影響を与えるか？
- RQ4 : ドメイン固有の単語埋め込みがASIMの性能向上に与える影響は？
- RQ5 : ASIMの一般化能力はどうか？

RQ1 | 同一クラスの知識ユニットペアの予測

○ (おそらく)同一クラスの知識ユニットペアを予測した際の精度の比較

Model/Metrics	Micro-F1	Precision	Recall
SOFTSVM	0.59	0.58	0.59
DOTBILSTM	0.75	0.75	0.75
ASIM	0.8228	0.8210	0.8228

○ アテンションメカニズムに基づくASIMはすべての評価指標で
先行研究の手法を上回った

RQ2 | 異なる関連性クラスの知識ユニットペアの予測

○ 異なる関連性クラスの知識ユニットペアを予測した際の精度の比較

Model/Classes	duplicate	direct	indirect	isolated	Micro-F1
SOFTSVM	0.53	0.57	0.44	0.79	0.59
DOTBILSTM	0.92	0.55	0.67	0.87	0.75
ASIM	0.9299	0.6845	0.7284	0.9437	0.8228

○ 直接的, 間接的なクラスは, ある程度意味的な関連性があるが,
2つのクラスに明確な違いがない → 精度が低い理由だと考える

RQ3 | アテンション機構の影響力

- アテンションメカニズムはASIMの改善にどの程度の影響を与えるか？
 - FL (Fusion Layer), SC (Shortcut Connections), Attn (Attention Layer)

Model/Classes	duplicate	direct	indirect	isolated	Micro-F1
ASIM	0.9299	0.6845	0.7284	0.9437	0.8228
ASIM (- FL)	0.9249	0.6552	0.7245	0.9404	0.8137
ASIM (- SC)	0.9222	0.6838	0.7175	0.9384	0.8165
ASIM (- FL - SC)	0.9177	0.6745	0.7097	0.9384	0.8124
ASIM (- Attn - FL - SC)	0.9153	0.6251	0.6697	0.9217	0.7846

- Attention Layer を削除すると性能が明らかに低下 (3.8%)
 - 質問の関連性を予測する上でアテンション機構は重要な役割である

RQ4 | ドメイン固有の単語埋め込みの影響力

- ドメイン固有の単語埋め込みがASIMの性能向上に与える影響は？
 - モデルはそのまま学習データのみを変えて比較を行う
 - Glove: ソフトウェア工学に関係のない学習データ (Shiraniらが使用)

Model/Classes	duplicate	direct	indirect	isolated	Micro-F1
GloVe word embeddings	0.9242	0.6667	0.7262	0.9365	0.8154
Doman-specific word embeddings	0.9299	0.6845	0.7284	0.9437	0.8228

- ソフトウェア工学ドメインに特化した単語埋め込みを使用することでASIMの精度が向上する

RQ5 | ASIMの一般化能力

○ 実験データセット:

- Rodriguesら[16]によって構築されたAskUbuntuの重複質問検出データセット

Models	Accuracy
Jcrd [34]	0.7291
SVM-bas [15]	0.7025
SVM-adv [16]	0.7587
CNN [15]	0.7450
DNN [20]	0.7865
DCNN [16]	0.7900
DOTBILSTM _[12]	0.87
SOFTSVM _[12]	0.90
ASIM	0.9625

○ ASIMは異なるソフトウェア工学タスクあるAskUbuntuの重複質問検出タスクでも良い性能を達成する

→ 一般化能力とロバスト性を証明

まとめ

- StackOverflowの質問の関連性を正確に特定できる
アテンション機構を用いたモデルASIM (Attention-based Sentence pair Interaction Model) を提案
- 提案手法は先行研究手法よりも精度良く質問の関連性を特定できる
- ソフトウェア工学ドメイン固有のコーパスを構築し、その有用性を示した

所感

- 最近の機械学習系ではアテンション機構が使われているため，取り入れることも考えて勉強しないといけないなと感じた．またアテンション機構の有無を比較する表を作れるので論文も書きやすいのかなとも．
- アテンション機構の各層の計算式等が丁寧に書かれていた(読み込めてないが)
- RQごとに動機・アプローチ・結果がまとめられている論文は読みやすい．
- ソフトウェア工学ドメイン固有のコーパスの構築15GBはやバそう