

単語区切りの違いによるQAサイトの質問回答ペアの分類

佐々木 稔 古宮嘉那子（茨城大学）

●はじめに

・QAサイト（QAサービス）

- ・利用者が質問や悩みを公開し、回答を募集する
- ・他のユーザーが回答したり、悩み相談にのる

・QAサイトの代表例

- ・Yahoo!知恵袋
- ・OKWAVE（オウケイウェイブ）
- ・発言小町

・利用者が質問を投稿する際に

- ・過去の投稿の中に、質問に対する回答が存在
- ・すぐに回答が得られると質問しなくてもよくなる
 - ・回答を待つ時間と質問時の労力を削減できる

・質問内容に対する望ましい回答を検索したい

- ・質問と回答のペアが正しいかどうかを判定する



●深層学習による言語モデルBERTを用いた質問・回答ペア判定

- ・これまでの質問回答判定
 - ・質問文と回答文に含まれる単語のマッチングで判定
 - ・適切な質問・回答ペアでも単語が一致しないことがある
- ・単語の一致だけでなく、文書の内容や単語の概念を利用したい
 - ・言語モデルBERTの学習済みモデルの利用が可能か検討
- ・単語区切りの違いによる分類精度の比較を行う
 - ・MeCab-ipadic-Neologd, JUMAN++

動物にかかわる仕事ってどんなものがありますか？？ 知ってる分だけお願いします！



少ないですけど、動物園の飼育員、トリマー、ブリーダー、盲導犬などのしつけ、ペットショップ。獣医。またはその・・・の助手？看護師？ペットホテルの従業員。くらいですかね？探せば他にもあるかもです。少なくとも本当にすみません。

●実験データ（3：1で訓練用とテスト用に分割）

- ・質問・回答の正しいペア500件と正しくないペア500件
- ・単語区切りとBERTの学習済みモデル
 - ・Mecab-ipadic-Neologd：日本語ビジネスニュース記事（300万件）
 - ・JUMAN++：黒橋研BERT日本語Pretrainedモデル（1800万件）
- ・質問回答ペアを判定できるように転移学習とFine-Tuningを行う
 - ・BERT出力を入力とする3層のFully Connected層を学習し、テストデータに対して判定結果を出力する

●実験結果

単語分割手法	Neologd	JUMAN++
判定精度	75.8%	86.0%

●Mecab-ipadic-Neologd

- ・QAサイトとは異なるジャンルの学習済みモデルを使用したことによる影響

●JUMAN++

- ・意味的に一貫性のある単語区切りの有効性と大規模な学習済みモデルを使用したことによる影響

●まとめ

- ・言語モデルBERTを用いた質問・回答ペアの判定を行った
- ・JUMAN++を用いた学習済みモデルを利用した場合に判定率が高くなった

●今後の課題

- ・その他の学習済みモデルの利用、学習データとして使用する効果的な質問・回答ペアの選び方の改良