文の接続性の定量化におけるクラスタリングの効果についての考察

前山 隼大　　指導教員：山村 毅

1 はじめに

自然言語は，人間がみな平等に使用できる一種のツールであり，使用にあたって特別な専門知識を必要としない。今日では，「機械と対話したい」という人間の純粋な好奇心から，自然言語を用いて行う機械と人間との対話，その中でも自然な対話 (人間にとって違和感なく繋がっている対話) に注目が集まっている。自然な対話を実現するためのアプローチとして，過去には自然言語の首尾一貫性や結束性，意図構造を意識したものが多く取り上げられてきたが，その中でも前山ら [1] は応答問題を対象に，ある一文とある一文が繋がる確率を考え，文の接続性を定量化を行った。

本研究では，この確率的手法に重きをおき，既存の文の接続性の定量化方法を改良するとともに，単語をクラスタリングすることで，文の接続性の定量化にどのような効果があるかについての考察を行う。

2 先行研究

著者ら [1] は，応答問題の質問文を $q = (q_1, q_2, \cdots, q_n)$，応答文を $r = (r_1, r_2, \cdots, r_m)$ と記述したとき (q_i, r_i はそれぞれ質問文・応答文の特徴)，以下の評価関数 (1) を最大化するものを適切な応答文として選択する手法を提案した。

$$f(q, r) = \prod_{i=1}^{m} P(r_i) \prod_{j=1}^{n} P(q_j|r_i) \tag{1}$$

3 評価関数の改良

3.1 評価関数の改良

先行研究の評価関数 (1) の $P(q_j|r_i)$ の部分は，端的に言えば単語同士の依存関係を表すが，全ての単語に依存関係があるわけではないので，$\prod_{j=1}^{n} P(q_j|r_i)$ と計算してしまうと単語同士の依存関係を過小に評価してしまう恐れがあった。これに対処するため，単語同士の相互情報量を利用し，新たな評価関数 (2) を作成した。

$$f(q, r) = \prod_{i=1}^{m} P(r_i) \prod_{j=f_1(r_i)}^{f_k(r_i)} P(q_j|r_i) \tag{2}$$

ここで，$q_{f_k(r_i)}$ は，r_i と k 番目に依存性の高い質問文中の単語を表す。$f_k(r_i)$ を計算するために，次に表す単語 r_i と q_j の相互情報量 $I(q_j, r_i)$ を用いる。

$$I(q_j, r_i) = P(q_j, r_i) \log \frac{P(q_j, r_i)}{P(q_j)P(r_i)} \tag{3}$$

3.2 実験と結果

先行研究と同じデータである，質問文 1 文に対して 3 つの応答文の選択肢がある 1038 例について，評価関数 (1)(2) を使用して応答文選択の比較実験を行った。特徴にはデータ中の全単語 2835 語を用い，分類器にはナイーブベイズ分類器，評価には 10 分割交差検定を用いた。また，ゼロ頻度問題への対応には加算スムージング ($\delta = 1.0 \times 10^{-3}$) を用いた。実験結果を表 1 に示す。

表 1　旧評価関数 (1) と新評価関数 (2) の性能比較

手法	正解	不正解	判定不能	正解率 (%)
先行研究 (評価関数 1)	429	609	0	41.3
本研究 (評価関数 2)	571	467	0	55.0

この結果は，正解率について先行研究 [1] に比べ，適合度検定 (有意水準 1%) で有意な差があった。

4 単語のクラスタリング

4.1 クラスタリングの意義

未知のデータに対する予測は，未知のデータの類似した学習データを見つけ，それを利用して予測することと捉えることができる。学習データが十分に存在する場合には，未知のデータに類似した学習データが見つかるが，データの数が不十分である場合には，そのような学習データが見つからず，誤った予測をしてしまう可能性がある。これに対処するために，学習データにおいて類似している単語をクラスタリングすることが考えられる。

4.2 Word Embedding

Word Embedding とは単語の分散表現のことである。単語は文字列により表されるが，それらの記号列から得られる情報は少ない。そこで単語に「この単語はこのような意味を持つ」といった情報を与えてベクトル表現することが考えられる。単語をベクトル表現するための方法の一つに，word2vec[*1]がある。word2vec では「同じような文で登場する単語は，類似した意味を持つ」という考えのもと，単語同士の関係性が数値化され，単語の意味が多次元ベクトル空間で表現される。つまり，ベクトル表現された単語同士のコサイン距離が近ければ類似した意味を持つ単語ということになる。

4.3 word2vec を用いたクラスタリング

先行研究と同じデータである，質問文 1 文に対して 3 つの応答文の選択肢がある 1038 例について，データ中の全単語のクラスタリング (閾値 $d = 1.00$〜0.70) を行い，その出力クラスを特徴として，評価関数 (2) で応答文選択実験を行った。分類器にはナイーブベイズ分類器，評価には 10 分割交差検定を用いた。また，ゼロ頻度問題への対応には加算スムージング ($\delta = 1.0 \times 10^{-3}$) を用いた。

各閾値での特徴数と，正解例題数の推移を図 1 に，最も正解例題数が多かった閾値 ($d = 0.83$) のときの結果を，実験結果を表 2 に示す。

表 2　実験結果 (word2vec)

手法	特徴数	正解	不正解	判定不能
クラスタリングなし	2835	571	467	0
word2vec($d = 0.83$)	2683	585	453	0

[*1] Google が開発したニューラル・ネットワークを利用した自然言語処理ツール

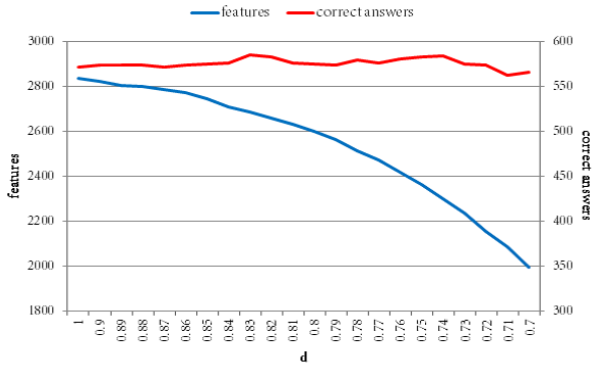

図 1　実験結果 (word2vec)

図 1 から，特徴数を大きく減らしながら正解率を維持できていることがわかる。しかし，最も精度が良かった $d = 0.83$ の場合の結果でも，正解例題数についてクラスタリングなしの結果に比べ，適合度検定 (有意水準 5%) で有意な差はなかった。

4.4　固有表現を利用した比較実験

何故有意差が出なかったのかを考えるために，比較実験として固有表現を用いてクラスタリングして同様の実験を行った。固有表現とは，特定の種類の何かを表す定名詞句であり，人名や地名，日付などがこれに該当する。固有表現の抽出には Stanford NER[*2]を用いた。実験結果を表 3 に示す。

表 3　実験結果 (Stanford NER)

手法	特徴数	正解	不正解	判定不能
クラスタリングなし	2835	571	467	0
Stanford NER	2497	597	440	1

この結果は，正解例題数についてクラスタリングなしの結果に比べ，適合度検定 (有意水準 5%) で有意な差はなかった。

4.5　Stanford NER の出力結果の手動訂正

表 3 の結果について検討したところ，Stanford NER の固有表現抽出結果には多くの誤りが見られることがわかった。そこで，これらの誤りについて手動訂正 (手動訂正 1 とする) を行い，併せて Stanford NER が固有表現として振り分けていないデータ中の数字の文字列 (“100“や “2017”) についても固有表現としてタグ付けを行った (手動訂正 2 とする)。再度同様の実験を行った結果を，表 4 を示す。

表 4　実験結果 (Stanford NER + 手動訂正)

手法	特徴数	正解	不正解	判定不能
クラスタリングなし	2835	571	467	0
Stanford NER	2497	597	440	1
手動訂正 1	2361	622	413	3
手動訂正 2	2290	628	407	3

手動訂正 1 及び手動訂正 2 の結果は，正解例題数についてクラスタリングなしの結果に比べ，適合度検定 (有意水準 1%) で有意な差があった。

[*2] “The Stanford Natural Language Processing Group” - http://nlp.stanford.edu/software/CRF-NER.shtml

4.6　品詞情報の考慮

word2vec では表層表現 (大文字・小文字区別なし) を基準に単語を認定し，一つの単語につき一つのベクトルが与えられている。従って、助動詞の “may” と 5 月の “May” を “同じ” と扱っている。しかし実際には，助動詞の “may” と 5 月の “May” とでは文中での役割や意味が大きく異なる。このように役割が異なる単語を，同じベクトル表現で扱うことには問題があると考えられる。そこで，文中での役割を区別するため，データ中の全単語に品詞情報を加え，word2vec を用いて再度学習を行い，同様の実験 (閾値 $d = 1.00$〜0.60) を行った。品詞情報の付加には NLTK[*3]を使用した。品詞情報を付加してクラスタリングをしなかったときの結果 ($d = 1.00$) と，最も正解例題数が多かった閾値 ($d = 0.68$) のときの実験結果を表 5 に示す。

表 5　実験結果 (word2vec + 品詞情報)

手法	特徴数	正解	不正解	判定不能
品詞情報 ($d = 1.00$)	3382	531	507	0
品詞情報 ($d = 0.68$)	2368	565	473	0

品詞情報を考慮した word2vec を用いた場合，正解例題数について，クラスタリングなしの結果に比べ，適合度検定 (有意水準 5%) で有意な差があった。

4.7　考察

Stanford NER の出力結果を手動訂正し，有意水準 1% で有意差を得たということから，単語をその役割をもとに適切に分類することができれば，正解率が飛躍的に向上するということがわかる。また，word2vec を用いた実験において，単語の表層表現のみを用いた場合には有意な差は得られなかったが，品詞情報を考慮した場合には有意差が得られたことからも，同様の結論が得られる。

5　まとめ

既存の文の接続性の定量化手法の評価関数について，相互情報量を用いることで改良を行った。ナイーブベイズ分類器によって実験を行った結果，最も高い正解率として 55.0% を得ることができ，先行研究 [1] との有意差を確認することができた。また，word2vec，固有表現，品詞情報を用いて単語のクラスタリングを行い，その実験結果と考察から，クラスタリングを行う際にはその単語の意味や役割を考慮した正確な分類が必要という結論を得た。今後の課題としては，品詞情報の解析誤りの手動訂正や，word2vec のニューラル・ネットワークの性質により複数回実験することなどが挙げられる。

参考文献

[1] 前山隼大, 秋田晃一郎, 山村毅: 応答問題を用いた英文の接続性の定量化, 電気・電子・情報関係学会東海支部連合大会講演論文集, L3-6, 2014

[*3] “Natural Language Toolkit - NLTK 3.0 documentation” - http://www.nltk.org/