概念識別子を語義とするニューラル語義曖昧性解消

情報科学科　森 大輔　　指導教員：山村 毅

1　はじめに

語義曖昧性解消は、文章に出現した多義語の意味の判別を行うタスクである。

単語の意味はその周辺単語から連想される [1]。そこで本研究では、対象となっている単語の意味を、その周辺単語を素性にして特定する。具体的には、周辺単語を BoW 表現し、これを入力として多層パーセプトロンによる語義曖昧性解消を試みる。また、周辺単語の BoW の構築の際に用いる単語数の適切な範囲も求める。

2　提案手法

単語の意味が周辺単語から連想されるという考えの下、周辺に存在する単語の情報を BoW 表現として入力し、対象単語の概念識別子を出力するニューラルネットワークモデルを構築することで、語義曖昧性解消を行う。ニューラルネットワークを用いることで、周辺単語を基にネットワークの値を更新していくことができ、意味を連想するモデルを実装可能と考えた。

3　BoW 表現

ニューラルネットワークには単語を直接入力することはできない。そこで、BoW と呼ばれる単語の出現回数を数える手法を用いる。例えば、

> 文 1 : 私 は とても 幸せ です
> 文 2 : 彼 は とても とても かっこいい

のとき、表 1 となる。このように表現した BoW の素性を各ニューロンに対応させる。

表 1　BoW の例

	私	は	とても	幸せ	です	彼	かっこいい
文 1	1	1	1	1	1	0	0
文 2	0	1	2	0	0	1	1

BoW の素性の選び方は、表 1 のように単語そのものを用いる以外にも様々なものがある。そこで本研究では、素性の選択方法を 3 通り提案する。(1) 周辺単語の表層形のみを用いるもの、(2) 周辺単語に加え、付与されている語義も単語とみなしたもの、(3) 周辺単語と付与されている語義を文字列結合したもの、の 3 通りである。以後は、これらを順に素性 1、素性 2、素性 3 と表す。

4　実験方法

4.1　実験データ

実験では、EDR 日本語コーパスおよび EDR 日本語単語辞書を用いる。EDR 日本語単語辞書には、概念識別子と呼ばれる単語の意味を表す 16 進整数が単語ごとに与えられている。このうち、同じ単語に対し複数の概念識別子が割り当てられているものを多義語とみなす。

一方、EDR 日本語コーパスは 207,802 文から構成されている。分かち書きが既に行われており、概念識別子や品詞が単語ごとに与えられている。このコーパス中の各文の中心付近から、上述の多義語に該当するものを 1 つ選定し、それを対象単語とする。

4.2　評価実験

対象単語ごとにニューラルネットワークを構築し、10 分割交差検定で分類を行い精度を検証する。ネットワークは合計 3 層で構成されており、中間層のユニット数は 900 である。また、BoW の構築の際に用いる単語数の範囲 (窓幅) を対象単語の前後 1〜8 の間で増減させ、窓幅を設けない場合との精度を比較する。窓幅の精度検証には、50 単語約 36000 文を利用する。機械学習には Chainer、BoW の構築には scikit-learn の CountVectorizer を用いた。

5　実験結果

窓幅の検証結果を表 2 に示す。

表 2　窓幅の検証結果

	平均精度 [%]				平均精度 [%]		
窓幅	素性 1	素性 2	素性 3	窓幅	素性 1	素性 2	素性 3
1	73.43	73.70	73.66	6	73.72	73.99	73.44
2	73.55	73.64	74.08	7	73.88	74.05	74.21
3	73.73	74.05	74.33	8	73.06	73.93	74.25
4	73.91	74.24	74.54	全体	72.81	79.54	81.51
5	73.86	73.94	74.21				

全体の精度検証は、合計単語数 2,137 個, 合計文章数 143,465 文, 1 単語あたりの平均語義数 2.61 個という条件の下行った。

多義語における語義には、使用頻度の偏りがある。MFS[2] はこの考えの下、用例で用いられた語義のうち、最も回数が多かった語義を選択する手法であるが、これは有効な手法として知られている。そこでこの MFS を実験における比較対象として採用する。

全体の精度検証結果を表 3 に示す。なお,† が付いているものは、マクネマー検定により統計的な有意差が得られたものである。

表 3　全体の精度検証

	MFS	素性 1	素性 2	素性 3
平均精度 [%]	72.78	72.80	79.18 †	81.39 †

6　考察と今後の課題

5 節より、窓幅の検証を行い、そのパラメータのもと、全体の精度検証を行った。素性 3 では、MFS を優位に上回り、平均精度が 80% を超えている。そのため、語義曖昧性解消における、意味情報の考慮の有効性を確認できた。

本研究では、対象の単語を文章当たり 1 つとして曖昧性解消を行っている。実際のシステムでは、文中に多義語が複数含まれる場合もある。そのため、文章中のすべての単語に対して曖昧性解消を行うことを今後の課題とする。

参考文献

[1] Yarowsky. David. ”One Sense per Collocation”, Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993.

[2] E. Agirre, and P. Edmonds. ”Word Sense Disambiguation: Algorithms and Applications.”, Springer, 2006.