大語彙の同義語集合からの文脈に応じた語彙選択

松原 勇介 †　辻井 潤一 †‡
† 東京大学大学院情報理工学系研究科
‡School of Computer Science, University of Manchester
National Centre for Text Mining, UK

文脈に沿った正しい語彙を認識し選択するタスクは、情報検索、機械翻訳、自然言語生成などの多くの自然言語処理の基礎となる。たとえば情報検索では検索対象のテキスト集合とクエリの間での表現の違いを効果的に吸収することが、高い再現率を達成するために必要とされる。

過去の語彙選択の研究は、2節に示すように、局所的な文脈だけを考慮に入れるもの、限定された語彙で評価されたものがほとんどだった。本論文では、大語彙での応用をめざして、統語構造など広い文脈を扱い、より大きな語彙での統計的語彙選択の手法を提案する。

1　タスク定義

[1] と [2] にしたがい、語彙選択を次のように定義する。

文、同義語候補集合、文中での候補が占めるべき位置（語彙ギャップ）が与えられたものとする。語彙選択のタスクは、ギャップの周辺の文脈に照らしたときそのギャップを埋めるのにもっとも適切な単語を候補集合の中から選ぶものである。

例として、文 "*Workers dumped large burlap sacks of the imported substance*$_{[p_1]}$ *into a huge bin*." が与えられたとする。この文中で $[p_1]$ の記号が指す *substance* の位置が埋めるべきギャップである。このギャップに対する同義語候補集合として *substance*, *stuff*, *material* が得られたとする。語彙選択システムはこの 3 者のうちから、ギャップを埋めるのにもっとも適切なものを選ぶ。

2　既存研究

語彙選択に用いる手がかりとして、ある単語と文脈がどれだけ強く関連するかを調べる様々な尺度が利用されてきた。たとえば t-score[1], 局所的統語文脈に対する条件付確率 [3], PMI スコア [2] などである。中でも PMI スコアは Edmonds [1] が提供するテストセットにおいてもっとも高い性能を示した。Inkpen [2] は、ある候補単語と、対象ギャップの周辺の固定単語数のウィンドウ内の関連度を測るために Point-wise Mutual Information (PMI) 基準 [4] を利用して Edmonds の t-score 法を改善した。その手法では、ある候補 t の文脈に対する適切さが、次の式に示される PMI スコアの和で測られる。

$$\mathrm{PMI}(w_{-k}^{-1}, w_1^k, t) = \sum_{w \in (w_{-k}^{-1} \cup w_1^k)} \frac{C(w,t)N}{C(w)C(t)}, \qquad (1)$$

ここで w_{-k}^{-1} は k 個の先行単語、w_1^k は k 個の後続単語、$C(\cdot,\cdot)$ はコーパスでの 2 つの単語の共起頻度、$C(\cdot)$ は単語の頻度、N はコーパスの総単語トークン数である。Gardiner and Dras [5] は、フルテキストではなく N-gram で提供されるコーパスで計算できる、 Inkpen の手法に対する近似を提案した。

Edmonds の実験設定を共有する研究 [1, 2, 5] でや人手で作成した言い換え文対で評価した研究 [6] では、語彙選択の評価の対象は語彙のサンプルに止まっており、全語彙に近いものではない。

我々はより大きな語彙での語彙選択を高精度に行うことが、情報検索をはじめとする応用により貢献すると考える。

3 提案手法

提案手法は、同義語候補集合の生成と、同義語選択の 2 つのモジュールからなる。入力テキストが与えられたとき、まず生成モジュールが言い換えすべき単語出現を選び、その単語出現に対する候補集合を、単純にシソーラスを用いて生成する。次に、同義語選択モジュールが個々の候補に確率を付与する。それぞれのステップの詳細は以下で説明する。

3.1 同義語候補の生成

同義語候補の生成では、入力文でのあるギャップに書かれていた単語を単純にシソーラスで引くことによって候補を取得する。ここで想定するシソーラスは、単語とその品詞から、同義語の集合を引くことができるものとする。この形式を仮定することにより、ドメインに特化して自動抽出されたシソーラスを作り利用するのが容易になる。ただし本論文の実験では WordNet から生成されたシソーラスを用いる。この際、多義語に関しては、単にすべての語義に関する同義語を合わせて同義語候補とし、同義語選択モジュールによって文脈に応じた選択をさせた。

3.2 同義語選択

文脈 C と同義語候補集合 S が与えられたとし、最大エントロピー法 [7] によって候補 $w \in S$ の文脈に対する条件付確率を下記のように表す。

$$P(w|S;C) = \frac{\exp\left[\Lambda \cdot F(w,C,S)\right]}{\sum_{v \in C} \exp\left[\Lambda F(v,C,S)\right]}, \qquad (2)$$

ただし、$F = \{f_1, f_2, \cdots, f_K\}$ は素性ベクトル、Λ は重みベクトルである。このモデルに基づいて、もっとも確率の高い同義語候補 $\hat{w}$ を

$$\hat{w} = \underset{w \in C}{\operatorname{argmax}}\, P(w|S;C) \qquad (3)$$

として選ぶ。

3.3 訓練

このモデルを訓練するため、ある単語がある文脈で使用できるのは、訓練コーパスで同じ文脈での使用があったときだけだとする、ナイーブな仮定をおく。3.1節と同じように訓練コーパスに同義語候補を付与し、実際にコーパスで用いられた単語を正例、それ以外の同義語候補を負例として訓練を行う。

表 1　List of features

名前	種類	定義
`frequency`(w)	baseline	w のコーパスでの頻度
`unigram_v`(w)	baseline	w が v と等しい
`pmi_v`(w)	window	先行p単語・後続qの PMI(v,w) の和
`pmiavg`(w,S)	window	$\frac{\sum_{v \in window} \mathrm{PMI}(v,w)}{window\ size}$
`cache_w`(w,S)	文書	w がそれ自身以外の S 中の場所に出現したか
`cache-l_w`(w,S)	文書	w と同じ語幹の語がそれ自身以外の S 中の場所に出現したか
`cache-o_w`(w,S;C)	文書	候補集合 C のいずれかと同じ語幹の語が S 中の他の場所に出現したか
`subtree_v`(w,S;C)	統語	w の S 中での統語木の兄弟ノードの品詞タグと v の組み合わせ
`unigram_pos`(w,S)	統語	w と v が品詞を含めて等しい

3.4 素性抽出

用いる素性の一覧を表 1に示す。ベースラインの素性に加え、文書ベースの素性と統語ベースの素性を用いる。

文書全体の bag-of-words に対して定義された文書ベースの素性により、長距離の文脈での単語の結びつきが捉えられると期待される。局所的な品詞パターンに対して定義された統語ベースの素性により、候補単語に対する統語的な制約が捉えられると期待される。

4 実験

下記に示すとおりの全単語を対象とした語彙選択タスクにおいて、ベースラインの素性に付加したと

きの文書ベースの素性と統語ベースの素性の貢献を調べる。

4.1 実験設定

mapping A	mapping B	target POS
NN,NNS,NNP,NNPS	NN	noun
JJ,JJR,JJS	JJ	adjective
VB,VBD,VBG,VBN,VBP,VBZ	VB	verb
others	others	*ignored*

表 3　Part-of-speech mappings

Part-of-speech	paraphrase candidates
adjective	*difficult*, *tough*, *hard*
noun	*error*, *mistake*, *oversight*
noun	*job*, *task*, *duty*
noun	*responsibility*, *commitment*, *obligation*, *burden*
noun	*material*, *stuff*, *substance*
verb	*give*, *provide*, *offer*
verb	*resolve*, *settle*

表 4　Edmonds' seven evaluation sets for lexical choice

二つのタスクで提案手法を評価する。ここで sample-words と呼ぶ最初のタスクは、[2] の設定をほぼ再現したものである。もう一つの all-words タスクは、すべての内容語を対象としたものである。表 2 にそれぞれのタスクの統計を示す。双方のタスクで、WordNet 3.0 を用いて同義語候補生成を行った。その際の品詞の変換法を 3に示す。

訓練コーパスとして、Penn Treebank corpus 3.0 と BLLIP 1987-89 WSJ Corpus の section W8_001 から W8_019 を用いた。語彙選択を行うに当たって、NN (nouns), JJ (adjectives), VB (verbs), RB (adverbs) の品詞を対象とした。

sample-words の評価では、Inkpen に合わせて、1987 年の Wall Street Journal における Edmonds の 7 つの同義語セット（表 4に示す） を用いた。all-words の評価では、BLLIP WSJ Corpus の W7_001 から W7_127 内のすべての内容語を対象とした。PMI スコアは W8_001 から W8_108 、W9_001 から W9_41 のテキスト中で 5 単語のウィンドウをとって計算した。

特に all-words の人手による評価は困難なため、[2] を踏襲して、元テキストで出現した単語のみを正解として評価した。

4.2 実験結果

表 4.2 に sample-words での結果、 表 4.2 に all-words の結果を示す。接尾辞 "-A", "-B" はそれぞれ表3 に示す異なる品詞マッピングを表す。"+Doc", "+Syn" はそれぞれ表 1に示す文書ベース、統語ベースの素性の導入を表す。

表 4.2 の sample-words の結果では、Proposed-A と Proposed-B は名詞と形容詞について Inkpen の手法を改善した。これらは本質的には Inkpen と同じ情報を使っているため、性能の差は、提案手法の PMI スコアの計算に使ったコーパスが Inkpen と異なり、対象ドメインのコーパスだったからだと考えられる。文書ベースの素性は名詞と形容詞についてさらに性能を改善したが、動詞については有効でなかった。

表 4.2の all-words の結果では、文書ベースと統語ベースの素性はともに結果を改善した。

表 5　Accuracy for sample-words task on BLLIP corpus of Wall Street Journal 1987

	Acc. (noun & adj.)	Acc.	#samples (noun & adj.)	#samples
Inkpen	70.01	65.16	18018	31116
A	70.95	58.29	25601	59676
B	72.03	65.03	17106	28645
A +Doc	71.09	58.18	25601	59676
B +Doc	72.19	64.63	17106	28645
A +Doc+Syn	65.96	50.73	25601	59676
B +Doc+Syn	64.13	45.49	17106	28645

		Source corpus	Number of word tokens
Sample words task	Training	BLLIP 1988 + Penn Treebank	16893445
	Testing	BLLIP 1987	22926540
All words task	Training	Penn Treebank	10778880
	Testing	(4-fold cross validation)	-

表 2　Tasks and corpora

表 6　Accuracy for All-words task on Penn Treebank

	Acc. (noun & adj.)	Acc.	#samples (noun & adj.)	#samples
A	79.23	75.07	86179	122131
A +Doc	81.20	77.00	86179	122131
A +Doc+Syn	81.59	77.47	86179	122131

5　考察

5.1　Sample-words と All-words の違い

我々のモデルは sample-words と all-words のタスクで異なる振る舞いをした。大きな違いとして、統語素性が all-words でのみ有効だったことが挙げられる。これは sample-words の同義語集合が all-words と比べて統語的にも用法の近い単語からなっているためだと思われる。一方で all-words の同義語集合は、多義語など比較的意味・統語的な差の大きい単語を含んでいる。

5.2　文書ベースの素性の貢献

文書ベースの素性が双方のタスクで結果を改善したことは、同義表現から適切なものを選ぶ際に、文書中での一貫性がひとつの重要な手がかりとなっていることを示唆する。ただし、この実験で我々はひとつのギャップについて単語を選択する際に他のギャップでは正しく単語が選択されているとみなしており、実用での有効性が同程度とは言えないことに注意が必要である。

6　結論

文書ベースと統語ベースの素性を導入した提案手法は、すべての内容語を対象とした all-words の語彙選択の性能を改善した。

参考文献

[1] Philip Edmonds. Choosing the word most typical in context using a lexical co-occurrence network. In *Proc. of the EACL*, 1997.

[2] Diana Inkpen. A statistical model for near-synonym choice. *ACM Trans. Speech Lang. Process.*, Vol. 4, No. 1, p. 2, 2007.

[3] S. Bangalore and O. Rambow. Corpus-based lexical choice in natural language generation. In *ACL'00*, 2000.

[4] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, Vol. 16, No. 1, pp. 22–29, 1990.

[5] Mary Gardiner and Mark Dras. Exploring approaches to discriminating among near-synonyms. In *Proc. of the Australasian Language Technology Workshop*, 2007.

[6] Michael Connor and Dan Roth. Context sensitive paraphrasing with a global unsupervised classifier. In *Proc. of the 18th ECML*, pp. 104–115, Berlin, Heidelberg, 2007. Springer-Verlag.

[7] Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. A maximum entropy approach to natural language processing. *Comput. Linguist.*, Vol. 22, No. 1, pp. 39–71, 1996.