

# 言い換え表現の曖昧さを調整可能な複合名詞の言い換えモデル

井上 翔太<sup>†</sup>      松吉 俊<sup>‡</sup>      内海 彰<sup>‡</sup>

<sup>†</sup> 電気通信大学 情報理工学部 総合情報学科

<sup>‡</sup> 電気通信大学大学院 情報理工学研究科

{s.inoue, matuyosi, utsumi}@uec.ac.jp

## 1 はじめに

複合名詞の構成語（修飾語 MOD と主辞 HEAD）の意味関係を明らかにし、解釈することは、特に言語理解が必要なタスクにおいて重要である。複合名詞の解釈に対するアプローチのひとつとして、複合名詞の言い換えがある。

ある複合名詞に対する言い換えは、ひとつではなく多数のものが考えられるが、その中でも多くの複合名詞に適用できる曖昧さの大きい一般的な言い換えと、曖昧さの小さい個別的な言い換えがある。前者は前置詞のみ用いた言い換えなどで、例えば *office desk* に対する *desk in office* である。後者は前置詞だけでなく動詞や関係詞なども用いた言い換えなどで、例えば *office desk* に対する *desk designed for office* や *desk which is available on office* である。一般的な意味をもつ言い換えは、個別的な意味をもつ言い換えよりも頻繁に使用される傾向があると考えられる [3]。つまり、高頻度な言い換えは曖昧さの大きい一般的な意味をもっており、低頻度な言い換えは曖昧さの小さい個別的な意味をもっていることが多いと考えられる。

複合名詞を言い換えるとき、多くの場合で一般的な意味の言い換えが用いられると考えられるが、例えば語彙の少ない非母語話者に対して複合名詞を説明するときは、曖昧さの小さい個別的な意味の言い換えをしたほうが伝わりやすいだろう。しかし、自由な言い換えて対象とする SemEval-2013 Task 4 [2] の評価指標は、低頻度な言い換えは評価値への影響が小さくなるよう設計されている。複合名詞の言い換えについての先行研究 [8, 5, 6, 9] でも、個別的な意味の言い換えて狙って出力する試みはされていない。

本研究では、曖昧さの大きい一般的な意味を表現する言い換えほど、事前に定義された意味関係を表すラベルに基づいて生成できると仮定する。この仮定のもと、高頻度な言い換えて生成するときは事前にその複合名詞に付与した意味関係ラベルを重視し、低頻度な

言い換えて生成するときは複合名詞の構成語（修飾語と主辞）を重視して学習するモデルを提案する。意味関係ラベルをどれだけ重視するか指定することで、このモデルは生成される言い換えての曖昧さを調整できる。

## 2 モデル

提案する複合名詞の言い換えモデルは、複合名詞の意味関係ラベルを必要とする。そのため、意味関係ラベルを予測するモデルと、言い換えてを生成するモデルの2つを実装した。

### 2.1 意味関係ラベル分類

複合名詞の構成語である修飾語 MOD、主辞 HEAD を入力として、その意味関係ラベルを予測する。モデルは、埋め込み層、biLSTM、平均プーリング層、MLP、Softmax で構成されたものを用いる。

### 2.2 言い換え

本研究では複合名詞の言い換えてを分類問題として扱う。そのために、複合名詞の言い換えてをテンプレート化し、そのテンプレートを複合名詞のクラスとみなす。言い換えてのテンプレートとは、言い換えて中に含まれる複合名詞の主辞と修飾語を、それぞれ  $H$ ,  $M$  というトークンに置き換えたものである。この手順は後述の 3.1.2 節に詳しい。

提案モデルの構造を図 1 に示す。この構造は 2.1 の意味関係ラベル分類モデルとほぼ同じだが、平均プーリング層が重み付き平均プーリング層となっている点で異なる。重み付き平均プーリングのための重みは、データから学習するものではなく、はじめから入力として与えられるものである。訓練時には、この重みは次式で決められる。

$$\begin{aligned} w_L &= \text{sigmoid}(5 - r) \\ w_H &= (1 - w_L)/2 \\ w_M &= (1 - w_L)/2 \end{aligned} \quad (1)$$

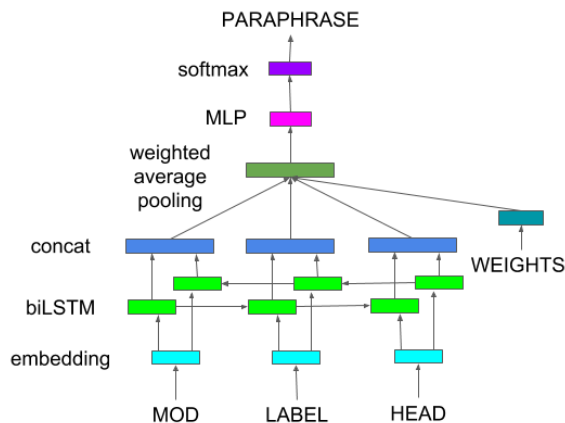


図 1: 提案する言い換え予測モデル

ここで  $r$  は、データ中で高頻度な言い換えには小さな値、低頻度な言い換えには大きな値が与えられる変数である (3.2 節参照)。提案モデルは、 $r$  の小さい言い換えを予測するときは意味関係ラベル LABEL,  $r$  の大きい言い換えを予測するときは複合名詞の構成語である修飾語 MOD, 主辞 HEAD を重視して学習する。これによって、予測時に  $w_L$  を大きな値にすると一般的な広い意味を持つ言い換えが出力されやすく、 $w_L$  を小さな値にすると個別的な意味を持つ言い換えが出力されやすくなる。

### 3 実験

#### 3.1 データ

##### 3.1.1 意味関係ラベルのためのデータ

本研究で用いた複合名詞の意味関係ラベルを学習するためのデータは、Tratz[7] のデータセットのうち、12 種類のラベルを定義したデータセット (coarse) である。この中でも、各事例を無作為に 75 : 5 : 20 (訓練データ : 開発データ : テストデータ) に振り分けたデータセットを用いた (random)。本研究で学習に用いたのは訓練データのみで、その事例数は 14,093 である。

##### 3.1.2 言い換えのためのデータ

SemEval-2010 Task 9 [1] と SemEval-2013 Task 4 [2] から提供されている訓練データを整形し、言い換えをテンプレート化した。この際、テンプレートの種類数を減らすため、冠詞 *a*, *an*, *the* を除くなど、細かい表現の揺れを消した。例えば、*car accident* に対して *accident happened by a car* という言い換えがあったとき、これを *H happend by M* というテンプレートとした。そして、データセット中で 2 回以上出現したテンプレートだけを残した結果、最終的な言い換えテ

ンプレートは 1782 種類となった。

また、SemEval-2013 のデータは各複合名詞に対する言い換えの最上位に前置詞のみを用いた言い換えが頻出するが、SemEval-2010 のデータは動詞と前置詞を用いるという制約があるため、前置詞のみを用いた言い換えは出現しない。この不均衡を解消するために、SemEval-2010 からの動詞+前置詞の言い換えデータに、前置詞のみの言い換えをいくつか頻度最上位の言い換えとして人手で追加した。追加する前置詞のみの言い換えは、文献 [4] にならって Netspeak<sup>1</sup> を用いて決めた。Netspeak は、クエリとして与えたパターンと合致する単語列を、頻度降順で提示する。これによって、複合名詞の構成語をつなぐ前置詞として何がよく使われているか分かる。提示された前置詞の中で、その複合名詞の言い換えとして適しているものを、テンプレート化して訓練データとして採用した。ここで訓練データは、複合名詞 422 個、各複合名詞に与えられている言い換え数の平均は約 38.9 個となった。

次に、訓練データ内の複合名詞に意味関係ラベルを付与した。この処理には、2.1 節の意味関係ラベル分類モデルで学習した分類器を用いた。このモデルには、埋め込み層 300 次元、biLSTM の出力 300 次元、MLP の出力 12 次元 (意味関係ラベルの種類数)、埋め込み層のドロップアウト率 0.3、MLP 内のドロップアウト率 0.1、バッチサイズ 128、最適化手法 Adam、エポック数 30 で学習を行い、開発データに対して最も性能の良かったエポック数のものを採用した。また、ラベルは語彙ではないが、事前学習済みの単語ベクトルとして扱えるように、*loc\_part\_whole* は *part* に、*owner\_emp\_use* は *use* に変えた。例えば、*computer hardware* に *loc\_part\_whole* というラベルが付与された場合、提案モデルの入力として *computer part hardware* という擬似的な単語列を与えた。

#### 3.2 評価手法

提案モデルは、埋め込み層 300 次元、biLSTM の出力 300 次元、MLP の出力 1782 次元 (言い換えテンプレートの種類数)、埋め込み層のドロップアウト率 0.3、MLP 内のドロップアウト率 0.1、バッチサイズ 128、最適化手法 Adam、エポック数 10 で学習を行った。埋め込み層は、6B トークンを事前学習した 300 次元の GloVe 単語ベクトル<sup>2</sup> で初期化し、学習時に更新しない設定とした。出力の各言い換えに対しては、

<sup>1</sup><http://www.netspeak.org>

<sup>2</sup><https://nlp.stanford.edu/projects/glove/>

表 1: 各システム出力の評価

手法	isomorphic	non-isomorphic
SemEval-2013 baseline[2]	0.138	0.406
SFS [9]	0.231	0.179
IIITH [6]	0.231	0.258
MELODI [8]	0.130	0.548
Shwartz & Dagan [5]	0.282	0.284
NC-wise ( $t = .000$ )	<b>0.285</b>	0.247
NC-wise ( $t = .005$ )	0.244	0.292
NC-wise ( $t = .050$ )	0.081	<b>0.619</b>
overall ( $t = .000$ )	0.271	0.227
overall ( $t = .005$ )	0.226	0.260
overall ( $t = .050$ )	0.089	0.567
ベースライン ( $k = 10$ )	0.136	0.406
ベースライン ( $k = 5$ )	0.095	0.553

表 2:  $r = 10$  のときのシステム出力の人手評価

手法	一般的 (高頻度)		個別的 (低頻度)	
	生成割合	適合率	生成割合	適合率
NC-wise	0.052	(1.000)	0.948	0.595
overall	0.000	-	1.000	0.456

Softmax で予測確率が与えられる。

また、訓練データに対して、各事例に対する  $w_L, w_H, w_M$  を決めるための  $r$  を、NC-wise 法と overall 法という 2 種類の方法で与えた。NC-wise 法では、各複合名詞に対する言い換え集合の中で、最も頻度の高い言い換えから順に  $r$  を 0, 1, 2, ... とした。overall 法では、まず訓練データ全体での各言い換え (テンプレート) の頻度を計算し、その頻度の降順で 0, 1, 2, ... と順位付けした。そして言い換えをその順位が 40 未満, 100 未満, それ以外の 3 つのグループに分け、その順に  $r$  を 0, 5, 10 とした。overall 法のデータにおいて、 $r = 0$  と与えられた言い換えが 40 種類、 $r = 5$  が 102 種類、 $r = 10$  が 1640 種類となった。NC-wise 法では同じ言い換えテンプレートでも同じ  $r$  が与えられるとは限らないのに対し、overall 法は同じ言い換えテンプレートには必ず同じ  $r$  が与えられる。

### 3.2.1 一般的な言い換えを狙った出力の評価

システム出力の言い換えとテストデータ中の言い換えを、n-gram 単位で比較する [2]。n-gram に含まれる各単語は、完全一致ではなく、先頭から何文字ぶん一致しているかで評価される。最終的な全体のスコアは、2 種類の方法で評価される。isomorphic は適合率と再現率の両方を評価し、システムは高ランクな言い換えから低ランクな言い換えまでをテストデータの順位通りに生成できればよい。non-isomorphic は適合率のみ評価し、システムは最上位の参照言い換えだけを生成できればよい。

提案モデル (NC-wise 法と overall 法) は、評価用の各複合名詞に対して、式 (1) で  $r = 0$  と設定して言い換えを生成した。各複合名詞に対する予測確率の最も高い 50 個の言い換えから、予測確率が閾値  $t$  以上の言い換えだけを残したものを最終的なシステム出力とした。ベースラインとして、各意味関係ラベルごとにそれと共起する言い換えの頻度を取り、意味関係ラベルに従って言い換えを頻度降順に  $k$  個並べて出力するモデルを用いた。テストデータは、複合名詞 181 個に対し、それぞれユニークな言い換えが平均 45.4 個が与えられている、合計 8,216 事例のデータである。

### 3.2.2 個別的な言い換えを狙った出力の評価

式 (1) で  $r = 10$  と設定したときの NC-wise, overall 法の出力を、それぞれ定義の異なる一般的な言い換えと個別的な言い換えに分けたうえで評価した。評価対象は、テストデータからランダムに選ばれた複合名詞 50 個に対する、それぞれ予測確率の最も高い 5 個の言い換えとした。

NC-wise 法における一般的な言い換えは SemEval-2013 のテストデータにおいて頻度 2 以上の言い換えで、個別的な言い換えはそれ以外である。NC-wise 法の一般的な言い換えは、テストデータに含まれるという定義上、その精度が必ず 1.000 となる。overall 法における一般的な言い換えはデータ中で  $r$  に 0 を与えられている言い換えで、個別的な言い換えはそれ以外である。次に、それぞれ決められた個別的な言い換えに対して、それが適切な言い換えであるかを人手で評価した。

## 3.3 結果

### 3.3.1 一般的な言い換えを狙った出力

表 1 に、先行研究と提案手法、ベースラインの評価値を示す。SemEval-2013 baseline は、10 種類の言い換えをすべての複合名詞に対して同じ順序で与えたモデルで、ベースライン ( $k = 10$ ) とほとんど同じ評価値である。ベースライン ( $k = 5$ ) では、適合率を評価する non-isomorphic での評価値が  $k = 10$  のときより大きく上がっている。学習データと閾値  $t$  を変えた 6 種類の提案モデルの中では、NC-wise 法 ( $t = .000$ ) の isomorphic における評価値が、同様に NC-wise 法 ( $t = .050$ ) の non-isomorphic における評価値が最も高くなった。

### 3.3.2 個別的な言い換えを狙った出力

表2に、提案手法における一般的／個別的な言い換えの生成割合とその適合率を示す。NC-wise法の出力では、一般的な言い換えはほとんど出現せず、個別的な言い換えは半分以上が適切であった。overall法の出力では、一般的な言い換えは全く出現せず、個別的な言い換えは半分弱が適切であった。

## 4 考察

### 4.1 一般的な言い換えを狙った出力

テストデータの各複合名詞に対して、NC-wise法 ( $t = .050$ ) が出力した言い換え数は平均3.82個、overall法 ( $t = .050$ ) は平均4.46個であった。これら提案モデルの出力と、ベースライン ( $k = 5$ ) の出力を比較する。ベースライン ( $k = 5$ ) の出力した言い換えは23種類、NC-wise法 ( $t = .050$ ) は48種類、overall法 ( $t = .050$ ) は29種類であった。提案モデルは  $w_M$  と  $w_H$  がほぼ0のときであっても、biLSTMの構造からMODとHEADの情報を考慮できるため、ベースラインより柔軟に入力された複合名詞に適した言い換えを生成できると考えられる。また、overall法は  $r$  が0 (つまり  $w_L$  が約1.0) のとき、データ中で  $r = 0$  と与えられている言い換えしかほとんど出力できなかったことから、NC-wise法と比べて性能が劣ると言える。

### 4.2 個別的な言い換えを狙った出力

NC-wise法の出力における各言い換えの頻度をとると、頻度上位は *H associated with M*, *H relating to M*, *H dealing with M* となった。これらの言い換えは、全体的な出現頻度は高いものの、個々の複合名詞に対する言い換えとしては頻度が低い。そのため、個別的な言い換えを生成したいとき、上述のような言い換えは偏って高い予測確率を与えられる傾向がある。

overall法の出力には、一般的と分類された言い換えは全く含まれていなかった。個別的な意味をもつ言い換えを適切に予測できた例として、*tax exemption* に対する *exemption given for tax payment*, *defence budget* に対する *budget spent on defence* などがあった。

### 4.3 意味関係ラベルへの重みと出力の傾向

$w_L$  の値を0から1まで0.1刻みで与え、各  $w_L$  の値におけるテストデータの複合名詞に対するoverall法による出力を得た。分析の対象は、各複合名詞に対するそれぞれ上位5個の出力に限った。

図2に、 $w_L$  各値における、各区分の言い換え出現

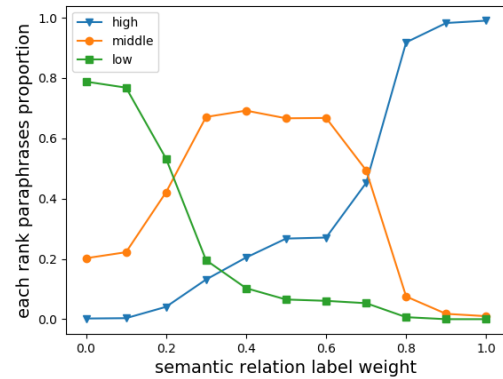


図2: overall法における、意味関係ラベルへの重み (横軸) と各区分の言い換え出現割合 (縦軸)

割合を示す。high (青線) が  $r = 0$ , middle (橙線) が  $r = 5$ , low (緑線) が  $r = 10$  を与えられた言い換えの出現割合である。 $w_L$  が大きいときはhighが大きく、中程度のときはmiddleが大きく、小さいときはlowが大きくなる。これより、overall法は複合名詞の出現割合をよく調整できると言える。

## 5 おわりに

本研究では、出力として得られる言い換え表現の曖昧さを調整可能な複合名詞言い換えモデルを提案した。実験の結果、曖昧さの大きい一般的な言い換えと曖昧さの小さい個別的な言い換えを分離して出力できることを示した。今後は、NC-wise法とoverall法の長所を合わせたモデルを考えたい。

## 参考文献

- [1] Cristina Butnariu, Su Nam Kim, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. Semeval-2010 task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions. *Proceedings of SemEval*, pages 15-16 (2010).
- [2] Iris Hendrickx, Zornitsa Kozareva, Preslav Nakov, Diarmuid O S eaghda, Stan Szpakowicz, and Tony Veale. SemEval-2013 task 4: Free paraphrases of noun compounds. *Proceedings of SemEval*, pages 138-143 (2013).
- [3] Paul Nulty, Fintan Costello. UCD-PN: Selecting general paraphrases using conditional probability. *Proceedings of SemEval*, pages 234-237 (2010).
- [4] Girishkumar Ponkiya, Kevin Pately, Pushpak Bhattacharyay and Girish K Palshikar. Treat us like the sequences we are: Prepositional paraphrasing of noun compounds using LSTM. In *Proceedings of COLING*, pages 1827-1836 (2018).
- [5] Vered Shwartz and Ido Dagan. Paraphrase to explicate: revealing implicit noun-compound relations. In *Proceedings of ACL*, pages 1200-1211 (2018).
- [6] Nitesh Surtani, Arpita Batra, Urmi Ghosh, and Soma Paul. II-ITH: A corpus-driven co-occurrence based probabilistic model for noun compound paraphrasing. *Proceedings of SemEval*, pages 153-157 (2013).
- [7] Stephen Tratz. Semantically-enriched parsing for natural language understanding. University of Southern California, (2011).
- [8] Tim Van de Cruys, Stergos Afantenos and Philippe Muller. MELODI: A supervised distributional approach for free paraphrasing of noun compounds. *Proceedings of SemEval*, pages 144-147 (2013).
- [9] Yannick Versley: SFS-TUE: Compound paraphrasing with a language model and discriminative reranking, *Proceedings of SemEval*, pages 148-152 (2013).