系列ラベリングによる自然言語文からの上位下位関係自動抽出

平松 淳 [1,a] 若林 啓 [2,b]

[1] 筑波大学 情報学群 知識情報・図書館学類，[2] 筑波大学 図書館情報メディア系

a. himkt@klis.tsukuba.ac.jp，b. kwakaba@slis.tsukuba.ac.jp

1 はじめに

意味を考慮した自然言語処理を行う上で，上位下位関係の抽出は重要な基盤技術である．上位下位関係とは，上位語・下位語の関係にある単語ペアに形成される関係であり．下位語を A，上位語を B として **isA (A, B)** と表される．上位下位関係を含んだ知識ベースの例として，WordNet[1] や DBpedia[2] が存在する．これらの知識ベースは辞書や知識源が人手で管理されており，信頼性の高いデータが蓄積されている．一方で，人手によるデータの管理には限界があり，自動化が求められている．

ウェブ上の資源から上位下位関係を抽出する手法は，半構造化データを活用する手法と自然言語文から抽出する手法の 2 種類に分かれる．半構造化データを活用する手法としては，Wikipedia の本文記事に含まれる階層構造から，機械学習を用いて上位下位関係を抽出する隅田ら [1] の研究がある．

自然言語文からの上位下位関係抽出は半構造化データからの抽出よりも難しいが，ウェブ上のデータの多くは自然言語で記述されているため，より多くのデータを抽出対象にできる．代表的な手法に表 1 に示すような定型表現を用いる安藤ら [2] の手法がある．

定型表現を用いる抽出手法には，定型表現の意味に関する曖昧性と複合語の抽出に関する問題がある．これらの問題を解決するために，本研究では，自然言語文からの上位下位関係抽出を系列ラベリング問題とみなす．抽出モデルとして，条件付き確率場を用いた手法と双方向 LSTM を用いた手法の 2 種類を提案する．

2 関連研究

単語ペアが入力されたとき，ペアが上位下位関係にあるかどうかを判定する分類器を構築する研究が活発に行われている [3][4]．分類器の素性には文の係り受け情報や単語の分散表現が用いられる．これらの手法は単語のペアを入力とするため，自然言語文から直接上位下位関係を抽出するためには，文から候補となる単語ペアを抜き出す必要がある．

自然言語文からの上位下位関係抽出手法として，上位下位関係を表す定型表現を用いてマッチングを行う手法が存在する [2][5]．例として「富士山などの山」という文には isA (富士山，山) という上位下位関係が含まれており，上位語と下位語の間を「などの」という定型表現が結んでいる．安藤ら [2] は定型表現の前後に出現する名詞を下位語・上位語とするルールによって上位下位関係を抽出している．

定型表現の前後の単語を抜き出す手法には，複合語の抽出ができないという課題がある．例えば「確率統計などの分野」という文からは **isA (統計, 分野)** という上位下位関係が抽出される．しかし，下位語は確率統計であるべきである．この問題に対しては，品詞に関するルールを加えることで対応することができる．品詞に関するルールとは，連続する名詞は 1 つの名詞句として，上位語・下位語の抽出対象を名詞ではなく名詞句とする，というものである．

また，名詞以外の品詞を含むような語を抽出するために，括弧で囲まれた形態素列は 1 つの名詞句とみなし，括弧を無視するルールも加える．これにより，「『君の名は。』という作品」という文章から **isA (君の名は。, 作品)** という上位下位関係を抽出できる．本研究では，定型表現ベースの抽出手法 [2] に上記の 2 つのルールを加えた手法をベースライン手法とする．

ベースライン手法は，ルールの追加により複合語の抽出が一部可能になった．しかし，括弧を含まず，名詞の連続ではないような複合語は依然として抽出できない．例えば，「トムとジェリーなどの作品」という文から **isA (トム，作品)**，**isA（ジェリー，作品）** という誤った上位下位関係を抽出してしまう．

複合語を抽出できる手法として，定型表現と統計的なフィルタリングを組み合わせた手法が存在する [6]．しかし，統計的なフィルタリングには大規模なデータが必要である．また，日本語の場合は上位下位関係を

[1] https://wordnet.princeton.edu
[2] http://wiki.dbpedia.org

表 1: 定型表現

下位語 **など** 上位語
下位語 **などの** 上位語
下位語 **という** 上位語
下位語 **のような** 上位語

表 2: 複合語を含む文のラベリング

確率	統計	など	の	分野
hypo_b	hypo_i	p_b	p_i	hyper_b

表 3: 上位下位関係を含まない文のラベリング

o	q_b	q_i	q_i	o	o	o
娘	の	よう	な	年齢	の	子供

表す定型表現に意味の曖昧性が存在する．「娘のような年齢の子供を見た」という文には定型表現「のような」が含まれるが，上位下位関係は含まれない．上位下位関係を表さない定型表現もコーパス中に大量に出現する場合，この問題は統計的なフィルタリングでは解決できない．

3 提案手法

3.1 ラベルの定義

表 1 で定義した定型表現は，上位下位関係以外の関係を表すことがあり，意味に曖昧性がある．実際に，「娘のような年齢」という文には上位下位関係が含まれない．ベースライン手法では定型表現の曖昧性を考慮することはできない．この問題に対して，提案手法では上位下位関係を表す定型表現と表さない定型表現を区別する．さらに，複合語を抽出するために自然言語文からの上位下位関係抽出を系列ラベリング問題とみなす．

上位下位関係を表す定型表現を p，上位下位関係を表さない定型表現を q と呼ぶ．また，上位語・下位語を hyper，hypo と呼び，対応するラベルを表 4 に定義する．o というラベルは上記のいずれにも該当しない形態素に付与されるラベルである．IOB タグセットの考えを用いて，フレーズの開始を表す b，フレーズの継続を表す i を付与している．

抽出対象の文が得られたら，まず形態素解析を行い形態素ごとに分割する．複合語を含む文に対するラベリングは表 2，上位下位関係を含まない文に対するラベリングは表 3 のようになる．抽出モデルとして，条件付き確率場によるモデルと双方向 LSTM によるモデルの 2 種類を提案する．

表 4: ラベル一覧

付与するラベル	対応する語
hyper_b	上位語の開始語
hyper_i	上位語の開始語以外
hypo_b	下位語の開始語
hypo_i	下位語の開始語
p_b	上位下位関係を表す定型表現の開始語
p_i	上位下位関係を表す定型表現の開始語以外
q_b	上位下位関係を表さない定型表現の開始語
q_i	上位下位関係を表さない定型表現の開始語以外
o	上記のいずれにも該当しない

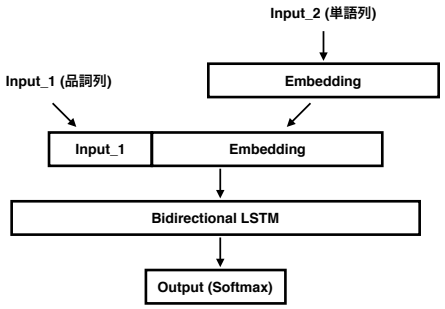

図 1: 双方向 LSTM を用いた提案手法のモデル構造

3.2 ラベルの推定法

3.2.1 条件付き確率場

条件付き確率場 [7] は系列ラベリング問題を解くための手法として有効な確率モデルである．条件付き確率場の素性には自分自身と前後 1 語の表層形と品詞タグを用い，実装には CRFsuite [8] を用いる．

3.2.2 双方向 LSTM

双方向 LSTM は深層学習の手法の 1 つであり，近年系列ラベリング問題で成果をあげている [9][10]．提案手法のモデルの構造を図 1 に示す．入力は品詞列と単語列であり，それぞれ Input_1 と Input_2 とする．Embedding 層と双方向 LSTM 層，出力層の 3 種類の層からなる．

Input_1 は品詞列を 1-of-K 符号化したものである．品詞タグは形容詞，助詞，名詞，接頭詞，助動詞，副詞，その他，連体詞，動詞，フィラー，記号，接続詞，感動詞および双方向 LSTM に入力する際に系列を揃えるために用いる padding の 14 種類である．

Embedding 層は Input_2 の単語列を受け取り，分散表現への埋め込みを行う層である．Embedding 層の重みには，実験コーパスに Skip-gram [11] を適用して得られる分散表現を用いる．また，Embedding 層の重みは学習によって更新しない．Skip-gram の実装は gensim [12] によって行い，negative sampling のパラメータは 5，分散表現の次元は 200 とする．

出力層は双方向 LSTM の出力を受け取りラベル列を 1-of-K 符号化したものを出力する．活性化関数は softmax とし，出力の次元は表 4 のラベルに padding

を加えたラベルの種類数を表す 10 である．学習の最適化アルゴリズムには Adam [13] を採用し，学習モデルの実装には Keras [14] を用いる．

3.3 ラベルの変換

文のラベリングを行ったのち，IOB タグセットの統合を行うことで上位下位関係に変換する．この処理により，表 2 のラベリング結果から **isA (確率統計，分野)** という上位下位関係を得る．q_b ラベルが付与された文からは上位下位関係の抽出を行わない．

4 実験

4.1 実験データ

まず，実験で使うデータについて説明する．オンライン百科事典 Wikipedia のアーカイブ[3]の 2016 年 7 月の本文記事 dump データを用いて実験コーパスを作成する．本文記事は MediaWiki 記法によってマークアップされているため，前処理として以下の要素を削除する．

- コメント文
- 見出し
- リンク
- 強調構文
- ブラケット
- HTML タグ

次に，教師データを作成する．得られた実験コーパスを文単位に分割すると，13,416,872 件の文を得る．得られた文から，定型表現を含み，文中の全単語がコーパス中に 5 回以上出現する文を無作為に抽出し，人手で表 4 のラベルを付与する．形態素解析器には MeCab[4]，辞書は IPA 辞書に Wikipedia の見出し語を追加したものを用いる．ラベルの付与は著者が一人で行い，上位下位関係を含む文を 314 件，上位下位関係を含まない文を 1,314 件含む教師データを作成した．

4.2 実験方法

定型表現を含む文に対して提案手法を適用し，表 4 で定義したラベルを推定する実験を行う．上位下位関係の抽出性能とラベルの推定性能を精度・再現率・F1 値によって評価する．評価する値は 10 分割交差検定のテスト平均とする．また，教師データの数と抽出性能の関係を明らかにするために，学習に用いる教師データの割合を変化させて同じテストデータに対する抽出実験を行う．上位下位関係を含まないデータの数が多いので，学習時はダウンサンプリングによって上位下位関係を含むデータの数と等しくなるようにする．

4.3 実験結果

ベースライン手法と提案手法の上位下位関係の抽出性能を表 5 に示す．提案手法がベースライン手法と比較して高い精度となっている．これは，上位下位関係を示さない定型表現に適切に q タグを推定することに成功していることが主な要因であるといえる．例えば，ベースライン手法は「次のような人物がいる」という文から **isA (次, 人物)** という誤った上位下位関係を抽出してしまうが，提案手法はこの誤抽出を避けることに成功している．一方，提案手法はベースライン手法と比較して再現率が低い．これは，上位下位関係を示す定型表現に誤って q タグを推定してしまう場合があることが主な原因であるといえる．例えば,「春と夏に行われる彼岸という行事」という **isA (彼岸, 行事)** という上位下位関係を含む文に対し，q タグを推定して上位下位関係の抽出に失敗した例がある．

複合語の抽出に成功しているかどうかを確認するた

表 5: 上位下位関係抽出の性能比較

手法	精度	再現率	F1 値
ベースライン手法	0.215	**0.361**	0.263
条件付き確率場	0.311	0.261	0.279
双方向 LSTM	**0.332**	0.308	**0.304**

表 6: 条件付き確率場のラベルの推定性能

ラベル	精度	再現率	F1 値
hyper_b	0.332	0.476	0.387
hyper_i	0.382	0.423	0.392
hypo_b	0.316	0.334	0.319
hypo_i	0.273	0.340	0.289
p_b	0.375	0.587	0.451
p_i	0.393	0.627	0.474
q_b	0.879	0.757	0.813
q_i	0.809	0.615	0.691

表 7: 双方向 LSTM のラベルの推定性能

ラベル	精度	再現率	F1 値
hyper_b	0.492	0.558	0.513
hyper_i	0.640	0.230	0.313
hypo_b	0.488	0.591	0.516
hypo_i	0.479	0.366	0.374
p_b	0.500	0.817	0.613
p_i	0.483	0.865	0.613
q_b	0.941	0.802	0.865
q_i	0.891	0.650	0.746

[3]https://dumps.wikimedia.org/
[4]https://taku910.github.io/mecab/

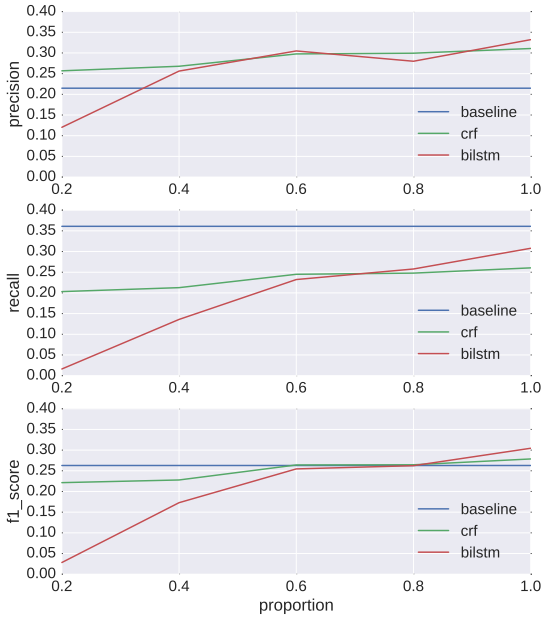

図 2: 教師データの比率と上位下位関係の抽出性能の変化 (0.2，0.4，0.6，0.8，1.0 は教師データの 20%，40%，60%，80%，100%用いて学習を行なったことを示す．)

めに，提案手法のラベルの推定性能を表 6 と表 7 に示す．双方向 LSTM を用いた手法のほうが条件付き確率場を用いた手法よりも良い推定性能を示している．どちらの提案手法も hyper_i と hypo_i の推定性能が悪く，複合語の抽出には依然として課題が残されている．また，q_b および q_i の推定には 80%の精度で成功している一方，他のラベルの推定精度は 40%程度にとどまっており，推定性能の向上が課題である．

教師データの数と上位下位関係の抽出性能の変化を図 2 に示す．提案手法は教師データを増加させると抽出性能が上昇する傾向にあることがわかる．

5 結論

本研究では，系列ラベリングを用いて自然言語文から上位下位関係を自動抽出する手法を提案した．自然言語からの上位下位関係自動抽出を系列ラベリング問題として定式化した．これにより，複合語を抽出できる問題設定となった．提案手法は条件付き確率場による手法と双方向 LSTM による手法の 2 種類を提案した．実験により，提案手法はベースライン手法よりも高精度な抽出が可能であることを示した．

また，教師データの数を変化させながら学習を行うことにより，教師データを増加させることで抽出性能が向上していることを明らかにした．教師データを増加させることでどこまで性能が向上するのかを確認することが今後の課題である．

謝辞

本研究の一部は，JSPS 科研費（課題番号 16H02904）の助成によって行われた．

参考文献

[1] 隅田飛鳥, 吉永直樹, 鳥澤健太郎. Wikipedia の記事構造からの上位下位関係抽出. 自然言語処理, Vol. 16, No. 3, pp. 3–24, 2009.

[2] 安藤まや, 関根聡, 石崎俊. 定型表現を利用した新聞記事からの下位概念単語の自動抽出. 情報処理学会研究報告, Vol. 2003, No. 98, pp. 77–82, 2003.

[3] Stephen Roller, Katrin Erk, and Gemma Boleda. Inclusive yet selective: Supervised distributional hypernymy detection. In *Proc. International Conference on Computational Linguistics*, pp. 1025–1036, 2014.

[4] Vered Shwartz, Yoav Goldberg, and Ido Dagan. Improving hypernymy detection with an integrated path-based and distributional method. In *Proc. Annual Meeting of the Association for Computational Linguistics*, pp. 2389–2398, Berlin, Germany, August 2016.

[5] Marti A Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proc. Conference on Computational linguistics*, pp. 539–545. Association for Computational Linguistics, 1992.

[6] Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q Zhu. Probase: a probabilistic taxonomy for text understanding. In *Proc. ACM SIGMOD International Conference on Management of Data*, pp. 481–492. ACM, 2012.

[7] John Lafferty, Andrew McCallum, and Fernando C N Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. International Conference on Machine Learning*, Vol. 8, pp. 282–289, 2001.

[8] Naoaki Okazaki. Crfsuite: a fast implementation of conditional random fields (crfs), 2007.

[9] Xuezhe Ma and Eduard Hovy. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *Proc. Annual Meeting of the Association for Computational Linguistics*, pp. 1064–1074, 2016.

[10] Peilu Wang, Yao Qian, Frank K. Soong, Lei He, and Hai Zhao. A Unified Tagging Solution: Bidirectional LSTM Recurrent Neural Network with Word Embedding. *CoRR*, 2015.

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In *Proc. International Conference on Learning Representations*, 2013.

[12] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50, 2010.

[13] Diederik P. Kingma and Jimmy Lei Ba. Adam: a Method for Stochastic Optimization. In *Proc. International Conference on Learning Representations*, pp. 1–15, 2015.

[14] François Chollet. Keras. `https://github.com/fchollet/keras`, 2015.