RNN 言語モデルを用いた日本語形態素解析の実用化

森田 一 †‡　　黒橋 禎夫 †‡
† 京都大学　‡ 科学技術振興機構 CREST

1　はじめに

入力文を単語に分割し，その品詞および活用を認識する日本語形態素解析は，言語処理を行う上で欠かせない重要な処理である．しかし現状では，後段の構文・格解析や文脈解析の誤りが形態素解析の誤りに起因するということが少なくない．

我々は，Wikipedia や Wiktionary 等から大規模語彙の獲得を行い，Recurrent Neural Network Language Model (RNNLM) の導入により，形態素解析の大幅な精度向上を達成した [1]．本稿ではさらに，RNNLM の学習に Noise Contrastive Estimation (NCE) [2] を用いることによる形態素解析の高速化と，解析誤りを修正するための部分アノテーションによる学習機構の追加を行った．解析誤りの詳細な分析・分類を行ったところ，後段の解析に悪影響を及ぼす誤りは 1-best 解で 1,000 文あたり 20 箇所程度，そのうち 5-best 解で正しい解釈が含まれないものは 10 箇所程度となり，実用上十分な精度が達成されたと考えられる．

2　RNNLM を用いた日本語形態素解析

RNNLM を用いた形態素解析のモデルでは $\mathbf{y}$ を単語列，s を入力文，$\mathcal{Y}(s)$ を入力文に対する全ての単語列の候補としたとき，次式を満たす単語列 $\hat{y}$ を求めることにより解析を行う，

$$\hat{y} = \underset{\mathbf{y} \in \mathcal{Y}(s)}{\operatorname{argmax}} \ \operatorname{score}(\mathbf{y}). \quad (1)$$

ここで用いるスコア関数 score($\mathbf{y}$) は，次式で表される:

$$\operatorname{score}(\mathbf{y}) = (1-\alpha)\Phi(\mathbf{y}) \cdot \vec{w} + \alpha \ \log(p_r(\mathbf{y})). \quad (2)$$

ただし，α を線形補間の重み，$\Phi(\mathbf{y})$ を単語列 $\mathbf{y}$ に対する素性ベクトル，$p_r(\mathbf{y})$ を RNNLM が単語列に与える確率，$\vec{w}$ を重みベクトルとし，$\vec{w}$ は訓練データを用いた教師あり学習により決定する．解析に用いる RNNLM は，自動解析した大規模ウェブコーパスで学習を行った後，自動解析による誤りの影響を軽減するため，こちらも訓練データを用いて再度学習を行う．素性等の詳細は文献 [1] を参照されたい．

我々の文献 [1] では，RNNLM の計算において Class-Based Softmax [3] を用いて確率の正規化を行っており，非常に大きな計算コストを必要としていた．本稿では，正規化項の計算を省略可能にする NCE[2] を用いることにより (実装は faster-rnnlm* を利用) RNNLM の計算を約 10 倍高速化した．これにより，後段の処理として想定する構文・格解析 (KNP†) と比べた解析時間が 1/5 程度となり，実用上必要な解析速度を達成したと考えている ‡．

3　部分アノテーションを用いた学習

実際に形態素解析を利用するうえで，解析誤りが生じることは避けられないが，解析の誤りは発見されしだい，随時修正されることが望ましい．その時，解析を修正するもっとも素朴な方法は，解析誤りのあった文に対して正しい単語列をアノテーションし，訓練データに追加する方法である．しかし，適切に文をアノテーションする作業は専門的な知識を必要とするため，高い人的コストがかかる．

本稿では，明らかな解析の誤りについては専門的な知識がなくとも修正できるように，部分アノテーションとして，解析を誤った箇所の単語境界のみを人手で与え，与えた単語境界を制約として解析した結果を訓練データに追加することにより，誤った解析を修正する仕組みを実装した．部分アノテーションにより解析を修正する効果を，次節でエラー分析とともに検証する．

4　評価およびエラー分析

ここでは，JUMAN§，MeCab¶，RNNLM を用いた形態素解析モデルについて精度による評価を行う．また，JUMAN および RNNLM を用いた形態素解析モデルの形態素解析の解析誤りを以下の 4 種に分類し，分析を行う ‖．

許容できる誤り

- **基準の違い**: コーパス・アノテーションと複合語の分割や品詞が違うが，解釈の誤りとはいえないもの．
  ex. | 北極/点 ← 北極点 |
  | 旧 (名詞 ← 接頭辞)| ソ連
- **意味的曖昧性に起因する誤り**: 文法的に問題のない単語列に分割されていて，形態素解析では区別しづらい

Development of Practical Japanese Morphological Analysis using Recurrent Neural Network Language Model
Hajime Morita †‡　Sadao Kurohashi †‡
† Kyoto University
‡ CREST, Japan Science and Technology Agency

*https://github.com/yandex/faster-rnnlm
†http://nlp.ist.i.kyoto-u.ac.jp/?KNP
‡しかし，従来の形態素解析器 (JUMAN) と比べると，正規化計算を省略した場合でも約 100 倍の解析時間を要しており，解析速度の改善は今後の重要な課題の 1 つである．
§http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
¶http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
‖誤りの例ではアノテーションと異なる箇所を | 解析結果 ← アノテーション | というフォーマットで表記する．

		1-best			5-best	
		JUMAN	+RNNLM	+部分アノテーション	+RNNLM	+部分アノテーション
許容できる誤り	基準の違い	203	139	138	–	–
	意味的曖昧性に起因する誤り	42	29	27	8	8
許容できない誤り	未知語，複合語の分割誤り	39(27)	12(3)	10(1)	12(3)	9(0)
	その他の誤り	28	13	8	3	1

表 2: 分析用データ (995 文) での解析誤りを分類した内訳．連続した形態素の解析誤りは 1 箇所の誤りとして数える．未知語，複合語の分割誤りでは，括弧内の値は未知語に起因する誤りを除いた内数を表す．

	分析用データ	精度評価用データ
JUMAN	97.89	97.91
MeCab	97.99	98.00
Base	97.77	97.58
+RNNLM	98.45	98.35
+部分アノテーション	98.52	98.44

表 1: 単語分割・品詞推定の精度 (F 値)

意味的な曖昧性がある場合．
ex. 単身赴任のようと | よく (形容詞 ← 副詞) | 言われる
さかのぼって | みる (接尾辞 ← 動詞) |

許容できない誤り

- **未知語，複合語の分割誤り**
  ex. | 北大/西洋 ← 北/大西洋 | 条約
  薄日が | 射/して ← 射して |
- **その他の誤り**
  ex. | おす/す/めな ← お/すすめ/な |
  増加の |一途(いちず)で(形容詞) ← 一途(いっと)(名詞)/で (助詞) |

実験には京都大学テキストコーパス [4]，京都大学ウェブ文書リードコーパス [5] を合わせて用い，訓練データ，分析用データと精度評価用データに分けて利用する *．以降，我々のモデルから RNNLM を除いたものを Base，RNNLM を用いた形態素解析を+RNNLM と表記する．比較する各モデルでは，Wikipedia 等から獲得した大規模語彙を辞書として用いた．

また，部分アノテーションによる解析誤り修正の効果を調べるため，分析用データ中の許容できない誤りのうち，+RNNLM の 5-best で正しく解析出来なかった誤りに対する部分アノテーションを行った．ただし，品詞の解析誤りと語彙の不足による解析誤りは単語境界を与えるだけでは正しく解析できないため，部分アノテーションの対象から除外した．部分アノテーションを与えた解析結果を訓練データに加え，再度学習を行ったモデルを+部分アノテーションと表記する．

分析用データと精度評価用データのそれぞれで評価を行った結果を表 1 に示す．JUMAN，MeCab, Base と比較して+RNNLM では大きく精度が向上している．さらに，+部分アノテーションを+RNNLM と比較すると，アノテーションを与えていない精度評価用データでも改善が見られ，精度を下げることなく解析誤りを修正できていることが分かる．

次に分析用データでエラーの分類を行った結果を表 2 に示す．後段の構文・格解析で解析結果の N-best を利用する場合を考え，5-best 出力を考慮した場合の解析誤りの数を示す（ただし，基準の違いによるものは N-best で解決するべきものでは無いためここでは省く）．1-best の JUMAN と+RNNLM を比べると許容できないエラーの数は大きく減少している．特に 5-best を考慮した時には，未知語よる解析誤りを除きほとんどの場合に正しい解析結果を提示できており，言語モデルを用いて解くべき問題は十分に解けているといえる．また，部分アノテーションを与えた箇所は 5-best の+部分アノテーションでは全て正しく解析できるようになっており，部分アノテーションが有効に機能した結果，許容できない誤りの数は更に減少している．

5 まとめ

本稿では RNN 言語モデルを用いた日本語形態素解析の実用化に向けた課題について述べた．解析誤りの詳細な分析・分類を行ったところ，後段の解析に悪影響を及ぼす誤りは 1,000 文あたり，5-best 解で正しい解釈が含まれないものは 10 箇所程度となり，実用上十分な精度が達成されたと考えられる．また現状のほとんどの解析誤りは未知語に起因していることが明らかになった．現在，解析の高速化に加え語彙のさらなる拡張を行っており，これらの誤りも解決される見通しである．

参考文献

[1] H. Morita, D. Kawahara, and S. Kurohashi. Morphological analysis for unsegmented languages using recurrent neural network language model. In *Proceedings of EMNLP 2015*, pages 2292–2297, 2015.

[2] X. Chen, X. Liu, M. J. F. Gales, and P. C. Woodland. Recurrent neural network language model training with noise contrastive estimation for speech recognition. In *Proceedings of ICASSP 2015*, pages 5411–5415. 2015.

[3] T. Mikolov, A. Deoras, D. Povey, L. Burget, and J.H. Cernocky. Strategies for training large scale neural network language models. In *Proceedings of ASRU 2011*, pages 196–201, 2011.

[4] D. Kawahara, S. Kurohashi, and K. Hasida. Construction of a Japanese relevance-tagged corpus. In *Proceedings of LREC-2002*, pages 2008–2013, 2002.

[5] M. Hangyo, D. Kawahara, and S. Kurohashi. Building a diverse document leads corpus annotated with semantic relations. In *Proceedings of PACLIC 2012*, pages 535–544, 2012.

*49,774 文を訓練データ，995 文を分析用データ，2,983 文を精度評価用データとした．