ニューラルネットワークに基づく 単語分割・品詞付与・構文解析の統合解析

栗田修平　　河原大輔　　黒橋 禎夫
京都大学大学院 情報学研究科
kurita@nlp.ist.i.kyoto-u.ac.jp
{dk, kuro}@i.kyoto-u.ac.jp

概 要

近年、ニューラルネットワークに基づく遷移ベースの依存構文解析モデルが提案されている。なかでも、単語の代わりに文字入力を用いたモデルが構文解析単独のタスクで高い精度を達成している。しかしながら、中国語のような単語分割が難しい言語では、文字情報を用いるのであれば、単語分割とは独立に依存構文解析を行うより、構文木作成を単語分割と同時に行うほうが、単語分割および構文解析の双方の結果がよくなることが、すでに知られている。本研究では、ニューラルネットワークに基づく手法で文字や文字列のもつ固有の情報を捉え、単語情報とも合わせて、単語分割、品詞付与、構文解析を同時に行うモデルを提案する。

1　はじめに

自然言語処理は、単語分割、品詞付与、構文解析、それらを利用する高度な処理というように流れ作業的に行われる。通常は、個別の処理ごとに高精度で処理を行うモデルが開発され、それらをパイプラインでつないで処理を行う。しかし、個別の処理の精度が高いときはこの手法でもうまく働くが、ひとたび前段の処理においてエラーが生じると、エラーは修復されずに後段の処理に伝播していくという性質を持つ。この問題は、スペースのような単語間の境界を示す記号を用いない言語において顕著である。中国語やアラビア語などの言語では、現在でも精度の良い単語分割は難しい課題であり、結果的に、品詞付与や構文解析、単語分割を利用するあらゆる自然言語処理タスクにおける誤りの原因の一つとなっている。とりわけ中国語の単語分割は、単語間の境界が明示されず、また中国語の単語そのものが明確に定義されていない点で、悪名高く難しいことが知られている。

エラー伝播問題に対する有効な解決策は統合モデルを使うことである。中国語に対する単語分割、品詞付与、構文解析の統合モデルは、Hatori *et al*., (2012) により提案された [1]。Hatori *et al*., (2012) によるモデルは、遷移ベースの構文解析アルゴリズムである Arc-Standard [2] を改良、文字に対する遷移を追加し、単語分割されていない入力文を、単語分割済みの構文木へと直接的に変換する。これにより、構文解析の精度が改善されるのみならず、構文解析の情報が単語分割の精度をも改善することを Hatori らは確認した。Zhang *et al*., (2014) は Hatori *et al*., (2012) のモデルを改良し、単語分割と構文解析の間に単語の “deque” を挟むことで統合構文解析モデルを作成した [3]。ただし、彼らのモデルでは、単語分割に関する遷移と構文解析に関する遷移のどちらを適用するかはルールによって決められており、構文木生成より先に単語の境界を定めている。これらのモデルはいずれも、単語および文字素性入力はバイナリの素性ベクトルで与えられている。

遷移ベースの構文解析の分類器として、近年はバイナリの素性入力や人手で作られた辞書情報の代わりに、単語や文字の分散表現を用いるニューラルネットワーク構文解析モデルが提案され、単語分割済みの文を入力文とする構文解析モデルとして成功を収めている [4, 5, 6, 7]。しかし、単語分割を含めた統合構文解析モデルは提案されておらず、特に中国語のような単語分割が困難な言語について、ニューラルネットワークに基づく統合構文解析手法の提案が望まれている。単語分割、品詞付与、構文解析の統合解析モデルは、構文解析開始時には、決定された単語境界を持つ単語の情報を使用できず、使用できるのは文字情報に限られる。一方で、利用できる単語情報は構文解析が進行するに従い増加し、同時に、単語分割の途中の部分単語や単語分割エラーによって生まれる不完全な単語も生じる。したがって、ニューラルネットワークモデル

の入力としては、文字情報と単語情報の双方が使用できるほか、不完全な単語の分散表現まで考慮する必要がある。

特に、中国語や日本語における問題として、これらの言語文はそれ自体が固有の意味を持つ大量の漢字により構成され、単語の中の部分文字列や、辞書にない不完全な文字列であっても、十分に固有の意味を持ちえる。例えば、「物理学」は中国語構文解析データセットである CTB において一単語として登録されているが、「物理」のみでも類似した意味を持つ中国語の単語になる。ゆえにニューラルネットワークに基づく統合構文解析を行うには、事前学習された単語や文字の分散表現の他、構文解析中に動的に生成される様々な文字の組み合わせの分散表現も利用する必要がある。

2 モデル

本論文のモデルは遷移ベースの構文解析モデルとニューラルネットワークによる分類器モデルに大別される。構文解析モデルは Hatori *et al*., (2012) によるものを使用し、ニューラルネットワークは Chen *et al*., (2014) および Andor *et al*., (2016) をベースにした。ニューラルネットワークの入力には、事前学習済みの分散表現リストに存在しない不完全な単語の分散表現を含む、あらゆる文字列の分散表現に対応できるようにした。このニューラルネットワークは、当然のことながら、一つの遷移ごとに前回の遷移に用いた情報をすべて消去し、次の遷移に関する手がかりを、ニューラルネットワークへの入力として受け取る。前回までの遷移の情報は、新たな入力して与えられ、自身の出力はニューラルネットワークの枠組みの外側で文に対する操作として実行され、操作に応じた状態の変更が生じる。すなわち、この種のニューラルネットワークは、外部に自身の情報を記録しておく仕組みを持ちえ、かつ、その記録単位である分散表現には、学習によって変更が加わる。

2.1 統合構文解析アルゴリズム

統合構文解析は Hatori *et al*., (2012) によるアルゴリズムを使用した。モデルは入力文字列からなる buffer と、word および word の child words が収められる stack からなる。buffer は文字が格納され、stack には単語、品詞、構文木が格納される。遷移は Arc-Standard に基づくが、以下のように単語分割に関する操作が加えられる。

- `SH(t)`：buffer の最初の文字を、新しい単語の始まりとして stack の先頭に置く。`t` は品詞をあらわし、この操作は品詞の数だけ存在する
- `AP`：buffer の最初の文字を、stack の先頭の単語に付け加える
- `RR`：stack の先頭 2 単語の右側の単語を消去し、左側の単語の子ノードとする
- `RL`：stack の先頭 2 単語の左側の単語を消去し、右側の単語の子ノードとする

品詞は `SH(t)` により付与されており、CTB-5 において品詞は 35 種類存在するため、遷移は 38 種類存在する。

本研究では、この遷移ベースの統合構文解析手法を、ビームサーチを使用しない場合と、使用した場合の両方について学習させた。ビームを用いる場合、Hatori *et al*., (2012) の統合解析モデルでは独特の並び替えモデルを使用している。遷移における `AP` のステップサイズを 2 とすることで、文字数 n の文の可能なすべての構文木に対する遷移数を ROOT に関する遷移を除き $2n-1$ にした。この研究でもそれを踏襲した。

2.2 分散表現

統合構文解析モデルでは、モデルの入力が既知の単語であるか不完全な単語であるかは事前にはわからず、構文解析中に動的に切り替えられる必要がある。そこで、事前学習された単語の分散表現に存在しない文字列については、本研究では動的に生成する。モデルは、入力された文字または文字列が分散表現の中に存在するか調べ、存在する場合はそれを使用し、存在しない場合は、文字列を構成する文字の分散表現の平均を使用する。これらはニューラルネットワークの計算グラフの中で切り替えられ、分散表現まで学習される。

単語と文字の分散表現の事前学習には、単言語の大規模コーパスを用いた。単語と文字の双方が埋め込まれたベクトル空間を生成するために、分散表現を学習するコーパスを単語分割したものと文字分割したものを用意し、それらを結合した上で word2vec を用いて学習した。また、学習セットに含まれる単語のうち、大規模コーパスに存在しない単語の分散表現は、ランダムに生成されて事前学習された分散表現に追加された。単語、および文字のベクトルサイズはいずれも 200 次元で、これらは同一のベクトル空間に埋め込まれた。単語と文字合わせて頻度順に 1M 語彙を利用した。なお、Hatori *et al*., (2012) が用いたような Wikipedia による辞書情報は使用していない。

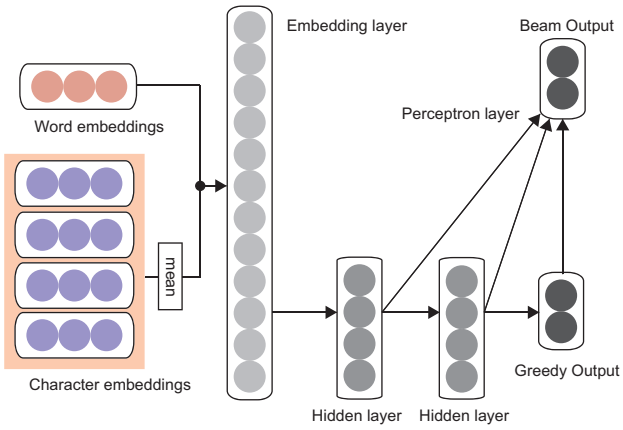

図 1: 本研究のニューラルネットワーク。再帰構造を含まないニューラルネットワークである。左端から文字列や単語の分散表現を受け取り、右端で構文解析アルゴリズムの遷移の確率を出力する。右端下側はビームを用いない greedy な学習の出力であり、右端上側はビームを用いる学習の出力である。ビームを用いる学習は、隠れ層からも入力を受け取る。

2.3 素性

素性は、Hatori *et al.*, (2012) が用いたものおよび Chen *et al.*, (2014) が用いたものを合わせたものを用いたが、その他、単語分割に関する素性を加えた。具体的には stack や buffer にまたがる素性や buffer に存在する単語を先読みする素性である。また、Hatori *et al.*, (2012) らが用いた素性は、複雑な素性共起に基づいているが、本研究では、素性の共起は取り除き、単語もしくは文字の分散表現の形で与えた。

2.4 ニューラルネットワーク

本研究では、ビームサーチを使用しない学習と、ビームサーチによる学習の両方を行う。ビームサーチを行わないニューラルネットワークモデルは、2 層の隠れ層をもつフィードフォワードニューラルネットワークである。モデルの概略を図 1 に示す。活性関数には ReLU を使用し、最終層には softmax を使用した。また、$L2$ 罰則項を dropout と併用した。最適化には Adagrad を用いた。学習に際し、トレーニングセット中の文は、順番をシャッフルした後に、いくつかの mini-batch にわける。一つの mini-batch 内の文は、全て同時に同じステップで処理される。テスト時にも、同様にテストセット中の文が全て同時に処理される。これにより、GPU メモリの容量が許す限り多くの文を同時に学習、処理することが出来るようになる。

CTB5	#snt	#wrd	#oov
Train	18k	494k	-
Dev.	350	6.8k	553
Test	348	8.0k	278

表 1: CTB5 data statistics.

ビームサーチは、ディープラーニング以前の中国語の統合構文解析において、とりわけ重要な役割を果たしていた。Hatori *et al.*, (2012) ではビームサイズが 64 において、最高のスコアを達成している。本研究のモデルの学習においても、ビームサーチを利用した。使用した手法は Andor *et al.*, (2016) のものである。このコスト関数は以下のようである。

$$L(d^*_{1:j};\theta) = -\sum_{i=1}^{j}\rho(d^*_{1:i-1}, d^*_i;\theta) + \ln\sum_{d'_{1,j}}\exp\sum_{i=1}^{j}\rho(d'_{1:i-1}, d'_i;\theta).$$

$d_{1,j}$ は遷移経路であり、$d^*_{1,j}$ は正解の遷移経路である. この学習は、ニューラルネットワーク全体に対して行うことができるが、実際の学習においては、GPU 上で再現できるモデルサイズの制約から、まず、最終層のみの学習を先に行う。次に全体を通した backprop を行う。場合によってはこれらを繰り返す。これらは、モデルサイズ及び学習時間の削減のためである。なお、ビームサーチの結果、batch が大きくなりすぎて GPU メモリに載らない文については、学習対象から外した。

3 実験

実験には CTB-5 を用いた。分割の統計を表 1 に示す。評価は Hatori *et al.*, (2012) と Zhang *et al.*, (2014) に従い、標準的な単語レベルの評価を F1 値で行った。品詞付与と依存構造は、関連する単語が正しく分割されていない限り正解とはならない。依存構造の評価には UAS を用いた。

実験では SegTag および SegTagDep,Dep の 3 種類の構文解析モデルを実験した。それぞれ、単語分割と品詞付与、フルジョイント、単独の依存構造解析である。

3.1 結果

3.1.1 統合単語分割、品詞付与、依存構造解析モデル

表 2 は SegTagDep の結果である。SegTagDep モデルは Andor *et al.*, (2016) の方法により、まず greedy な

Model	Seg	POS	Dep
Hatori+12	97.75	94.33	81.56
M. Zhang+14 EAG	97.76	94.36	**81.70**
SegTagDep (greedy)	98.24	94.49	80.15
SegTagDep	**98.37**	**94.83**	81.42

表 2: 統合単語分割、品詞付与、構文解析モデル。Hatori *et al*., (2012) の CTB5 スコアは Zhang *et al*., (2014) のものを用いた。Zhang *et al*., (2014) の EAG は Arc-Eager 解析器を示す。先行研究ではビームサイズは 64 である。

Model	Seg	POS	Dep
Hatori+12	97.75	94.33	81.56
M. Zhang+14 STD	97.67	94.28	81.63
M. Zhang+14 EAG	97.76	94.36	81.70
Y. Zhang+15	98.04	94.47	82.01
SegTagDep	98.37	94.83	81.42
SegTag+Dep	**98.41**	**94.84**	**82.36**

表 3: SegTag と Dep のパイプラインモデル。End-to-end モデルではない Zhang *et al*., (2015) の結果も示す [8]。

手法で学習された後に、ビームサーチを用いて学習された。表 2 の値は Zhang *et al*., (2014) から引用した。提案手法によるモデルは、単語分割と品詞付与において、既存の end-to-end モデルを上回る性能をみせているが、依存構造解析のスコアはこれらの既存手法に僅かに劣る。

3.1.2 統合解析モデルのパイプラインモデル

次に統合単語分割品詞付与モデルと依存構造解析モデルのパイプライン、すなわち、SegTag と Dep モデルのパイプラインについて結果を示す。このモデルについて SegTag は簡単のため greedy な手法で学習され、対し、Dep はビームを用いて学習された。ビームサイズは 4 である。結果、SegTag+Dep モデルは依存構造解析とその他のスコアにおいて最高のスアを達成した。SegTag+Dep モデルがフルジョイントモデルである SegTagDep よりも優れたスコアを達成したのは、構文木構造を含むビームの順位付けの困難さにあると思われる。同様に、Zhang *et al*., (2014) によるモデルも、“deque” を設けて SegTag と Dep モデルをモデル内で分離することで、高いスコアを得ている。本研究では、stack や buffer にまたがる素性や buffer に存在する単語を先読みする素性を SegTagDep に加えたが、SegTag+Dep のスコアの方が優れていた。

4 結論

中国語の統合構文解析において、単語列の分散表現を提案し、SegTagDep モデルの有効性を確かめ、SegTag と Dep モデルのパイプラインによって、既存手法よりも優れた解析精度を達成した。一方、モデルはあくまで素性抽出に依存し、文字列の意味を捉えるためには単純な算術平均を用いたが、これらは LSTM などを用いたモデルに置き換えることでよりよいスコアが期待できる可能性がある。これについては将来の課題とする。

参考文献

[1] Jun Hatori, Takuya Matsuzaki, Yusuke Miyao, and Jun’ichi Tsujii. Incremental joint approach to word segmentation, pos tagging, and dependency parsing in chinese. In *Proceedings of ACL*, 2012.

[2] Joakim Nivre. Incrementality in deterministic dependency parsing. In *Proceedings of the ACL Workshop*, 2004.

[3] Meishan Zhang, Yue Zhang, Wanxiang Che, and Ting Liu. Character-level chinese dependency parsing. In *Proceedings of ACL*, 2014.

[4] Danqi Chen and Christopher Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of EMNLP*, 2014.

[5] David Weiss, Chris Alberti, Michael Collins, and Slav Petrov. Structured training for neural network transition-based parsing. In *Proceedings of ACL and IJCNLP*, 2015.

[6] Miguel Ballesteros, Chris Dyer, and Noah A. Smith. Improved transition-based parsing by modeling characters instead of words with lstms. In *Proceedings of EMNLP*, 2015.

[7] Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. Globally normalized transition-based neural networks. In *Proceedings of ACL*, 2016.

[8] Yuan Zhang, Chengtao Li, Regina Barzilay, and Kareem Darwish. Randomized greedy inference for joint segmentation, pos tagging and dependency parsing. In *Proceedings of IJCAI*, 2015.