係り受け木上の不要な情報にマスクを行う関係分類

辻村 有輝　　三輪 誠　　佐々木 裕
豊田工業大学
{sd18602, makoto-miwa, yutaka.sasaki}@toyota-ti.ac.jp

1 はじめに

文中のエンティティ間に成立している関係を識別する関係分類は，知識グラフの構築や質問応答システムなど，より上位の言語処理アプリケーションに応用される．関係分類を行う際に重要な情報の多くは，係り受け木における対象エンティティペアを結ぶ最短経路上に存在するという知見がある．これに基づき，最短経路上の情報のみを利用することで，不要な情報に対する過学習を回避し，高性能な関係分類を実現できることが報告されている．しかし，この知見は経験則であり，最短経路外に重要な情報が存在する場合に対応できないという問題点がある．

そこで，本研究では，関係分類に不要なトークンの決定則をデータから自動獲得することを目指し，不要なトークンの判断を学習する機構を提案する．この機構は関係分類予測に対する損失を元に学習を行い，Gumbel softmax を用いることでトークンの利用・不利用という離散的な挙動を表現する．提案手法を SemEval-2010 task 8 データセット [1] で評価したところ，最短経路の経験則を用いる場合と同程度の識別性能となった．また，学習されたマスクは，最短経路とよく似たものとなり，最短経路の有用性を数値的に再確認する結果となった．

2 関連研究

2.1 最短経路を用いた関係抽出

Bunescu らは文における関係分類において，重要な情報はその文の係り受け木上における，対象とするエンティティペア間を結ぶ最短経路上によく現れるという経験則に基づいた，最短経路カーネル [2] を提案し，これを用いてサポートベクターマシンによる関係分類の性能向上を達成した．近年提案されたニューラルネットワークによる関係分類モデルにおいても，最短経路上のトークンのみを利用することによって性能の向上が報告されている [3, 4]．三輪らの提案した LSTM-ER モデル [4] は，このような最短経路の経験則を用いるニューラル関係分類モデルのひとつである．このモデルは，系列 LSTM (Long Short-Term Memory) に木構造 LSTM を積み上げたモデルとなっており，木構造 LSTM において最短経路上のトークンのみで計算を行うことで，全体木を用いる場合よりも性能が向上することが報告されている．

2.2 Gumbel softmax

Jang らは，確率変数についての勾配を計算できる離散的な値をサンプリングする手法である Gumbel softmax[5] を提案した．これによって，訓練中もネットワーク内で離散的なサンプリングを行いつつ，勾配法による学習を可能にしている．

3 提案手法

最短経路の経験則を利用せず，文中の全ての位置から重要なトークンを学習を通じて選択し利用するニューラル関係分類モデルを提案する．このモデルは三輪らの LSTM-ER モデルをベースモデルとし，木構造 LSTM 層で利用するトークンを選択するマスク機構と，マスクされた入力を元に計算された木構造 LSTM 層からの出力を集約するアテンション機構を導入したモデルとなっている．マスク機構は Gumbel softmax を利用しており，これによって学習中に利用する・しないという離散的な判断について勾配法による学習を行う．提案モデルの概観を図 1 に示す．

以降，まず 3.1 節において全体木をそのまま用いるベースモデルについて述べる．次に 3.2 節でマスク機構と，そのマスク量の増進のためのノイズの加算について説明し，最後にアテンション機構について述べる．

3.1 ベースモデル

ベースモデルは，入力文の不要な情報の判断を行わず，全体木をそのまま用いるモデルとなっており，三輪らの LSTM-ER モデル [4] のうち，関係分類を行う

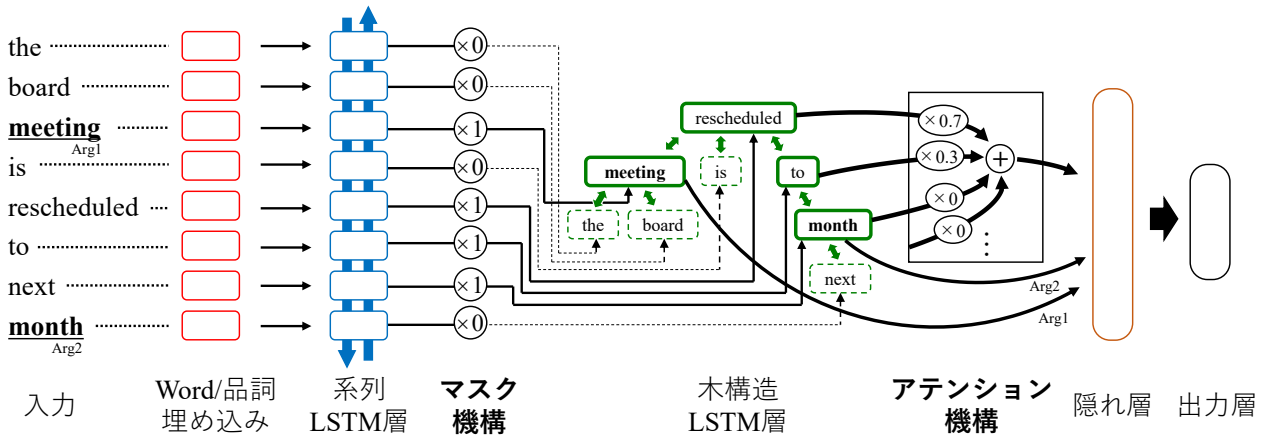

図 1: 提案モデルの概観.

部分のみのモデルを元にしている．このベースモデルは，入力層・双方向系列 LSTM 層・双方向木構造 LSTM 層・隠れ層・出力層の 5 層からなるニューラル関係分類モデルとなっている．LSTM-ER モデルとの違いとして，次の 4 点が挙げられる．まず，利用する木構造 LSTM には Child-Sum Tree-LSTM を採用する．これは，三輪らが用いた木構造 LSTM は，最短経路かどうかに応じて計算に用いるパラメータを変える構造となっており，これをそのまま使ってしまうと最短経路の経験則を利用してしまうためである．次に，LSTM-ER モデルでは，論文中で Pair として参照される，系列 LSTM 層からの出力を木構造 LSTM 層をスキップして直接隠れ層へ入力する結合があるが，本研究で用いるモデルではこの結合を利用しない．これは，提案手法が木構造 LSTM 層に影響するものであり，その影響をより直接観測するためである．さらに，系列 LSTM 層では peephole 結合を備えた LSTM を用い，モデル内の全層に対してドロップアウトを適用する．最後に，LSTM-ER モデルでは系列 LSTM 層の出力は順方向と逆方向 LSTM のそれぞれの出力を結合したベクトルであったが，ベースモデルではこれに加え，各トークンの親トークンへの係り受けラベルの埋め込みと，対象エンティティからの相対位置の埋め込みも併せて結合したベクトルとする．

3.2 マスク機構

対象エンティティペア間の関係分類において重要なトークンを判断し，不要なトークンを除去するマスク機構を提案し，前節のベースモデルに導入する．この機構は木構造 LSTM 層への入力に対しマスクをかけることで不要な情報をマスクする．初めに i 番目のトークンがどの程度対象エンティティペアの関係分類において重要かを表すスコア c_i を $[0, 1]$ の範囲で計算する．このスコアの計算式は Vaswani らのマルチヘッドアテンション [6] の構造を参考にしており，以下の計算式によって行う．

$$
\begin{aligned}
\boldsymbol{v}_{mq} &= \frac{\boldsymbol{v}_{\mathrm{arg1}} + \boldsymbol{v}_{\mathrm{arg2}}}{2} \\
\boldsymbol{h}_{mq} &= \mathrm{LayerNorm}(W_{mq}\boldsymbol{v}^{mq}) \\
\boldsymbol{h}_{mk} &= \mathrm{LayerNorm}(W_{mk}\boldsymbol{v}_i) \\
c_i &= \sigma\left(\frac{\boldsymbol{h}_{mq}{}^{\top}\boldsymbol{h}_{mk}}{\sqrt{d_m}}\right)
\end{aligned}
\tag{1}
$$

ここで $\boldsymbol{v}_i$ は i 番目のトークンの系列 LSTM 層からの出力で，arg1 と arg2 はそれぞれ対象エンティティを表す．W_{mq}，W_{mk} は重み行列である．LayerNorm は Layer Normalization である．

訓練中は，i 番目のトークンに対するマスク m_i を，Gumbel softmax によって以下の計算式でサンプリングする．

$$
\begin{aligned}
u_{i0}, u_{i1} &\sim \mathrm{Uniform}(0, 1) \\
l_{i0} &= \frac{\log(1 - c_i) - \log(-\log(u_{i0}))}{\tau} \\
l_{i1} &= \frac{\log(c_i) - \log(-\log(u_{i1}))}{\tau} \\
m_i &= \frac{\exp(l_{i1})}{\exp(l_{i0}) + exp(l_{i1})}
\end{aligned}
\tag{2}
$$

ここで τ は正の値をとる温度パラメータで，0 に近いほどサンプリングされる m_i も 0 か 1 に近付きやすくなり，訓練中は徐々に減少させ 0 に近づける．サンプリングされたマスク m_i は $[0, 1]$ の範囲の値となり，0 に近いほどマスクされたことを表す．ここで，Gumbel

softmax によってサンプリングされた値は勾配を保つため，勾配法によってスコアの学習を行うことが出来る．テスト中は，Gumbel softmax を用いる代わりに，スコアが 0.5 を超えるかどうかによってマスクを決定的に計算する．

$$m_i = \begin{cases} 1, & c_i \geq 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

木構造 LSTM 層には，系列 LSTM 層からの出力 $\boldsymbol{v}_i$ の代わりに，それに計算されたマスク m_i を適用した $\boldsymbol{v}_i^t = m_i \boldsymbol{v}_i$ を入力する．

3.2.1 ノイズによるマスクの促進

上述のマスク機構を導入することだけでもマスクの表現は出来るものの，この機構の導入だけでは不要なトークンをマスクするようには学習が進まない．そこで，全体でトークンがどの程度マスクされなかったかを表す生存率をもとに，その生存率が大きくなるほどスケールが大きくなるようなノイズを系列 LSTM 層からの入力に加えることで，ノイズ減少のためにマスクされるトークンが増えるように学習が進行することを目指す．具体的には，まずマスク機構で計算されたスコア c_i をもとに，生存率 r を以下の通り計算する．

$$r = \frac{\sum_{i=1}^{n^s} c_i}{n^s} \tag{4}$$

ここで n^s は入力文のトークン数である．この生存率 r を係数とした正規分布に従うノイズを加えたものを実際の木構造 LSTM への入力 $\boldsymbol{v}_i^t$ とする．

$$\begin{aligned} \boldsymbol{p}_i &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \\ \boldsymbol{v}_i^t &= m_i(\boldsymbol{v}_i + \alpha r \boldsymbol{p}_i) \end{aligned} \tag{5}$$

ここで α はノイズの大きさを制御するハイパーパラメータである．ここで加えるノイズは，平均 0，分散 $\alpha^2 r^2$ の正規分布に従うノイズと捉えることもでき，利用する情報が多い（生存率が高い）ほど入力の分散が大きくなるということをモデルしている．このノイズは学習中のみ加算し，テスト時は利用しない．

3.3 アテンション機構

ベースモデルは，上向き木構造 LSTM の出力のうち，係り受け木上でルートノードとなるトークンの出力を利用するが，ノイズ機構を導入したモデルにおいては，そのトークンがマスクされる可能性がある．マスクされたノードを出力として利用すると，不要な情報を包含してしまうため，ルートのトークンの出力の代わりに，上向き木構造 LSTM アテンション機構によって集約された出力を，隠れ層への入力に用いることとした．

4 実験

本研究では，ベースモデルと，それにマスク機構及びアテンション機構を導入した際の性能比較を行った．

4.1 実験設定

実験では，SemEval-2010 task 8 データセット [1] を用いた．このデータセットは 8,000 文の訓練セットと 2,717 文のテストセットで構成されており，それぞれの文には関係分類の対象となるエンティティペアが 1 つずつ示されている．訓練事例のうち 9 割を実際の訓練セットとし，残り 1 割を開発セットとした．ハイパーパラメータは開発セットにおける Micro-F 値を最大にするようにチューニングした．チューニングしたハイパーパラメータを用いて，訓練セットと開発セットを合わせて再度学習を 5 回行い，学習後のモデルをテストセットについて公式の評価指標である Macro-F 値によって評価した．

前処理として，Stanford parser[1]を用いて品詞タグと係り受け関係木を得た．モデルの単語表現ベクトルの初期値には，LSTM-ER モデルで用いられたものと同じ事前学習済みベクトルを利用し，ファインチューニングを行った．LSTM-ER モデルと同様に，重みの平均化を行った．モデルの実装には Tensorflow バージョン 1.8[2]を利用した．

4.2 実験結果

各モデルにおける Macro-F 値を表 1 に示す．マスク

モデル		開発	テスト
マスク	アテンション	データ [%]	データ [%]
		81.43 ± 0.28	83.17 ± 0.16
✓		82.89 ± 0.13	83.53 ± 0.29
	✓	82.84 ± 0.64	82.95 ± 0.43
✓	✓	83.98 ± 0.65	83.89 ± 0.64
LSTM-ER (最短経路, Child-Sum) [4]		83.8	-

表 1: 各モデルにおける 5 回の試行での Macro-F 値．ベースモデルに対し導入した機構をチェックで表した．

[1]https://nlp.stanford.edu/software/lex-parser.shtml
[2]https://www.tensorflow.org/

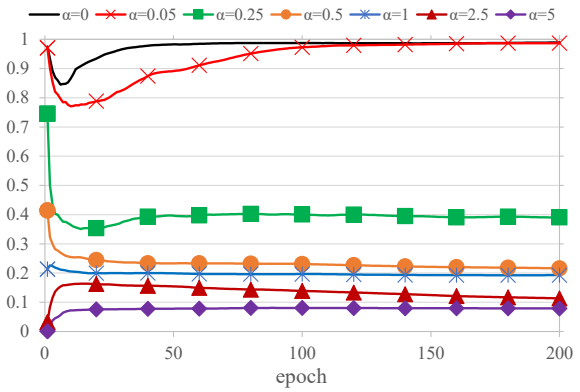

図 2: Learning curve of non-masking rate

機構とアテンション機構のどちらも導入しなかったモデルがベースモデルである．表には，5 回の実行における平均値と標準偏差を示した．表のとおり，マスク機構とアテンション機構を両方導入したモデルが最も高い平均 Macro-F 値となった．これは，LSTM-ER モデルにおいて木構造 LSTM に Child-Sum Tree-LSTM を利用し，最短経路の経験則を利用したとして三輪らが報告したスコアと同程度のスコアとなっている．また，どちらかの機構を単独で導入した場合でも，ベースモデルより高い平均 Macro-F 値となった．マスク機構とアテンション機構を両方導入したモデルと，ベースモデルについて，それぞれの 5 回のテストデータの予測結果のうち中央値の Macro-F 値を記録したものについて Approximate Randomization [7] によって有意差検定を行った際の p 値は 0.0197 となり，提案手法によって全体木をそのまま用いるモデルから有意に性能が向上したことが示された．

4.3 マスクされるトークン

図 2 に，マスク機構のみを導入したモデルについての，ノイズの係数 α ごとのマスク機構内のスコア c の平均値の学習曲線を示す．図の通り，ノイズの係数を大きくするほどマスクされる量が促進された．一方，$\alpha = 0$ のノイズを一切加えない場合では，ほとんどのトークンがマスクされておらず，このことからノイズのマスクへの必要性がわかる．

このモデルのうち開発データで最も高い Micro-F 値を記録した時の平均生存率は 20.2%となった．文内における最短経路に属するトークンが占める割合の平均値は 19.3%であり，これと近い値となっている．

学習されたマスクと，最短経路がどの程度異なるかを調べるために，各入力文について，マスク機構とアテンション機構両方を導入したモデルの，開発データで最も高い Micro-F 値を記録したモデルで学習されたマスクを並べたベクトル $[m_1, m_2, ..., m_{n^s}]$ と，最短経路に含まれるトークンを 1，含まれないトークンを 0 としたベクトルのコサイン類似度を計算した．訓練データと開発データの平均コサイン類似度は，それぞれ 0.931 と 0.933 となり，最短経路上のトークンをよく残すように学習される結果となり，最短経路の関係分類への有効性を再確認する結果となった．

5 おわりに

本論文では，最短経路の経験則を使用せず，学習を通じて不要な情報をマスクするマスク機構と，そこから計算された出力を集約するアテンション機構を備えたニューラル関係分類モデルを提案した．結果としては，提案モデルは最短経路と同等の Macro-F 値を記録し，学習されたマスクは最短経路上のトークンをよく残すものとなり，最短経路の有効性を再確認する結果となった．

参考文献

[1] Iris Hendrickx, et al. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the SemEval*, 2010.

[2] Razvan Bunescu and Raymond Mooney. A shortest path dependency kernel for relation extraction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 2005.

[3] Yan Xu, et al. Classifying relations via long short term memory networks along shortest dependency paths. In *Proceedings of EMNLP*, 2015.

[4] Makoto Miwa and Mohit Bansal. End-to-end relation extraction using lstms on sequences and tree structures. In *Proceedings of ACL*, 2016.

[5] Jang, et al. Categorical reparametrization with gumble-softmax. In *ICLR*, 2017.

[6] Ashish Vaswani, et al. Attention is all you need. In *Advances in NIPS*. 2017.

[7] Eric W Noreen. *Computer-intensive methods for testing hypotheses*. Wiley New York, 1989.