明示的な非文との識別による言語モデルの文法能力の向上

能地 宏　　高村 大也
産業技術総合研究所人工知能研究センター
{hiroshi.noji,takamura.hiroya}@aist.go.jp

1 はじめに

言語は階層構造を持ち、自然文の正しい理解、生成のためには、構文解析など階層構造を扱う技術が必要と思われてきた。近年 RNN を始めとするニューラルネットワークの発展によって、この状況が変わりつつある。これらは系列モデルであるが、長く流暢な文の生成を可能にするなど、明示的な教師信号なしに言語の階層構造をある程度獲得できているようにみえる。

近年、特に言語モデルに着目し、系列に対するニューラルネットワークがどれほど文法を獲得できているかを明示的に調べる研究が進んでいる (Linzen et al., 2016; Gulordava et al., 2018; Marvin and Linzen, 2018)。これらは作為的に選択もしくは生成した文に対するモデルの振る舞いを評価する。多くは英語や他の言語の名詞/動詞の数の一致に着目し、例えば *The keys to the cabinet are/*is* という列に対し、言語モデルが正しい動詞である *are* に *is* よりも高い確率を付与するかを評価する。動詞に近い名詞は単数の *cabinet* であるため、*are* と言い当てるためには *keys to the cabinet* が名詞句をなすという階層構造をモデルが認識する必要がある、と考えるのである。過去のいくつかの研究を総括すると (第 2 節参照) 生の文のみから学習したニューラル言語モデル、特に LSTM は、従来の n グラムモデルよりもこれら統語現象に敏感であるが、完璧ではなく、大量のテキストのみから系列モデルが文法を獲得するか否かに関しては、はっきりした答えは得られていない。

ニューラル言語モデルの学習は、通常訓練文の出現確率を最大化することにより行われる。これは、当たり前であるが、モデルは自然文に関する正例のみしか観測しないということである。しかし、上述した込み入った現象に対する文法能力など、言語のあらゆる側面が正例のみから学習できるのか、また学習する必要があるのかに関しては、検証の余地がある。例えば複雑な長距離依存関係の解決に関しては、コーパス中で現象の頻度が相対的に少ないため、生文のみからでは十分な学習の手がかりが得られない可能性がある。

本稿では、正例だけでなく、明示的に負例を用いることにより、特定の現象に対する言語モデルの性能を向上させられることを示す。言語モデルにとっての負例とは、非文である。本稿では特に、これまで研究が進んでいる動詞の数の一致に焦点を当て、観測した文に対し、その中の動詞の単数/複数を反転させた文を負例として生成する。学習の際は、これら二文間の生成確率が離れるようにマージンに基づく損失を設計し、観測文の確率を上げるだけでなく、生成した明らかな非文に対する確率を下げることにより、学習を行う。

テストデータでの評価指標が主に動詞の数の一致であるため、このように数の一致を特別視した学習で評価性能が上がるのは、当たり前と思われるかもしれない。つまり、精度が上がったとしても、それは単に訓練データで出現した名詞/動詞のペアを丸覚えしたため、という可能性が捨てきれない。本研究で特に注目するのは、特定の名詞/動詞のペアに寄らず、数の一致という現象をモデルが一般化した形で獲得しているか否かである。様々な実験を通じ、モデルは個別の事例でなく、現象一般に対する頑健性を獲得していることを明らかにする。

2 関連研究

本研究は、言語モデルから文法能力を切り離して評価を行う一連の研究を基礎に置く。言語モデルは通常パープレキシティ、すなわちモデルがどれほどテスト文に高い確率を付与したかにより評価される。これは単一の比較しやすい指標ではあるが、その数値の解釈が問題となる。文中の単語の選択には、選択選好のような意味に関わる側面、文法的な側面など、様々な要因が関わるが、パープレシティのより低いモデルが具体的に言語のどの側面に対する能力を向上したのかに関して解釈を与えることができない。

これらの研究は主に、特定の現象に着目した新たなデータセットの構築とそれらに対する既存の言語モデルの評価によって進んできた。初期の Linzen et al. (2016) は、英語 Wikipedia 中から上記 *The keys to ...* のような名詞とそれに対応する動詞が離れて出現する文を抽出し、二単語間の距離に応じて難易度を分けて評価を行なっている。評価は数の異なる対象動詞 (*are* vs. *is* など) の尤度を比較することで行われる。彼らの研究では LSTM は複雑なケースでは予測が行えないという結果であったが、後の Kuncoro et al. (2018) で、Linzen らは LSTM を過小評価しており、より大きなネットワークを用いることで性能が向上することが明らかになった。その上で、 Kuncoro らは文法を明示的に扱う言語モデル (RNNG) により性能が向上することを示しており、 LSTM の限界も指摘している。

本データセットは Wikipedia から抽出したものであ

るから、自然に出現する文に対するモデルの反応を調べることができる。一方、モデルの文法能力を様々な側面から調べる際、複雑な統語現象はコーパス中で頻度が少なく、自然文からでは現象を収集しにくいという問題がある。Marvin and Linzen (2018) はより詳細な文法能力の評価のために、テンプレートと限定された語彙を用いて、異なる統語現象毎に分かれたデータセットの作成を行なった。例えば以下の文は長い動詞区の並列 (*Long VP coordination*) に属する例である。

(1) The manager writes in a journal every day and likes/*like to watch television shows.

基本的に例は上記 *likes* など一語のみが異なった文対からなる。その他の現象として、関係節の種類 (主格/目的格) などによる分類が存在する (表 1)。

最後に Gulordava et al. (2018) のデータセットを紹介する。これは言語モデルの文法能力の評価という側面を突き詰めたものであり、次のような文対が含まれる。

(2) The absolute payable lighting I have contributed operates/*operate that in the suggestion of ...

この文は既存の文の内容語をランダムに入れ替えることで生成されたものであり、そのため意味不明である。これは有名な Chomsky の *Colorless green ideas sleep furiously.* のように、文法的に適格であるが意味的に不適格な文を意識して生成されている。本データセットで特徴的なのは、数を一致すべき名詞/動詞のペアがランダムに生成されたものであるから、訓練データには (ほぼ) 直接存在しない、ということである。そのため、目的の統語現象に対するモデルの一般化性能を真に測ることができると考えられる。

本研究の実験では、様々な統語現象に対する性質を調べるために Marvin and Linzen (2018) のデータセットを主に用い、また Gulordava et al. (2018) のデータセットも用いることで、モデルの現象そのものに対する汎化能力を調べる。

次に手法の観点から、言語モデルの学習への非文の利用に関する関連研究について述べておく。関連が深いのは、Okanohara and Tsujii (2007) などの識別言語モデルである。これは文/非文を分類する分類器を構築するもので、その学習にはルールもしくは統計的に生成された非文が利用される。本研究での学習はこれと似ているが、モデルが識別モデルでなく、あくまで文に確率を与える生成モデル (LSTM) であるという点が異なる。我々は文/非文に与える確率値の差が大きくなるように学習を行う。

ニューラル言語モデルの学習へのマージンの導入は本研究が初ではなく、Huang et al. (2018) でも検討されている。我々がマージンを取る負例をルールで生成しているのに対し、彼らは MT などの出力のリランキングを目的に、別の MT モデルの出力を擬似的な負例とみなし、それらより参照文の生成確率が高くなるように言語モデルの学習を行なっている。彼らの負例はこのように別のモデルから自動で生成されるため、特定の言語現象に特化したものとはなっていない。本研究の焦点は、特定の言語現象に特化してマージンに基づく学習を行うことで、モデルがその現象一般に対する頑健性を得られるか否かである。

3 非文を用いた学習

本研究で用いる学習法は非常に単純である。訓練文の集合を $\mathbf{x}$ とし、ある訓練文 $x \in \mathbf{x}$ に対し、その動詞を変化させた非文の集合を $Y_x = \{y\}$ とする。この時、以下の損失関数を最小化する。

$$L_{neg} = \sum_{x \in \mathbf{x}} \sum_{y \in Y_x} \max(0, \delta - (\log p(x) - \log p(y))).$$

δ は損失を操作するマージンであり、文 x と負例 y に対する対数尤度の差が δ を超えたとき損失は 0 となる。

これは単純であるが、ニューラル言語モデルに対しては通常のクロスエントロピー損失 $L_{lm} = \sum_{x \in \mathbf{x}} p(x)$ が非常に有効に働くことが分かっている。そこで我々は、二つの損失関数の和を取り、マルチタスク学習で学習することとした。両者の重みは共に 1.0 とする。各エポックで、二つのタスクそれぞれでミニバッチの集合を得た後、それらを合わせてシャッフルした。

4 実験設定

まず学習データの構築について述べる。Gulordava et al. (2018) は 1 億語程度の文の集合を Wikipedia から収集し公開しており、Marvin and Linzen (2018) も同じデータを用いている。しかし、本データは低頻度語の置き換えなどを含む前処理が既になされてしまっており、後に述べる品詞タグ付けが困難となっている。そこで我々は彼らの方法に従い、同程度の量の文を Wikipedia から新たに収集し、8:1:1 の割合で訓練 (81,967,702 語)、開発 (10,250,099 語)、テストデータ (10,261,911 語) とした。実験では、文法能力の評価に加え、このテストデータでのモデルのパープレキシティを報告する。

各文 $x \in \mathbf{x}$ に対して、次の方法で負例の集合を得る。まず各文を Stanford CoreNLP を用いて品詞タグ付けし、VBZ (三人称単数現在形) の語を複数形に、他の現在形 (VBP) を逆に変換する。ある文に複数の変換候補がある場合は、一つの動詞の数が異なるペアを一事例とし、複数の訓練事例を作成した。VBZ もしくは VBP が 3 語ある文については、3 事例の非文が作成されることになる。非文を得た後、訓練データ中の上位 100,000 語を語彙とし、残りは全て同一の未知語で置き換えた。

実験は全て同一のパラメータを持つ LSTM を用い、通常の L_{lm} に基づく学習と、非文を用いる $L_{lm} + L_{neg}$ による学習を比較する (第 3 節)。ただしモデルの LSTM として、Gulordava et al. (2018) などより強力なもの

	LSTM		+margin			n-gram	# sents
	M&L18	Ours	$\delta = 1$	$\delta = 3$	$\delta = 10$		
SUBJECT-VERB AGREEMENT:							
Simple	0.94	0.99	0.99	0.99	1.00	0.79	280
In a sentential complement	0.99	0.93	0.99	0.99	0.99	0.79	3360
Short VP coordination	0.90	0.88	0.97	0.96	1.00	0.51	1680
Long VP coordination	0.61	0.83	0.93	0.97	0.97	0.50	800
Across a prepositional phrase	0.57	0.90	0.95	0.97	0.99	0.50	44800
Across a subject relative clause	0.56	0.85	0.98	0.99	1.00	0.50	22400
Across an object relative clause	0.50	0.86	0.87	0.88	0.88	0.50	44800
Across an object relative (no that)	0.52	0.74	0.79	0.81	0.78	0.50	44800
In an object relative clause	0.84	0.90	0.85	0.99	1.00	0.50	44800
In an object relative (no that)	0.71	0.85	0.73	0.88	0.96	0.50	44800
REFLEXIVE ANAPHORA:							
Simple	0.83	0.93	0.86	0.95	0.90	0.50	560
In a sentential complement	0.86	0.81	0.82	0.87	0.83	0.50	6720
Across a relative clause	0.55	0.67	0.69	0.67	0.64	0.50	44800
NEGATIVE POLARITY ITEMS (NPI):							
Simple	0.40	0.98	0.52	0.70	0.63	0.06	792
Across a relative clause	0.41	0.58	0.43	0.60	0.45	0.60	31680
Perplexity	-	60.46	69.64	69.88	70.60	-	

表 1: Marvin and Linzen (2018) データセットでのモデルの比較。彼らの LSTM (M&L18) と n-gram の結果は論文から抜粋した。+margin は非文を用いた学習を追加した場合で、δ がマージンの大きさを表す。

を用いる。Kuncoro et al. (2018) が指摘するように、LSTM の能力はモデルサイズによって大きく異なり、LSTM 自身の限界や学習法の効果を論じるには、より高い性能をもつベースラインを利用することが重要であると考える。具体的には Merity et al. (2018) で提案された各種正則化法を含んだ LSTM を用いた。ただし単語を削減するドロップアウトは用いず、他のドロップアウトの重みは 0.1 とした。モデルサイズは、400 次元の単語埋め込み、1,150 次元の 3 層 LSTM で、入出力層は重みを共有する。平均 SGD を用い、学習率は 10.0、バッチサイズは 128 とした。元のモデルは文書単位で単語の予測を行うが、これを各文を独立に扱うよう変更した。これらのパラメータは、この設定で、単一タスクの LSTM が比較的良いパープレキシティを示したものである。非文を用いる場合でも、早期終了は言語モデルの開発データ上でのパープレキシティのみで判定した。最大 18 エポックの学習を行い、最もパープレキシティが低かったモデルを選択した。

5 実験結果と議論

Marvin and Linzen (2018) のデータセットは必要とする文法能力によって分かれている (表 1)。本実験は先行研究にならい、対象動詞だけでなく文全体の尤度を比較し、正しい動詞の文に高い尤度を割り当てた場合を正解と見なす。まず注目すべき点として、先行研究の LSTM (M&L18) と比較し、我々の LSTM はほとんどの場合に精度が向上している。M&L18 は 2 層 LSTM など本研究より小さいモデルを用いている。この結果により著者らは LSTM の限界について論じていたが、精度の低かった主格の関係節を挟む例 (*The author that likes the guard is/*are tall.* など) でも大きく精度が向上している。これより元論文での LSTM の性能は過小評価されていた、と言ってよいだろう。

次にマージンを導入した場合の傾向をみる。主語/動詞の一致 (表の上部) に着目すると、小さい値 ($\delta = 1$) では良くなる項目もあるが、関係節の中 (In ... clause) など、悪くなる項目もあった。$\delta = 3$ で全体の予測性能は向上し、$\delta = 10$ では一部を除き完璧に近い精度を達成できている。精度の向上が小さいのは、目的格の関係節に続く動詞の予測であり、特に次のように代名詞が省略された場合 ($\delta = 10$ で 0.78) である。

(3) The movie the guards like is/*are unpopular.
(-61.33/-55.61)

数値は $\delta = 10$ での各文の対数尤度を表す。なぜこの現象での精度が低いのか、本研究の範囲では明らかなことは言えない。ただし、目的格の関係節は主格より人間にとってより難しく、その難しさの一部は日常での出現頻度の少なさで説明できるという主張は心理言語学で古くから存在する (Levy, 2008)。訓練文中に出現する統語現象の頻度を詳細に調べ、本学習法の特性と限界点を論じることは、今後の課題であるといえる。

表の次の 3 行は再帰代名詞に関するものである (*The senators embarrassed themselves/*herself.* など) が、ベースラインと比較し大きな性能の変化は見られなかった。再帰代名詞に関しては今回は負例の作成は行なっておらず、またその解決方法は数の一致の場合とはかなり異なるといえる。ここから、特別視しなかった現象と離れた現象については今回の手法では本質的な改善は

	G18	LSTM	$\delta = 1$	$\delta = 3$	$\delta = 10$	# sents
Orig.	0.81	0.90	0.90	0.87	0.87	41
Nonce	0.74	0.78	0.87	0.85	0.89	369

表 2: 対象動詞の予測確率に基づく Gulordava et al. (2018) のデータセットでのモデルの比較。Orig. はコーパス中に出現する元の文、 Nonce は内容語を入れ替えた意味的に不適格な文である。G18 は原論文での LSTM の性能である。

行われず、逆に大きな悪影響も及ぼさない、ということがいえそうである。

その下の NPI の解釈に関してはもう少し複雑であるため、詳しく説明する。マージンを導入した場合、Simple NPI に対して LSTM (0.98) と比較し大きく精度が低下している (0.5-0.7)。Simple NPI は次の *few/ever* など、呼応する語の扱いを調べるものである。

(4) Few$_{-9.12}$ farmers have ever$_{-6.14}$ been popular.

(5) *Some$_{-5.44}$ farmers have ever$_{-9.26}$ been popular.

添字は各単語に $\delta = 10$ が割り当てた対数尤度である。これは典型的な例であり、まず先頭の語は頻度の影響を受けやすく、モデルは *few* の方が頻度が低いことからより低い確率を付与する。従って問題は、ここで生じた差を後の呼応する *ever* で取り返せるか否かとなる。モデルは正解の *ever* に高い確率を与えるが、先頭の *few* で生じた差を取り返すほどにはならず、文単位の尤度は微妙に低くなる。なお正解のテスト文の先頭は *few* と *no* のみであり、いずれも対となる語 (*some*, *many* など) と比較し頻度が低い。モデルの性能は内容語に対しては大きく変わらず、どの場合に *ever* で確率を取り返せるかはほぼ先頭の語の種類で決まってしまうため、モデル毎に大きく変動しやすい傾向が生じる。

とはいえパープレキシティをみると、マージンを導入することで約 10 ポイント上昇しており、単語全体の予測精度は悪化していることがみてとれる。問題はこの差が応用にとって意味のあるものかどうかである。上記 NPI の例では、呼応する *ever* には正例の方が高い確率が割り当てられている。言語モデルを逐次的な文生成技術の基礎とみなした場合これは望ましい動作といえるだろう。本研究の範囲では、このパープレキシティの悪化が数の一致に強くなることの代償として他の本質的な能力を失った結果であるのか、判断はできない。しかしこのような言語モデルに対するある種のユニットテストに対して性能の向上が見られ、同時にパープレキシティが悪化するという結果自体は興味深い。第 2 節冒頭で述べたように、パープレキシティは様々な影響を受けるため解釈しにくい。現在のデータセットは特定の統語現象のみに特化したものであるが、これをより広範な現象に拡張し、実験を通じて事例についても改善を加えていくというのが、言語モデルの評価の本質的な改善に繋がるのではないだろうか。

これまでの結果は、対象の名詞/動詞のペアについては制限された語彙から選択されており、訓練文中にも直接共起しうるという点で、モデルが単語によらず数の一致という現象に対し頑健になっているかは正確に評価できなかった。この点を評価するため、意味的には不適格な文 (nonce) の集合での評価を表 2 と表 3 に示す (第 2 節参照)。表 2 は原論文に従い対象の動詞だけを比較した場合、表 3 は先ほどと同様に文の尤度で比較した場合の結果である。どちらでも LSTM から精度が向上する[*1]が、nonce を見ると興味深いことに特に文単位で判定した方が精度向上が大きく、$\delta = 10$ では 0.94 に達した。これらの文の処理は訓練データの丸覚えでは対応できず、従ってモデルは特定の単語でなく、名詞/動詞の数に関するより一般的な知識、能力を獲得しているといえそうである。これは、 LSTM は適切な学習を用いることで、述語が未解決の主語に対して、対応する動詞の数を一致させるといった特定の文法処理に対する能力を獲得しうる、ということを示唆する。

	LSTM	$\delta = 1$	$\delta = 3$	$\delta = 10$	# sents
Orig.	0.95	0.97	0.95	0.97	41
Nonce	0.76	0.86	0.90	0.94	369

表 3: 対象動詞だけでなく文全体の予測確率の比較に基づく Gulordava et al. (2018) のデータセットでのモデルの比較。

6 おわりに

特定の言語現象に対し頑健な言語モデルの学習法として、現象を違反する “絶対に生成したくない” 文を学習に利用する方法を示し、文法能力の観点からその有効性を検証した。本稿では文法に着目し、非文の生成を行なったが、何が生成したくない文かは応用によって異なり、今後異なる問題への適用も考えられる。

謝辞

この成果は、国立研究開発法人新エネルギー・産業技術総合開発機構 (NEDO) の委託業務の結果得られたものです。

参考文献

K. Gulordava, P. Bojanowski, E. Grave, T. Linzen, and M. Baroni (2018) “Colorless Green Recurrent Networks Dream Hierarchically,” in *NAACL-HLT*, pp. 1195–1205.

J. Huang, Y. Li, W. Ping, and L. Huang (2018) “Large Margin Neural Language Model,” in *EMNLP*, pp. 1183–1191.

A. Kuncoro, C. Dyer, J. Hale, D. Yogatama, S. Clark, and P. Blunsom (2018) “LSTMs Can Learn Syntax-Sensitive Dependencies Well, But Modeling Structure Makes Them Better,” in *ACL*, pp. 1426–1436.

R. Levy (2008) “Expectation-Based Syntactic Comprehension,” *Cognition*, Vol. 106, No. 3, pp. 1126–1177.

T. Linzen, E. Dupoux, and Y. Goldberg (2016) “Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies,” *TACL*, Vol. 4, pp. 521–535.

R. Marvin and T. Linzen (2018) “Targeted Syntactic Evaluation of Language Models,” in *EMNLP*, pp. 1192–1202.

S. Merity, N. S. Keskar, and R. Socher (2018) “Regularizing and Optimizing LSTM Language Models,” in *ICLR*.

D. Okanohara and J. Tsujii (2007) “A discriminative language model with pseudo-negative samples,” in *ACL*, pp. 73–80.

[*1] 表 2 Orig. の 0.90 と 0.87 の正解数の差は 1 である (37→36)。