自己学習による all-words WSD の半教師あり学習

新納 浩幸　　曹 鋭　　古宮 嘉那子
茨城大学大学院 理工学研究科 情報工学専攻

{hiroyuki.shinnou.0828, 18nd305g, kanako.komiya.nlp}@vc.ibaraki.ac.jp

1　はじめに

本論文では自己学習を利用した all-words WSD の半教師あり学習を試みる。

語義曖昧性解消 (Word Sense Disambiguation, 以下 WSD) は文中の多義語の語義を識別する処理であり、意味解析の必須要素技術である。しかし一般の WSD は対象単語を限定した形で行われるため、実際のシステムで利用することは難しい。そのため対象単語を限定せず文中の全ての多義語に対して語義を付与する all-words WSD が期待されている。

通常 WSD は教師あり学習で解決される。つまり WSD の解決には対象単語毎にラベル付きの訓練データ（語義付きの用例文）が必要となる。all-words WSD の場合、対象単語が多義語全体となるために、必要となる語義付きの用例文は膨大であり、そのため教師なし学習のアプローチも取られてきたが（[1] など）、この場合、精度に問題がある。一方、徐々に all-words WSD のためのラベル付きデータ、つまり語義タグ付きコーパスも整備されてきたために、近年、all-words WSD を教師あり学習の枠組みで解くことも試みられている [6][8]。ただし現在利用できる語義タグ付きコーパスは小規模であり、十分な精度が得られているとは言えない。そのため我々は高精度の all-words WSD システムの構築のために、半教師あり学習を検討している。半教師あり学習とは少量のラベル付きデータと大量のラベルなしデータから分類器を学習する手法である。all-words WSD の場合、ラベルなしデータはプレーンなコーパスに対応し、プレーンなコーパスを大量に入手することは比較的容易であるため、all-words WSD に対する半教師あり学習は、有望なアプローチと言える。

ここでは半教師あり学習の中で、最も簡易な手法である自己学習を all-words WSD に対して試みる。自己学習とはその時点で利用可能な分類器を利用して、ラベルなしデータに確率付きでラベルを付与し、その確率の高いものは正しいラベルと考えて、ラベル付きデータを増強して分類器を再構築する学習手法である。自己学習は簡易な半教師あり学習であるが、all-words WSD の場合、単一の分類問題を解いているのではないため、その実現方法には工夫が必要である。本論文では all-words WSD を実現するために双方向 LSTM を用いる。双方向 LSTM における学習では信頼度の低いラベルからの損失を無視することで、自己学習を実現できる。また双方向 LSTM のモデルの自己学習による改善手順として、ラベル付きデータを増やして、ゼロからモデルを作り直す他に、追加されるデータを利用して fine-tuning を行うアプローチも考えられる。本論文では自己学習により得られたラベル付きデータから学習できたモデルを、本来の持っていたラベル付きデータを使って fine-tuning するという新たな形も試みる。

2　関連研究

分類器に対する半教師あり学習に対しては多くの研究がある。古典的には Co-training と EM アルゴリズムの利用がよく知られている。Co-training は 2 つの独立した観点から相互に分類器を改善してゆく手法であり、EM アルゴリズムを利用する手法は、生成モデル $p(x;\theta)$ を設定し、ラベル z を潜在変数と見なして、$p(z|x)$ を構築する。またアイデアの観点からは半教師あり学習は大きく 2 つに大別できる。一つは、ラベル付きデータから得られる分類器を使って、ラベルなしデータに確信度付きのラベルを付けて、それを利用して分類器を改善してゆくタイプの手法である。自己学習やラベル伝播などがこのタイプである。もう一つの手法がデータをある空間[1]へマップするタイプの手法である。まずラベルなしデータをうまく分離できるような空間にマップし、次にラベル付きデータもその空間にマップし、その空間上で分類器の学習と識別を行うタイプである。通常、低次元にマップできれば、

[1] 一般にもとのデータの次元よりも低次元の空間。

クラスを分ける境界を推定するためのラベル付きデータは少量で済むので、半教師あり学習が成立する。多様体論の手法や生成モデルを利用した手法がこのタイプである。深層生成モデルを利用した半教師あり学習は、生成モデルを利用した半教師あり学習と枠組み的には同じである。ラベルなしデータをうまく分離するような空間にマップする手法にネットワークを利用していると見なせる（[2] など）。

系列ラベリング器に対する半教師あり学習としては事前学習が代表的である [4][5]。これは入力となる素性ベクトルに識別に有効な素性を、大量のラベル無しデータを利用して予め学習しておき、それを訓練データとテストデータに追加するものである。

また all-words WSD に関して言えば、トピックモデルを利用した教師無し学習がいくつか提案されているが [1][3]、これは生成モデルを構築するために、容易に半教師あり学習に拡張できるはずである。

3 双方向 LSTM による all-words WSD

all-words WSD は入力となる単語列の各単語にラベル（語義）を付与する系列ラベリング問題とみなすことができる。系列ラベリング問題をニューラルネットワークで扱う場合、LSTM を使用する。LSTM では時刻 t の中間層の内容を時刻 $t-1$ の入力に使い状態を保持しながら学習することで時系列に対応している。LSTM は時系列データを扱うものであり、自然言語処理では文や文書の単語列を時系列データとみなして利用する。そのため通常、注目している時刻 t 以降の単語も利用できるため、データを逆方向からも解析できる。順方向の LSTM と逆方向の LSTM を同時に用いて、時刻 t での出力を求めるのが双方向 LSTM である (図 1 参照)。

4 双方向 LSTM の自己学習

自己学習では現在の分類器を利用してラベルなしデータに確率付きでラベルを付け、高い確率を持つラベルを正しいラベルと考えて、そのデータをラベル付きデータ（訓練データ）に追加することで分類器の精度を徐々に上げてゆく。系列ラベリング器に対して自己学習を行う場合、系列ラベリング器はラベルなしの単語列を入力として受け取り、各単語に対して確率付きでラベルを付与する。このとき確率の高いラベルと確率の低いラベルが混在するために、単純にその単語列を訓練データに追加することができない。このため系列ラベリング器に対する自己学習では、(1) どのように訓練データを増やしてゆくのか、(2) 増やした訓練データをどのように利用するのか、という二つの問題が存在する。

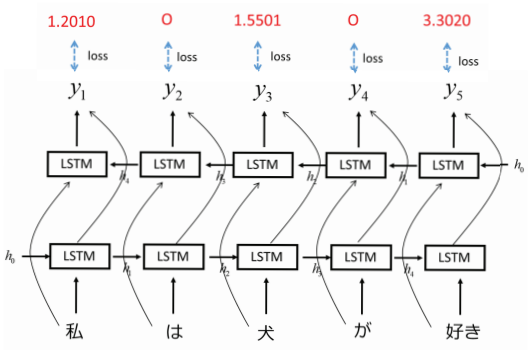

図 1: 双方向 LSTM による語義の付与

4.1 確率の低いラベルからの学習の回避

上記 (1) の問題に対して、ここでは確率（信頼度）の低いラベルからの学習を行わないことにする。つまり、系列ラベリング器がラベルなし単語列の各単語に対して確率付きでラベルを付与するが、その確率の大小に関わらず、その単語列を訓練データに加える。そして学習時に確率の低い部分からの学習を行わないことにする。LSTM でこの処理を行うのは容易である。LSTM では各単語に対する出力値とその単語 w_i のラベルとの差から損失 $loss_i$ を求め、その損失を累積してゆき、文末まで処理が終わったときに、累積された損失 $\sum_i loss_i$ を基にネットワークのパラメータを更新する。信頼度の低いラベルがあった場合には、$loss_i = 0$ と設定すればよい。

4.2 増補されたラベル付きデータ利用

上記 (2) の問題に対しては、以下のような 3 つのアプローチが考えられる。なお、ここでは最初にあったラベル付きの訓練データを D とし、自己学習により得られた確率付きのラベル付きデータを A とする。

(a) $D \cup A$ を新たな訓練データとして双方向 LSTM を学習

(b) D から作られた双方向 LSTM のモデルを A を利用して fine-tuning する

(c) A から作られた双方向 LSTM のモデルを D を利用して fine-tuning する

本論文では上記 3 つのアプローチを試し、最も効果のあるアプローチを考察する。

5 実験

ここでは分類語彙表の語義番号を語義と見なしている。BCCWJ のコアデータに分類語彙表の語義番号が付与されたデータが公開されている。このデータの約 1 割をテストデータ T とし、残りをラベル付きの訓練データ D とした。文の数では D は 12,482 文、T は 1,498 文である。また自己学習に利用するラベルなしデータ U としては、毎日新聞の 1993 年から 1999 年の文の中からランダムに 10 万文を取り出したものを利用した。

双方向 LSTM のモデルとしては、2 階層を用い、単語から分散表現に変換する部分は学習を行わずに既存の日本語分散表現データである nwjc2vec [7] を用いた。

まず D から双方向 LSTM を学習し、T を用いて評価を行った。T をシステムで単語分割した場合、36,263 単語に分割された。そのうち 2,212 単語は正解データとの単語分割が異なったので、残り 34,051 単語に付与された語義が評価対象である。またこの中で多義語になっているものは 18,522 単語であり、この 18,522 単語に付与された語義の正解率を all-words WSD の正解率として評価する。結果を図 2 に示す。横軸は双方向 LSTM の学習における epoch 数、縦軸は上記の正解率を示す。18 epoch 後に構築されたモデルの正解率が 0.799 で最も良い値であるが、ここでは論文 [8] のシステムを利用しているため 20 epoch 後に構築されたモデルの正解率 0.796 をベースの正解率とする

次に 20 epoch 後に構築されたモデルを利用して U に確率付きでラベルを付与した。信頼度を表す確率 0.8 以下のラベルは -1 のラベルに変えて増補版のラベル付きデータ A を作成した。

(a) $D \cup A$ から双方向 LSTM を学習

$D \cup A$ を新たな訓練データとして双方向 LSTM を学習させ T を用いて評価を行った。結果を図 3 に示す。この手法の場合、正解率は 0.798 まで向上した。

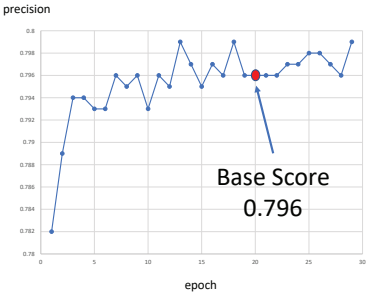

図 2: D から双方向 LSTM を学習

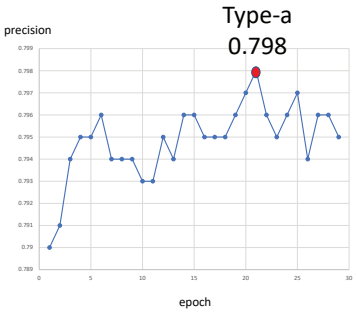

図 3: $D \cup A$ から双方向 LSTM を学習

(b) $D \rightarrow A$ の fine-tuning

D から 20 epoch 後の双方向 LSTM のモデルを作り、それを A を利用して fine-tuning し、T を用いて評価を行った。結果を図 4 に示す。この手法の場合、正解率は 0.794 と下がってしまった。

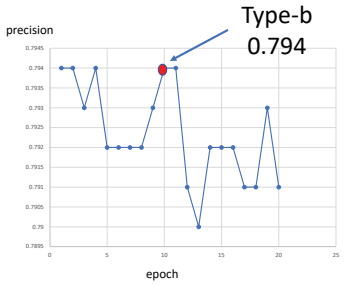

図 4: $D \rightarrow A$ の fine-tuning

(c) $A \rightarrow D$ の fine-tuning

A から 20 epoch 後の双方向 LSTM のモデルを作り、それを D を利用して fine-tuning し、T を用いて評価を行った。結果を図 5 に示す。この手法の場合、正解率は 0.799 まで向上した。

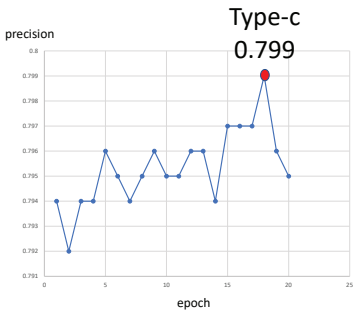

図 5: $A \rightarrow D$ の fine-tuning

6 考察

増補されたラベル付きデータの利用法として、増補されたラベル付きデータからベースのモデルを作り、それを本来のラベル付きデータで fine-tuning するという (c) のアプローチは、図 5 をみると徐々に正解率が向上していく形であり、しかもベースとなる系列ラベリング器の正解率を改善しているので、手法としては有望だと考えている。

ただし正解率の改善はわずかであり、本実験に限れば、自己学習に効果があったとは言えない。本来、識別器の自己学習では、増補した訓練データの中に新たな知識を獲得できる情報が含まれていないので、半教師あり学習としては効果が出ないと考えられる。系列ラベリング問題の場合、ラベルの組み合わせに多様性が出るために、運良く効果が出ることを期待したが、本実験ではうまくいかなかった。ただしラベルなしデータの量（本実験では 10 万文）や疑似ラベルを正しいラベルとする閾値（本実験では 0.8）のパラメータを変更することで効果が現れる場合もあると思われるので、今後、それらの適切な値を調べていくつもりである。

また LSTM の学習時に各単語の損失に重みを付ける工夫は有効だと予想している。本実験では信頼度である確率が 0.8 以下のものを重み 0 にし、それ以上のものを重み 1 として学習した重み付き学習と捉えられる。信頼度を重みとすればより適切に自己学習が行えると考えられる。この点の調査も今後の課題である。

7 おわりに

本論文では自己学習を利用した all-words WSD の半教師あり学習を試みた。all-words WSD を系列ラベリング問題とみなし、双方向 LSTM を用いて all-words WSD を実現する。学習できたモデルからラベルなしデータに信頼度に対応する確率を付与して、訓練データを増加させる。LSTM の学習において、確率が高くないラベルでは、その部分の損失を累積しない学習を行うことで自己学習を実現した。また増補したラベル付きデータからモデルを作り、真のラベル付きデータを用いてそのモデルを fine-tuning する学習法を試し、わずかながらベースの系列ラベリング器の正解率を改善できた。ただし改善はわずかであり、自己学習の効果があったとは言えない。信頼度に対応する確率を損失の重みとして学習する手法を試すことを今後の課題とする。

参考文献

[1] Jordan Boyd-Graber, David Blei, and Xiaojin Zhu. A Topic Model for Word Sense Disambiguation. In *EMNLP-CoNLL-2017*, 2007.

[2] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pp. 3581–3589, 2014.

[3] Kanako Komiya, Yuto Sasaki, Hajime Morita, Minoru Sasaki, Hiroyuki Shinnou, and Yoshiyuki Kotani. Surrounding Word Sense Model for Japanese All-words Word Sense Disambiguation. In *PACLIC-29*, pp. 35–43, 2015.

[4] Matthew E Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*, 2017.

[5] Yanjun Qi, Pavel Kuksa, Ronan Collobert, Kunihiko Sadamasa, Koray Kavukcuoglu, and Jason Weston. Semi-Supervised Sequence Labeling with Self-Learned Features. In *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*, pp. 428–437. IEEE, 2009.

[6] Hiroyuki Shinnou, Kanako Komiya, Minoru Sasaki, and Shinsuke Mori. Japanese all-words WSD system using the Kyoto Text Analysis ToolKit. In *PACLIC-31*, pp. 392–399, 2017.

[7] 新納浩幸, 浅原正幸, 古宮嘉那子, 佐々木稔. nwjc2vec: 国語研日本語ウェブコーパスから構築した単語の分散表現データ. 自然言語処理, Vol. 24, No. 5, pp. 705–720, 2017.

[8] 新納浩幸, 鈴木類, 古宮嘉那子. 双方向 LSTM による分類語彙表番号を語義とした all-words WSD. 国語研言語資源活用ワークショップ, P2-04-E, 2018.