構文解析の分野適応における訓練データの自動生成法

張本佳子　Kenji Sagae　辻井潤一
東京大学理学部情報科学科
E-mail: {harry, sagae, tsujii}@is.s.u-tokyo.ac.jp

1 はじめに

構文解析における分野適応とは、既存の解析済みコーパスで訓練された構文解析器を、生物学論文など別分野のテキストに適応させるタスクである。特に、解析済みコーパスの開発は多大な労力を必要とすることから、対象分野の解析済みコーパスを利用せず、生データのみを用いる分野適応手法が注目されている。実際、CoNLL 2007 の共通タスクとして依存構造解析の分野適応が採用され、様々な手法が提案されている。

CoNLL 2007 の分野適応タスクにおいて最高精度を達成した Sagae & Tsujii (2007) の手法では，対象分野の生データを性質の異なる2つのモデルで構文解析し、それぞれの解析結果が一致した文だけを選び出すことで対象分野の訓練データを自動生成する。彼らの手法は簡潔で非常によい結果を得ているが、選ばれた文の文長の分布が本来の分布と大きく異なるなどの問題があり、またモデルの組み合わせ方などでも改善の余地が考えられる。

本研究では、それらの問題点を解決した訓練データの自動生成手法を提案する。具体的には、最終的な解析精度に影響する要因として、文長の分布、モデルの組み合わせ、訓練データ選択の際の一致率、訓練データのサイズ、の4つについて考察する。4つの要因が構文解析精度に与える影響を実験により調査し、その結果から、訓練データを自動生成する最適な手法を実験的に決定する。

訓練、テストデータはそれぞれ CoNLL 2007 のタスクで与えられた Wall Street Journal (WSJ) 訓練データと生物学の生データを使用し、精度評価には CoNLL 2007 で用いられた評価方法を用いた。

2 背景

混乱を避けるためにまず今後使う用語をいくつか紹介したい。

ME: 最大エントロピー法
SVM: サポートベクトルマシン
後向き: 文の後方から解析すること
前向き: 文の前方から解析すること
関係ラベル: 係り先との具体的な係り関係
トークン: 言語の最小単位。句読点も含む。

2．1 依存構造解析

依存構造解析と各トークンの係り受け先の場所や係り受け先との係り関係などを示す重要な処理である．その手法として近年ではスタックを用いる Shift-reduce 法なども注目されている。

依存構造解析の精度評価には、CoNLL 2007 で用いられた Labeled Attachment Score (LAS) と Unlabeled Attachment Score (UAS) の二種類を用いる。LAS は係り先と関係ラベルがともに正しいトークンの割合を示し、UAS は係り先が正しいトークンの割合を示す。

2．2 既存手法

本論文の提案手法のベースとなった分野適

応手法 (Sagae & Tsujii, 2007) は以下のようなものである。

1. 前向き ME モデルと後向き SVM モデルの両方を WSJ コーパスで訓練する。
2. ステップ 1 で訓練した2つのモデルを用いて生物学論文の生コーパスを構文解析し、2つのモデルによる解析結果が完全一致した文だけを選び出し、訓練データとして WSJ コーパスに加える。
3. ステップ 2 で生成した新しい訓練データを用いて再訓練し、得られたモデルでテストデータを解析して精度を測定する。

この手法は対象分野の解析済みコーパスを用いず、既存の解析済みコーパスと対象分野の生データだけから訓練データを生成するため、どの分野に対しても容易に適応が可能で汎用性が高い。また CoNLL での生物のテストデータでテストした結果、精度は WSJ 単独を使用する場合の 79.37%から 80.23%に上がった。

3 提案手法

我々は前節で紹介した Sagae & Tsujii (2007) の手法をベースに、それを改善する手法を提案する。我々はまず構文解析の精度に影響する要因として、以下の4つに注目した。

1. 文長の分布:Sagae & Tsujii (2007) の手法では解析結果が完全一致する文のみを選択するため、短くて構造が簡単な文が選択されやすく、結果的に訓練データの文長の分布は対象テキストの本来の文長の分布と大きく異なることになる。実際、対象テキストの平均文長が 25 単語であるのに対して Sagae & Tsujii (2007) の手法で生成した訓練データの平均文長は 14 単語である。この分布の違いは解析精度を悪化させる原因の一つと考えられる。
2. モデルの組み合わせ:Sagae & Tsujii (2007) では前向き ME と後向き SVM の組み合わせが用いられたが、本研究では前向き/後向きの ME/SVM の計4種類のモデルの全ての可能な組み合わせを検証する。4種類のモデルはそれぞれ F_ME、B_ME、F_SVM、B_SVM と表記する。また、3種類のモデルの組み合わせについても検証を行う。
3. 解析結果の一致率:異なるモデルによる解析結果の間の一致率は、入力文中の単語の係り先及び関係ラベルが一致している割合として定義する。以降では、例えば「一致率 90%を基準にする」といった場合は、解析結果が 90%以上一致した文を訓練データとして選ぶということを意味する。データ量が同じの場合、より高い一致率でを設定した場合はよりよい解析精度が得られると予想される。
4. データサイズ:通常は訓練データのサイズが大きいほど解析精度が高くなると期待されるが、我々の手法では自動生成した訓練データを使用するため、訓練データ自身に誤りが含まれ、データが大きすぎると逆に解析精度を悪化させる可能性がある。

本研究では、実験によりこれら4つの要因が構文解析精度に与える影響を調査し、その結果から、それぞれの要因の最適値を決定することで、Sagae & Tsujii (2007) の手法を改良する。

4 実験

訓練及び評価データは、CoNLL 2007 の共通タスクで与えられた解析済み WSJ コーパス

(1万5千文)と生物学論文からなる生コーパス(27 万文)を用いた。

図 1 は解析対象テキストの文長の分布を表す。長さが 21~40 単語の文が主要部分を成し、文長が 1~20、21~40、41~60 の文の比率は 7:11:2 になることが分かる。またデータサイズを固定し(6 万 5 千語)、文長の分布によって訓練データを分けた解析結果を表1で示す。

図 1　文長の分布

表1:　文長別訓練データによる解析精度

Sentence Length	LAS	UAS
1~20	77.8	79.0
21~40	80.0	81.4
41~	78.4	80.1

表 2 は生物学論文の生コーパスで実験した場合の一致率と構文解析の精度及び訓練データサイズの関係を示す。モデルの組み合わせとしては F_ME&B_SVM を使用した。

表2　一致率と構文解析精度

Degree of coincidence	Sentence Number	LAS	UAS
100	9237	80.33	81.86
95	23878	79.51	80.96
90	42897	79.15	79.15
85	96191	77.24	79.07

表3:モデルの組み合わせと訓練データサイズ及び構文解析精度

モデルの組み合わせ	訓練データサイズ	LAS	UAS
B_ME&F_SVM	1159k	78.67	80.23
F_ME&B_SVM	533k	79.51	80.96
B_ME&B_SVM&F_ME	337k	79.65	81.16
B_ME&B_SVM&F_SVM	279k	80.27	81.78
F_ME&F_SVM&B_ME	619k	79.63	81.04
F_ME&F_SVM&B_SVM	275k	80.27	81.78

表3はいくつかのモデルの組み合わせで訓練データを生成した場合に得られた訓練データのサイズと構文解析精度を表す。ただし、本実験では一致率は 100%、生データのサイズは 750 万語に固定している。表3より、モデル2つの組み合わせでは F_ME & B_SVM、3つの組み合わせでは F_ME & F_SVM & B_SVM が最適であることが分かる。

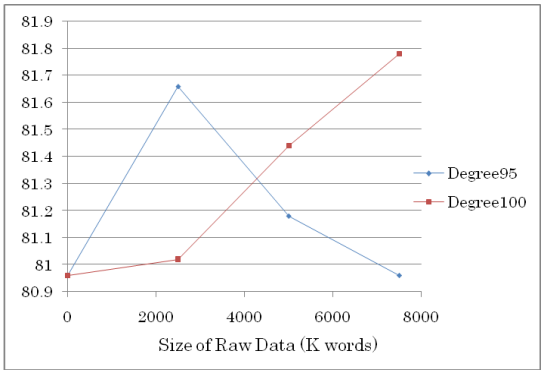

図2:一致率と生コーパスのサイズと構文解析精度(F_ME&F_SVM&B_SVM)

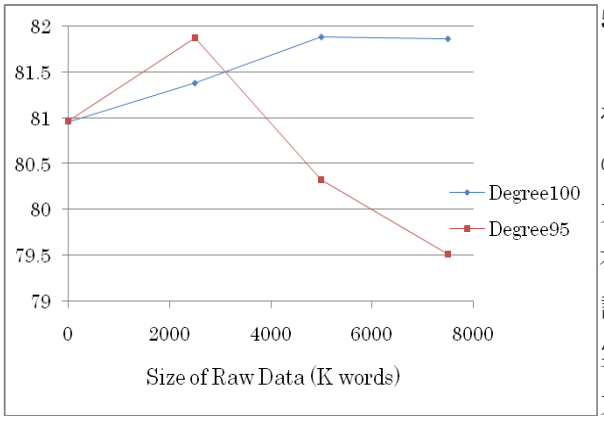

図3： 一致率と生コーパスのサイズと構文解析精度（F_ME&B_SVM）

図2と図3はそれぞれF_ME ＆ F_SVM & B_SVMとF_ME&B_SVMの組み合わせを用いて、一致率が 95%以上の文と 100%の文を訓練データとして WSJ に加えて再訓練した場合のテスト結果を表す。横軸は訓練データの自動生成の元となった生データのサイズを表す。この図により、最適な一致率は生データのサイズと大きく関係していることが分かる。生データが十分にある場合は、モデル3つの組み合わせを用い、一致率 100%の文だけを選択した方がよい。これは、生成される訓練データが十分大きいため、より高い精度の訓練データを得られる設定がよいと考えられる。一方、生データが十分でない（例えば300万語以下）場合は、2つのモデルの組み合わせを用い、また一致率を下げることで、訓練データのサイズをある程度確保した方が適切だと考えられる。

Sagae&Tsujii(2007)では生物のテストで評価した場合は LAS80.23%、UAS81.74%となっているのに対して、F_ME&B_SVM を用いた場合では最高で LAS80.33%、UAS81.86%、F_ME&F_SVM& B_SVMを用いた場合では最高でLAS80.45%、UAS82.04%という結果が得られ、0.2～0.3%の上昇が見られた。

5 結論

本論文では、Sagae & Tsujii (2007) の手法をベースとして、分野適応における訓練データの自動生成手法を提案した。我々は、まず構文解析精度に影響する要因として、文長の分布、モデルの組み合わせ、解析結果の一致率、訓練データのサイズ、の4つについて調査した。実験結果より、これら4つの要因は最終的な構文解析精度に影響することが示され、またこれらの要因のそれぞれの最適値は生データのサイズによることが明らかとなった。生データが十分にある場合は訓練データの精度を、生データが十分にない場合は訓練データのサイズを重視した方が適切だということが示された。また、いずれの場合も文長の分布を対象テキストと同等にした方がよいことが判明した。また既存の結果より0.2～0.3%の上昇が見られた。

本論文では生データのサイズを 750 万語に固定したが、生データをさらに増やし、より高い精度の訓練データが得られるようなモデルの組み合わせを用いることで、さらに解析精度を向上させることができると考えられる。

参考文献

K. Sagae, J.Tsujii. 2007. Dependency Parsing and Domain Adaptation with LR Models. EMNLP-CoNLL'07 Shared Task. http://acl.ldc.upenn.edu/D/D07/D07-1111.pdf

N.Joakim,H.Johan,N.Jens,C.Atanas . 2007 MaltParser:Alanguage-independent system for data-driven dependency parsing Natural Language Engineering 13(2): 95-135. http://w3.msi.vxu.se/~nivre/papers/nle07.pdf