構文解析の分野適応における精度低下要因の分析及び分野間距離の測定手法

張本佳子　宮尾祐介　辻井潤一
東京大学大学院情報理工学系研究科コンピュータ科学専攻
E-mail: {harry, yusuke, tsujii}@is.s.u-tokyo.ac.jp

1.はじめに

分野適応とは新聞コーパスなど既存の正解つきコーパスを用いて訓練された構文解析器を化学、物理など訓練データと異なった分野に適応させることである。分野別コーパスを開発するには多大な労力と金銭が費やされるため分野適応の重要性がますます認識されるようになってきた。特に CoNLL2007 では共通タスクとして採用され、Sagae ら（2007）に代表される様々な手法が提案された。

訓練データと異なる分野に対しての構文解析精度が低下する原因は主に未知語や文長の分布など構造上の違いによるものだと考えられてきたが、現在まで精度低下の原因を定量的に分析する実験がなかった。そこで本研究ではまずその原因を分析し、実際には未知語による影響が意外に少なく、いくつかの品詞の高頻度語の使われ方の違いこそ大きな要因であることを示す。さらにこの分析結果に基づき分野間の距離の測定手法を提案する。具体的にはまず分野ごとに重要品詞の高頻度語の頻度や文長の分布、品詞の分布を調べ、それらの差に基づいて距離を求める。本手法により、正解データのない分野に対し、距離の近い分野の既存コーパスを用いることで当該分野の構文解析精度が向上することを示す。

2.背景

分野適応については現在までにたくさんの手法が提案された。Steedman ら（2003）は co-training の手法が分野適応においても有効だということを発見し、Clegg ら（2005）は voting など様々な構文解析器を同時に利用することで生物分野に対する解析精度が改善されることを示した。その他に McClosky ら（2007）は Reranking や Self-Training の手法を分野適応に応用することで解析精度が大きく向上したことを発見し、Sagae ら（2007）は対象分野の生データを様々な解析器で解析した結果を統合することによって自動的に訓練データを構築する手法を提案した。

しかし解析精度の低下原因を定量的に調べる実験は今までに行われていない。また、これらの研究は対象テキストの分野が予め分かっていることを前提としており、そもそも対象テキストがどの分野のテキストであるかを自動認識する手法が別途必要である。

3.データ

本実験では Penn Treebank（Marcus ら, 1994）を用いる。具体的には WSJ （Wall Street Journal）という既存の新聞コーパスと BROWN という既存の分野別コーパスを使用する。BROWN コーパスの各分野は表1に示す。以降では分野名を省略記号で表すことにする。単純比較しやすいために各分野のサイズを 6 万語に合わせ、そのうち 5 万 5 千語を訓練に、5 千語をテストに用いた。また CF、CL、CP の 3 分野を実験に、CG、CK、CN の 3 分野を最後の検証に使用した。また、構文解析器は Sagae ら（2007）の依存構造解析器を用いた。

表 1：BROWN コーパスの各分野

省略記号	分野	平均文長（語）
CF	伝説	24.43
CG	手紙	26.56
CK	フィクション	18.24
CL	ミステリー	16.72
CN	冒険物語	17.01
CP	恋愛物語	18.45

図1:構文解析精度に影響する要素の分解

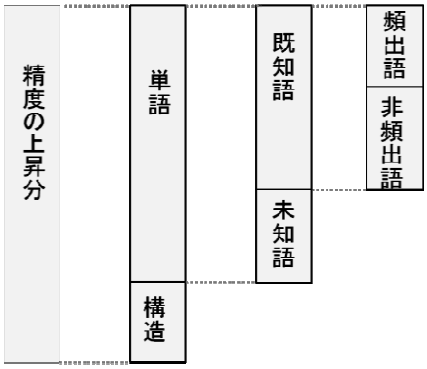

表 2:要素別解析精度への影響力

	上昇	構造	単語	未知語	既知語	非頻出語	頻出語
CF	1.44	0.71	0.73	-0.06	0.79	0.3	0.49
CL	3.47	1.72	1.75	0.42	1.33	-0.12	1.45
CP	3.19	1.81	1.38	-0.08	1.46	0.22	1.24

表 3:品詞別解析精度への影響力

品詞	割合 (%)	精度変化 (%)
名詞	18.46	-0.01
動詞	16.81	-0.51
代名詞	9.66	0
前置詞	9.30	-2.01
冠詞	8.74	-0.04
.	7.01	0.18
副詞	6.42	-0.65
形容詞	4.91	-0.07
,	4.81	-0.6
接続詞	2.95	-0.64
TO不定詞	2.22	-0.09
助動詞	1.45	-0.04

4.精度に影響する要素の分析

精度に影響を与える各要素の影響力を定量的に調べるためにまず図1のように精度の上昇に貢献しうる要素を分ける。各要素の影響力は WSJ だけを訓練データに用いたときに比べ、BROWN の中でその要素に関する情報を訓練データに加えたときの精度の上昇度で測る。

この実験では訓練には WSJ(99 万語)と Brown の注目分野以外のデータ(6 万語×5 分野≒30 万語)を用い、テストには BROWN の注目分野 のデータ(6 万語)を使用した。注目分野をそれぞれ CF、CL、CP とした時の結果を表 2 で示す。ただし頻出語とは頻度が上位 500 位以内に入っている単語であり、非頻出語はそれ以外の単語を指す。

また品詞別の単語情報の影響力をも調べた。各分野の訓練データを品詞ごとに単語情報を抜いた時の精度の低下度合で各品詞の単語情報の解析精度への貢献度を測る。

表 3 より、品詞によって精度への影響力がかなり異なり、特に読点、前置詞、接続詞の影響力が大きいことが分かる。

5.分野間の距離

5.1 距離の基準の作成

分野 AB 間の距離の基準を以下のように決める。分野Aと分野Bの各々の学習データで訓練された構文解析器が分野 A のテストデータに対して出した解析結果をそれぞれ結果A、結果Bとすると、結果Aに対する結果 B の一致率で距離を決める(図 2)。両結果の一致率が高いほど、分野Bの訓練データが対象分野 A の訓練データと似た傾向があると考えられる故、一致率が高いほど距離が短いとする。

この手法による距離基準は表 4 のようになる。ただし各列は対象分野を表し、各行は訓練データに用いる分野を表す。各数字は対象分野の学習データで訓練したモデルが出した結果との一致率を表す。

図 2:距離基準の作成方法

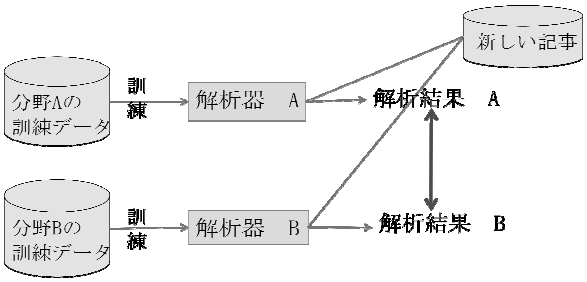

表 4:距離の基準

	CF	CG	CK	CL	CN	CP
CF-MODEL		84.48	83.04	82.84	83.32	82.96
CG-MODEL	84.48		82.56	82.1	82.15	82.41
CK-MODEL	83.04	82.56		83.87	84.15	84.35
CL-MODEL	82.84	82.1	83.87		84.23	84.18
CN-MODEL	83.32	82.15	84.15	84.23		84.38
CP-MODEL	82.96	82.41	84.35	84.18	84.38	

表 5:全単語の単語別頻度による距離の測定結果

	CF	CL	CP
CF		17013	16765
CG	10657	19098	18773
CK	12155	9776	9435.4
CL	12314		8512.9
CN	12920	8741	9340.2
CP	12463	8586	

表 6:CF、CL、CP と他分野の測定距離

	CF	CL	CP
CF		1172	1182
CG	353	1477	1479
CK	969	262	257
CL	1155		274
CN	1243	248	277
CP	1173	278	

5.2 距離の測定法

5.1 節で距離の基準の作成方法を紹介したが、本実験の目標は生データだけの情報でこの基準にできるだけ一致した分野間の距離を定義することである。ここでは分野 X と Y の各 feature(特徴成分)を正規化した後の値を Xi, Yi とすると、分野 X と Y の距離は以下の式のように Xi と Yi の平方根の差の和とする。

$$\sum_{i=0}^{i=n}(\sqrt{Xi}-\sqrt{Yi})$$

これは各分野全体を一つのベクトル、各 feature をそのベクトルの要素と見なした時正規化したベクトルの大きさが等しいという性質に注目した式である。たとえば全単語の単語別の頻度に注目した場合は以下の手順で距離を測る。

(1)分野ごとに各単語の出現回数を総単語数で割り、各単語の出現頻度を計算する

(2)分野ごとに各単語の頻度リストを作り、分野 X と Y の距離を測るには X と Y の単語別頻度の平方根の差を足して距離を出す

この手法による距離の測定結果を表 5 で示す。

5.3 距離測定法の評価

本実験では分野間の距離測定に全単語の単語別頻度の差、頻出語の単語別頻度の差、文長の分布の違い、品詞の分布の違いなどに着目した様々な手法を用いた。それぞれの測定距離と 5.1 節で定義した距離基準との相関係数で採用すべき手法を実験的に決める。精度と距離が逆の相関関係にあるため相関係数の絶対値で評価する。

例えば全単語の単語別頻度による手法では CF、CL、CP において測定された距離と距離の基準の相関係数の絶対値はそれぞれ CF:0.91、CL:0.82、CP:0.87 であり、平均値は 0.87 である。この値が大きいほどその距離測定手法がよいと考える。

様々な手法で距離を測り各測定手法を評価したところ、二つの連続した品詞(例:名詞＋名詞、形容詞＋名詞)の頻度の差に着目した手法が一番よいことが実験的に分かり、その時の相関係数絶対値の平均は 0.91 である。この手法による距離を表 6 に示す。またこの距離による分野の近さは以下のようになる。

CF: CG<CK<CL<CP<CN

CL: CN<CK<CP<CF<CG

CP: CK<CL<CN<CF<CG

これは距離基準とかなり近い結果である。

6.検証

各分野についてその分野の訓練データ(5 万語)とその分野と距離の近い順の三つの分野(15 万語)と遠い順の三つの分野(15 万語)をそれぞれ訓練データとして使った時の解析精度を比較する。その結果を表7で示す。表7より、訓練データの選択において本稿で提案した距離測定法は有効な指標になることが分かる。

表 7: 応用手法の測定結果(実験用)

	CF	CL	CP
近い三つで	81.15	84.24	81.47
遠い三つで	79.34	83.52	80.84
自分で	82	80.62	78.55

表 8:CG、CK、CN と他分野の測定距離

	CG	CK	CN
CF	350	979	1259
CG		1287	1589
CK	1269		243
CL	1448	263	249
CN	1562	239	
CP	1457	263	280

表 9:CG、CK、CN と距離基準の相関係数

CG	CK	CN	AVERAGE
0.89	0.93	0.97	0.910264

表 10:応用手法の測定結果(検証用)

	CG	CK	CN
近い三つで	80.15	83.84	83.01
遠い三つで	78.54	83.42	82.5
自分で	79.7	82.15	80.01

検証のために CG、CK、CN の 3 分野に対しても同様な距離測定法を用い距離を測り、その結果を表 8 に、距離基準との相関関係を表 9 で示す。この 3 分野においても距離基準と高い相関性を示した。またこの 3 分野に応用方法を適用させたところ表 10 のような結果となり、応用方法の有効性が検証された。

7.結論

本論文では訓練データと異なる分野に対して構文解析精度が低下する要因を定量的に調べた。その結果、精度の低下は従来考えられた未知語などによるものではなく、いくつかの品詞の高頻度語の使われ方の違いによるものだということが分かった。また分野間の距離を測定するために特定品詞の単語の頻度、全単語の頻度、文長の分布、品詞の分布など様々な手法を試したところ、実験的に品詞と品詞の連続の頻度の違いを用いた手法が一番よい距離を出すことが分かり、検証でも同様な結果を得られた。

また本稿で提案した手法で測定した距離を用いて訓練データのない分野に対して既存コーパスより訓練データを選択したところ、確かに対象分野との測定距離の近い分野の訓練データが対象分野の学習に相応しいことが検証された。更に距離が近い分野のコーパスが十分に存在している場合、少量の対象分野の訓練データよりも良い解析精度を出すこともあった。これは分野適応において重大な意味を持つと考えられる。

参考文献

A. B. Clegg and A. Shepherd. 2005. Evaluating and integrating treebank parsers on a biomedical corpus. In Proceedings of the ACL Workshop on Software.

M. Marcus, G. Kim, M. A. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger. 1994. The Penn Treebank: annotating predicate argument structure. In Proceedings of the Workshop on Human Language Technology.

D. McClosky, E. Charniak, and M. Johnson. 2006. Reranking and self-training for parser adaptation. In Proceedings of ACL-2006.

K. Sagae and J.Tsujii. 2007. Dependency parsing and domain adaptation with LR models. In Proceedings of EMNLP-CoNLL'07 Shared Task.

M. Steedman and M. Osborne. 2003. Bootstrapping statistical parsers from small datasets. In Proceedings of ACL-2003.