CRF を用いたブログからの固有表現抽出

齋藤 邦子† 鈴木 潤‡ 今村 賢治†

NTT サイバースペース研究所† NTT コミュニケーション科学基礎研究所‡

{saito.kuniko,imamura.kenji}@lab.ntt.co.jp jun@cslab.kecl.ntt.co.jp

1 はじめに

ブログの普及が進む現在、ブログの情報源としての価値が高まっており、ブログを対象とした情報抽出の研究が盛んになってきた。中でも、ブログ上に書かれた個人の意見、例えば、企業や新製品に対する評価や TV 番組・映画・書籍等のメディアについての感想などのいわゆる評判情報に対しては、別のユーザの購買行動に影響を与えたり、市場のトレンドをいち早く表していたりするということで、近年注目が集まっている。NTT でも goo においてブログ検索やブログ上の意見評判などを抽出する検索サービスなどを一部実験的ではあるが行っている[1]。評判抽出サービスを実現するには、抽出したい評判情報の対象が人名・地名・店名・商品名などの固有表現(Named Entity:NE)であることが多いため、ブログに対して精度良く固有表現抽出を行う技術が必要になる。しかしブログなどの CGM(Consumer Generated Media)は新聞と比較すると、文体は書き手に応じて様々であり、話題も幅広く移り変わりも早いため、新聞を対象とした固有表現抽出よりも難しいタスクであることが予想される。

一方、近年、固有表現抽出などの自然言語処理では HMM(Hidden Markov Model)などの生成モデルから CRF(Conditional Random Fields)などの識別モデルが成功を収めている[2]。そこで本稿では、CRF を用いたブログからの固有表現抽出を評価し、新聞およびブログドメインの比較、HMM との性能差、更に固有表現分類を拡張した場合の固有表現抽出結果について報告する。

2 ブログと新聞のドメイン比較

HMM や CRF など、統計的言語モデルを用いて固有表現抽出を行う場合、処理対象となるテキストとモデルのドメインがどの程度適合しているかは処理性能を左右する。直感的には処理対象のテキストと同時期で同ドメインのテキストを学習データとして言語モデルを作成すれば高い解析精度を達成すると期待できる。しかし、今 Web 上に存在するブログを処理対象とする場合、同時期で同ドメインの学習データを維持するのは時間とコストがかかるため、実際のシステム利用においては固定した時期やドメインでの学習データを作成し、モデルを学習するのが実情である。そのため、異なるドメインや時期のモデルでどの程度性能が出るのかは常に懸案事項である。そこで、固有表現抽出という観点から見て、ブログドメインは新聞ドメインと比較してどのような違いや難しさがあるのか調査した。

我々は、独自に新聞ドメイン：5.3 万文、ブログドメイン：3.5 万文の固有表現抽出正解データを収集・整備した。固有表現抽出は IREX 日本語固有表現抽出タスクの定義に準拠した[3]。なお、本稿では IREX 定義の固有表現のうち、人名・地名・組織名・固有物名に注目し、それぞれ PSN・LOC・ORG・ART と表記する。新聞ドメインは 2004 年 7 月～2005 年 5 月の新聞を中心に、更にそれ以前の古い記事を収録してあり、ブログドメインは 2004 年 9 月～2005 年 12 月のブログを中心に収録してある。

この正解データを利用して、ブログおよび新聞ドメインに登場する固有表現分布を調べると、上位から順に

ブログ PSN:23% LOC:22% ART:18% ORG:17%

新聞 LOC:27% ORG:23% PSN:16% ART:6%

のようになった。

ブログでは新聞と比較して PSN・ART が増えており、特に PSN・ART について、ブログドメインデータで出現頻度の多いものを上位 5 件挙げると PSN:{剛・中居・伊東美咲・大黒・ジョニー}、ART:{電車男・スターウォーズ・新健康磁化水カップ・iPod}となっており、ブログ上では芸能人や TV 番組、商品名などが多く登場していることが窺える。

続いてブログおよび新聞ドメインの固有表現としての語彙の重なりを調べるため、両ドメインの正解データに登場する固有表現をそれぞれリストアップし、ブログドメインの固有表現のうち、新聞ドメインでも登場する固有表現の割合を調べた。すると、ラベル全体では 50%であり、ラベル別には LOC:70%

ORG:40% PSN:30% ART:12%であった。即ち、両ドメインでは、LOC の重なり率は高いが、その他は 4 割以下であり、特にブログでよく登場する ART はきわめて低い。実際、ブログでは前述の通り、芸能人や TV 番組・商品名が頻出するが、新聞ドメインで出現頻度が高いものは、PSN では{ブッシュ・小泉・森・松井秀喜}の政治家やスポーツ選手、ART では{NASDAQ・政治改革関連法案・イラク復興支援特別措置法・アカデミー賞、政治資金規正法}のように政治経済関連の用語が目立つ。このように、両ドメインではトピックが必ずしも一致しておらず、ブログで出現する固有表現を新聞でカバーするのは難しい。

なお、3 章以降の実験では、両ドメインとも 1000 文ずつランダムにテストデータを抽出し、残りをモデルの学習データとして用いる。そこで、この評価実験のタスクの難易度を学習データの固有表現語彙という観点で調べてみた。表 1 は各ドメインの学習データについて、テストデータに対する固有表現カバー率を示す。つまり、テストデータに出現する固有表現のうち、それぞれのドメインでの学習データにも登場するものの割合を示したものである。新聞ドメインでは同じドメインの新聞学習データであれば 9 割近く固有表現をカバーできるのに対して、ブログドメインでは同じブログ学習データでも 6 割強、異なるドメインの新聞学習データでは 4 割強しかカバーできない。つまり同じブログドメインの学習データを使っていてもテストデータに対する固有表現カバー率は低く、まして異なる新聞ドメインの学習データでは更に固有表現カバー率が低いという難しさがある。

続いて、データの収集時期の影響を調べるために、テストデータをモデル学習データの収集時期よりも 3 ヵ月新しいデータにして同様の評価をしたところ、ブログ学習データのブログテスト(新)に対するカバー率は 50%に落ちた。一方、新聞学習データの新聞テスト(新)に対するカバー率は 72%と高い水準を保っており、このことからも、ブログからの固有表現抽出は学習データの収集時期の違いで更にモデルの性能が落ちるリスクがあり、新聞ドメインと比較して難しいタスクである。

表 1. 学習データの固有表現カバー率

	学習データドメイン	
	新聞	ブログ
ブログテスト	44%	63%
新聞テスト	87%	51%

3 固有表現抽出実験

2 章で述べたように、両ドメインの学習データから CRF および HMM の固有表現抽出モデルを学習し、テストデータを利用して open test 方式で評価した。

なお、形態素解析器として JTAG を利用した[4]。また、CRF の学習には学習誤り最小法に基づく CRF 学習モジュールを用いた[5]。本稿では詳細な説明を省略するが、これは誤り最小化の枠組みを利用した学習法であり、誤りを推定するあらゆる評価関数を学習時の目的関数として利用可能である。本実験では F 値を目的関数として設定した。HMM の学習には我々が開発した多言語固有表現抽出エンジンを利用した[6]。

3.1 CRF モデルの学習曲線

図 1 にブログ・新聞両ドメインについての学習曲線を示す。凡例はテストドメイン-モデルドメインを示しており新聞-新聞は新聞テストを新聞ドメインモデルで処理したことを意味する。なお、HMM の結果を各ドメインともに学習データを全て利用した時のみ示した。

新聞モデルは新聞テストに対して F 値 92%の高精度で処理するが、ブログテストでは F 値 60%に低下する。一方、ブログモデルは新聞・ブログテストともそれほど精度に差が無く F 値 75%であった。この現象は表 1 で示した固有表現カバー率に由来すると思われる。また、CRF はどちらのドメインでも HMM で全学習データを投入した場合と比べ 1/3～1/5 である 1 万文程度で同等の性能を達成し、HMM より低コストで学習できている。しかし、新聞テストを新聞モデルで解析する場合は 92%と高性能であった CRF もブログテストではブログモデルを用いても 75%と、ブログドメインのタスクの難しさが現れている。実際のシ

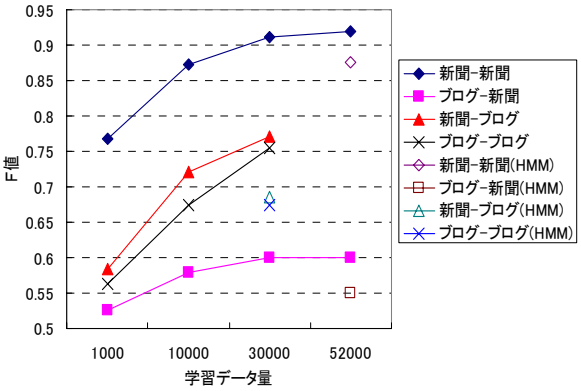

図 1. 学習曲線

ステムでは時期の古いモデルやドメインの異なるモデルを利用することが想定され、更なる精度の低下につながる。新聞ドメインと同程度の精度を達成する単純なアプローチは新聞ドメインより多くの学習データを集めることである。しかし、荒く見積もると、実験結果では3万文程度のブログモデルでは1000文の新聞モデルと同程度の精度であり、且つ、新聞ドメインでは 3 万文の新聞モデルで精度がほぼ上限に達することから、ブログモデルではざっと 90 万文程度の学習データが必要となる可能性がある。

3.2 混合モデル

実際のシステムでは少しでも沢山の学習データを集めるために、単独のドメインのみではなく複数のドメインのデータを統合して利用することも多い。特にブログの正解データを大量に整備することはコストもかかるので既存の新聞コーパスに少しずつブログデータを加えてモデルを充実させることがある。そこで新聞モデルにブログモデルを段階的に混合し、それぞれの性能を評価した。新聞モデルは図 1 の結果を踏まえて、性能がほぼ上限に達する 3 万文規模をベースラインとし、そこにブログモデルを 1000 文、1 万文、3.4 万文と追加し、新聞・ブログテストを解析した。

図 2 に各テストでの F 値を示す。新聞テストに対しては高精度を維持しつつブログテストの解析精度を向上させ、新聞モデルのみよりも約 23%改善した。ブログに対しては性能向上の効果が出ている。

ただし、図 1 よりブログモデル単体では F 値が 76%程度であり、図 2 の混合モデルの 74%は単体モデルのレベルに達していない。この原因について詳細はまだ明らかではないが、まず考えられるのは、元々固有表現カバー率が低い状態で異質のモデルを混ぜるのは性能を下げるリスクがあるということである。一方、新聞モデルに異質のモデルを混ぜても新聞ドメインに対する性能評価には影響がでなかった。これは新聞モデルが新聞ドメインに対しては既に高精度に学習ができていたためであると考えられる。また、別の可能性としてブログと新聞での固有表現の書かれ方には違いがあると考えられる。ブログは CGM ならではの傾向として、登場する固有表現の種類や表記にも多様性がある。例えばブログでしばしば登場する芸能人の名前はカタカナや英語と漢字が混じったものがあるが、そこに新聞のように整然としたテキストを学習したモデルが加わることで、漢字の部分だけを認識してしまう。また本来固有表現ではないが、事物を擬人化して○○ちゃんと表記したものを誤って人名と認定した例もあった。その他、新聞モデルを加えることで、ORG を新聞ドメインで優勢な LOC へ認定しがちという現象もある。いずれにしても実システムでは異なるドメインモデルの混ぜ方に注意が必要である。

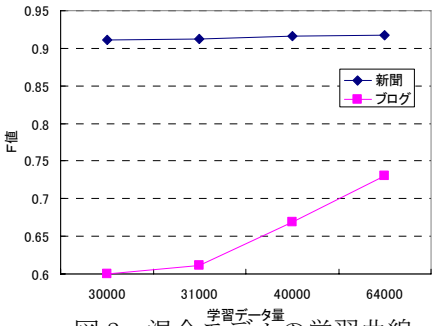

図 2. 混合モデルの学習曲線

3.3 未知語の影響

以降は、ブログテストをブログドメインモデル（3.4 万文）で解析するという条件で話を進める。

2 章で述べたように、ブログドメインでは学習データの固有表現カバー率が 6 割強しかなく、更に処理対象が学習データの収集時期よりも 3 ヵ月新しくなると 5 割に落ち込むため、ざっと半分は未知の固有表現を抽出しなければならない状況であると言える。また固有表現抽出においては、固有表現として未知語であるだけでなく、形態素解析の誤り、特に未知形態素の出現の影響も受ける。本稿では前者を NE 未知語、後者を形態素未知語と呼ぶことにする。新聞テストでは全形態素中の形態素未知語の出現率は 0.7%だがブログテストでは 3%に上昇し、また形態素未知語が固有表現の一部を構成している割合は、新聞・ブログいずれの場合も 6-7 割であった。つまり形態素未知語が出現する箇所は固有表現部分に集中する。更に、形態素未知語を含む固有表現はそれ自体が NE 未知語となる傾向があり、実際ブログテストでは全固有表現の 2 割に形態素未知語が含まれ、そのうち 95%が NE 未知語であった。以上を踏まえ形態素未知語または NE 未知語が解析精度にどのような影響を与えているかを分析した。

分析には、テスト正解データの固有表現が、NE 既知語・未知語であるか、およびその固有表現に含まれる形態素が既知語・未知語であ

るかの状況によって、システム出力が正しく固有表現と認定できたか、すなわちシステムの正解率（再現率）で評価した。表 2 はブログテストについて CRF と HMM で比較したものである。NE・形態素ともに既知語であれば CRF と HMM では差がないが、いずれかが未知語になった時に CRF が優勢となる。特に形態素未知語の場合は NE 既知語・未知語ともに CRF の優位性が際立っている。この背景として、CRF ではモデル学習で利用する素性、即ち言語的特徴の種類が HMM よりも柔軟で豊富であることが考えられる。本実験では、HMM の学習には連続する 2 つの形態素と固有表現ラベルの情報を用いるが、CRF では連続する 2 つの固有表現ラベルの他に、現位置の形態素の前後 2 単語の範囲、すなわち 5 つの連続する形態素での表記や品詞を単独で、または組み合わせた形で利用した。そのため形態素未知語が出現した際、HMM では確率が付与できなくても CRF ならば多くの素性を利用して確率を付与し、正解を出力できる。これは HMM も形態素既知語・NE 既知語であれば CRF と遜色ないことからも裏付けられる。また CRF は NE 未知語でも形態素既知語であれば 6 割弱の正解率であるから、形態素解析器の辞書を充実することで固有表現抽出を補うことができることを示唆しており、この点も HMM にはない強みである。しかし CRF でも NE・形態素ともに未知語であると正解率は 50%に満たず、今後の改善の鍵である。

表 2. NE および形態素の既知・未知別正解率

	NE 既知語		NE 未知語	
	CRF	HMM	CRF	HMM
形態素既知語	85%	86%	57%	45%
形態素未知語	70%	38%	46%	29%

3.4 固有表現分類の拡張

実際にブログを対象とした評判検索システムに固有表現抽出を適用すると、IREX 定義より詳細な固有表現分類が必要となることが多い。拡張固有表現については既に関連研究があるが[7]、我々は主に評判抽出を目的として表 3 に示す体系を新たに定義し 4.2 万文（ブログ 3.4 万文、新聞 8000 文）の正解データを整備した。このデータから 1000 文をテストデータとし、残りを学習データとして CRF・HMM の評価をしたところ、F 値は CRF：0.64、HMM：0.55 と、CRF が高精度であり、ラベル別では特に ART:タイトル 14%、PSN:芸能人 16%と、F 値で HMM との差が大きく出た。これは 3.3 節でも一部述べたが、タイトルは長い固有表現が多く、広い形態素範囲で素性を使う CRF が有利であり、また芸能人は未知語形態素も多いため CRF の優位性が出たと考えられる。

表 3. 拡張固有表現分類

従来	拡張後
ART	イベント、,タイトル、製品名、,その他
LOC	施設、GPE、地形、その他
ORG	企業、教育機関、政治、スポーツ機関、その他
PSN	政治家、スポーツ選手、芸能人、その他

4 まとめ

CRF に基づくブログからの固有表現について評価・分析し、新聞ドメインとのタスク比較、HMM と CRF の性能差分析、固有表現分類の拡張などを行った。ブログは新聞と比べて同じドメインでの学習データの固有表現カバー率が低く、また時期のずれによってもカバー率の低下があることから新聞ドメインより難しいタスクである。しかし CRF は HMM の 1/3～1/5 程度の学習データ量でも同程度の性能を発揮し、且つ HMM よりも形態素未知語や NE 未知語に強いという特徴がある。特に CRF は HMM よりも形態素未知語に強いという傾向があり、また NE 未知語であっても形態素解析辞書を強化すれば固有表現抽出精度を向上できる可能性もあることを確認した。これは CRF が HMM よりも豊富で柔軟な素性を扱えることが要因と考えられる。今後は NE 未知語や形態素未知語の事例に対してより効率的に学習するアプローチ、例えば active-learning や semi-supervised learning 等の検討が課題である。

5 参考文献

[1] http://www.goo.ne.jp/ , http://labs.goo.ne.jp/

[2] Lafferty, J., McCallum, A. and Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, Proc. of ICML, pp.282-289, 2001.

[3] IREX 実行委員会（編）. IREX ワークショップ予稿集, 1999. http://nlp.cs.nyu.edu/irex/index-j.html

[4] Fuchi, T. and Takagi, S.: Japanese Morphological Analyzer using Word Co-occurrence-JTAG, Proc. of COLING-ACL, pp.409-413,1998.

[5] Suzuki, J., McDermott, E. and Isozaki H.: Training Conditional Random Fields with Multivariate Evaluation Measures, Proc. of COLING-ACL, pp.617-624,2006.

[6] Saito, K. and Nagata, M.: Multi-Language Named Entity Recognition System Based on HMM,. Proc. of ACL Workshop on Multilingual and Mixed-language Named Entity Recognition, pp.41-48, 2003.

[7] 新納浩幸, 関根聡:拡張固有表現タガーの作成とその問題点の考察, 言語処理学会第 12 回年次大会 発表論文集, pp.105-108, 2006.