固有表現抽出における事後確率を用いた能動学習

齋藤 邦子　今村 賢治
NTT サイバースペース研究所
{saito.kuniko, imamura.kenji}@lab.ntt.co.jp

1　はじめに

近年、様々な言語処理のタスクにおいて、大量の正解コーパスから学習した統計的言語モデルを解析に用いるアプローチが広く普及している。このアプローチでは言語の文法的な知識を統計的な特徴量として捉えることができ、形態素解析や固有表現抽出、機械翻訳など、多様な自然言語処理で活用されている。

このような統計的手法に基づく言語処理では、モデルを学習するための正解コーパスをいかに効率よく低コストで収集・作成するかが常に課題となってきた。正解コーパスを全て人手で作成すれば高精度なモデルを得られるが、膨大な時間とコストがかかる。一方、デコード結果をそのまま正解コーパスとして流用すると、時間もコストもかからないが、コーパスに解析誤りを多く含むため、モデルの精度は向上しない。また、一旦ある程度の量の正解コーパスを人手で作成しても、ドメインやデータ収集時期が異なるテキストに対しては解析精度が低下することがしばしばある。そのため、モデルを常に高品質に維持するためには、正解コーパスを追加してモデルを適宜更新することも必要になる。その際、追加するコーパスには今までのコーパスに出現していなかった新しい知識が多く含まれるほど、効果的な学習が期待できる。

本稿では、固有表現抽出タスクにおいて、事後確率を利用してデコード結果の解析誤りを識別する手法を提案し、その効果について報告する。この手法により、現モデルでは解析を誤る事例を効率よく発見できるため、学習効果の高いデータを優先的に選択する能動学習が実現できる。またこの特徴を利用した新規固有表現語彙獲得の実験についてもあわせて報告する。

2　固有表現抽出

固有表現抽出とは、テキストに含まれる人名、地名、組織名などの固有表現を抽出するタスクであり、n 語からなる単語列 $W = w_1 \cdots w_n$ に対して固有表現の種類を表す固有表現タグ列 $T = t_1 \cdots t_n$ を付与する系列ラベリング問題として考えることができる。CRF(Conditional Random Fields)[1]は、系列ラベリング問題を解く識別モデルとして成功を収めてきた。固有表現抽出に CRF を用いると、固有表現タグ列の確率は以下の式で表される。

$$P(T \mid W) = \frac{1}{Z(W)} \exp\{\sum_{i=1}^{n} \sum_{a} \lambda_a \cdot f_a(t_i, w_i)\} \quad (1)$$

$$Z(W) = \sum_{T} \exp\{\sum_{i=1}^{n} \sum_{a} \lambda_a \cdot f_a(t_i, w_i)\} \quad (2)$$

$Z(W)$ は正規化項、w_i と t_i は位置 i における単語と固有表現タグ、$f_a(t_i, w_i)$ は当該単語および固有表現タグがある条件を満たす時に 1 となる素性関数、λ_a は素性関数の重みである。

本稿では IREX[2]で定義される８種の固有表現を抽出対象とし、以下のように IOB2 方式[3]で表現し計 17 種のタグを想定する。

単語列 W	NTT	持株	会社	の	齋藤	です
タグ列 T	B-ORG	I-ORG	I-ORG	O	B-PSN	O

なお、形態素解析器として JTAG[4]、固有表現タグのラベリングには学習誤り最小法に基づく CRF 学習を利用した固有表現抽出モデルを用いた[5]。

3　全体構成

まず本稿が提案する学習スキームの構成を説明する（図 1)。本スキームでは、人手で作成した小規模正解コーパスからベースラインモデルを学習する。そして大規模平文データを形態素解析し、ベースラインモデルを利用してデコーダーで固有表現抽出を行う。その後、入力の各単語に対して付与されうる全タグのタグ信頼度を計算するが、このタグ信頼度の算出には事後確率を利用する。なお、この事後確率は事後確率最大化（MAP）で用いるような文単位のものではなく、各単語に付与されるタグ単位のものであるところが重要である。続いてリジェクターにて各単語のタグ信頼度を参照し、ベースラインモデルでデコードした固有表現タグをアクセプトするかリジェクトするかを判定する。アクセプトされたタグはそのまま正解とみなし、リジェクトされたタグのみ人手でチェック・修正し、新しい大規模タグ付きコーパスとする。そして、従来の小規模正解コーパスに追加して再学習し、ベースラインモデルを更新する。

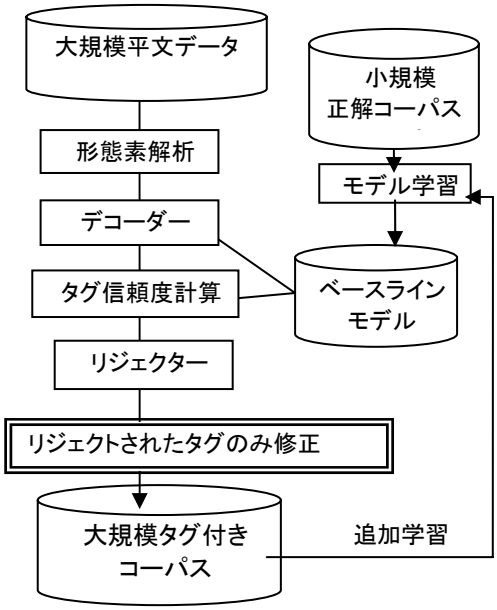

図 1. 学習スキームの全体構成

以下、事後確率を利用したタグ信頼度、リジェクターについて詳しく説明する。

3.1 事後確率を利用したタグ信頼度

タグ信頼度とはベースラインモデルで推定したタグの信頼度を表す確率値で、タグ毎の事後確率を計算して求める。

図 2 にタグ信頼度計算の模式図を示す。入力単語列 $W = w_1 \cdots w_n$ について、単語 w_i のタグ $t_{i,j}$ の信頼度は以下の式で計算する。なお、$j = 1, \cdots, k$ はタグの全種類に対応するものであり、本稿では 17 種である。

$$P(t_{i,j} \mid W) = \sum_{T} P(t_{i,j}, T \mid W) \tag{3}$$

これは、単語 w_i のタグが $t_{i,j}$ である全てのタグ列 T の事後確率を総和したものである。入力単語列 W の全単語について、取りうる全てのタグ列の事後確率を計算し、タグ信頼度を得る。

タグ信頼度の計算では、文頭から現位置のタグに至るまでの全ての経路の和 $\alpha_{i,j}$ と、現位置のタグから文末に至るまでの全ての経路の和 $\beta_{i,j}$ を使って以下のように分解できる。

$$P(t_{i,j} \mid W) = \frac{1}{Z(W)} \alpha_{i,j} \cdot \beta_{i,j} \tag{4}$$

$\alpha_{i,j}$ と $\beta_{i,j}$ は再帰的に以下のように計算する。

$$\alpha_{0,j} = 1 \quad , \quad \beta_{n+1,j} = 1 \tag{5}$$

$$\alpha_{i,j} = \sum_{k} \{\alpha_{i-1,k} \cdot \exp\{\sum_{a} \lambda_a \cdot f_a(t_i, w_i)\}\} \tag{6}$$

$$\beta_{i,j} = \sum_{k} \{\beta_{i+1,k} \cdot \exp\{\sum_{a} \lambda_a \cdot f_a(t_{i+1}, w_{i+1})\}\} \tag{7}$$

本稿では、式(4)の対数を取って

$$\log P(t_{i,j} \mid W) = \log(\alpha_{i,j}) + \log(\beta_{i,j}) - Z(W) \tag{8}$$

として計算した。

以上のようにして、通常は文単位で考える事後確率をタグ単位に適応することで、各単語に付与されるタグの信頼度が得られる。

3.2 リジェクター

リジェクターでは、デコーダーの出力とタグ信頼度を利用して、各単語でデコーダーが出力したタグを正解としてアクセプトするか不正解としてリジェクトするかを以下の手順で判定する。

(1) デコーダーの推定タグと信頼度 1 位のタグが不一致ならばデコーダーの推定タグをリジェクトする
(2) でアクセプトされた場合、信頼度 1 位のタグが閾値 θ 以下ならリジェクトし、そうでなければアクセプトする
(3) 最終的にリジェクトされたタグのみを人手で修正し、アクセプトされたタグはデコード結果をそのまま追加学習に用いる。

この手順では、信頼度 1 位の絶対値のみに着目して閾値を設定したが、例えば信頼度 1 位と 2 位の差や対数差など要するに信頼度 1 位が十分高いことを示す指標であれば良い。

閾値は実験的に設定するが、閾値が高ければリジェクトされるタグ数が増え、即ち人手のチェック・修正量が増える。本稿では、学習のコスト＝

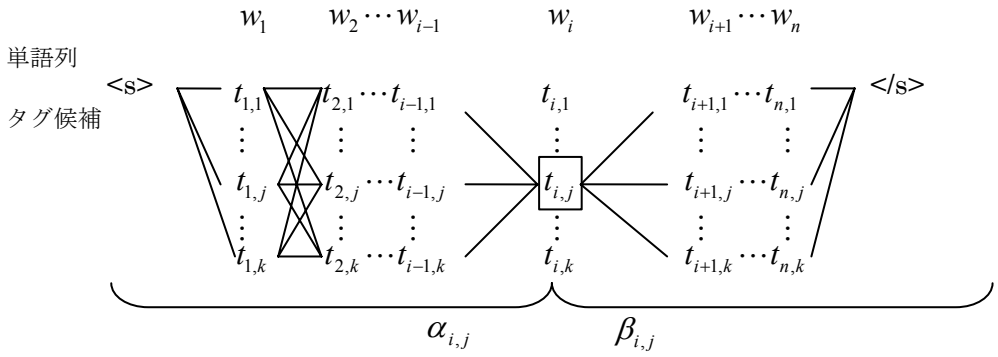

図 2. タグ信頼度計算の模式図

人手でタグをチェック・修正した語の割合（WRR: Word Replace Rate）とみなした。WRR が高ければ、正解の割合が増えてコーパスの精度が向上するが、学習コストも増大する。次章では、閾値を段階的に変化させて WRR を変え、追加学習したモデルの性能を評価した。

このように全ての正解タグを人手で付与するのではなく、リジェクターがリジェクトしたタグについてのみ人手で付与するため、従来の正解コーパス作成作業と比較してコストを抑えることができる。

4 実験

タグ信頼度とリジェクターを利用する学習の効果をブログドメインで評価した。実験データとして、ベースラインモデル学習のための正解コーパス約 1 万文、追加学習のための正解コーパス約 3.4 万文を準備した。いずれも予め人手で全ての正解タグが付与されているコーパスである。追加学習のための正解コーパスから 1000 文を test set とし、残りを kept set とした。

kept set をベースラインモデルでデコードし、更にタグ信頼度計算およびリジェクター処理を行った。そして、リジェクターでリジェクトされたタグについては予め付いていた正解タグと置き換えるが、アクセプトされたタグはデコード結果をそのまま採用した。これは、リジェクトされた箇所は人手で修正する状況を模した操作である。以上のようにして kept set のデコード結果を部分的に正解へ変換したものをベースラインモデルの学習データに追加してモデルを更新した。更新モデルで test set を解析し F 値で評価した。

リジェクターの閾値は 0.1 から 1.0 の範囲で 0.1 ずつ段階的に変化させ、WRR を変えながら F 値の変化を調べた。この追加学習との比較のため、通常の教師あり学習として kept set（全て正解タグ）を 5000 文、1 万文、2 万文、3.3 万文と、ベースラインモデルの学習データに段階的に追加して評価した。このときの WRR は、追加したデータ全てに人手の作業が入っているとみなし、WRR={追加したデータの全単語数}/{kept set 全体の全単語数}とする。

図 3 に本手法と教師あり学習の学習曲線を示す。教師あり学習の精度は WRR の増加に伴ってなだらかに上昇するのに対し、本手法の学習では初期の精度向上の立ち上がりが急激である。これは、本手法ではリジェクターによってデコーダーの解析誤りを効率よく発見していること、またそれらを優先的に修正することで学習効果の高い箇所を集中的に選択する能動学習ができたことを示唆している。

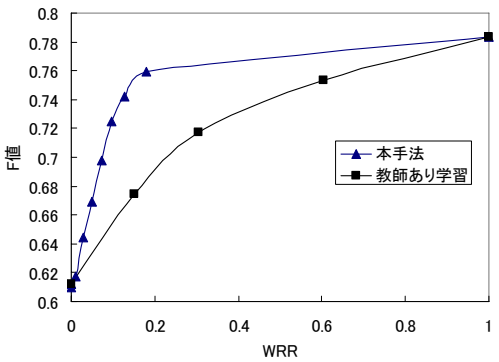

図 3. 学習曲線

5 新規固有表現語彙の獲得

4 章の実験により、タグ単位の事後確率から求めたタグ信頼度を利用すると、ベースラインモデルが解析を誤る箇所を優先的に発見して選択する能動学習が実現できることが示唆された。この特徴を利用してベースラインモデルの学習データには出現しなかった新規固有表現語彙を獲得する実験を試みた。手順は以下の通りである。

(1) ベースラインモデルを利用して各単語のタグ信頼度を 2 位まで求める
(2) リジェクターでタグ信頼度 1 位のタグが閾値以下であれば 2 位のタグまでをタグ候補とし、そうでなければ 1 位のみを候補とする。
(3) 残ったタグ候補から長さ優先で固有表現を収集し、固有表現を構成する語でタグ信頼度を平均して固有表現信頼度を求める
(4) 1～3 を全テキストに対して実行し、同じ固有表現が複数回登場した場合は固有表現信頼度を加算して、最終的な信頼度スコアとする
(5) 4 で収集できた固有表現から、ベースラインモデルを学習した正解データに出現しなかった固有表現を信頼度スコアの高い順に取り出して新規固有表現リストとする

表 1 を用いて 2～4 の処理を補足する。表 1 は、1 位のタグ信頼度が閾値 0.5 以下である時に 2 位のタグまで候補として残った状態である。「905」「i」の単語のみ 2 位まで残っている。このようにリジェクターによって部分的に 2 位までのタグ候補が残っている状態で取りうる固有表現を収集する。その際、長さ優先とするので、「905i/ART」は採用するが「905/ART」のような部分的固有表現は採用しない。そして「905i/ART」の固有表現信頼度 (0.208+0.201)/2=0.2045 を得る。この処理を全ての記事で行い、「905i/ART」が複数回得られた場合は、固有表現信頼度を加算していき、信頼度スコアとする。

表 1. リジェクター処理後のタグ信頼度

	タグ信頼度	
	1 位	2 位
やがて	NIL/0.846	
905	NIL/0.474	B-ART/0.208
i	NIL/0.417	I-ART/0.201
が	NIL/0.997	
でる	NIL/0.995	

この語彙獲得の精度を評価するため、同じベースラインモデルを利用して Nbest(N=2)解析の固有表現抽出結果からの語彙獲得と比較した。Nbset 解析の場合は固有表現単位の確信度が得られないため、単純に出現頻度をスコアとして利用した。4 章の実験と同じデータを用いて、kept set からの新規語彙獲得を行った。正解タグの付いた kept set から予め固有表現を収集した正解リストと、前述の 2 通りの手法で獲得した新規固有表現リストを比較し、新規固有表現リストにある固有表現が、元の正解リストにあれば正解、無ければ不正解とする。そして新規リストの上位 M 個の固有表現について、M を変化させながら precision-recall curve を測定した。なお

recall=全正解数/正解リストの総語彙数

precision=全正解数/M

で計算する。また、収集できた新規固有表現の総語彙数は本手法が 8694 語、2best 解析の手法が 5479 語であった。

図 4 で示す通り、本手法では 2best 解析の手法より 5-10%程度精度良く新規語彙を収集できた。その第 1 の要因は、本手法ではタグ単位の事後確率を計算してタグ毎の信頼度を得ている点が考えられる。そしてこのようにタグ単位で信頼度が得られることは 2 つの効果をもたらす。1 点目は、同じ 2 位までの候補を考慮する条件でも、文単位で 2best 解析の場合は、せいぜい 2-3 語程度でしか複数のタグ候補が存在しないことが多いが、本手法ではタグ単位で独立に 2 位までの候補を扱えるため、より幅広く固有表現の出現可能性を考慮できることである。2 点目は、ある固有表現が様々な文脈で出現する現象を、固有表現の信頼度という客観的な数値で評価できることである。これにより、文脈的により確実に固有表現と判断されるものが優先されるだけでなく、1 つ 1 つの出現での信頼度が低くても何度も出現するものはそれなりに確からしいという多数決の効果も加わる。

当初の狙い通り、事後確率に基づくタグ信頼度を利用することで、能動学習では学習効果の高い知識を優先的に選択できたように、新規語彙獲得でも新しい固有表現を効果的に発見収集できた。

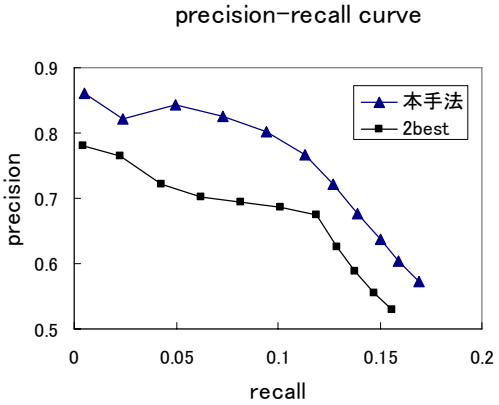

図 4. 新規語彙獲得の精度比較

新規固有表現語彙の獲得技術は、固有表現抽出処理そのものの精度向上や、ブログからの話題語抽出などに利用できる。

6 まとめ

本稿では、固有表現抽出タスクにおいて、各単語に付与された固有表現タグの信頼度を計算し、デコード結果の解析誤りを識別する手法を提案した。タグ信頼度はタグ単位の事後確率を用いて計算する。本手法を利用するとデコード結果の解析誤りを効率よく発見できるため、デコード結果を人手でチェック・修正する学習コストを通常の教師あり学習よりも大幅に削減できた。また、ベースラインモデルで解析を誤る箇所を優先的に発見できるため、学習効果の高いデータを効果的に選択する能動学習が実現できた。

更に、この能動学習の特徴に注目して、タグ信頼度を利用した新規固有表現語彙獲得法を評価した。Nbest 解析による語彙獲得法と比較して 5-10%収集精度が向上することを確認した。

いずれの場合も、タグ単位の事後確率に基づくタグ信頼度を利用することで、学習すべき新たな知識を優先的に発見して選択する能動学習の効果が発揮された。

7 参考文献

[1] 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.

[2] 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

[3] Sang, E.F.T.K. and De Meulder, F.: Representing text chunks, Proc. of EACL, pp.173-179, 1999.

[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.