対象，属性，評価語の相互依存関係を考慮した三つ組抽出

森田　一　　高村　大也　　奥村　学
東京工業大学，東京工業大学 精密工学研究所
morita@lr.pi.titech.ac.jp，{takamura，oku}@pi.titech.ac.jp

1　はじめに

Web 上の自然言語文から何らかの対象についての評価を調べる評判情報の抽出を目的としたときに，評判情報の中心となる評価表現の対象語，属性語および評価語の三つ組を抽出することは，非常に重要なタスクである．本研究では，一文中に存在する三つ組の抽出を系列ラベリングの一種として各語に対して対象語，属性語，評価語もしくはその他の 4 ラベルを付与する問題として扱う．小林ら [6]，鈴木ら [8] の研究では，後続する助詞や品詞により評価表現として選ぶ語を限定しているが，本研究では，より一般化された範囲の評価表現を得るため，評価表現に選ぶ内容には制限を加えていない．また，三つ組に含まれる表現も単語に限らず，任意の長さの単語列を扱う．これにより，「特製のタレが(属性語)とろけるほど美味しい(評価語)焼き鳥(対象語) が出てくるお店」のような，より複雑な評価表現を対象に含むことができる．五十嵐ら [7] も，三つ組抽出に枝分かれ同時確率を用いたモデルを提案しているが，三つ組の要素は単語に限定される．

近年，一般的な分類学習を拡張した枠組みである構造学習 [5] が提案され，様々な用途に用いられている．構造学習とは，一般的な二値分類の学習と異なり，ある事例が与えられたときに構造を持った出力をとる学習を行う機械学習の枠組みである．これにより従来では機械学習の適用が難しかったタスクや，より問題の構造を反映した手法に対して学習の効用を得ることができるようになった．三つ組抽出では，評価表現の各要素はそれぞれが強く関連しており，この相互の依存関係を利用した分類を行うことで，三つ組の各要素を判別する際の性能向上を見込むことができる．そのため，本研究では評価表現の要素単体についての素性のみならず，要素間で組み合わせた素性を分類と学習に用いる．一文中に存在する三つ組を同時に分類することで，この依存関係を捉えた三つ組抽出を行うための手法を提案する．

2　提案手法

2.1　手法概要

提案手法では，文から三つ組を抽出する問題を対象語，属性語，評価語もしくはその他の 4 ラベルを付与する制約付きの系列ラベリングとして解く．図 1 が示すように，各文は複数の三つ組を含む場合がある．このため，本稿でははじめに文が含む三つ組の数を判定し，その数に応じた三つ組の抽出を行う．三つ組の数は，三つ組の内識別しやすい評価語の数を評価語主辞の識別器を用いて判定する．識別器が出力した評価語の数 n に従い，モデル $\mathbf{w}$ に対してスコアが最大となる n-best の系列 $(\mathbf{y}_0, \ldots, \mathbf{y}_{n-1})$ を 2.3 節で述べるアルゴリズムにより探索し，抽出する．

系列を評価するモデル $\mathbf{w}$ は，各文について探索によって得られる出力が正解に近づくように学習を行う．学習の過程では，各文 $\mathbf{x}$ における探索の結果 $\bar{\mathbf{y}}$ と 正解ラベル系列 $\mathbf{y}$ からそれぞれ素性ベクトルを ψ 関数を用いて生成する．生成したベクトルの差分 $\psi(\mathbf{x}, \mathbf{y}) - \psi(\mathbf{x}, \bar{\mathbf{y}})$ が適切なマージンを持つようオンライン学習アルゴリズムを用いてモデルのパラメータを更新する．これを繰り返し行うことで正しい系列が最もスコアが高く評価されるようにモデルを訓練する．評価語主辞の識別器は，二値の分類器として同様のオンライン学習アルゴリズムと，同様の素性セットを用いて学習を行なっている．

2.2　素性表現

2.2.1　ラベル表現

本稿のタスクでは図 1 が示すように，文中の各文節に対象語，属性語，評価語，その他というようにそれぞれラベルを付与する．系列ラベリングでは BIO タグを用いてラベル系列を表すことが一般的だが，ここではラベルの境界で文を区分することでラベル系列 $\mathbf{y}$ を表す．文には複数の三つ組が含まれるため，一つのラベル系列 $\mathbf{y}$ が一つの三つ組を表し，複数の系列を出

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

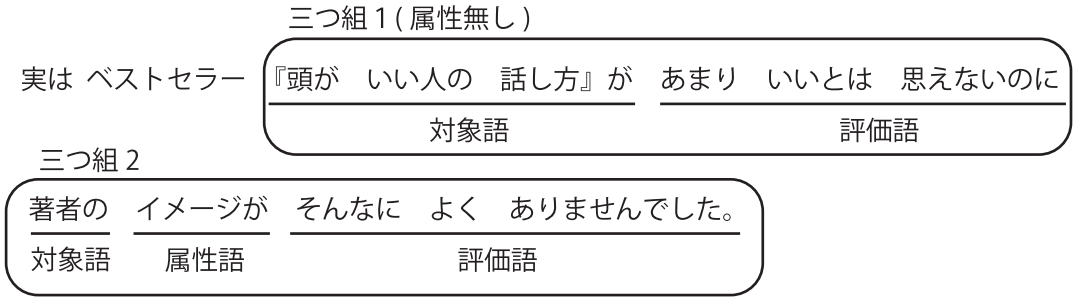

図 1: 複数の三つ組をもつ文の例

力することで対応する. これは, 三つ組の各要素が文内に複数存在するときに, 複数の対象語, 属性語, 評価語が混じって一つの三つ組として出力されることがないようにするためである. 付与するラベルによって文を対象語とその前後, 属性語とその前後, 評価語とその前後, 及び未分類のいずれかに属す文節の集合として表す (図 2 を参照). この区分した文節の集合をエリア a_i と呼び, それぞれ付けたラベルとの位置関係によりラベル前エリア, ラベル付きエリア, ラベル後エリア, 未分類エリアと呼ぶこととする. 分類および学習時にはこの区分の組み合わせごとに素性ベクトル $\psi_{area}(a_i, a_j)$ を生成する.

例を挙げると, $a_{\text{対象語}-\text{前}}$ は文の先頭から対象語の一つ前までの範囲の文節を示し, $a_{\text{対象語}-\text{未分類}}$ は文に対象語ラベルが付与されていない場合に文の全ての文節を示す. このようにラベル系列を表現することで, 必要な組み合わせを考慮した列挙を行いやすくすることができる.

2.2.2 文節の組を表す素性ベクトル

まず, 文節 b_p と文節 b_q の組み合わせを表すラベルに依存しない素性ベクトル $v_{p,q}$ を, b_p を含む 1-3gram と b_q を含む 1-3gram の組み合わせを素性 $v^i_{p,q}$ とするベクトルとする. $dist(p, q)$ は, 文節 b_p から b_q に係り受け関係がある場合の, 係り受け木上でのパスの長さを表す. 係り受け木上で兄弟関係にあるような, パスをたどる際に折り返す必要がある場合には係り受けなしとする. この文節間の 1-3gram の組み合わせ $v^i_{p,q}$ に付ける素性値を, この $dist(p, q)$ を用いて以下の式により与える:

$$v^i_{p,q} = \lambda^{dist(p,q)} + \text{lower bound}. \tag{1}$$

ただし, λ は係り受け距離による重みの減衰率で, 本稿では 0.5 とする. lower bound は素性値の下限で, 同様に 0.5 とする. 文節間に係り受け関係が無い場合は, 素性値は lower bound のみとなる. これは本研究が Web 上のテキストを対象としているため係り受け解析に失敗しやすく, 解析結果に必要以上に依存することを避けるためである. この係り受け関係は文節間に定義されるので, 同じ文節の組み合わせが含む素性はすべて同じ素性値を持つ.

2.2.3 文とラベル系列に対する素性ベクトル

文節の組に対応する素性ベクトル $v_{p,q}$ を用いて, エリア a_i, a_j に対する素性ベクトルはエリアに含まれる文節の組について $v_{p,q}$ 素性ベクトルを足し合わせることで与えられる:

$$\psi_{area}(a_i, a_j) = \sum_{\{(b_p,b_q)|b_p,\in a_i,b_q\in a_j\}} v_{p,q}.$$

これを用いて, ラベル系列と文に対する素性ベクトルを以下の式で示す:

$$\psi(\mathbf{y}, \mathbf{x}) = \bigoplus_{\{(a_i,a_j)\in PairSet\}} \psi_{area}(a_i, a_j). \tag{2}$$

ただし $\bigoplus$ は添字についてベクトルを連結する演算子とし, $PairSet$ は予め決定したエリアの組み合わせの集合を表す. 組み合わせることで有効と考えられるエリアの組み合わせは限定されるため, $PairSet$ はラベル付きエリア同士の組と, あるラベルについてのラベル付きとその前後を表すエリア, 各エリア自身の組み合わせとした. ここで, 文節に対応する素性ベクトル $v_{p,q}$ はラベル系列に依らず決定でき, 高々文節数の二乗の組み合わせしか存在しない. このため予め生成しておくことが可能で, 文全体とラベル系列に対応する素性ベクトルを生成する際には, 必要な文節の組み合わせを列挙するだけでよい.

2.3 探索

ラベル系列に対応する素性ベクトルを与える $\psi(\mathbf{y}, \mathbf{x})$ が得られたので, 次にこの素性ベクトルを評価した際のスコアが最大となる n-best 出力 $(\mathbf{y}_0, \ldots, \mathbf{y}_{n-1})$ を

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

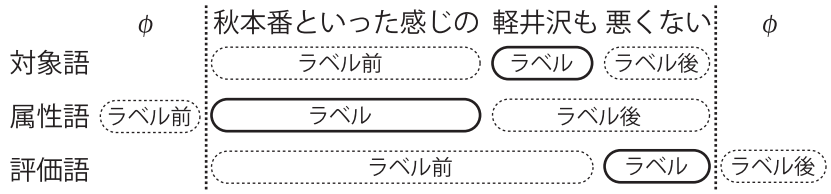

図 2: 文節を区分によるラベル系列の表現

探索するアルゴリズムを考える. 一般的に, 系列ラベリングでは Viterbi アルゴリズムが用いられるが, ラベル間に依存関係が存在する場合には, このアルゴリズムを用いても必ずしも良い結果は得られない. また, この素性表現では各ラベルが文中に連続しない形で二箇所以上存在することを許していないので, その点を考慮した探索アルゴリズムを用いる必要がある.

本稿では, Viterbi ベースと Beam search ベースの２つの探索アルゴリズムを用意し, 実験を行なっている. Viterbi ベースの探索では, ラベル系列は開始時点ではすべて三つ組に含まれないその他として初期化され, 文節ごとにラベル系列を文末から文頭に向けて順に一つずつ探索し, 探索地点の各ラベルごとにスコアが最大となる系列を文末まで決定していく. ただし, 今回の設定ではラベル間に遠距離の依存関係が存在するため Viterbi アルゴリズムにより最適解が選ばれる保証はない. Beam search ベース の探索でも初期値をすべてその他のラベル系列として探索を行う. Beam search では Beam 幅を最大として複数のラベル系列と探索の進行度合いを保持しておき, 保持している系列に対して探索を一つ進めた結果を再び保持する系列に加えて Beam 幅内で保持するものを決める. それぞれ, 非連続な同一ラベルを許さない制約下での探索が行えるように, 探索地点と同一で連続しないラベルをその他に書き戻す処理を加えている.

2.4 学習手法

本研究で学習に用いる, Passive Aggressive Algorithm(PA) について説明する. PA は Crammer ら [2] により提案されたパーセプトロンに似た, オンライン最大マージンアルゴリズムの一種である. このアルゴリズムでは各反復において正しい三つ組 $\mathbf{y}$ をのぞくスコアが最大となる出力 $\bar{\mathbf{y}}$ を探索し, 正解と出力の対について損失関数 $\mathrm{loss}(\mathbf{y}, \bar{\mathbf{y}})$ 以上のマージンを確保しつつ, モデル $\mathbf{w}$ のノルムを最小化することで学習を行う.

文 $\mathbf{x}$ のラベル系列探索では, モデルと素性ベクトルの積が最大となる n-best のラベル系列を分類結果として選ぶ. 学習時には n の値は実際の三つ組の個数と事前に指定したパラメータのいずれか大きい方が与えられる. 本稿では文に対する正しい三つ組が複数存在するため, n-best の系列に全ての三つ組を含むことができるよう学習する必要がある. このため, 単一の三つ組に出力が近づくよう学習するのではなく, 複数の正解となる三つ組と n-best 出力の間でアライメントをとり, それぞれの三つ組に対してアライメントされた出力が近づくように学習を行う. アライメントは全ての正しい三つ組が, 最も類似する出力に対してそれぞれ一対一となるよう行う. 出力が正しい三つ組の個数よりも多く一対一となるアライメントが行えない場合には, できる限りアライメントを行った後に対応の無い出力を最も近い三つ組と対応付ける. パラメータの更新は PA(I) の通常の更新式を用いて, アライメントされた組ごとに順に行う.

2.5 損失関数

学習で用いる損失関数として, 間違えたラベルによって非対称な重みを付ける損失関数を用いる. 対象語, 属性語, 評価語のそれぞれで出現数は異なるため, 間違った箇所ごとに $\{j_{\text{対象語}}, j_{\text{属性語}}, j_{\text{評価語}}\}$ の３つの重みパラメータを以下のように設定する:

$$loss(\mathbf{y}, \bar{\mathbf{y}}) = \frac{1}{\text{文節数}} \left\{ err_o + \sum_{t=\{\text{ 対象語, 属性語, 評価語 }\}} j_t \cdot err_t \right\}.$$

ここで, err_t は t について正しい三つ組 $\mathbf{y}$ と出力されたラベル系列 $\bar{\mathbf{y}}$ の間でラベルを間違えた数を表す.

3 実験設定

本研究の提案手法を, CRF をベースラインとして比較する実験を行う. ベースラインでは一般的な系列ラベリングの問題として, 各文節ごとに評価表現の要素にラベルをつける問題として解いた. 実験用のコーパスは Web 上から収集した blog に対して, 評価表現の各ラベルがアノテートされたものである. 各ラベルは連続した単語列に対してそれぞれタグ付けされており, 必ずしも一文節だけを抜き出せば良いわけではない. 属性語は特に難しい場合が多く, 大畑のように (対象

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

語) どんくさく駆けずり回ってトライ奪うって姿、(属性語) かっこいいよね。(評価語), というように極端に長く境界を見つけることが困難な例もある程度存在している. 全データを訓練事例：テスト事例を９：１に分割し, それぞれから三つ組の要素のいずれかを含む文のみを抜き出し実験に使用した. 結果, 訓練事例が1900 文, テスト事例が 211 文がそれぞれ実験に用いるデータとして得られた. 各文に対して, MeCab[4] を用いた形態素解析と, CaboCha[3] を用いた係り受け解析を行っている. ベースラインと提案手法では共に, 素性として単語と品詞のみを用いている. 実験に用いた提案手法のパラメータは 正則化パラメータ C=1, 損失関数のパラメータを属性語のみ $j_{\text{属性語}} = 5$, それ以外のラベルでは 1 に設定した. イテレーション回数は 6 に固定し, 学習時の n-best 出力の規定値には 2, Beam search を行う際の Beam 幅は 15 を用いている. 一つの文に対して複数の正しい三つ組が存在するため, 評価は n-best 出力と三つ組の間で一致ラベル数によりアライメントをとり, 各三つ組と出力との組ごとにラベルの一致/不一致を計算している.

4 実験結果及び考察

実験の結果を表 1 に示す. ２つの探索手法による違いを比較すると, Viterbi ベースの探索を行った場合には CRF に比べ目立った改善は見られない一方で, Beam search をベースとした手法では全ての指標においてベースラインの CRF を上回る結果となった. これは, 組み合わせ素性と Beam search を組み合わせることによって, ラベル間の依存関係を捉えることが出来たためであると考えられる. しかし, 他の対象語と評価語が向上したのにも関わらず依然として属性語の Recall は低く留まっている. これは文の６割に含まれる対象語や全文に含まれる評価語に比べ, 属性語は全文の２割程度と出現数が少ないことに加えて, 非常に属性語が長い例や, 境界が曖昧であるような難しい例が多いためと考えられる.

5 結論及び今後の課題

構造学習の枠組みを用いて文とラベル系列の対応を学習し, 対象語, 属性語, 評価語の結びつきの強さを含めて学習を行う枠組みを提案した. 提案手法を Web 上のテキストに対して適用し, ベースラインと比較し有効性を示した. 今回は素性として ngram 素性のみを用いているが, 評価表現の辞書など外部知識の導入および文法的な情報の活用を行っていきたいと考えている.

表 1: 実験結果

		提案手法		CRF
		Viterbi	Beam search	
対象語	Precision	50.7	**76.5**	64.4
	Recall	59.1	**71.0**	42.5
	F-value	54.6	**73.7**	51.2
属性語	Precision	53.7	**92.0**	56.5
	Recall	40.7	**42.6**	25.5
	F-value	46.3	**58.2**	35.1
評価語	Precision	55.5	**75.3**	61.7
	Recall	64.8	**84.6**	62.3
	F-value	59.8	**79.7**	62.0

参考文献

[1] Yasemin Altun, Ioannis Tsochantaridis, Thomas Hofmann, Hidden Markov Support Vector Machines. In *proceedings of ICML 2003*, pp.3–10, 2003.

[2] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, Yoram Singer, Online Passive-Aggressive Algorithms. *Journal of Machine Learning research*, Vol.7, Mar, pp.551–585, 2006.

[3] Taku Kudo and Yuji Matsumoto. 2000. Japanese dependency structure analysis based on support vector machines. In *Proceedings of EMNLP/VLC-2000*, volume 13, pp.18–25.

[4] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of EMNLP 2004*, pp.230–237.

[5] Ioannis Tsochantaridis, Tohomas Hofmann, Thorsten Joachims, Yasemin Altun, Support Vector Learning for Interdependent and Structured Output Spaces. In *Proceedings of ICML 2004* , pp.823–830.

[6] 小林 のぞみ, 乾 健太郎, 松本 裕治, 立石 健二, 福島 俊一, テキストマイニングによる評価現象の収集, 情報処理学会研究報告. 自然言語処理研究会報告, pp.77–84, 2003.

[7] 五十嵐力, 藤本浩司, 但馬康宏, 小谷善行, 枝分かれ同時確率モデルを用いた 対象-属性-属性値関係の抽出. 情報処理学会研究報告, pp.21–26, 2009

[8] 鈴木 泰裕, 高村大也, 奥村 学, Weblog を対象とした評価表現抽出. 人工知能学会, セマンティックウェブとオントロジー研究会, SIG-SWO-A401-02, 2004.

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved