部分的アノテーションコーパスから学習可能な固有表現認識器

笹田 鉄郎 †, 森 信介 †, 河原 達也 †, 山肩 洋子 ††
† 京都大学学術情報メディアセンター †† 京都大学大学院情報学研究科

1 はじめに

自然言語処理の重要な応用の一つとして、用語抽出がある。自然言語処理の研究分野においては、新聞記事を対象とした用語抽出 (人名、地名、組織名等) の研究が広く行われている [1] [2] 。このような用語は固有表現 (Named Entities; NEs) と呼ばれ、また、自動的にこのような用語を認識するタスクは固有表現認識 (Named Entitie Recognition; NER) と呼ばれる。固有表現認識タスクは系列ラベリング問題として解くことが一般的であり、サポートベクターマシン (Support Vector Machine; SVM)、条件付き確率場 (Conditional Random Fields; CRFs) などを用いた手法が提案されている [3] [4] 。

近年、自然言語処理技術はより一層その適用対象を広げており、多言語の様々な文書を対象とした研究が行われている。特に、固有表現認識や用語抽出を行なうにあたっては、一般分野ではなくある特定の応用・目的に応じて独自の定義を行なう研究も見られ、テキストマイニングの分野では医療用語や企業の製品名などに適用されている [5] 。本論文において対象とするドメインはレシピであり、文脈ごとに食材を正しく認識することが重要となる。例えば、「ステーキハウスのハンバーガー」における「ステーキ」を食材として認識することは適切ではなく、「ハンバーガー」のみを食材として認識する必要がある。このような特定の応用・目的においては、利用可能な少量のアノテーションコーパスを用いて迅速かつ低コストで固有表現認識システムを開発することが求められている。

このような背景の下、我々は部分的アノテーションコーパスから学習可能な固有表現認識器を提案する。固有表現認識に用いる部分的アノテーションコーパスとは、一部の単語にのみ固有表現タグが付与されており、他の単語には付与されていないコーパスを指す。少量のフルアノテーションコーパスでは出現する固有表現の量が少ないと考えられるため、我々の提案する手法では (1) 単語ごとに固有表現タグ確率を独立に推定する、(2) タグ確率を利用して固有表現系列の探索を行なう、の 2 段階推定を行なう。評価実験では、フルアノテーションコーパスと部分的アノテーションコーパスが混在する複数の状況において既存手法との比較実験を行ない、我々の提案する手法の優位性を示した。

2 関連研究

本論文で取り扱う固有表現認識は系列ラベリング問題として解かれることが一般的で、多くの手法が提案されており [6] [7] 、中でも CRF による系列ラベリングは最も精度の高い手法の一つである [8] 。これらの手法においては、BIO 方式を用いて各単語にタグを付与したコーパスが用いられる。BIO の B (Begin) は最初の単語、I (Intermediate) は同一種の固有表現の継続、O (Other) はいずれの固有表現でもないことを意味する。J 種類の固有表現タグ T_1, T_2, ..., T_J がある場合、単語列 w_1, w_2, ..., w_n に対して固有表現 T_j のタグを付与する際には w_1/T_j-B, w_2/T_j-I, $\cdots$, w_n/T_j-I のようなアノテーションが行われ、いずれの固有表現にも属さない単語に対しては w/O のようなアノテーションが行われる。J 種類の固有表現タグに対応する BIO タグの種類数は $2J+1$ 種類であり、各単語にはいずれか 1 種類の BIO タグが付与される。また、BIO タグ系列には接続制約があり、例えば $(w_i/\mathrm{O}, w_{i+1}/\mathrm{T}_j\text{-I})$, $(w_i/\mathrm{T}_j\text{-B}, w_{i+1}/\mathrm{T}_k\text{-I})$, $(w_i/\mathrm{T}_j\text{-I}, w_{i+1}/\mathrm{T}_k\text{-I})$ $(j \neq k)$, のような接続は定義されない。CRF に基づく固有表現認識器は、自動的にこのような接続制約を学習し入力単語列に対する系列ラベリングを行なうことができるが、SVM [3] やロジスティック回帰 [9] (Logistic Regression; LR) に基づく点予測の識別器を用いたいくつかの研究では、動的計画法 (DP) による経路探索を併用することで接続制約を適用しており、本論文の評価実験においても利用する。

本論文における固有表現認識の評価実験では、調理レシピを対象ドメインとしている。固有表現認識に関する初期の研究では、新聞記事を対象とした実験が多く行われていたが、現在ではさまざまなテキストが対象となっており、医療分野のテキストなどを対象とした研究 [5] が行われている。レシピ固有表現認識は、一般分野や医療分野の固有表現認識に比べて利用できる言語資源が少ないため、そのような状況における固有表現認識の良いテストになる。なお、本論文で我々が提案する手法は、レシピだけではなく他のドメインにも適用可能である。

表 1: レシピ固有表現の一覧

クラス	意味	クラス	意味
F	食材	Ac	調理者の動作
T	道具	Af	食材の変化
D	継続時間	Sf	食材の様態
Q	分量	St	道具の様態

3 レシピ固有表現

本論文における評価実験では、レシピテキストに対してレシピ固有表現 (表 1 参照) [10] を付与したコーパスを用いる。レシピ固有表現は、単語あるいは単語列に対して付与され、入れ子を許さないものとする。また、単語内の一部文字列に対して固有表現を付与することはなく、各単語は最大でも 1 つの固有表現タグを持つ。レシピ固有表現はレシピテキスト中に表現する調理者の行動、食材、食材の様態といったものを認識することを目的として定義される。このようなレシピ固有表現は、レシピ検索 [11] やその理解 [12] 、あるいは調理動画と言

表 2: ロジスティック回帰の素性一覧

種類	素性テンプレート
文字 n-gram	$x^{-1}, x^{+1}, x^{-2}x^{-1}, x^{-1}x^{+1}, x^{+1}x^{+2}$
	$x^{-2}x^{-1}x^{+1}, x^{-1}x^{+1}x^{+2}$
文字種 n-gram	$c(x^{-1}), c(x^{+1}), c(x^{-2})c(x^{-1}),$
	$c(x^{-2})c(x^{-1}), c(x^{-1})c(x^{+1}), c(x^{+1})c(x^{+2}),$
	$c(x^{-3})c(x^{-2})c(x^{-1}), c(x^{-2})c(x^{-1})c(x^{+1}),$
	$c(x^{-1})c(x^{+1})c(x^{+2}), c(x^{+1})c(x^{+2})c(x^{+3})$

語表現の照合によるシンボルグラウンディング [13] といった応用に有用である。

4 2 段階固有表現認識

本論文で我々が提案する固有表現認識手法は、以下の 2 段階に分けられる。

1. 各単語に対して独立に BIO タグとその信頼度の組を推定し、列挙する。
2. BIO タグとその信頼度を用いて、最適な BIO タグ列の探索を行なう。

本節では以上の 2 つの手続きについて詳細を述べる。

4.1 点予測によるタグ信頼度の推定と列挙

我々の提案する固有表現認識システムでは、入力として単語列が与えられると、1 段階目のモジュールが各単語に対して BIO タグとその信頼度を推定し、2 段階目のモジュールに渡す。1 段階目のモジュールとしては点予測に基づく識別器を用いる。点予測の素性としては入力単語列に含まれる情報のみを用い、推定結果を素性として用いないものとする。

上述の設計から明らかなように、1 段階目のモジュールはアノテーションされた単語とその文脈情報のみを学習に用いるため、部分的アノテーションコーパスからモデルを学習することが可能である。以下に示す部分的アノテーションの例:

ex.) 黒/F-B 胡椒/F-I を ふりかけ る

では、「黒」と「胡椒」の 2 単語にのみ BIO タグが付与されており、それぞれの単語に対応する文脈情報とタグは以下のように表される。

left context	word	right context	tag
$\langle BOS \rangle$	黒	胡椒を	F-B
$\langle BOS \rangle$ 黒	胡椒	をふり	F-I
黒胡椒	を	ふりか	–
胡椒を	ふりかけ	る $\langle EOS \rangle$	–
りかけ	る	$\langle EOS \rangle$	–

上述のようなデータを訓練データとして、SVM あるいは LR のような点予測に基づく識別器を構築する。

一般的な固有表現認識システムとは異なり、このモジュールは各単語に対して全ての可能なタグとその信頼度を列挙する。信頼度としては、SVM における分離平面からのマージンやロジスティック回帰の確率を用いることが可能である。本論文においては入力単語列中の各単語 w_i に対し、各 BIO タグ t_j ごとのロジスティック回帰確率 $s_{i,j}$ を以下の式によって推定し、信頼度として用いる。

$$s_{i,j} = P_{LR}(t_j | \boldsymbol{x}^-, w_i, \boldsymbol{x}^+).$$

ここで $\boldsymbol{x}^-, \boldsymbol{x}^+$ は文脈素性であり、ロジスティック回帰で用いている素性の一覧を表 2 に示す。$c(\cdot)$ は文字 x に対応する文字種 (漢字、平仮名、片仮名、数字等) を得る関数である。ロジスティック回帰により、各単語に対するタグとその信頼度の組 $(\langle t_1, s_{i,1} \rangle, \langle t_2, s_{i,2} \rangle, \cdots, \langle t_{2J+1}, s_{i,2J+1} \rangle)$ が得られる。ここで、$2J+1$ は J 種類の固有表現タグに対応する BIO タグの総数である。

$P_{LR}(t\|w)$	黒	胡椒	を	ふりかけ	る
			w		
F-B	**0.40**	0.40	0.00	0.15	0.00
F-I	0.00	**0.60**	0.00	0.15	0.00
Ac-B	0.00	0.00	0.00	**0.70**	0.00
t T-B	0.50	0.00	0.01	0.00	0.01
⋮	⋮	⋮	⋮	⋮	⋮
O	0.10	0.00	**0.99**	0.00	**0.99**

図 1: BIO 制約と動的計画法による最適経路探索 (太字)

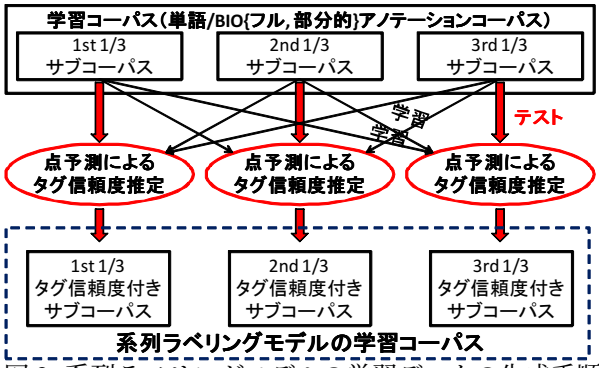

図 2: 系列ラベリングモデルの学習データの生成手順

4.2 最適系列の探索

2 段階目のモジュールでは、与えられた単語列と 1 段階目で推定されたタグ・信頼度の組から最適な系列を探索する。本節では、2 段階目のモジュールとして動的計画法と系列ラベリングを用いる場合について述べる。

4.2.1 動的計画法による探索

動的計画法によって最適系列を選択する場合、以下の式で最尤のタグ列を得られる。

$$\hat{\boldsymbol{t}}_1^m = \underset{t_1, t_2, \ldots, t_m}{\mathbf{argmax}} \prod_{j=1}^{m} s_{i,j}.$$

図 1 に最適経路選択の具体例を示す。動的計画法による探索の際には、2 節で述べた BIO タグの制約を適用しながら経路探索を行なう。例えば、図 1 において、“黒/T-B 胡椒/F-I” は BIO タグ系列として不適切な系列であり、経路探索の候補とならない。本手法は、後述する評価実験においてベースライン手法の一つとして用いる。

4.2.2 系列ラベリング

我々の提案する手法では、2 段階目のモジュールとして系列ラベリングを用い、最適系列の探索を行なう。このモジュールの入力は、1 段階目のモジュールでタグと信頼度の組が付与された単語列である。本手法ではフルアノテーションコーパスのみを学習コーパスとして用いる場合と、それに加えて部分的アノテーションコーパス

表 3: CRF の素性一覧

種類	素性テンプレート
単語 n-gram	$w^{-1}, w^{+1}, w^{-2}w^{-1}, w^{-1}w^{+1}, w^{+1}w^{+2}$
文字種 n-gram	$c(w^{-1}), c(w^{+1}), c(w^{-2})c(w^{-1}),$ $c(w^{-1})c(w^{+1}), c(w^{+1})c(w^{+2}),$ $c(w^{-2})c(w^{-1})c(w^{+1}),$ $c(w^{-1})c(w^{+1})c(w^{+2})$
タグと信頼度の組 (LR+CRF)	$\langle t_1, s_{i,1}\rangle, \langle t_2, s_{i,2}\rangle, \cdots, \langle t_{2J+1}, s_{i,2J+1}\rangle$

表 4: コーパスの詳細

用途	文数	レシピ固有表現数	単語数	文字数
学習	2,946	17,243	54,470	82,393
テスト	371	1,996	6,072	9,167
(合計)	3,317	19,239	60,542	91,560

を用いる場合に対応するため、1 段階目のモジュールを構築する段階で学習コーパスを分割し、以下の手続きによって全ての学習コーパスに対するタグと信頼度の組を推定する。

(i) 学習コーパスを N 個のサブコーパスに分割する。
(ii) N 個のサブコーパスのうち、i 番目のサブコーパスを除いた $N-1$ 個のサブコーパスを学習コーパスとして点予測による識別器を構築する。
(iii) $N-1$ 個のサブコーパスから学習した点予測による識別器を用いて、i 番目のサブコーパス中の単語に BIO タグと推定された信頼度の組を付与する。

図 2 に上述の手続きを図にしたもの ($N = 3$ の場合) を示す。以上の手続きにより、元々BIO タグ付与をされていない部分的アノテーションコーパス中の単語を含む、全てのサブコーパス中の単語に対してタグと信頼度の組 ($\langle t_1, s_{i,1}\rangle, \langle t_2, s_{i,2}\rangle, \cdots, \langle t_{2J+1}, s_{i,2J+1}\rangle$) を推定し、系列ラベリングの学習データとして用いることができる。系列ラベリングの実装として、本手法では CRF [8] を用いているが、SVM など [14] 他の手法を用いることも可能である。表 3 に CRF で用いている素性の一覧を示す。

5 評価

本節では、提案手法ならびに既存手法の評価実験を行なう。フルアノテーションコーパスと部分的アノテーションコーパスを固有表現認識の学習コーパスとして用いることを想定し、複数の状況をシミュレーションして各手法の比較を行なった。

5.1 実験設定

本実験ではドメインを調理レシピとし、3 節で述べたレシピ固有表現を認識する実験を行なう。表 4 にレシピ固有表現が付与されたコーパスの詳細を示す。

本実験では言語資源が少ない状況における固有表現認識について評価するために、全体のフルアノテーションコーパス (1/1 FULL) を 2 つに分割し、片方のコーパスをフルアノテーションコーパス (1/2 FULL)、残りのコーパスを擬似的な部分的アノテーションコーパス (1/2 PART) とした。上述の 1/2 PART は、1/1 FULL から 1/2 FULL を除いた残りのコーパスにおけるアノテーションのうち、同じ単語に対するアノテーションを最大 3 回までに制限することで擬似的な部分的アノテーションコーパスとしている。本実験では、以下に示す 3 種類のコーパスセットを用いて評価実験を行う。

- 1/2 FULL: フルアノテーションコーパスの 1/2 を学習コーパスとする。
- 1/2 FULL + 1/2 PART: フルアノテーションコーパスの 1/2 と、残りの 1/2 を擬似的な部分的アノテーションコーパスとして用いる。
- 1/1 FULL: すべてのフルアノテーションコーパスを学習コーパスとする。

表 5 に上述の各実験設定における文数、レシピ固有表現数、BIO タグ数を示す。

本実験では、上述の 3 種類のコーパスセットを学習コーパスとし、以下の 4 種類の手法による固有表現認識器を構築して比較する。

- **CRF**: 部分的アノテーションコーパスからモデル学習を行なうことが可能な CRF の実装を用いてタグの推定を行なう。
- **LR**: ロジスティック回帰による点予測の識別器を用いてタグの推定を行なう。
- **LR + DP**: ロジスティック回帰によるタグと信頼度の組の推定と動的計画法の 2 段階推定 (4.2.1 項参照) を行なう。
- **LR + CRF**: ロジスティック回帰によるタグと信頼度の組の推定結果を CRF の素性として利用する 2 段階推定 (4.2.2 項参照) を行なう。

提案手法の **LR + CRF** では、学習コーパスを 3 分割として実験を行なう (4.2.2 項、図 2 参照) 。

本実験では、部分的アノテーションコーパスから学習可能な CRF の実装として、partial-crfsuite [1] [15] を用いる。また、点予測に基づくロジスティック回帰として、KyTea [2] [16] を用いる。手法 **CRF** と **LR** で用いた素性は表 3 と表 2 に示した通りである。手法 **LR + CRF** の 2 段階目のモジュールである CRF を利用する際には、追加の素性としてタグと信頼度の組を用いる。

表 5: 実験用コーパスセットの詳細

コーパスセット	文数	レシピ固有表現数	BIO タグ数
1/2 FULL	1,473	8,543	27,119
1/2 FULL + 1/2 PART	2,946	10,810	31,770
1/1 FULL	2,946	17,243	54,470

表 6: 実験結果 (1/2 FULL)

手法	BIO 精度	適合率	再現率	F 値
CRF	0.8949	0.8491	0.8372	0.8438
LR	0.8930	0.8441	0.8407	0.8424
LR+DP	0.8951	0.8397	0.8477	0.8437
LR+CRF	**0.8989**	**0.8591**	0.8402	**0.8495**

5.2 実験結果

表 6、7、8 に 5.1 節で述べた各コーパスセット、各手法ごとの実験結果を示す。また、図 3 に実験結果をグ

[1] https://github.com/ExpResults/partial-crfsuite
[2] http://www.phontron.com/kytea/

表 7: 実験結果 (1/2 FULL + 1/2 PART)

手法	BIO 精度	適合率	再現率	F 値
CRF	0.8990	0.8612	0.8452	0.8531
LR	0.8995	0.8559	0.8452	0.8505
LR+DP	0.9012	0.8539	0.8552	0.8546
LR+CRF	**0.9112**	**0.8773**	**0.8632**	**0.8702**

表 8: 実験結果 (1/1 FULL)

手法	BIO 精度	適合率	再現率	F 値
CRF	0.9065	0.8759	0.8627	0.8693
LR	0.9056	0.8713	0.8582	0.8647
LR+DP	0.9069	0.8696	0.8652	0.8674
LR+CRF	**0.9157**	**0.8853**	**0.8742**	**0.8798**

ラフにしたものを示す。図 3 より、提案手法 **LR + CRF** による固有表現認識の精度が既存手法 **CRF**、**LR**、**LR + DP** を上回っていることがわかる。以下に実験結果についての詳細を述べる。

1/1 FULL(表 8 参照) のように多くのフルアノテーションコーパスを利用可能な状況では **CRF** が **LR** や **LR + DP** に比較して良い精度を示していることがわかる。しかしながら、1/2 FULL (表 6 参照) や、1/2 FULL + 1/2 PART (表 7 参照) のように学習コーパスとして利用可能なフルアノテーションコーパスの分量が少なくなると、**LR** や **LR + DP** の精度が **CRF** の精度を上回っていることがわかる。

実際のアノテーション作業においては、1 箇所のアノテーションによって向上する精度を最大化することが望ましい。単純なアノテーション戦略として、新しいレシピ固有表現のアノテーションを数回にとどめ、カバレージの向上を優先することが考えられ、1/2 FULL + 1/2 PART のコーパスセットはこの状況をシミュレートしている。我々の提案手法である **LR + CRF** の精度は 1/2 FULL + 1/2 PART のコーパスセットを用いた場合に特に高く、1/1 FULL における **CRF** の精度と同程度となっている (図 3 参照)。1/2 FULL のコーパスを 1/1 FULL のコーパスにするには (8,700 = 17,243 - 8,543) 個のレシピ固有表現をアノテーションする必要があるが、1/2 PART のコーパスを追加する際にはその 4 分の 1 程度の (2,267 = 10,810 - 8,543) 個のレシピ固有表現のアノテーションを行なうだけでよく (表 5 参照) 、さらに提案手法 **LR + CRF** を適用することで 1/1 FULL のコーパスにおける手法 **CRF** と同程度の精度を達成可能であることがわかる。以上より、少量のフルアノテーションコーパスが利用可能な状況においては、部分的アノテーションコーパスを追加し、提案手法 **LR + CRF** を適用する戦略が有効であると結論できる。

6 結論

本論文では、部分的アノテーションコーパスから学習可能な固有表現認識について述べた。提案手法は、柔軟な言語資源を許容する点予測の部分と、最尤のタグ系列を探索する系列ラベリングの部分からなる。実験の結果、提案手法は、特に言語資源が少ない状況下において

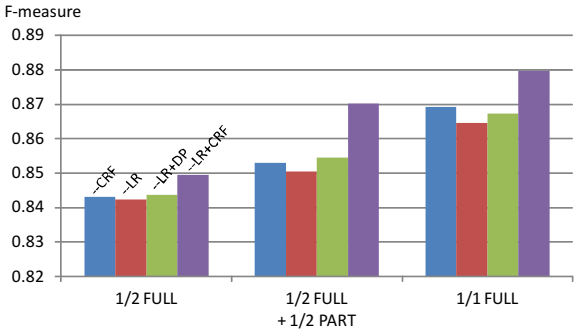

図 3: 実験結果

高い精度を達成可能であることが示された。

参考文献

[1] Chinchor, N. A.: Overview of MUC-7/MET-2, *Proc. of the MUC-7* (1998).

[2] Ratinov, L. and Roth, D.: Design Challenges and Misconceptions in Named Entity Recognition, *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, Boulder, Colorado, Association for Computational Linguistics, pp. 147–155 (2009).

[3] Finkel, J. R., Grenager, T. and Manning, C.: Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling, *Proc. of the ACL05*, pp. 363–370 (2005).

[4] McCallum, A. and Li, W.: Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons, *CoNLL03* (2003).

[5] Ben Abacha, A. and Zweigenbaum, P.: Medical Entity Recognition: A Comparaison of Semantic and Statistical Methods, *Proceedings of BioNLP 2011 Workshop*, Portland, Oregon, USA, Association for Computational Linguistics, pp. 56–64 (2011).

[6] Borthwick, A.: *A Maximum Entropy Approach to Named Entity Recognition*, PhD Thesis, New York University (1999).

[7] Sang, E. F. T. K. and Meulder, F. D.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition, *Proc. of the CoNLL2003*, pp. 142–147 (2003).

[8] Lafferty, J., McCallum, A. and Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proc. of the ICML01* (2001).

[9] Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. and Lin, C.-J.: LIBLINEAR: A Library for Large Linear Classification, *Journal of Machine Learning Research*, Vol. 9, pp. 1871–1874 (2008).

[10] Mori, S., Maeta, H., Yamakata, Y. and Sasada, T.: Flow Graph Corpus from Recipe Texts, *Proc. of the LREC14*, pp. 2370–2377 (2014).

[11] Wang, L., Li, Q., Li, N., Dong, G. and Yang, Y.: Substructure Similarity Measurement in Chinese Recipes, *Proc. of the WWW08*, pp. 978–988 (2008).

[12] Bollini, M., Tellex, S., Thompson, T., Roy, N. and Rus, D.: Interpreting and Executing Recipes with a Cooking Robot, *Proceedings of The 13th International Symposium on Experimental Robotics*, pp. 481–495 (2013).

[13] Rohrbach, M., Qiu, W., Titov, I., Thater, S., Pinkal, M. and Schiele, B.: Translating Video Content to Natural Language Descriptions, *Proc. of the ICCV13* (2013).

[14] Tsochantaridis, I., Joachims, T., Hofmann, T. and Altun, Y.: Large Margin Methods for Structured and Interdependent Output Variables, *Machine Learning*, Vol. 6, pp. 1453–1484 (2005).

[15] Liu, Y., Zhang, Y., Che, W., Liu, T. and Wu, F.: Domain Adaptation for CRF-based Chinese Word Segmentation using Free Annotations, pp. 864–874 (2014).

[16] Neubig, G., Nakata, Y. and Mori, S.: Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis, *Proc. of the ACL11*, pp. 529–533 (2011).