Support Vector Machines を用いた英語の文区切りの同定

西光 雅弘 †* 渡辺 太郎 ‡ 隅田 英一郎 ‡ 河原 達也 †‡

† 京都大学 情報学研究科 知能情報学専攻
‡ATR 音声言語コミュニケーション研究所
e-mail: saikou@ar.media.kyoto-u.ac.jp

1 はじめに

近年、コンピュータ技術の発展にともなって、大規模コーパスが構築され、様々な自然言語処理技術に応用されている。コーパスを有効に利用するには、大量に収集されたテキストを解析し、形態・構文・意味などの様々な情報を付与 (タグ付与) することが重要である。タグ付与は膨大なコストを必要とする作業であるため、自然言語処理技術を用いた、計算機によるテキストの自動解析が不可欠である。

計算機を利用したテキスト解析技術の多くは文を入力の基本単位としている。したがって収集された大量のテキストを文に区切らなければならない。そして、その文区切り精度がテキスト解析の精度に大きく影響を及ぼす。これは、テキスト解析技術が正しい文を入力の前提としているため、非文が入力された場合、正しい解析結果が得られないからである。

英語の文区切りは、文区切りを意味する標識を参照することで、容易に区切ることが可能であると考えられがちであるが、実際にはそれほど容易ではない。例えば、“Sometimes those aims diverge from U.S. goals.” という文の文末単語がU.S. とgoals. のどちらであるかを、文区切りを意味する標識であるピリオドのみで同定することはできない。

これに対して、決定木を用いた手法 [1] や Maximum Entropy(ME) を用いた手法 [2] などが提案されている。本研究では英語の文区切りの同定手法として、機械学習の１つである Support Vector Machines(SVM) を用いた手法を提案する。

2 文区切りの同定手法

本研究では文区切りの同定を text chunking 問題として扱う。すなわち chunk を文と定義し、テキストの中から文区切りを意味する chunk の末尾単語を同定する。本手法は NP Chunking に代表される text chunking 問題を、単純に拡張することで実現可能である。text chunking を行なう際に必要な単語は、単語間に存在する空白を用いてテキストより切り出した。したがってピリオドなどの標識は単語に付着した状態となっている。

表 1: Start/End 法で用いるチャンクタグ

B	1 つ以上の単語からなる chunk の先頭単語
E	1 つ以上の単語からなる chunk の末尾単語
I	1 つ以上の単語からなる chunk の先頭・末尾以外の中間単語
O	単独で 1 つの chunk を構成する単語
S	chunk に含まれない単語

単語の素性は、以下の 10 種類を用いる。これらの素性は全て、単語自体から得られる基本的な素性である。

- 単語
- 単語に記号が含まれているか否か (Y/N)
- 単語に数字が含まれているか否か (Y/N)
- 直後の単語の先頭文字が大文字か否か (Y/N)
- 単語の語頭文字列 (1〜3 文字目までの 3 種類)
- 単語の語尾文字列 (1〜3 文字目までの 3 種類)

ラベリングスキームには、チャンクタグを表１の５種類とする Start/End 法 [3] を用いた。タグ付与の実例を表 2 に示す。

テキストチャンカーには SVM ベースの YamCha[3] を用いる。YamCha における具体的なパラメータは、多項式カーネルの次数を 3、解析方向を Left to Right、多値クラス識別を Pairwise 法とした。

*本稿は、第一著者が ATR 音声言語コミュニケーション研究所にて夏季実習を行った際の研究成果に基づいている。

表 2: タグ付与の例

単語	記号	数字	大文字	語頭文字列			語尾文字列			チャンクタグ
				(1 文字)	(2 文字)	(3 文字)	(1 文字)	(2 文字)	(3 文字)	
It	N	N	N	I	It	Null	t	It	Null	B
is	N	N	N	i	is	Null	s	is	Null	I
aimed	N	N	N	a	ai	aim	d	ed	med	I
at	N	N	Y	a	at	Null	t	at	Null	I
Japan.	Y	N	Y	J	Ja	Jap	.	n.	an.	E
The	N	N	N	T	Th	The	e	he	The	B
federal	N	N	N	f	fe	fed	l	al	ral	I

3 実験と評価

3.1 実験対象コーパス

実験には Linguistic Data Consortium(LDC) から提供されている Wall Street Journal(WSJ) コーパス [4] を用いた。WSJ コーパスは約 200 万文で構成されており、基本的に文単位でタグ付与が行なわれている。本研究では新聞の見出しにあたる HL タグと記事にあたる S タグの付与された文を対象とした。HL タグの付与された文は基本的に文末にピリオドが存在しない。S タグの付与された文は「.」「!」「?」などで終わる文が含まれている。従来研究 [1][2] では新聞の記事にあたる S タグの付与された文のみを対象としているが、本研究では WSJ コーパス以外の様々なコーパスへの適用なども考慮し、新聞の見出しにあたる HL タグの付与された文も対象とした。

実験と評価を行う際に必要となるチャンクタグについては、WSJ コーパスが文単位でタグ付与が行なわれていることを利用し、WSJ コーパスに付与されているタグに基づいて自動的に付与した。

3.2 素性の組み合わせ

YamCha はモデル学習の特徴として比較的自由に素性の組合せが可能である。本論文ではモデル学習の特徴として以下の 4 種類を用いた。

- 直前単語の素性 (F:-1..0)
- 前後 1 単語の素性 (F:-1..1)
- 前後 2 単語の素性 (F:-2..2)
- 前後 3 単語の素性 (F:-3..3)

3.3 学習セット

本研究で用いる学習セットは表 3 の 5 種類である。WSJ コーパスは 3 年分のテキストから構成されており、年度毎にタグ付与の定義が若干異なっている。そのため、年度毎の学習セット (1987, 1988, 1989) を用意した。学習セット (Mix) は年度毎の学習セットを合わせたものであり、学習セット (Large) は学習セット (Mix) にデータを追加して、学習データ量を大きくしたものである。

表 3: 学習セットの仕様

学習セット	総文数	総単語数	HL 率 (%)
1987	8,119	164,192	12.5
1988	7,072	158,825	5.3
1989	7,401	148,951	15.4
Mix	22,592	461,968	11.2
Large	37,964	785,235	10.6

表 4: テストセットの仕様

テストセット	総文数	総単語数	HL 率 (%)
1987_1	7,988	164,652	13.7
1987_2	7,798	160,472	13.5
1988_1	7,200	161,341	4.9
1988_2	7,243	162,497	5.3
1989_1	7,779	150,227	14.5

表 3 にある HL 率は学習セットに含まれる HL タグの付与された文の割合、つまり文末にピリオドが存在しない文の割合である。表 3 に示す 5 種類の学習セットと 3.2 で示した 4 種類の特徴を組み合わせ、計 20 種類のモデルを作成し実験を行なった。

3.4 テストセット

テストセットとして表 4 の 5 種類を用いた。テストセットは open なテストとなるように、学習セットとは異なるデータを WSJ コーパスより選択した。ここでも年度毎にタグ付与の定義が異なっているため、年

表 5: 各モデルごとの再現率・適合率・F 値

	1987		1988		1989		Mix		Large	
F:-1..0	89.03	93.92	86.63	97.56	89.21	94.99	89.40	96.21	89.40	96.21
	91.41		91.77		92.01		92.68		92.71	
F:-1..1	92.56	95.95	88.56	98.47	92.26	97.91	94.64	97.68	95.10	97.68
	94.23		93.26		95.00		96.14		96.37	
F:-2..2	92.30	96.36	88.15	98.32	91.82	97.92	94.46	98.14	95.05	97.96
	94.29		92.96		94.77		96.26		96.48	
F:-3..3	91.95	96.27	87.85	98.13	91.43	97.77	94.26	98.10	94.91	98.10
	94.06		92.71		94.49		96.14		96.48	

再現率	適合率
F 値	

表 6: 従来手法との比較結果

Rule		ME		提案法	
81.21	74.52	87.04	97.93	95.05	97.96
77.72		92.16		96.48	

度毎にテストセットを用意している。

3.5 評価方法

評価尺度として以下の式で表される再現率・適合率・F 値を用いる。表 4 に示した 5 種類のテストセットそれぞれについて評価尺度を求め、その平均を評価量とする。

$$\text{再現率} = \frac{\text{正解と一致した } E \text{ タグの総数}}{\text{正解である } E \text{ タグの総数}} * 100$$

$$\text{適合率} = \frac{\text{正解と一致した } E \text{ タグの総数}}{\text{推定された } E \text{ タグの総数}} * 100$$

$$F \text{ 値} = \frac{2 * \text{再現率} * \text{適合率}}{\text{再現率} + \text{適合率}}$$

3.6 実験結果

実験結果を図 1 と表 5 に示す。また、本研究のベースラインとして、単純な規則による手法 (Rule) と ME による手法 (ME)[2] の結果も表 6 で示す。ベースラインとして用いた規則は、単語の最終文字が次の 4 種類 (「.」「,」「?」「!」) で直後の単語の先頭文字が大文字であれば文区切りとするというものである。

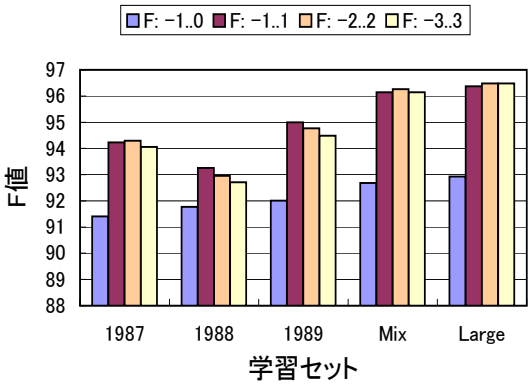

図 1: 学習セットに対する F 値

図 1 より、学習データ量の増加によって F 値が改善することがわかる。各モデルを比較すると、直前単語の素性のみを特徴とするモデル (F:-1..0) は他のモデルに比して総じて F 値が低く、学習データ量の増加による改善も小さいことから、用いた特徴が不十分であるといえる。学習データ量が少ない場合 (1987, 1988, 1989) は、前後 1 単語の素性を特徴とするモデル (F:-1..1) で最も高い F 値が得られた。一方、学習データ量が大きくなると、より多くの特徴を持ったモデル (F:-2..2, F:-3..3) が最も高い F 値を示した。これは、特徴を多く持ったモデルの学習が、(1987, 1988, 1989) のデータ量では不十分であるためと考えられる。すべてのモデルで実験を行なった結果、学習セット (Large) で前後 2・3 単語の特徴 (F:-2..2, F:-3..3) を用いた時、F 値で最大約 96.5%を得た。

表 6 より、提案法は従来手法と同等以上の結果を得ていることが確認できる。2 つの従来手法は本実験で用いた WSJ コーパスに適応させたモデルではないため、単純に比較することは難しいが、ME はコーパスに出現する省略語のリストを素性として利用しているのに対して、提案法は単語自体から得られる基本的な

素性のみを利用しており、MEによる手法よりも汎用性が高いことから、提案法が英語の文区切り手法として有効であると考える。

4 誤りの解析

4.1 文区切りとして推定された誤り

文区切りとして推定された誤りを調べると、省略語とチャンクタグ自動付与の影響によるものの2つに大別された。前者の誤りは、英語文において文区切りを意味するピリオドと省略を意味するピリオドを誤判別したことによると考えられる (図2(a))。後者の誤りは、本実験で用いた正解チャンクタグの誤りによるものである (図2(b))。本実験では、実験と評価に用いるチャンクタグを、WSJコーパスのタグに基づいて自動付与した。WSJコーパスは、基本的に文単位でタグ付与が行なわれているが、注釈を含む文などは必ずしも文単位でタグ付与されているわけではない。その影響で、正しく推定されているにもかかわらず誤りと判定された。

4.2 文区切りとして推定されなかった誤り

文区切りとして推定されなかった誤りを調べると、そのほとんどがHLタグの付与された文であった (図2(c))。つまり、誤りのほとんどが文末にピリオドがないものであった。したがって、文末にピリオドを持つ普通文に関しては、ほぼ全て抽出できていると言える。この原因として、表3のHL率より、文末にピリオドを持たない文が学習セットに約10%しか含まれおらず、見出しにあたる文の学習が不足していることが考えられる。

5 おわりに

本研究ではSVMを用いた英語の文区切りの同定手法を提案した。また、提案法の有効性を示すためにWSJコーパスを用いた評価実験を行い、F値で約96.5%の正解率を得た。提案法は単語自体が持つ基本的な素性のみを用いて学習していることから、従来手法よりも汎用性が高く、様々なコーパスへの応用が期待できる。

今後の課題としては、誤りの主要な原因である省略語と見出しにあたる文に対して有効な素性や、Conditional Random Field(CRF) による手法の検討などが挙げられる。

(a)“D.C.” を文区切り単語と誤って推定した。

The back rooms of the D.C. Circuit Court are still buzzing with an anecdote.

(b)“transactions.” を文区切り単語として正しく推定したが、WSJコーパスより自動付与した正解が誤っているため不正解となった。

The ruling effectively precludes Sumitomo from directly learning how Goldman crafts transactions. (As an insurance company, Nippon Life has no such strictures limiting its investment in Shearson.)

(c)“Report” を文区切り単語として推定できなかった

International Corporate Report FUJITSU LTD., a leading Japanese computer maker, said it will start production mid-month of printers for personal computers in Spain.

図 2: 誤りの実例

参考文献

[1] David D. Palmer and Marti A. Hearst. Adaptive multilingual sentence boundary disambiguation. *Computational Linguistics*, Vol. 23, No. 2, pp. 241–267, 1997.

[2] J. Reynar and A. Ratnaparkhi. A maximum entropy approach to identifying sentence boundaries. In *Proc. of the 5th Conference Natural Language Processing*, pp. 16–19, Washington D.C., 1997.

[3] T. Kudo and Y. Matsumoto. Chunking with support vector machines. In *Proc. of the 2nd Meeting North American Chapter of the Association for Computational Linguistics*, 2001.

[4] http://www.ldc.upenn.edu/Catalog/LDC2000T43.html.