サポートベクターマシンによる英語俳句の抽出

檀 裕也† 和田 武‡ 墨岡 学†
松山大学 経営学部† 愛媛大学 総合情報メディアセンター‡

1．はじめに

情報爆発の時代と呼ばれる現在、Web 上には膨大な情報が溢れている。その中から目的の情報を得るための情報検索として様々な工夫が提案されている。本研究は、Web 文書から英語俳句を抽出することを目的とする。すなわち、英語俳句とそうでない英語表現の特徴について機械学習によって分類する自然言語処理の手法を提案する。その際、訓練データとして著者らによって運用されている英語俳句サーバに投稿され、蓄積された作品を用いる。本研究の成果は、検索エンジンを併用することによって、特定のキーワードを含む英語俳句を抽出することに応用できると期待される。英語表現の特徴を学習させやすいと考えられる英語俳句についてサポートベクターマシン(SVM)による文書分類の手法を適用し、提案手法の有効性と限界を示す。

2．問題の背景

著者らは、俳人・正岡子規が生まれた松山において、世界の俳句愛好家と英語俳句に関する情報を交換するため、俳句サーバ SHIKI を運用して 15 年が経過した。SHIKI では、Web サイト HaikuSphere[1]をはじめ、英語俳句メーリングリスト NOBO List やフォーラム Shiki Haiku Forum と、それらのアーカイブ(Shiki old archives)などのサービスを提供している。メールによって投稿される英語俳句や Web ページに掲載される HTML 形式のファイルなど、これまでに多くの情報を文書として蓄積してきた。これらの文書はインターネットで一般向けに公開しているが、他に有効な利用方法はないか模索していた。[2][3]

そのとき、情報処理技術の応用として、過去ログを機械学習させることによって、コンピュータに英語俳句を理解させることは可能かという問題が提起された。もし、機械的に英語俳句の内容と形式を判定することが可能だとしたら、現在も投稿されている NOBO List におけるメールの内容について、英語俳句のみを投稿時のリアルタイムに抽出することができるようになる。また、検索エンジンを併用することによって、任意の場所に存在する Web ページから英語俳句を自動的に抽出する道も開けることになる。

2.1 英語俳句の特徴

英語俳句は、自由な形式で記述された英語による創作的な表現である。日本語の俳句にある季語や 5・7・5 の音節などの制約はなく、英語の文法規則に厳密である必要はない。原則 3 行の中で、自然の存在や人間の感情を詠む。次の英語表現は英語俳句の一例である：

```
between spring field
and blue sky
birdsong
```

Shiki HaikuSphere (2007) by Shiki Team

この句は、「春の草原 心地よい季節 青空 小鳥のさえずり」というイメージを表現している。

2.2 サポートベクターマシン

文書などのデータを機械的に分類するとき、それぞれのデータを多次元空間におけるベクトルで表し、グループごとのまとまりを超平面で切断できる手法として、サポートベクターマシン(SVM)がある。SVM は Vapnik[4]によって提案された統計的パターン認識法で、汎化能力が高いという特徴がある。

いま、特徴ベクトル $x=(x_1,x_2,\ldots,x_n)$ は n 次元ユークリッド空間 R^n の要素として考える。また、m 個の特徴ベクトル $d_1,d_2,\ldots,d_m \in R^n$ には、それぞれ正解と不正解が対応しているものとする。特徴ベクトルが線形分離可能のとき、正解と不正解の特徴ベクトルを分割する超平面 $w\cdot x+b=0$ を決定するために、与えられた訓練データから最適なベクトル $w=(w_1,w_2,\ldots,w_n)$ と実数 b を求めることができる。

また、問題が非線形の場合であっても、カーネルトリックの手法を適用することによって、2つの集合に分割する超曲面を決定することが可能である。[5]

Extraction of English Haiku using Support Vector Machine
†Yuya Dan, Manabu Sumioka;
Faculty of Business Administration, Matsuyama University
‡Takeshi Wada;
Center for Information Technology, Ehime University

3．実験

英語俳句サーバShikiはメーリングリストNOBO Listに投稿された英語俳句を蓄積している。その文書を訓練データとして機械学習を行った。ここで、英文の形態素解析にはTreeTagger[6]、SVMによる学習と分類にはLIBSVM 2.9[7]を用いた。

3.1 英語俳句の収集

英語俳句用のWebサイトShiki Haikusphereにおけるメーリングリスト NOBO Listに投稿された英語俳句を収集する。メールで投稿された記事は、NOBO Listに登録した利用者にメール配信されるとともに、過去ログとしてWebサーバに蓄積され、一般に公開されている。メールによる投稿には、英語俳句そのものでない挨拶やイベントの案内、SPAMメールなどが入っているが、投稿記事ごとにそれぞれHTMLファイルに保存されている。そこで、Javaプログラムを使って過去ログから英語俳句を含む文書を収集した。

3.2 文書の特徴ベクトルの生成

HTMLファイルには、投稿記事がpre要素として記録されている。そのため、前処理としてpre要素からメールで投稿された本文のみを取り出し、形態素解析ツールTreeTaggerを使って英単語ごとに品詞に分解した。その結果から、文書ごとに各品詞成分の出現頻度を要素とする特徴ベクトルを生成した。この文書の特徴ベクトルを使って、LIBSVM 2.9による学習と分類を行った。なお、SVMのアルゴリズムにはソフトマージンを適用し、線形のカーネル関数を使った。

3.3 精度

SVMにかけられた全文書のうち、英語俳句を正しく分類できた割合を精度(accuracy)という。本研究では、分類の性能について精度を使って評価する。

4．実験結果

過去3年間に投稿された作品を訓練データとして与えた。総数1,276件の内訳を**表1**に示す。

表1．学習に使った訓練データ

	文書の件数		
	英語俳句	その他	総数
2007年	56	386	442
2008年	67	480	547
2009年	65	222	287

英語俳句とその他の英語表現の判定は文書を一つずつ目視によって確認した。この点については、人間の主観的な判断に基づく。その他の英語表現には、イベントの案内や、新年やクリスマスにおける時候の挨拶、商業広告などのSPAMメールが含まれる。

2007年～2009年の各年に投稿されたメールの本文を使ってSVMに学習させ、同じ年の訓練データをSVM分類器にかけると、それぞれ87.3%、87.8%、77.4%の精度を得た。ただし、すべての文書がその他の英語表現と判定されたことが理由である。

次に、すべての訓練データを使って学習させ、2010年の投稿記事に適用した結果は**表2**の通りである。なお、実験で使った文書の総数は13で、そのうち2件が英語俳句である。

表2．抽出実験の結果

文書数	英語俳句	精度	再現率
13	2	0.846	0

5．まとめ

本研究では、メールの本文やWeb文書の中に英語俳句が記述されているとき、一定の精度で判定することが期待されたが、文書の品詞成分を要素とする文書ベクトルを生成するだけでは、期待した効果が得られなかった。実験では、文書ベクトルの構成に単語の語幹を使わなかったが、訓練データのうち正解文書を増やす、メール本文のうち対象とする英語俳句の位置を絞り込むといった工夫の余地があると考えられる。

今後は、英語俳句の特徴から最適な文書ベクトルを構成し、抽出の精度を上げるとともに、作者ごとに英語俳句の特徴をとらえ、句風による分類や作品の良し悪しを判定する手法の開発が課題である。

References

[1] Shiki Haikusphere http://haiku.cc.ehime-u.ac.jp/

[2] 和田武、檀裕也、墨岡学「HAIKUサーバのアクセスログの解析と運用管理」情報処理学会第69回全国大会 2D-4

[3] 和田武、墨岡学「インターネット俳句サーバSHIKIの運用と効果」大学情報システム環境研究, Vol. 7, pp.71-74 (2006)

[4] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag (1995)

[5] 大北剛「サポートベクターマシン入門」共立出版 (2005)

[6] TreeTagger http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[7] LIBSVM http://www.csie.ntu.edu.tw/~cjlin/libsvm/