コーパスを用いた中国語ネット語の判定システム

竇　梓瑜（東京農工大学　工学府　情報工学専攻）
古宮 嘉那子(東京農工大学 工学研究院 先端情報科学部門)
小谷 善行(東京農工大学 工学研究院 先端情報科学部門)

A Detection System of Chinese Netspeak Using Text Corpus

Ziyu Dou(Graduate School of Engineering, Tokyo University of Agriculture and Technology)
Kanako Komiya (Institution of Engineering, Tokyo University of Agriculture and Technology)
Yoshiyuki Kotani (Institution of Engineering, Tokyo University of Agriculture and Technology)

１．研究背景

現在、中国では、インターネット利用者が爆発的に増え、それに伴って大量な中国語ネット語（以下、ネット語と書く）が現れた。ネット語の独特な言葉やその使い方は、インターネットだけではなく、徐々に人々の生活にも浸透してきている。しかし、中国の人口はおよそ１３．５億であるのに対して、中国のインターネット利用者は５．１３億と言われており、中国人の半分以上はインターネットを利用していない。そのような人々にとって、ネット語は理解しにくく、意味が分からなかったり、または意味の誤解から、トラブルになったりすることがある。こうした事態を避けるため、我々はコンピューターで自動的にネット語か書き言葉かどうかを区別するシステムを作成した。本システムは任意の中国語の一つ以上の文の入力に対して、ネット語であるかどうかの判断結果を出力する。

２．中国語ネット語特徴の検出システムの構成

中国語ネット語特徴検出システムの構成を、図１に示す。

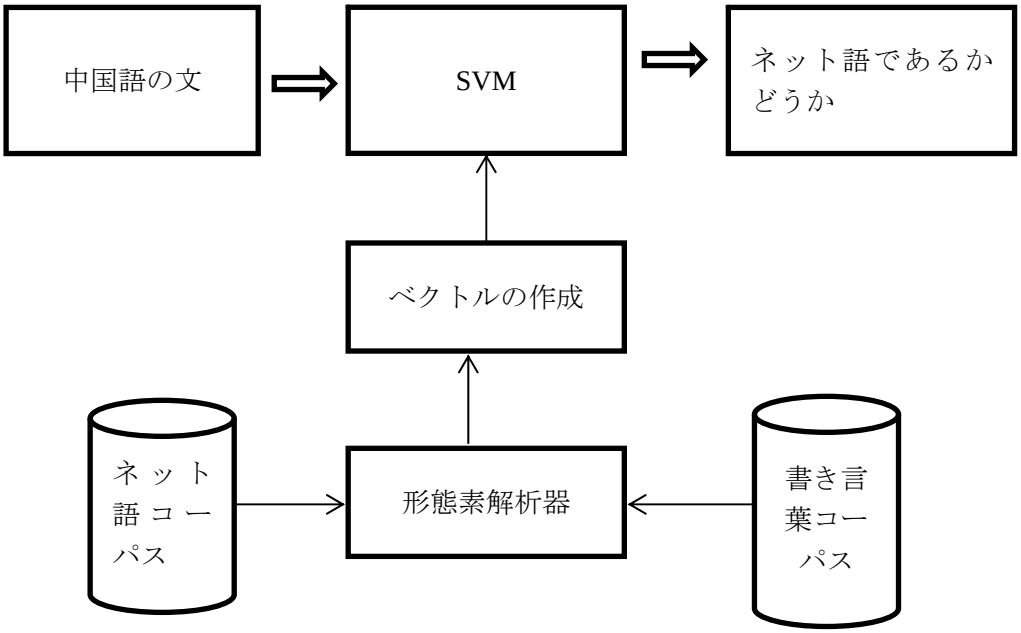

図１　中国語ネット語特徴の検出システムの構成

システムの入力は中国語文であり、出力は、ネット語であるかどうかである。
まず、ネット語コーパスと書き言葉コーパスの形態素解析を行い、それを特徴としてベクトルを作成し、サポートベクターマシン（SVM）（http://otndnld.oracle.co.jp/document/products/oracle11g/111/doc_dvd/datamine.111/E05704-02/algo_svm.htm）を用いて機械学習を行う。次に、入力の中国語文に対して同様に形態素解析を行い、ベクトルを作成し、先ほど学習した学習器を利用して、入力の中国語文がネット語かどうかを判定する。以下に各部分について順次述べる。

２．１　形態素解析

形態素解析の部分はICTCLASという中国科学院計算技術研究所の無料形態素解析プログラム（http://ictclas.org/）を使用した。このプログラムはHMMを基本解析方法として構成されている。プログラムはC言語版とC++言語版が存在するが、本システムで使われているのはそのC++言語版である。なお、この形態素解析プログラムの基本機能は単語分割、品詞つけ、未知語表記で、ユーザ自身でも辞書に単語を入れることが出来る。

このICTCLASプログラムの役割は、書き言葉コーパスの文、ネット語コーパスの文及び入力の文の形態素解析を行い、品詞タグを付けることである。また、ここで使用される品詞タグには、計算所漢語詞性標注集（http://icl.pku.edu.cn）が使われている。

２．２　ベクトルの作成とSVM

SVMにはSVM－lightという無料プログラムを使用した。SVMの入力はベクトルであるため、前処理としてベクトルを作成した。まず、ネット語コーパスと書き言葉コーパスの形態素解析の結果を合わせ、出現した全ての単語を統計し、単語毎に番号を付ける。ただし、ここでは、単語が同じでも、品詞が違う場合には、異なる単語として違う番号を付与した。特にここで統計するとき、アルファベットを除くため、タグ　/x　が付いている単語を全部除いた。ベクトル作成は、コーパスの文ごとに行った。素性は単語であり、素性値はコーパス中の頻度である。

３．実験と結果

３．１　　データ

実験用のデータは、全てインターネットから収集したものである。

このうち、ネット語コーパスは、新浪微博（http://www.weibo.com）から収集した。この新浪微博は、現在２０１２年１月までに、２．５億を超えるユーザを持つ、中国最大のミニブログである。このミニブログはツイッターと同様、一発言として入力できるのは１４０文字までという制限がある。政府などの公式機関のユーザも多数あるが、ほとんどの発言はインターネット利用者の日常的な呟きなので、典型的なネット語があると考えられる。

これらのインターネットユーザ発言を収集するとき、火車採集器（http://www.locoy.com）という無料ウェブデーター収集プログラムを利用し、無作為にユーザを選択してデータを取得した。最初に収集されたデータは既にHTMLタグを全部除いた文である。このような文が一行一文という形で、テキストの中で記録されている。新浪微博から収集したネット語コーパスの文の数は５０００文である。このほかに、百度貼バ（BBS サイト）（http://tieba.baidu.com）から１００文を、ネット語のコーパスとしてテストに利用した。

書き言葉コーパスは前述のように中国国家文字委員会の現代中国語コーパス（http://www.cncorpus.org/）の中の新聞と社論というカテゴリのコーパスから取ったも

のである。このコーパスは、ネット語コーパスと同じく一行一文でテキストの中に記録されている。現代中国語コーパスから利用する文の数は２０００文である。このほかに、SOHU ニュース（http://www.sohu.com/）から１００文を、書き言葉のコーパスとしてテストに利用した。

３．２　　実験設計及び結果

ネット語コーパスと書き言葉コーパスのどちらの文かを判定する制度を見るために、実験１～実験４の四種類の実験を行った。以下にそれぞれについて述べる。

（1）実験１　CLOSED　テスト

まず、ネット語コーパスと書き言葉コーパスのどちらの文かを判定する制度を見るための CLOSED テストを行った。CLOSED テストでは、テストデータは訓練データとして利用したものである。また、CLOSED テストでは、顔文字などの符号を削除していない（略：符号あり）実験を行った。以下に、CLOSED　テストにおける、訓練データ、テストデータの種類と数、また結果の求め方を説明する。さらに実験の結果を表１に示す。表１から、CLOSED テストでは、全ての文がネット語であるかどうか、正しく判定できていることが分かる。

訓練データ：ネット語コーパス　５０００文、書き言葉コーパス　２０００文
テストデータ：ネット語コーパスからの１０００文と書き言葉コーパスからの４００文
結果：ネット語の率＝SVM がネット語として認識した文/１０００
　　書き言葉の率＝SVM が書き言葉語として認識した文/４００

ここで、ネット語におけるネット語の率はネット語の再現率であり、書き言葉における書き言葉の率は書き言葉の再現率となる。

表１　CLOSED　テスト　結果

	ネット語の率	書き言葉の率
ネット語コーパスからの１０００文（符号あり）	１００％	０％
書き言葉コーパスからの４００文（符号あり）	０％	１００％

（2）実験２　OPEN　テスト（符号あり及び符号なし）

次に、訓練データとテストデータの重複を許さない OPEN テストの二つを行った。OPEN テストでは、符号と英文字によって構成され顔文字や略語の結果への影響を実証するため、符号および英文字がある実験と符号および英文字がない（略：符号なし）実験を行って比較した。以下に、OPEN テストにおける、訓練データ、テストデータの種類と数、また結果の求め方を説明する。さらに実験の結果を表２に示す。

訓練データ：ネット語コーパス　４０００文、書き言葉　１６００文
テストデータ：ネット語コーパスの上記４０００文を除いて残った１０００文、書き言葉

コーパスの上記１６００文を除いて残った４００文
結果：ネット語の率＝SVM がネット語として認識した文/１０００
　　書き言葉の率＝SVM が書き言葉語として認識した文/４００

表２　OPEN テスト　結果

	ネット語の率	書き言葉の率	正解率
ネット語コーパスからの１０００文（符号あり）	９８．４％	１．６％	９２．６％
書き言葉コーパスからの４００文（符号あり）	２２．０％	７８．０％	
ネット語コーパスからの１０００文（符号なし）	９８．９％	１．１％	８４．９％
書き言葉コーパスからの４００文（符号なし）	５０．２％	４９．８％	

（3）実験３　ネット語１００文と書き言葉１００文　テスト（符号あり及び符号なし）
次に、訓練データとして使用したコーパスとは異なるコーパスとして、ネット語コーパスに BBS、書き言葉コーパスに新聞を利用した際の OPEN テストを行った。以下に、その際の訓練データ、テストデータの種類と数、また結果の求め方を説明する。さらに実験の結果を表３に示す。この実験の際にも実験２と同様、符号ありと符号なしの実験を行って比較した。
訓練データ：ネット語コーパス　５０００文、書き言葉コーパス　２０００文
テストデータ：　百度貼バ　（BBS サイト）から取った１００文をネット語とし、SOHU ニュース　から取った１００文を書き言葉とし、テストを行う
結果：ネット語の率＝SVM がネット語として認識した文/１００
　　書き言葉の率＝SVM が書き言葉語として認識した文/１００

表３　ネット語１００文と書き言葉１００文　テスト　結果

	ネット語の率	書き言葉の率	正解率
ネット語１００文（符号あり）	８７％	１３％	７１．５％
書き言葉１００文（符号あり）	４４％	５６％	
ネット語１００文（符号なし）	６９％	３１％	８３．５％
書き言葉１００文（符号なし）	２２％	７８％	

（4）実験４　アンケート（符号ありと符号なし）

最後に、比較対象として、人間に符号ありと符号なしの際、どの程度ネット語を判定できるかのアンケートを行った。以下に、アンケート実験における、訓練データ、テストデータの種類と数、また結果の求め方を説明する。さらに実験の結果を表４に示す。この実験の際にも実験２、実験３と同様、符号ありと符号なしの実験を行って比較した。

テスト方法：数人の中国人インターネット利用者にアンケート

テスト内容：上記の　百度貼バ　（BBS サイト）から取ったネット語５０文を、SOHU ニュース　から取った５０文を符号ありと符号なし二回、被験者に判断してもらった。顔文字で簡単に人間がネット語を判断し、その結果を覚えてしまう可能性を排除するため、先に符号なしのアンケートを行い、その後符号ありのアンケートを行った。また、

結果：ネット語の率＝SVM がネット語として認識した文/５０

　　書き言葉の率＝SVM が書き言葉語として認識した文/５０

として計算した。

表４　ネット語５０文と書き言葉５０文の人工判定テスト結果

	ネット語の率（符号あり）	書き言葉の率（符号あり）	正解率（符号あり）	ネット語の率（符号なし）	書き言葉の率（符号なし）	正解率（符号なし）
２０代中国人男性学生	４４％	１００％	７２％	４４％	１００％	７２％
２０代中国人女性学生	２０％	１００％	６０％	１６％	１００％	５８％
平均	３２％	１００％	６６％	３０％	１００％	６５％

４．考察

まず、表２の OPEN テストの結果が示すように、本研究のネット語コーパスおよび書き言葉コーパスは、確実に区別が存在する。特にネット語コーパスは、符号ありと符号なしの場合、それぞれ、９８．４％と９８．９％の再現率となった。しかし、書き言葉コーパスに対する Open テストは、符号ありのとき、再現率が７８．０％で、符号なしの場合の再現率は４９．８％まで下がった。これは、訓練に使用したネット語コーパスの量が多いためと、ネット語コーパスの中でも、書き言葉のような文が多数存在するためであると思われる。また、符号がある場合とない場合の再現率の差から、機械学習において、符号の影響が大きいことが分かる。

続いて、実際な文に対するテストの結果（表３）を分析する。まず、符号がある場合とない場合と比べると、ネット語１００文に対する認識正解率は１８ポイント上がり、８７％まで達成した。これに対し、書き言葉は符号なしの場合のほうが２２ポイント上回り、７８％まで達成した。これによって、符号がある場合、文がネット語として認識される傾向が強まり、符号がない場合には、書き言葉として認識される傾向が強まることが分かる。

アンケート（表４）の結果を見ると、全部書き言葉の判定は１００％正解したが、ネット語の判定はいずれも４４％と２０％まで止まったことが分かる。また、符号がある場合の、符号がない場合と比べた正解率の上昇はわずかであった。実験４と実験３の正解率を

比べると、機械のほうが、正解率が上回ることがわかった。とこれは、ネット語といっても、BBS では、書き言葉的な表現も多数存在することが、判定の結果に大きく影響したためだと思われる。それに対し、書き言葉は、１００％の正解率で、人間が書き言葉を認識するのは簡単だったことが分かる。ネット語の定義を人間が判断できるものとすれば、再現率にも変化があるだろう。

最後に表３と表４から機械学習と人間の正解率を比べる。表３から、機械学習は最高８３．５％、表４から、人間の判断は最高６６％であるため、作成したシステムの性能が人間に上回ることがわかる。

５．結論

本論文では、文の入力に対してネット語かどうかの判定を行うシステムを作成した。入力の文は形態素解析を行い、ベクトル化したあと、SVM を使ってネット語かどうかを判定した。

実験結果から、本システムは、符号がある場合、ネット語に対する判定の正解率が上がり、符号がない場合、書き言葉に対する判定再現率が上がることが分かった。また、人間に対するアンケートの結果から、人間でもネット語かどうかの判定が難しいことが分かる。特に、BBS やミニブログなどの情報には、ネット語的な特徴がない言葉も多数存在するので、難しかったようである。また、機械学習と人間の正解率を比べると、機械学習は最高８３．５％、人間の判断は最高６６％であるため、作成したシステムの性能が人間に上回ることがわかった。

文　献

佐藤敏紀　「Perl で自然言語処理」　東京工業大学奥村研究室
http://www.slideshare.net/overlast/perl-5460697

谷岡　広樹、丸山　稔(2005)「形態素解析に基づく SVM を用いたアスキーアートの識別」電子情報通信学会技術研究報告. PRMU, パターン認識・メディア理解 「104:670」, pp.25-30

黒橋禎夫　機械学習に基づく自然言語処理　京都大学情報学研究科
http://nlp.ist.i.kyoto-u.ac.jp/member/kuro/lecture/LIP10/LIP09.pdf

Jin'ichi Murakami, HMM(Hidden Markov Model, 隠れマルコフモデル)
http://unicorn.ike.tottori-u.ac.jp/murakami/doctor/node7.html

語料庫在線　http://www.cncorpus.org/

SVM-light　http://www.cs.cornell.edu/People/tj/svm_light/

ICTCLAS http://ictclas.org/

情報と通信のハイパーテキスト http://www.yobology.info/text/index.htm

Shogo Computing Laboratory http://sora-blue.net/~shogo82148/memo/algorithm/svm/