Vol.2009-NL-193 No.4
2009/9/28

中国語インターネット用語コーパスの構築及び分析について

丁育華[†]　　任福継[††]

近年、インターネットの普及に伴い、従来の伝統言語形式と異なるインターネット用語は出現している。インターネット用語は特に若者達に愛用され、その自己表現の手段の 1 つとして流行されている。本稿は、従来の自然言語処理ではあまり扱わなかった中国語インターネット用語のコーパスを構築、分析することを目的とし、中国語インターネット用語をその用途や形式、出典に基づいて 8 種類に分けた。また、各類別の分布を分析し、さらにインターネット用語に反映される若者達の心理を解析した。

Constructing Chinese Internet Terminology Corpus

Yuhua Ding[†]　and Fuji Ren[††]

In recent years, internet widely used in the internet industry with a different form of the traditional languages i.e. internet terminology has been emerged. The internet terminology is very popular among young peoples, and it is used for self-expression. A few researchers have studied the Chinese internet terminology language processing but exact meaning of the words is not completely understood yet. Therefore an attempt has been made to build the corpus of Chinese internet terminology. In this paper, the Chinese internet terminology is divided into eight types based on the purpose, format, and source. We have carried out the analysis of the distribution of each category, and the psychology of young people reflected in the Chinese internet terminology.

1. はじめに

1.1 背景と目的

近年、インターネットの普及に伴い、従来の伝統言語形式と異なるインターネット用語は出現している。The New York Times によると、アメリカ国家創作委員会（The National Commission on Writing）とピューネット、アメリカライフプロジェクト（the Pew Internet & American Life Project）との合同研究では、700 名の 12～17 歳の学生を対象として調査を行った結果、約 2/3 の被験者は彼らの宿題やレポートの中でインターネット用語を使用していることがわかった。約 1/2 の被験者は正確な文章記号を省略したり、誤って大文字を使ったりすることがよくあるそうである。また、約 1/4 の学生は宿題の中で記号からなる顔文字を使用し、1/3 の学生は略語（例 : LOL=laugh out loud）を使ったことがあるとわかった。

一方、中国では、ネットユーザーは主に若者を中心とし、平均年齢は 27 歳ぐらい、学歴は大半大卒以上、そして相対的に高収入の人々となっている。こうした若年層ユーザーの存在により、新しい言語手段としてインターネット用語は作り出された。これらのインターネット用語は伝統の中国語と異なり、方言・俗語・外国語・略語・諧音（xieyin、同じ発音をする異文字）等から変形し、さらにこれらの組合せにより、新たな言語表現が形成されている。

インターネットという非現実の世界では、身分、年齢、性別などによる制限がなく、誰でも自分の経歴、知識、趣味からオリジナルの考えを主張することができる。これらのオリジナルの考えを生き生きとした言語で表現するために、インターネット用語は汎用化されてきた。その中で、若者達は中高年と比べ、コンピュータ技術に馴染みやすく、またコミック等の影響で自分の感情をより簡単、形象に表現することが好ましい。Blog といったネット日記や BBS（Bulletin Board System）、チャットといった各種の同時交流ソフトウェア等はちょうど若者達にこうした自己表現の場を提供している。若者達にとってのインターネット用語は、新鮮感、高効率だけではなく、ファッションの印ともなっている。一方、インターネットはすべてのユーザーに公平な創作の場を提供しているが、必ずしもすべてのユーザーのオリジナル表現がインターネット用語として定着していくとは限らない。その僅かの一部は簡潔さ、使いやすさ、または面白さに優れ、ほかのユーザーにも転用されながら広がり、時間経つとともに特定した物事を指す一種の公用語として定着してきたのである。インターネット用語は簡潔な表現、迅速な反応、また書きやすいといった特徴は、若者達の反抗心理も反映

[†] 徳島大学大学院ソシオテクノサイエンス研究部国際連携教育開発センター
Institute of Technology and Science the University of Tokushima
[††]徳島大学大学院ソシオテクノサイエンス研究部知能情報工学科
Institute of Technology and Science the University of Tokushima

している。匿名という非現実の世界で若者達の抑えられた感情・個性の解放のコミュニティ手段としてインターネット用語は伝播されてきた。

第 19 回「中国インターネット発展状況統計報告」では、2006 年まで中国のインターネットユーザー数は 1.37 億人に達し、総人口の約 10.5%に占め、近い未来では中国のユーザー数はさらなるスピートで増加すると報告された。こうして、少なくても近未来でも存続しそうなインターネット用語をコーパスとして構築するのは、今後のインターネットライフに役立つと言うまでもなく、若者達の心理を解析することにも情報を提供することができると考えられる。

そこで本研究は、ユーザーが利用しやすい日中インターネット用語と若者用語の対訳システムの構築を最終目的として念頭に置き、現段階ではまず中国語インターネット用語に着手する。本稿は、従来の自然言語処理ではあまり扱わなかった中国語インターネット用語のコーパスを構築、分析することを目的とする。

1.2 既存研究との関連および本研究の特徴

ここでは、本研究に関連する中国語インターネット用語の分析や情報処理分野でのコーパスに関する研究を概観し、本研究の特徴を述べる。

まず、中国語インターネット用語については、張らは、インターネットで発生した交流現象をネット交流、現実社会で発生した交流現象を現実交流と定義し、また文化的背景や心理的差異といった要素は人々の交流に制約作用を行い、こうした制約作用的な要素の集合を「語境」(言語環境、言語的脈絡)と意味付け、インターネット用語をめぐる語境要素について論じた[1]。林は、言語学の視点からインターネット用語の類型およびその特徴について述べた[2]。

情報処理分野でのコーパスに関する研究をみると、中川らは、専門家の語感を信用して、中心的概念から距離感を考慮した用語選択を行うことにより、少人数でも妥当性の高いがん用語集合の作成が可能であることを示した[3]。このほか、慣用句の曖昧性解消技術の確立に向けて、日本語慣用句コーパスを構築した橋本らの研究[4]や、1 万文の製品レビュー文に対して、人手による注釈付けを行い作成した評判情報コーパスについて、その特徴を分析した宮崎らの研究[5]もある。

これらの研究は、言語学から中国語インターネット用語について論じたものがあるが、自然言語処理では中国語インターネット用語を扱うものがほとんどみられなかった。また本稿は、中国語インターネット用語コーパスを構築することを通して、現在の若者達の心理について解析を行い、これは現代社会に起る様々な若者問題の解決にも一助になると確信している。

以下では、２章で中国語インターネット用語の定義、特徴及び自然言語処理における困難点を述べる。次に、３章では中国語インターネット用語コーパスの構築、４章ではコーパスに対する分析を記述する。最後に、５章で本論文のむすびと今後の課題を述べる。

2. 中国語インターネット用語について

2.1 定義

インターネット用語はコンピュータを通じてインターネットコミュニケーション領域で使用されている特定非正式言語である。これは日常生活の交際活動と緊密に関わりながらも、明確な区別がある。一般的に区分すると、広義のインターネット用語を 3 種類に分けることができる（表 1)。第 1 類は、インターネットと関わる専門術語で、つまりインターネットに関する術語を中国語で表記したものである。第 2 類は、インターネットと関わる特別用語である。これはインターネットから生じた様々な物事・現象に対して称するものである。第 3 類は、BLOG や BBS 等で使われている慣用語である。これは前述したように、その多くは方言・俗語・外国語・略語・象形文字・誤字・諧音等から変形したものであり、または BBS でよく転用される一種の特定な事象を表すフレーズもこれに含まれる。狭義のインターネット用語は第 3 類を指している。本論文ではこの狭義の意味での中国語インターネット用語を扱うことにする。

表 1　インターネット用語の分類と用例

類別	用例	日本語訳
インターネットと関わる専門術語	鼠標	マウス
	硬件	ハードウエア
	軟件	ソフトウエア
	病毒	ウィルス
	寛帯	広帯域ネットワーク
	在線	オンライン
	聊天室	チャットルーム
	防火墙	セキュリティ
インターネットと関わる特別用語	网民	ユーザー
	网吧	インターネットカフェ
	黒客	ハッカー
	短信息	オンラインメッセージ
	電子商務	インターネット上の商務活動
	虚擬空間	Virtual Host Virtual Server
BLOG、BBS等で使われる慣用語	美眉	綺麗な女の子
	斑竹	BBS等の管理員
	恐竜	容貌の醜い女性に蔑視して称する
	菜鳥	不器用な人、またはインターネット経歴の短い人
	東東	「東西」と同義、物事に対する可愛らしい呼び方
	B4、BS	「鄙視」と同義、軽蔑する
	GG	「哥哥」と同義、兄のこと

2.2 特徴

インターネット用語は第 4 の媒体、即ちインターネットの普及により興った時代の産物で、インターネットという限定環境の中で使用される言語である。これは公用語や日常生活用語と違って、以下の特徴が挙げられる。

(1) 更新の迅速性

公用語といった正式言語は大抵長い歴史の中で発展、変遷してきたため、その語彙、文法等は系統的に一致して伝承されてきた。これに対して、インターネット用語は出現してから今日に至るまで十数年に過ぎないにも関わらず、その語彙の内容・数量・類型は迅速なスピートで更新されてきた。

(2) ユーザーの閉鎖性

前述したように、インターネット用語は新鮮な表現、迅速な反応、また書きやすいといった特徴が特に若者達に好まれている。また、更新の速さより、利用頻度の高いユーザーのみ使いこなせることに決まっている。利用頻度の低いユーザーはその更新を知らない場合、意味を理解できないという結果につながる。これは本研究の目的にも反映され、最新のインターネット用語コーパスの構築はこういったユーザーのインターネットライフに貢献できると考えられる。

(3) 非系統性

公用語や日用生活用語は、文字、語彙、文法等の面では厳密な系統を形成し、その固定規範をもっている。一方、インターネット用語は元々普通言語から由来したが、語彙や文法の面ではもっと利便性や簡潔性を求め、普通言語を修飾した一種の新しい言語になっている。そのゆえ、語彙や文法において独立した言語体系に至らない。多くのインターネット用語は普通言語と結合しないと、完全の意味を表せないことがある。

(4) 生き生きとした表現とユーモア性

例えば、「菌男」と「霉女」はそれぞれ容貌の醜い男性、女性を意味し、本来は公用語の「俊男」と「美女」から変遷した諧音文字である。「俊男」と「美女」は容貌の美しい男女に対する呼称であることと反して、「菌」「霉」(「カビ」の意味）は賞味期限の切れた食品や変質といったイメージに連想しやすく、諷刺的な逆意味になっている。また､顔文字などを用いて生き生きとした感情や心理を表すこともよくみられる。

(5) 簡潔性

スピートを重視するインターネット環境では、言語表現の簡潔さ、早さが求められている。例えば、「MM」は「美眉」（MeiMei、表 1）の発音の最初のアルファベットを取り、綺麗な女の子を意味する。「GF」は英語の「Girl Friend」の略語となっている。「1314」は中国語の「一生一世」の発音と似た数字諧音で、一生の意味を指す。このように、ユーザーは入力のスピートを上げるため、様々な略語や顔文字等を使用している。

2.3 自然言語理解における困難点

前述したように、インターネット用語の多くは、方言・俗語・外国語・略語・象形文字・誤字・諧音等から変形したものであるため、自然言語理解において以下のような困難点がある。

(1) 公用語にない方言・俗語の理解不可

インターネット用語には親しみを表すには、方言や俗語等はよく使用されている。例えば､「粉」は福建の方言で、中国語の「很」と同じに「とても」の意味をする。「素」は標準語の「是」に対する台湾の発音で、意味は「是」と同じに「～は～だ」と表す。また､「白烂」は福建の俗語である「白卵」を指し、愚かで煩わしい人を意味する。このように、元々公用語にない方言や俗語は出現している場合、自然言語処理の分野で理解不可と扱われることになる。

(2) 略語の理解不可

略語の使用もスピートを重視する BBS ではよくみられる。例えば、「JJ」は「姐姐」（JieJie、姉)、「DD」は「」（Didi、弟）の略語である。また「PF」は「佩服」（PeiFu、感心・敬服する）の略語として、「PL」は「漂亮」（PiaoLiang、綺麗）の略語として使われている。こういった略語の使用は入力のスピートを大幅に高めることができるが、自然言語処理の場合では理解不可の結果となってしまう。

(3) 象形文字・諧音・誤字の誤解

一方、前文で述べたように、現時的なインターネットコミュニケーションを行うとき、生き生きとした表現、または可愛らしい表現は特に若いユーザーに好まれる。これで、わざと間違った誤字、本来の意味と関係ない象形文字、また同じ発音する異文字の諧音などを用いる傾向がある。例えば、你」(ni、あなた）の誤字として「泥」(ni、泥）を使ったり、「喜歓」(xihuan、好き）の代わりに諧音の「稀飯」(xifan、お粥）を使ったりすることがみられる。また､「囧」「粽」といった古文字を新たにアレンジし、本来の意味と全く関係のない、一種の新しい象形文字として用いる傾向がみられる。「囧」(Jiong）は本来「照らす、光明」の意味であるが、見かけが困った顔をした人に見えることから､ユーザーは､「憂鬱､悲しい､どうしようもない､言葉で言えない」といった新しい意味をつけた。「粽」(Mei）の本来の意味はともかくとして、呆けた状態を意味する「呆」が 2 つ一緒になっていることから、他人に対して「馬鹿、純粋すぎる」という時に用いるとみられる。しかし、自然言語理解の分野では、以上のような象形文字や諧音、誤字の由来についての認識がなければ、処理上においてもちろん誤解されやすいと考えられる。

(4) 同一文字の違う表現

また、自然言語分野において理解を阻むもう 1 つの壁は、「稀飯」や「泥」､「斑竹」「恐竜」(表 1）といった単語は、本来のもつ意味を全く違う意味に転用される場合である。これらの単語は、一般的な言語環境の中では本来の意味と解釈するが、BLOG

や BBS といったインターネット環境の中ではその転用される特定した意味と解釈しないと、文章の脈絡が通らなくなる。しかし、どんな手法でどのように区別するのは、インターネット用語だけではなく、他種類の自然言語処理においても未だに難点の 1 つである。

3. コーパスの構築

3.1 データ収集

本研究はインターネット用語コーパスの構築を目的とするため、データの収集はインターネット上で行った。主に BLOG や BBS、ネット掲示板を対象とし、その上に書かれている特定した意味をもつ用語を抽出しコーパスにまとめた。

3.2 中国語インターネット用語の分類

収集したインターネット用語を、その用途や形式、出典に基づいて 8 種類に分けることができる。以下では、この 8 種類を例に示しながら説明する。

(1) BBS における日常用語

BBS でよく使用され、一般の標準語の意味と異なる用語の例を表 2 に示す。これをみると、「马甲」（中国服のそでなし）、「水手」（水夫）、「灌水」（水を灌ぐ）といった用語は標準語の中でも独自の意味をもつが、BBS の中で使われる場合、BBS 用語としての特定した意味に解釈されなければ、文脈は理解できなくなる。

表 2　BBS における日常用語の用例

斑竹	「版主」とも言う。BBS等の管理員。
马甲	同じ登録ユーザーは他のIDも登録し、その他のIDを指す。
大蝦	「大侠」と同音。インターネット経験の長い人に対する尊称。
灌水	元々はあまり意味のないスレッドを書く行為を指していたが、 その後スレッドやレスポンスを書くことと派生した。
水手	スレッドやレスポンスを書くのを好きな人。
拍砖	スレッドやレスポンスに対して違う意見を発表すること。
楼主	最初にスレッドを書いた人。
盖楼	1つのスレッドに対してコメントを書くこと。
顶	一般的に1つのスレッドの中にレスポンスが書かれると、そのスレッドは話題の最上層にあがることになる。このレスポンスを書くことを「顶」という。
路过	真面目にレスポンスを書きたくないが、ポイントまたは経験値が欲しいため、「通りかかる」とコメントする。

(2) 諧音、誤字、方言、俗語、象形文字等からの転用

諧音、誤字、方言、俗語、象形文字等から転用してきたインターネット用語（表 3）は、一見してその自身の意味をもつが、BBS 用語と同じように、その文脈の中で前後と呼応してこの特定した意味と解釈しないと、文章の意味を理解できないことになる。

表 3　諧音、誤字、方言、俗語、象形文字等の転用の用例

表	「不要」の中国語の発音を早く言った感じ。「不要」と同義、してはいけない。…するな。
偶	「我」の方言。私の意味。
酱紫	「這様子」の中国語の発音を早く言った感じ。こんなに、このように、こんなふうに。
果酱	「过奖」の諧音。ほめすぎる。過分に褒める。
油墨	「幽默」の諧音。ユーモア。ユーモラスな。
虾米	「なに」の意味。福建の発音。
青蛙	容貌の醜い男性。
筒子	「同志」の諧音。同志。人、…さん。 または「知らない人に呼びかけるときに単独で用いる」もしもし。
新蚊连啵	「新聞聯播」の諧音。たくさんの蚊に噛まれること。
人参公鸡	「人身攻撃」の諧音。人身攻撃。

(3) 中国語からの略語

中国語からの略語（表 4）は、中国語漢字の発音（拼音）の最初のアルファベットを取って組み合わせたものである。よって、拼音を熟知することが好ましい。しかし、出現率の高い用語は、その拼音略語が大抵定着し、拼音を熟知しないユーザーにとっても障碍なく利用されやすいという利点があり、BLOG や BBS の中でよくみられる。

表 4　中国語からの略語の用例

略語	中国語単語	意味
BT	变态(BianTai)	変態
BC	白痴(Baichi)	白痴、馬鹿。
FB	腐败(FuBai)	行いが堕落する。または制度・組織などが乱れる、腐敗する。
BH	彪悍(BiaoHan)	猛々しい、剽悍である。
FQ	愤青(FenQing)	「愤怒青年」を短縮した、中国語で「怒れる若者」を意味する言葉。
CJ	纯洁(ChunJie)	純潔である、汚れのない。
HC	花痴(HuaChi)	アイドルなどにすごく夢中になっている女の人
PAP	拍马屁(PaiMaPi)	お上手を言う。媚び諂う。ごまをする。
XDJM	兄弟姐妹(XiongDiJieMei)	姉妹兄弟
PFPF	佩服佩服(PeiFuPeiFu)	感心する。敬服する。頭が下がる。

(4) 英語からの略語

英語からの略語は、中国語拼音からの略語と同じに、英語に対するある程度の熟知度が要求されるが、その内容はほとんど表 5 のような日常会話であるため、ユーザーの英語力にあまり制限されなく定着してきた。

表 5　英語からの略語の用例

Q	Cute
GF	Girl Friend
BF	Boy Friend
CU	See You
DIT	Do It Yourself
LOL	laugh Out Loud
BTW	By The Way
BRB	Be Right Back
SOHO	Small Office Home Office
TTYL	Talk To You Later

(5) 数字の略語

数字略語は、中国語漢字の諧音を利用し、煩瑣な漢字文章を簡単な数字で表すことで、現時的なネットコミュニケーションに面白さと便利さを添えている。また、その内容の多くはロマンチックなフレーズであるため、ファッションを追い、愛の宣言を恐れない若者達に愛用されている。

表 6　数字の略語の用例

略語	漢字諧音	意味
065	原谅我	私を許してください。
1314	一生一世	一生のこと。
2037	为你伤心	貴様のことで悲しむ。
3399	长长久久	長い間、久しい間。
4242	是啊是啊	そうだそうだ。
520	我爱你	愛している。
687	对不起	申し訳ない、すみません。
737420	今生今世爱你	この一生愛する。
88	拜拜	Byebye、さようなら。
98	酒吧	バー、酒場。

(6) 漫画やアニメ、ゲームより出典のもの

漫画やアニメ、ゲームより出典のもの（表 7）は、前述した 5 種類と比べ、ユーザーの属性によって偏ることがある。例えば、漫画やゲーム等によく接触しているユーザーにとってこれらは欠かせない表現の手段だが、一般のユーザーにとって知られていないものが多いと考えられる。

表 7　漫画やアニメ、ゲームより出典のものの用例

幼齿	「素人」ともいう。幼くて、あまり経験のない人。
达人	日本語の「達人」と同義。
兄贵	筋肉マンのような男の人。
姐贵	筋肉マンのような女の人。
轰杀	人を殺す行為を指す。
破天	天を破りひらくようなこと。激怒するときに使う。
收声	「黙れ」との意味。
废柴	能なし、碌で無し。
热血	本来は攻撃力を2倍に高める能力、その後、激昂した情緒を指すことに派生した。
恶趣味	変な癖、変わった嗜好。

(7) 耽美文化に関する術語

第 7 種類の耽美文化に関する術語（表 8）は、ほとんど日本の耽美文化から由来し、またはそこから派生したものである。そのため、こういった用語の漢字は日本語の漢字を踏襲し、意味も大体日本語と似ている。字面で見ると中国語では全く意味不明のものとなるが、日本語での意味と関連付けないと分からないことになる。

表 8　耽美文化に関する術語の用例

BL	Boy's Love、男性の同性愛。
GL	Girl's Love、女性の同性愛。
SM	Sadism & Masochism、それぞれ相手を肉体的・心理的にいじめることで、快感を得るタイプと、いじめられることで快感を得るタイプ。
18禁	18未満の人は見てはいけないものを指す。
耽美	日本語の「耽美」から由来し、その後、BLのことを指すことになった。
耽美狼	耽美文化が好きな女の人。
同人女	言葉通りの「同人活動に関わっている女性」という意味よりも、やおい、ボーイズラブなどの同人誌に対する強い趣味を持つ女性を差別的に呼称する意味合いの強い俗語である。
萝莉控	（英語: Lolita complex、短縮形：ロリコン）とは、幼女・少女に対する（主に成人）男性の性的または恋愛的関心・性嗜好をいう。

(8) **顔文字**

顔文字（表 9）は実際に文字とは言えず、記号の組み合わせることによって、人間の顔表情を生き生きとして表わすものである。これは系統的に整う文字体系と異なり、人間の主観的な感情、嗜好によって作り出されたものであり、作り出す主体の個人差があると言うまでもない。こうして様々な顔文字は違う見る人に違う意味と解釈されることがあるが、前述したように、人間の顔表情を表しているので、理解の個人差があっても全く反対の意味と解釈される可能性が低い。本研究では最も多い方の解釈に基づいて、顔文字の説明をコーパスにまとめることにした。

表 9　顔文字の用例

:-)	普通の笑顔
:-D	非常にうれしそうに笑う
8-)	笑っている人の目は大きい
:-O	びっくりした顔
:-<	悲しい時の苦笑い
#-)	徹夜して、目を開けられない表情
?_?	茫然した顔
=^_^=	恥ずかしげに笑っている顔

3.3 インターネット用語コーパスの構築

本研究で構築した中国語インターネット用語コーパスに対する使用例を図 1 に示す。これは、前文で述べた 8 種類に基づいて、中国語インターネット用語の意味を検索するときの手順を示しているものである。これをみると、このコーパスの利用を通して、中国語インターネット用語の意味を簡単に検出することが可能であると考える。

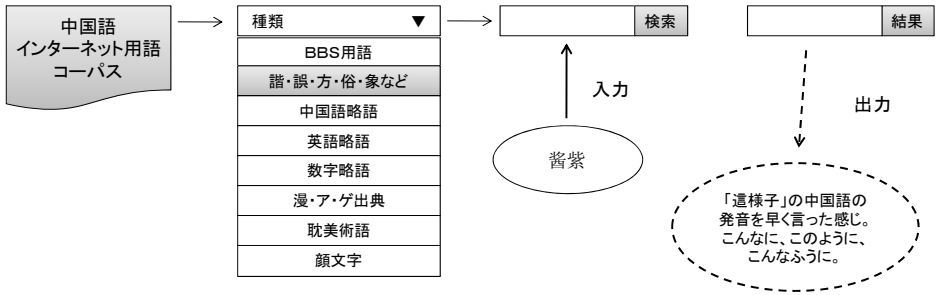

図 1　インターネット用語コーパスの使用例

4. 分析

4.1 中国語インターネット用語の類別分布

本節では、前文で述べた 8 種類のインターネット用語をコーパスにまとめた結果について分析する。まず、これらのインターネット用語の類別分布を表 10 に示す。

表 10　中国語インターネット用語の類別分布（個）

BBS用語	諧･誤･方･俗･象など	略語（中）	略語（英）	略語（数）	漫･ア･ゲ出典	耽美術語	顔文字
81	111	79	61	246	41	35	147

これをみると、中国語インターネット用語の中では、類別分布の高い順は略語（数）、顔文字、諧・誤・方・俗・象など、BBS 用語、略語（中）、略語（英）、漫・ア・ゲ出典、耽美術語となっている。実際の出現頻度、つまり使用頻度については統計的なデータがないため、集計できなかったが、筆者がデータを収集している感覚では、諧・誤・方・俗・象などと、BBS 用語といった種類の用語は、使用頻度が他の種類と比べより高くみられると感じた。これはやはり他の種類と比べ、新鮮・ユーモアなアイディにより新たに作り出されたものが多いため、ファッションを追う若者達に愛用されているだろうと考えられる。

4.2 中国語インターネット用語の意味解析

非現実的なインターネット世界においては、言葉がもたらす責任などを負う必要はない。こうした気楽で自由な環境は人間の言語に対する潜在的な創造力を刺激し、他人からの注目を呼ぶために知恵を絞り尽くさせる。若者達は情熱にあふれ、自由や個性を追求し、大胆な独創精神をもつといった性格から、どんどん新しく、面白いインターネット用語を創出している。そこで、これらのインターネット用語より若者達の心理を窺うこともできる。

まず、中高年と相反する若者の価値観をインターネット用語から読み取れる。西洋文化の影響などにより、重々しい中国伝統文化と離れつつある若者達は、自由や個性を追い、自己表現を重んじ、そして、注目を呼ぶことが好き、「好きなままにする」ことを強調する。たとえば、彼らはあまり謙遜せず、他人からの褒め言葉に対して堂々と受け取る。時々、自己陶酔することもあり、英語を引用する略語の出現やユーモアの溢れる諧音の転用などには、こういった表現欲は含まれていると考えられる。

また、中高年の落ち着いた生活態度と違い、若者は気楽で活発な生活雰囲気を作ることが好きである。彼らはインターネット上ですべての人やモノをからかい、気楽なネットライフを通じて、現実社会でたまった精神的なストレスを解消し、爽やかな心理状態を整える。たとえば、容貌の醜い女性を「恐竜」と呼んだり、意味のないスレッドを書くことを「灌水」に皮肉を言ったりするのはこういう面白い雰囲気作りのた

めである。また、多くの若者は面白いページや画像、動画を見ることを通してストレスを解消できるともいわれる。

心理的な面では、若者達の心理状態は中高年より露出しやすく、心理上では外の世界に対する一種の占有欲がある。この心理的な需要は所有する物資的なものを通して自己存在や位置を確認する。現在の若者達にとっては、高強度の仕事や競争の激しい社会環境などは全部ストレスとなっている。若者達は、高圧的な現実社会から気楽なインターネットに逃げ、そこで自己表現・陶酔することで平衡の心理状態へと整えるのである。

4.3 自然言語処理への応用

インターネット用語の大量出現と広く使用に従って、自然言語処理に大きな困難を与えている。本研究で開発される中国語インターネット用語コーパスはこのような困難に対して、幾つかの解決手法が考えられる。

(1) 意味解析

本研究で構築したコーパスを通して、従来の自然言語処理ではあまり扱わなかった中国語インターネット用語の意味を簡単に検索することが可能となった。例えば、従来の手法で「楼主」は「建物のホスト」を意味するが、インターネットでは、最初にスレッドを書いた人を指す。このコーパスと従来の意味辞書を融合して、用語の意味解析、さらに文章の意味解析が可能である[6,7]。

(2) 感情認識

本研究のコーパスに収録したインターネット用語は、現代の若者心理・感情に対する自然言語理解の技術向上には統計的な研究データを提供することができた。例えば、「青蛙（カエル）」は「容貌の醜い男性」を意味し、「嫌い」という感情を持つ。これらの感情情報を活かし、人間感情の認識及び機械感情の創生をもつ感情インターフェースの開発に貢献できると考える[8,9,10]。

(3) テキストマイニングへの応用

テキストマイニング技術の多様的な応用に新しい項目を添え、人間の感情・態度・心理の分析技術の 1 つとして、今後の意識評価や若者問題の解決などに貢献できると考えられる。

(4) 人間らしい会話システムの開発

インターネット用語コーパスを活かし、時代特色を持たせる会話システムを構築することが可能であると考えられる。このコーパスは動的に常に新しい用語を収集し、分類と解析を行い、コンピュータ会話システムの知識ベースに組み込むことによって、人間らしい会話システムを構築することが期待されている。

5. おわりに

本研究は、ユーザーが利用しやすい日中インターネット用語と若者用語の対訳システムの構築を最終目的として念頭に置き、現段階ではまず中国語インターネット用語に着手している。本稿は、従来の自然言語処理ではあまり扱わなかった中国語インターネット用語のコーパスを構築、分析した。そこで得られた成果として、まず、中国語インターネット用語をその用途や形式、出典に基づいて 8 種類に分けることができた。また、各類別の分布を分析し、さらにインターネット用語に反映される若者達の心理を解析した。これらの成果は今後の対訳システムの構築において大きな利便性をもたらすことができると考えられる。

今後の課題として、ユーザーが使いやすい中国語インターネット用語コーパスのインターフェースを開発し、Web で提供すると同時に、様々な応用システムに組み込んで、新しい自然言語処理技術を開発することが挙げられる。

参考文献

1) 張洪超, 林綱: 試析网絡用語的語境因素, 徐州師範大学学報(哲学社会科学版), Vol.29, No.3, pp.59-62 (2003).
2) 林綱: 网絡用語的類型及其特征, 修辞学習, Vol.109, No.1, pp.26-27 (2002).
3) 中川晋一, 内山将夫, 三角真, 島津明, 酒井善則: コーパスに基づくがん用語集合の作成と評価, 自然言語処理, Vol.16, No.2, pp.3-44 (2009).
4) 橋本力, 河原大輔: 日本語慣用句コーパスの構築と慣用句曖昧性解消の試み, 情報処理学会研究報告, Vol.2008, No.67, pp.1-6 (2008).
5) 宮崎林大郎, 森辰則: 製品レビュー文に基づく評判情報コーパスの作成とその特徴の分析, 情報処理学会研究報告, Vol.2008, No.90, pp.99-106 (2008).
6) Jiajun Yan, Bracewell B. David, Shingo Kuroiwa *and* Fuji Ren : Chinese semantic dependency analysis: Construction of a treebank and its use in classification, *ACM Transactions on Speech and Language Processing*, Vol.4, No.2, pp.1-20, 2007
7) Caixia Yuan *and* Fuji Ren : Accurate Learning for Chinese Function Tags from Minimal Features, *Proceedings of the ACL-IJCNLP 2009 Student Research Workshop*, pp.54-62, 2009
8) Fuji Ren, Recognizing Human Emotion and Creating Machine Emotion, Invited Paper, Information, Vol.8, No.1, pp.7-20, 2005.
9) Changqin Quan *and* Fuji Ren : Construction of a Blog Emotion Corpus for Chinese Emotional Expression Analysis, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp.1446-1454,2009
10) Fuji Ren : Affective Information Processing and Recognizing Human Emotion, *Electronic Notes in Theoretical Computer Science*, Vol.225, No.2, pp.39-50, 2009.