「不自然言語処理コンテスト」第１回開催報告

萩原正人 (楽天)　大原一輝 (フリー)　水野貴明 (バイドゥ)
橋本泰一 (東京工業大学)　荒牧英治 (東京大学)　竹迫良範 (サイボウズ・ラボ)
masato.hagiwara@mail.rakuten.co.jp, math.empress@gmail.com, takaaki.mizuno@gmail.com,
hashimoto.t.ab@m.titech.ac.jp, eiji.aramaki@gmail.com, takesako@gmail.com

1　はじめに

近年のインターネットの普及とウェブサービスの発展にともない，言語処理技術に対するニーズが増加している．それに対して，形態素解析や係り受け解析をはじめとする言語処理技術が実用化されるようになっている．しかし，学術論文においては，形態素解析が99%以上，係り受け解析が90%以上の解析精度として報告されているが，実際の場面ではそれほど高い精度の解析結果を残していないように思われる．

多くの自然科学において理論を実用化するときには，新たな問題が起きる．最も多い問題は観測器や解析器におけるノイズの問題である．例えば，音声認識における雑音の問題や画像認識におけるノイズなどがこれにあたる．言語処理においても，言語理論に沿わないノイズというべき言語表現が実際のテキストには存在し，その影響で実データでの解析精度が論文で報告に遠く及ばないのではないかと考えられる．

この言語的ノイズとは，どのようなものであるだろうか．まず，言語処理とは記号列の意味理解だと考えられる．これまでの言語処理は，入力が文法的な言語表現を前提に研究が行われていた．例えば，新聞記事の文章などである．しかし，実際のテキストは文法的に言語表現ではない表現も多数ある．例えば，顔文字「**(*ﾟ－ﾟ)v ｵﾊﾖ♪**（おはよう)」やギャル文字「**T=〃レヘ ﾑ (≠**（だいすき)」などの記号列が言語化するケース，「○○容疑者（３６）が」「**1.** はじめに」のような意味を持った記号列が挿入されるケース，アスキーアートのように記号列が言語的意味を持たないケースなどがある．このような表現は，これまでの自然言語処理においては形態素解析も構文解析も意味解析もできないノイズとして扱われてきたが，実際には人と人とのコミュニケーションにおいて非常に重要な役割を担っており，これらを上手く処理できることがこれからの言語処理に不可欠であると考えられる．

我々は，このように文法的には例外な言語表現や記号列を処理するための言語処理技術を不自然言語処理と呼ぶ．本論文では，不自然言語処理に焦点を当てたコンテスト「第1回不自然言語処理コンテスト」について報告する．

2　開催概要

2.1　応募要項・方法

第1回不自然言語処理コンテストは，バイドゥ株式会社が，研究者，及びエンジニアへのプレゼンスを高めるための活動の一環として開催された．開催に当たっては特設ページ[1]を用意し，Twitterアカウント（@baidu_unlp）を使って応募状況を伝える，質問に答えるなどの対応を行った．応募自体はメールを通じて行い，応募作品はそのメールに添付するか，公開したURLを記述する形とした．グランプリ，準グランプリには賞金を用意し，その賞金はそれぞれ10万円，3万円とした．さらにコンテスト授賞式を企画し，応募者にはその場でのLightningTalkを依頼した（任意）．そしてLightningTalk賞を設け，その賞金は1万円とした．またこれに合わせて「不自然言語」を題材とした壁紙やバイドゥの公開するIME向けの「2ちゃんねる辞書」などの関連データも同ページにて公開した．

2.2　お題

コンテストの題目は「不自然言語を使ったコミュニケーションを豊かにするサービスや作品，プログラムなどを作ること」であり，未発表の作品であればその形式や形態については自由であり，複数の作品をひとりで応募することも可能とした．またコンテスト開催と同時に『Baidu絵文字入りモバイルウェブコーパス』を公開したが，これについては必ず利用しなければならないものではなく，また他のデータやWeb API等を自由に使って良いものとした．本コーパスの詳細は，3節において述べる．

2.3　コンテスト授賞式

コンテストには26作品の応募があり，これを予め審査員（荒牧，竹迫，萩原，水野）で審査し，グランプリ，準グランプリを決定した．コンテスト授賞式は2010年7月25日にバイドゥ株式会社の会議室にて行われ，予め特設ページ上で募集が行われていた参加者とコンテストの応募者をあわせて約40名が参加，12のLightningTalkが行われた．グランプリとなったの

[1] http://www.baidu.jp/unlp

は「空目したことをつぶやく」という Twitter 上の流行をグラフによって可視化した「Soramegraph」，準グランプリは入力したテキストが類似した「誤字」に次々と置き換わっていく「誤字ェネレータ」が受賞した．LightningTalk 賞はすべての発表が終わったあとで会場出席者の拍手によって決定され,「文字の最初と最後だけ合っていれば，その間の文字の順序が異なっていても人は読める」という研究 [1] に基づいた掲示板の検索避けシステム「ケンブッリジ大学」が受賞した．なお，当日の様子についてはバイドゥ株式会社の公式ブログにて報告が公開されている[2]．

2.4 反響

コンテストの特設ページははてなブックマークにて 300 近いブックマーク数を集めた．また，開催時には #unlpcon という Twitter のハッシュタグを用意，コンテスト開催告知から約 250 の Tweet（ReTweet をのぞく）が記録されている．コンテスト授賞式では Ustream によるストリーム配信も行っている．現在もその録画は公開されており[3]，その総視聴数は約 450 である．

2.5 スイカ割り

エンジニア間での親睦を深めるために，コンテスト授賞式では参加者全員によるスイカ割り大会も行われた．スイカは小ぶりではあったが，大変甘いものが用意された．

2.6 リクルーティング効果

第 1 回の不自然言語処理コンテストはバイドゥ株式会社の単独開催であり，その目的にはバイドゥ株式会社のプレゼンス向上のほかに，ウェブ検索という「不自然言語」の集合体を相手にするにあたっての，優秀な研究者やエンジニアの獲得という目的もあった．その具体的な効果については非公開となっているが，一定の効果があった．

3 Baidu 絵文字入りモバイルウェブコーパス

本コンテストの開催に合わせて，筆者らは『Baidu 絵文字入りモバイルウェブコーパス』（以下,「本コーパス」と呼ぶ）を公開した．本コーパスは，モバイル Web 上の日本語形態素 n-gram ($1 \leq n \leq 5$) の統計情報であり，従来の Web 検索エンジンでは活用されてこなかった，モバイル Web に存在する絵文字を，通常の形態素と同等に扱った点が特徴である．

バイドゥモバイル検索向けに 2010 年 6 月までにクロールした Web ページからランダムサンプリングしたもののうち，絵文字の含まれるページのみを使用している．形態素 n-gram の情報はキャリア（携帯電話の通信事業者）別に提供しており，各ページがどのキャリア向けかは，絵文字データの表現方法，およびコード範囲より判定している．形態素解析には mecab-0.98 および mecab-ipadic-2.7.0- 20070801 を用いているが，絵文字トークン <EMOJI_XXX> は 1 形態素となるように分割して絵文字トークン <EMOJI_XXX> は 1 形態素となるように分割している．その他のコーパス構築時の技術的詳細（文字エンコーディング，本文抽出，文抽出，頻度カットオフ等）については，コーパス付属の ReadMe ファイル[4]を参照のこと．本コーパスの規模（異なり形態素 n-gram 数）は，表 1 の通りである．

表 1: Baidu 絵文字入りモバイルウェブコーパスの規模

異なり n-gram 数	docomo	au	softbank
$n = 1$	116,283	45,174	5,084
$n = 2$	1,663,344	252,911	8,648
$n = 3$	2,617,067	252,371	6,047
$n = 4$	2,123,263	158,151	4,453
$n = 5$	1,244,631	99,999	3,652

本コーパスを用いた作品としては，4. 応募作品において紹介したもの以外にも，統計情報を一般のコーパスと比較したもの[5]，文を自動生成するもの[6]，バイグラム情報を利用して絵文字を挿入するもの[7]などがある．

4 応募作品

4.1 応募作品傾向

応募作品全体の傾向としては，メッセージに絵文字を挿入するものや，文字・文を操作してより「不自然な」ことばに書き換えるものが多かった．開発言語はまちまちであるが，Perl, Ruby, Python などの軽量言語 (LL) が多い印象である．また，Web サービスとしての発表形態が一般的であった．

表 2 に，受賞作品の概要を示した．以下では，各受賞作品の詳細について述べる．

4.2 グランプリ「Soramegraph」

概要 Twitter 上で,「○○を××に空目した」というような，類似した単語を「空目」したことをつぶやくことがある．この関係をグラフ化して可視化するツールである (図 1) [8] [9]

[2] http://staffblog.baidu.jp/2010/07/26/unlp/
[3] http://www.ustream.tv/user/baidu_unlp/videos
[4] http://www.baidu.jp/corpus/mobile/readme.txt
[5] http://d.hatena.ne.jp/nokuno/20100720/1279636678
[6] http://d.hatena.ne.jp/n_shuyo/20100720/unlp
[7] http://www.slideshare.net/moaikids/lt-4833540
[8] http://aaatxt-gae.appspot.com/soramegraph
[9] 図 1 は http://aaatxt.blog57.fc2.com/blog-entry-65.html から引用

表 2: 受賞作品の概要

作品名	プログラミング言語	材料・コーパス	ベースとなる言語処理技術
Soramegraph	Java	Twitter	パターンによる関係抽出
誤字ェネレータ	Ruby	常用漢字の文字画像	画像類似度
感情のこもった 返答テンプレ生成君	Perl	Twitter メッセージ Baidu コーパス	ベクトル空間モデル，クラスタリング
ケンブッリジ大学	C#	なし	形態素解析

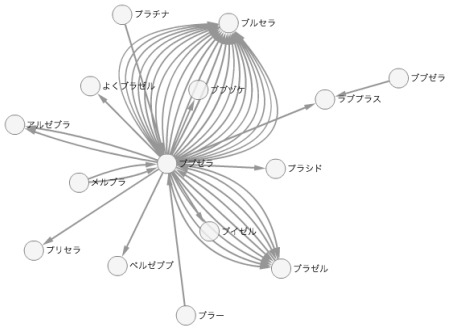

図 1: Soramegraph

制作動機　空目し易い紛らわしい単語を把握し，誤解を避けたり，あえて誤解を狙ったコミュニケーションを補助する．また，Tweet を可視化することにより，自分と感性の近い人を発見することもできる．

アプリケーション説明　リアルタイムの Twitter 解析結果と，蓄積したデータを解析し，「空目関係」にある単語を特定，グラフにして表示する．対応する Tweet を表示したり，キーワード検索を実行する機能もある．

用いたプログラミング言語　Java，Cytoscape Web, Google App Engine, Twiiter4J

材料・コーパス　Twitter のメッセージ

ベースとなる言語処理技術　「〜を〜に空目」といったパターンによる関係抽出，グラフの可視化

4.3　準グランプリ「誤字ェネレータ」

概要　文字列を入力すると，その一部が「誤字」すなわち類似した文字に置き換わるウェブアプリケーションである (図 2)[10]．

制作動機　誤字によって意味が喪失するさまを視覚化する．

[10]http://goji.polog.org/

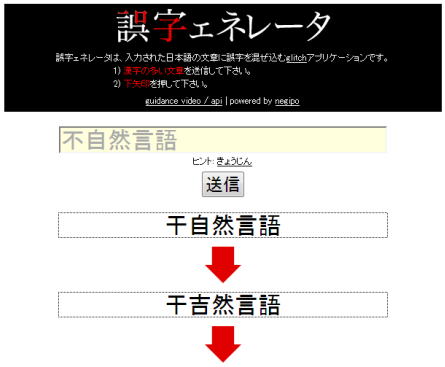

図 2: 誤字ェネレータ

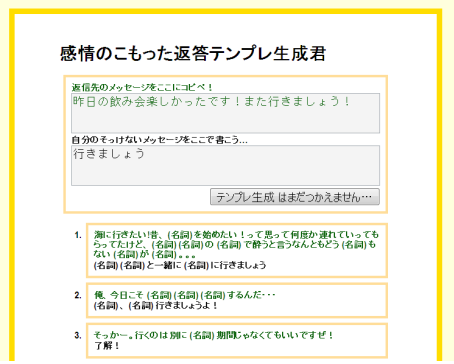

図 3: 感情のこもった返答テンプレ生成君

アプリケーション説明　文字列の変形を繰り返すことで，「不自然言語」から「干自然言語」そして「干吉然言語」というように次々と変化させていくことができる．誤字の選択は，それぞれの文字のラスタ画像を用いて計算した類似度をデータベース化し，それに基づいて行われている．

用いたプログラミング言語　Ruby

材料・コーパス　常用漢字 1945 文字の明朝体フォント

ベースとなる言語処理技術　ラスタ（ビットマップ）画像類似度

4.4　審査員特別賞「感情のこもった返答テンプレ生成君」

返信先のメッセージと返信先のメッセージと自分のそっけないメッセージを入力とすると，そっけなくな

こんちには みさなん おんげき ですか？　わしたは げんき です.
この ぶんょしう は いりぎす の ケブンッリジ だがいく の
けゅきんう の けっかにんんげ は もじ を にしんき する とき
その さしいょ と さいご の もさじえ あいてつれば
じばんゆん は めくちちやや でも ちんやと よめる という けゅきんう に
もづいとて わざと もじの じんばゅん を いかれえて あまりす.
どでうす？　ちんゃと よゃちめう でしょ？
ちんやと よためら はのんう よしろく

図 4:「ケンブッリジ大学」コピペ

いメッセージのテンプレを生成してくれるツールである (図 3)[11].

製作動機　テンションの高いメールを返すのが面倒である.

アプリケーション説明　メッセージ対をあらかじめクラスタリングしておき，入力のメッセージ対に対して類似したクラスタをランキングして出力する．テンプレートを出力するため，df の値が小さい形態素は削除する.

用いたプログラミング言語　Perl

材料・コーパス　Baidu コーパス, bayon, Twitter の Streaming API で取得した日本語 tweet 約 160,000 件

ベースとなる言語処理技術　ベクトル空間モデル，クラスタリング

4.5　LT 賞「ケンブッリジ大学」

概要　入力文字列を，人間には読めるが，検索エンジンには認識しづらい「ケンブッリジ大学難読化」画像に変換する.

作成動機　検索エンジン等に拾われたくない文章をブログや掲示板に投稿するため.

ケンブッリジ大学難読化とは　「人間は最初と最後の文字で単語を認識しているため，単語の途中の文字の順序が間違っていても，文章をストレス無く読める」という内容が，ケンブッリジ大学の研究の結果であるとして，インターネットで一時流行した (図 4).

アプリケーション説明　検索エンジンやクローラーに収集されたくないという意図で，略語や当て字を用いて文章を難読化したり，文章を画像化することが一部の掲示板等で行われているが，難読化や画像化は面倒なだけではなく，難読化した文章を人間が読み取れないという本末転倒な事態が生じることも多々ある．本アプリでは「ケンブッリジ大学難読化」，さらにそれを画像化することにより，システムでは読み取りづらいが，人間には問題無く読めるよう文章を加工し，検索エンジンやクローラーを気にせず，円滑なコミュニケーションを可能にする

用いたプログラミング言語　C#

材料・コーパス　材料・コーパス無し

ベースとなる言語処理技術　MeCab による形態素解析

4.6　応募作品の総評

言語処理というと，文法的な文章を入力として扱うことが多く，不自然な言語を扱うという本コンテストは，その取り組みそのものが非常にユニークで興味深い試みである．どのような作品が寄せられるか予想がつかない中，グランプリとなった Soramegraph をはじめ，用途に広がりを感じる斬新なサービス，夢のあるアプリケーションが多数集まった.

全体的に，絵文字を扱った作品が多く，従来，対象対象とされていた絵文字が高い関心を集めており，新しい言語現象となりつつあることが伺える．また，言語処理研究者，言語処理研究室の学生でない方の参加が大半で多く，言語処理自体の知名度が高まっていることが実感された．IME など言語処理アプリケーションの普及や，入門書[12]の発売などによるアウトリーチ活動のたまものだと思われる.

5　おわりに

本稿では，文法的には例外な言語表現や記号列を処理するための言語処理技術である「不自然言語処理」の概念を導入し，この不自然言語処理に焦点を当てたコンテスト「第 1 回不自然言語処理コンテスト」について報告した．従来処理対象外とされていた絵文字等を扱った作品をはじめ多数の応募があり，言語処理の新しい対象としての「不自然言語」の広がりを感じさせる反響であった．本年次大会のテーマセッションをはじめ，今後の「不自然言語処理」の発展が,「コミュニケーションを豊かにする」というコンテストの本来の目標に貢献できることを願っている.

参考文献

[1] Richard Shillcock and Padraic Monaghan. An anatomical perspective on sublexical units: The influence of the split fovea. *University of Edinburgh*, 2003.

[11] http://tokuota.ddo.jp/extext/

[12] Steven Bird, Ewan Klein, Edward Loper 著, 萩原 正人, 中山 敬広, 水野 貴明 訳『入門 自然言語処理』オライリージャパン, 2010.