

言語処理学会へ  
遊びに行ってきたよ！

～不自然言語処理へのお誘い～

# 自己紹介

- 金融機関で金融工学の研究者
- 大学院でテキストマイニングを学ぶ
- 言語処理を用いてコミュニケーションの活性化を図りたい！
- toilet\_lunch, todesking達とすき焼きしてたら、いつの間にかテキストマイニング勉強会発足してた

# 本発表の目的

1. 学会で得た最新の情報の中で、実務に使えるような内容・レベルのものを紹介
  - ・ 新しいサービス提案の切っ掛けに
  - ・ 実践のプロセスを学ぶ
2. 不自然言語処理へのお誘い

# 学会へ遊びに行こう！

- 専門の学生か、GとかYとかIとか、ごく一部の企業に所属していないと、最新技術動向は掴めない
- 学会に行けば、最新の情報がわんさか手に入る！
- すごい人達と知り合いになって、仕事して貰ったり仕事貰ったりする！
- 自分の疑問点や手法について議論できる！
- 学会参加費はそんなに高くないよ！
- そうは言っても中々敷居が高く感じられるので、まずはテキストマイニングマスター達のブログでキャッチアップしよう

# 必ずチェックすべき10のブログ

1. コーパスいぢり (langstat)
2. あらびき日記 (a\_bicky)
3. 睡眠不足？ (sleepy\_yoshi)
4. EchizenBlog-Zwei (echizen\_tm)
5. Overlasting::Life (overlast)
6. おとうさんの解析日記 (isseing333)
7. はやしのブログ (phosphor\_m)
8. nokunoの日記 (nokuno)
9. ぬいぐるみライフ(仮) (mickey24)
10. Mi manca qualche giovedì (shuyo)

# 発表論文目次

1. Webからの飲食店舗の評判情報抽出
2. Wikipediaのカテゴリ階層を利用したTwitterユーザのカテゴリライズ
3. 大規模Web情報分析のための分析対象ページの段階的選択
4. マイクロブログの分析に基づくユーザの嗜好とタイミングを考慮した情報推薦手法の提案
5. 不自然言語処理コンテスト第一回開催報告
6. 文頭固定法による効率的な回文生成
7. 顔文字情報と分の評価表現の関連性についての一考察

# Webからの飲食店舗の評判情報抽出

高尾美代子他

- 目的
  - 適当にブログ等をクローリングしても評判情報を得難い
  - 効率的な評判情報抽出の手法を提案しよう！

# 既存の評判情報抽出とその問題点

- 手順
  1. 店舗名を含むテキストを取得する
  2. テキストから評価部分を抽出
  3. 抽出した評価情報から店舗の評判を得る
- 問題点
  - 評価部分を抽出することが難しい
  - 全テキスト参照すると評価と関係無いノイズが増える
  - 逆に抽出部分が狭すぎると、評価を得られない
- **上手く評価部分のテキストだけ抽出したい！**

# 本稿の提案

- 評判情報を得やすいページとそうでないページに分類することで、より良い評判情報抽出が可能になる
- 評判情報を得やすいページに分析対象を絞ろう
- テキストのどの部分を参照すれば、評判情報を得やすいのかを調べよう

# 実験の手法と手順

1. 共起表現抽出範囲, 素性選択をパラメタとする
2. 各パラメタごとに、対象ページが評判情報を含むか否かを判定した分類精度を出す
  - Yahoo!検索APIを用い、評判情報を含む/含まないページ100件ずつ用意
  - 分析ツール: SVMLight
3. 各パラメタの抽出結果を比較し、最適な組み合わせを得る

# 効果的な共起表現抽出範囲

- なぜ評判分析で共起表現を抽出するか
  - 評価を表す単語は店舗名の周辺に集中しているから
- 抽出範囲18パターン
  - 店舗名の前方/後方/前後の3パターン
  - 2~7単語の6パターン
- 結果
  - 平均精度: 後方83.3%, 前後60%, 前方57%
  - 評価は店舗名の後方に集中する
  - 共起語数は4~6単語が最適
  - 3以下は評判情報を含み難く、7以上はノイズが多い

# 効果的な素性パターン

- 品詞パターン

1. 動詞＋形容詞
2. 動詞＋助動詞
3. 形容詞＋助動詞
4. 形容詞＋助詞＋動詞
5. 名詞＋助詞＋形容詞
6. 名詞＋助詞＋動詞
7. 形態素nグラム
8. 単語nグラム

- 結果は店舗によってまちまち

- 平均して7, 8の精度が比較的高い

# まとめ

- 評判分析をするには、適切な評価情報を含んだページの取得が必要
- 評価は店舗名の後方4~6単語に集中する
- 評判分析をする際、本研究を参考にして評価情報を取得してみよう！

# Wikipediaのカテゴリ階層を利用した Twitterユーザのカテゴリライズ

放地宏佳他

## 背景

- Twitterのカテゴリは8種類と少なすぎる
- 情報抽出する際、適切なカテゴリライズは有用

# 提案手法

- 前提
  - Wikipediaのカテゴリライズを使おう
  - 適切なカテゴリライズは日々のメンテナンスが必要であり、高コスト。Wikipediaのカテゴリライズを流用して自動化出来れば非常に有用である
- 手順
  - ツイートから各ユーザーの特徴語抽出
  - Wikipediaから特徴カテゴリ抽出

# 特徴語とは

- ユーザが用いる頻度高い単語 ≠ ユーザの特徴語
- 頻度の高い単語は皆も使っているモノが多い
- 特徴語とは、比較的他と比べてそのユーザだけが用いる頻度高い単語

# 特徴語抽出

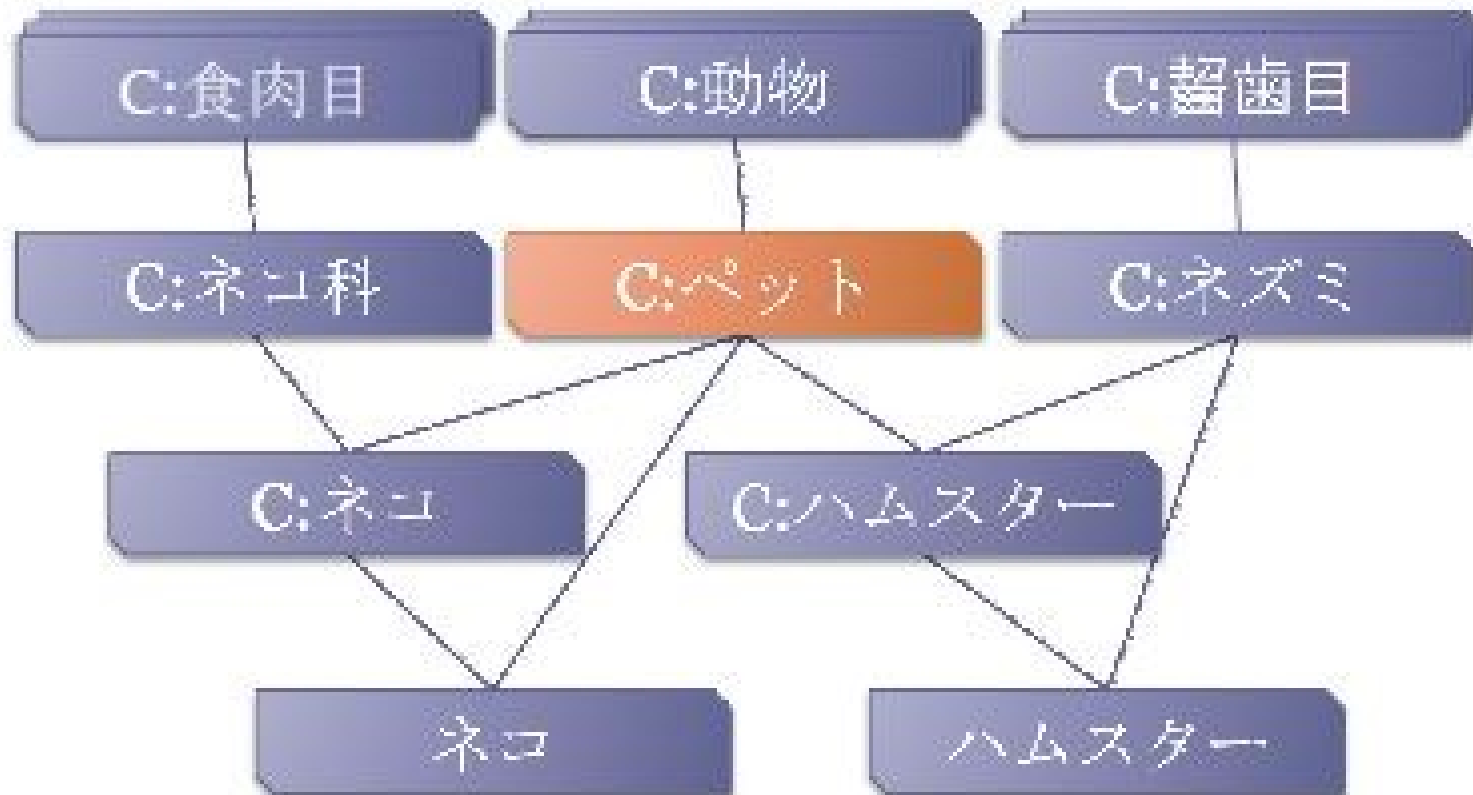
1. 各ツイートの正規化 (@username, RT・QT文, URL, ハッシュタグの除去)
2. Wikipediaの記事名と一致する語を抽出し、出現回数とする
3. 2で得られた語をツイートに含むユーザ総数を出現頻度とする
4. 出現回数 $>2$ ,  $1/\text{出現頻度} > 0.5\%$ を満たす語を特徴語とする

# 特徴カテゴリ集合抽出

- 各特徴語の最上位カテゴリまでのパス集合を取得
- 全特徴語のパス集合から共通カテゴリを取得
- 共通カテゴリを割り当てられたユーザの総数を出  
現頻度とする
- 最上位カテゴリから共通カテゴリまでの距離をパス  
の大きさとする
- パスの大きさ/同一共通カテゴリの数 $>2$ ,  $1/\text{出現頻度}$   
 $>0.005$ を満たす共通カテゴリを特徴カテゴリとす  
る

# パス集合

特徴語がネコとハムスターの場合の共通カテゴリ



# 評価実験

- ランダムに選択した20ユーザ、各ユーザの最大発言数2000とする
- 特徴カテゴリがそのユーザのカテゴリとして適切か人手で判断
- 実験結果

取得された特徴カテゴリ数	88
適切である特徴カテゴリ数	43
不適切である特徴カテゴリ数	26
ユーザ分類において不適切である特徴カテゴリ数	19

# 結果の考察

- 「スポーツ」「コンピュータ」などは直感的なツイートが多くわかりやすい
- 「物理」「心理学」など専門用語が日常用語と被るカテゴリは判別しづらい
  - 「反射」「振動」を多用する人は音響の人かも？
- reply, RT, 実況は特徴が掴みづらい

# まとめ

- カテゴリー化を行う場合、replyやRT、実況などのツイートを削除する必要がある
- 専門用語と日常用語を切り分ける手法が必要
- 自動化が適用できるカテゴリとそうでないカテゴリの選別が必要

# 大規模Web情報分析のための 分析対象ページの段階的選択

赤峯享他

- 目的と背景
  - 情報分析の処理は重いため、処理をかける前に不要なページを対象から外したい
  - Webには低品質のページが多い
  - 通常の検索では検索結果上位の高品質なページしか見ないためあまり意識されないが、クローラを回すとゴミばかり集めてしまう

# 選択の方針: 質の高いページとは

- テキスト情報が豊富なページ
  - 人気のあるページ≠テキスト情報が豊富なページ
  - 絵画・動画サイトではテキスト情報少ない
  - ページランクの高いページとテキストマイニングにテキスするページは異なる
- 多様な発信者/サイトを含むページ集合

# ページの選択

- フィルタリングでスパム、ミラーページを対象から除外
- ページランクや高品質ページに出やすい特定単語の出現頻度などの属性を用いた重み付きサンプリング
- サイト単位でページの品質を考える。同一サイトのページの品質は似ているため、低品質なページを含むサイトを丸ごと対象から除外

# ページ選択に利用する属性

ページ中のテキスト内容	テキスト量	文の数・長さ・密度
	文体	助動詞, 感動詞, 終助詞, 絵文字 の種別と出現数
	専門性 (名詞)	病名, 専門用語の出現数
	具体性 (固有名詞)	組織名, 人名の出現数
	高品質ページに出やすい単語	「検証」, 「証明」等
	低品質ページに出やすい単語	「死ね」, 「おまえ」等
	アダルトページに出やすい単語	
	高品質ページに出やすい構文を作る単語	意見, 原因・理由, 比較
	ページの種別	ニュース, ブログ, 商品販売, リンク集
ページ中の情報の有無	広告量	アフィリエイトサイトへのリンク数
	連絡先	住所, 電話番号, メールアドレス の有無
	プライバシーポリシーの有無	
メタ情報	ページランク	
	OutLink の数	
	ページのサイズ	
	更新日	現在の時間からの差
	URL	階層, 長さ, クエリ

# まとめ

- Webから収集した10億ページを、先程のフィルタリングなどにかけて1億ページまで分析対象を絞ることに成功した
- ランダムサンプリングしたものより分析精度は高い

# マイクロブログの分析に基づく ユーザの嗜好とタイミングを考慮した 情報推薦手法の提案

向井 友宏他

- 目的
  - twitterのリアルタイム性を利用し、ユーザに最適なタイミングで情報推薦を行いたい

# 提案手法

- 各ユーザのRTの名詞からユーザのプロファイルを作成する
- プロファイルを用いてクラスタリングを行う
  - Wikipediaのカテゴリ情報を利用し、類似した嗜好のユーザをクラスタリングする
  - {サッカー|フットサル}文字列は違うが嗜好は似ている
- 最適なタイミング発見のため、バーストを用いる

# バーストとは

- 時系列における投稿数の急激な変化
- バースト判定値Bの評価式

$$B = \frac{N}{\sqrt{\bar{A}}} \cdot \frac{N-A}{N+A} \quad \dots(1)$$

N : その区間におけるつぶやき数

A : 直前 X 区間のつぶやき数の平均

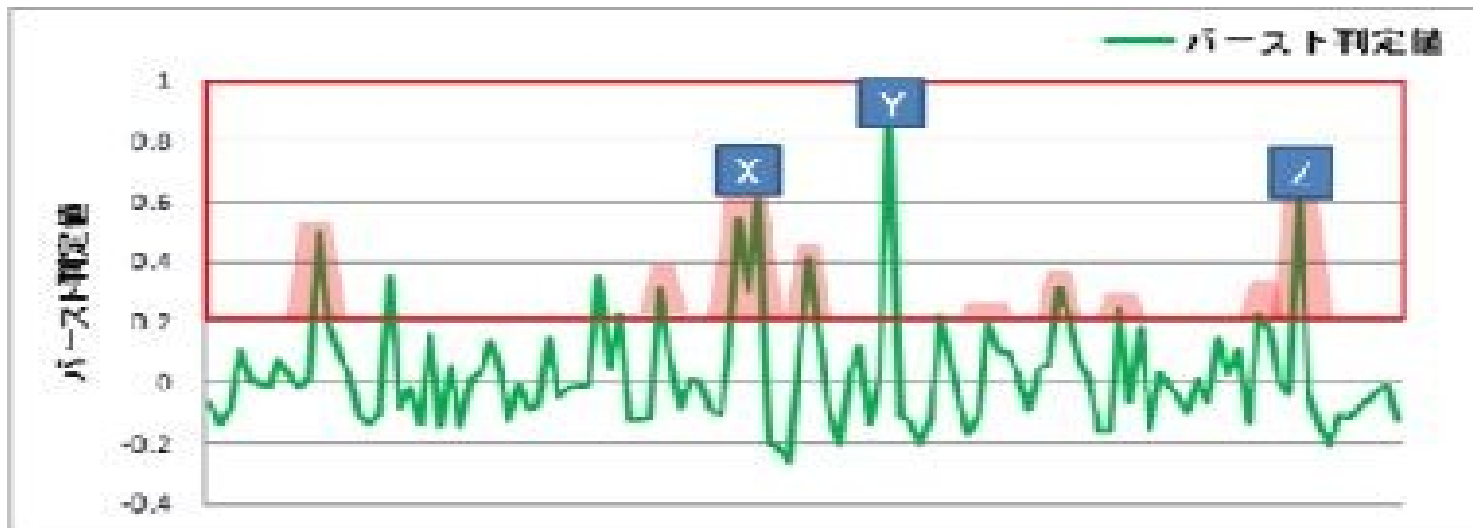
$\bar{A}$  : 直前 Y 区間のつぶやき数の平均

# 評価実験準備

- 2010年度日本シリーズのロッテファン524人20万以上のツイートを収集。11/7分を訓練に利用
- 極性評価の準備
  - 極性評価用の手がかり語を手手で収集
  - P:ポジティブ語数、N:ネガティブ語数とする
  - ポジティブバースト:  $P/(P+N) > 0.7$
  - ネガティブバースト:  $N/(P+N) > 0.7$

# バーストの検出

- ヒューリスティックに以下のパラメタを利用
  - $X=3, Y=30$
  - 閾値 $\alpha=0.2$
- バースト区間
  - 判定値 $B$ が $\alpha$ を超え、再び $\alpha$ を下回るまでの区間



# 商品とユーザとのマッチング

- 楽天商品データ1000件の各商品説明から特徴語を抽出
- 各商品の特徴語とユーザカテゴリをマッチング
  - スポーツクラスタにはサッカー商品を薦めるなど

# まとめ

- 最適な商品を推薦するだけではなく、バーストを利用して、最適な推薦のタイミングまで考えよう！
- 結果は正直かなり悪かった
  - RT数が少なくて学習が不十分
  - カテゴリに即した商品がないことも
  - 噺・落語クラスタに何薦めればいいのか？
  - Wikipediaのカテゴリと楽天のカテゴリのミスマッチ

# 総評・雑感

- Wikipediaを利用してコーパス作成、カテゴリライズするのが流行している
- twitter特有のソーシャル性、即時性を使おう
- これらは各データに階層構造やタグなど、高品質なメタデータが人手で付与されている
- しかし、実際の利用は困難っぽい。BOWは無理。ゼロ照応解析、共参照解析、談話解析等が必要
- FOBOSやpLSAを学部生が使ってる…

# 不自然言語処理とは

- そもそも「**自然**言語処理」の言う自然とは？
- 「MeCabで分析できる言語＝自然言語」
- そんなもの自然言語じゃない！
- 実際の言語は誤字、脱字、略字、隠語、顔文字、絵文字、AA、数式・化学式、等々が溢れている！
- 従来のテキストマイニングでは、顔文字などはゴミとして除去していた
- 顔文字にこそ書き手の思いが宿っているのでは？
- 顔文字等を有効活用するのが次世代マイニング

# 不自然言語処理コンテスト

- baiduの「不自然言語」専門の言語処理コンテスト
- なぜか発生するスイカ割り
- 参加してLT賞頂きました
- コンテスト受賞作と言語処理学会の不自然言語セッションで発表された論文を紹介します
- 不自然言語処理を楽しもう！

# Soramegraph

- 概要

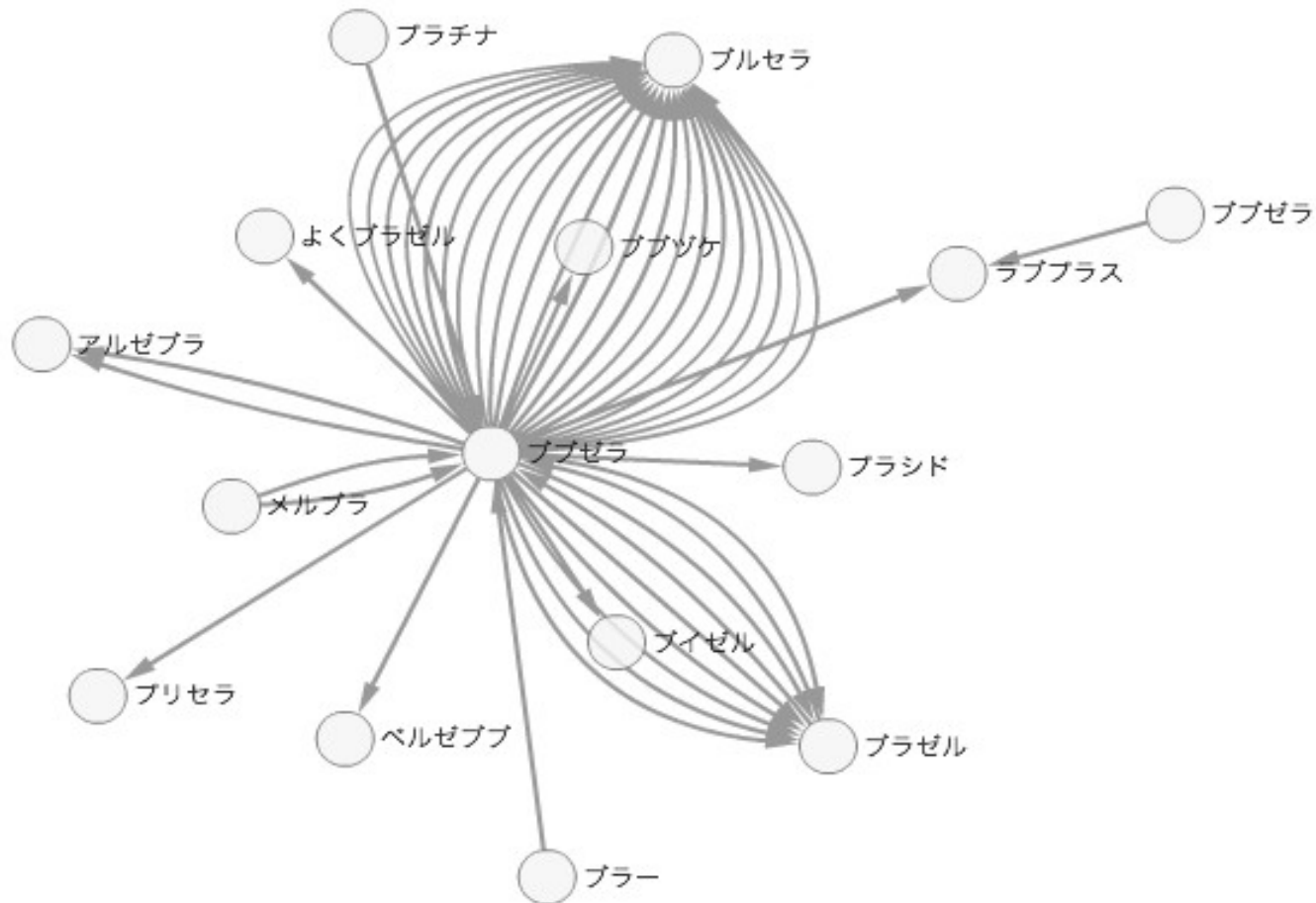
- Twitter上で、「〇〇を××に空目した」というような、類似した単語を「空目」したことをつぶやくことがある。この関係をグラフ化して可視化するツール

- 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

# デモ

- <http://aaatxt-gae.appspot.com/soramegraph>



# 誤字エネレータ

- 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

- 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

# デモ

- <http://goji.polog.org/>

**誤字エネレータ**

誤字エネレータは、入力された日本語の文章に誤字を混ぜ込むglitchアプリケーションです。

- 1) 漢字の多い文章を送信して下さい。
- 2) 下矢印を押して下さい。

[guidance video](#) / [api](#) | powered by [neepo](#)



# 感情のこもった返答テンプレ生成君

- 概要

- 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、そっけなくないメッセージのテンプレを生成してくれるツール

- 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.

# デモ

- <http://tokuota.ddo.ip/extext/>

## 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をここにコピペ！

昨日の飲み会楽しかったです！また行きましょう！

自分のそっけないメッセージをここで書こう...

行きましょう

テンプレ生成 はまだつかえません...

1. 海に行きたい!昔、(名詞)を始めたい!って思って何度か連れていってもらってたけど、(名詞)(名詞)の(名詞)で酔うと言うなんともどう(名詞)もない(名詞)が(名詞)。。。 (名詞)(名詞)と一緒に(名詞)に行きましょう
2. 俺、今日こそ(名詞)(名詞)(名詞)するんだ... (名詞)、(名詞)行きましょうよ!
3. そっかー。行くのは別に(名詞)期間じゃなくてもいいですぜ! 了解!

# ケンブリッジ大学

- 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「ケンブリッジ大学難読化」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.

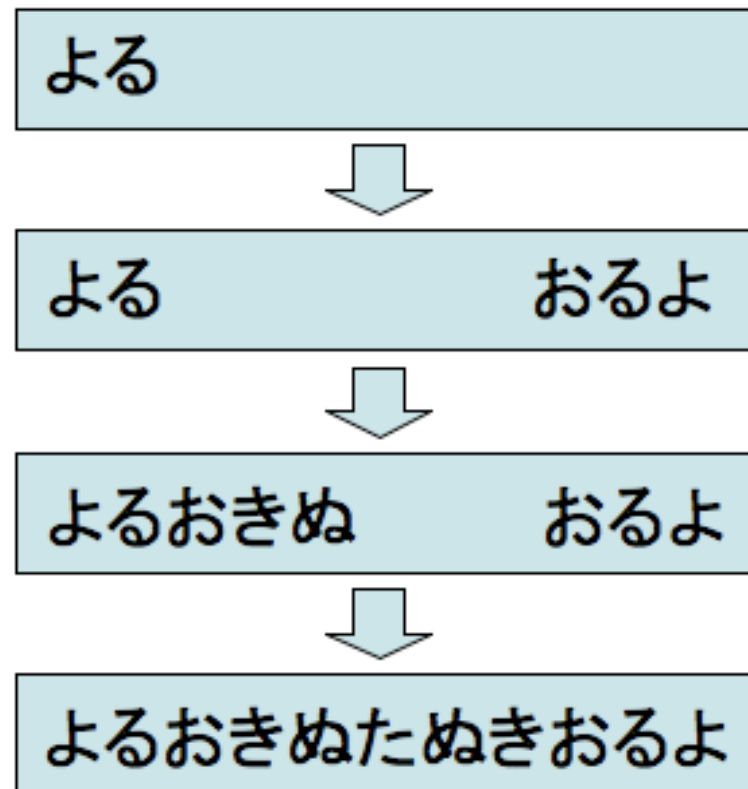
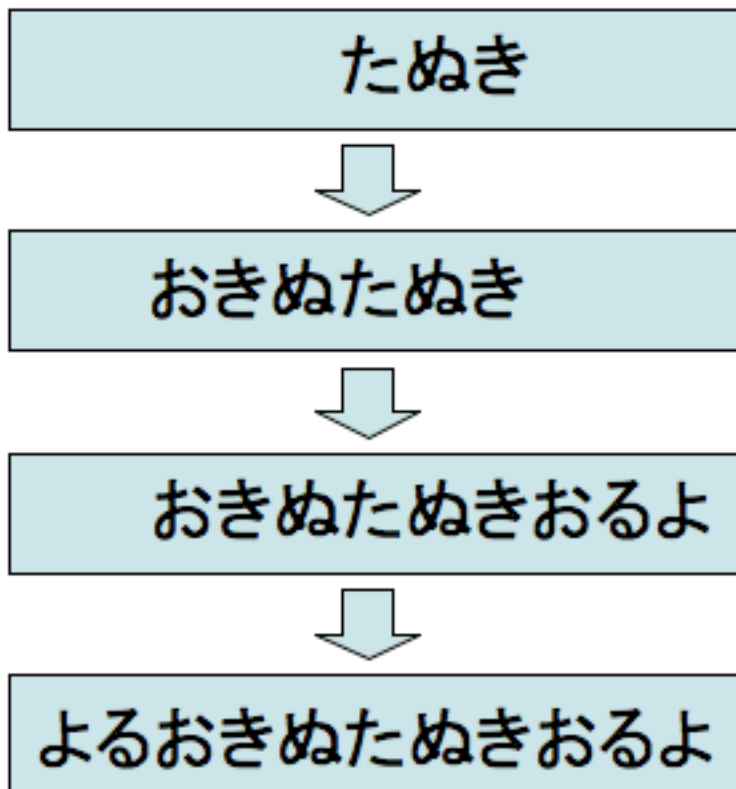
# ケンブリッジ大学コピペ

こんにちはみさなんおんげきですか？ わしたはげんきです。  
このぶんようはいりぎすのケブンツリジだがいくの  
けゆきんうのけっかにんんげはもじをにしんきするとき  
そのさしいよ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ゃやでもちんやとよめるというけゆきんう  
にもづいとてわざともじのじんばゆんをいかれえてあまりす。  
どでうす？ ちんやとよやちめうでしょ？

# 文頭固定法による効率的な回文生成

鈴木啓輔他

- 回文候補生成法：折り返し固定法と文頭固定法



# 速度比較実験とその考察

文節数	折り返し固定法	文頭固定法
3	21:41	0:42
4	198日 20:34	17日 14:10

- シード文節から出現する初期状態数が少ない
- 不足文字列の短い初期状態が出現しにくい
- 回文を使って面白いキャッチコピーを作ろう！

# 顔文字情報と文の評価表現の 関連性についての一考察

村上浩司他

1. 顔文字は周辺言語的要素を持つ
2. 顔文字単体の極性だけではなく、文脈把握が大切
3. (^\_^;), (;;)などは回答者によって快・不快バラバラ
4. 極性が異なるのに同じ顔文字が使われる事も
5. クラス分類ではなく、複数の感情軸を併せ持つ
6. 自身は意味を持たず、強調、緩衝材としての顔文字利用
  - 飲み会来るなよ～(^\_^)←冗談だと示している

# もっと不自然言語で遊ぼう！

- どんなとき不自然言語を使う？
  - 仲の良い人同士だと砕けた表現や隠語使いやすい
  - 他の人より頻繁に不自然言語を用いて会話する相手＝仲が良いのでは？ソーシャルネットワーク抽出出来る
  - 不自然言語の利用度合いが親密さを表すかも
- 顔文字は非言語的な情報まで伝達出来るかも
- 誤字・脱字から精神状態などを読み取れるかも
- 誤った語の使い方から年齢等が推定できるかも
- やってみよう！！！！

# 終わりに: 学会での関根先生の言葉

(楽天 & ニューヨーク大学)

- 不自然言語処理こそ真の自然言語処理であり、超自然言語処理と改名すべき！
- 10年前の技術が今も楽天で有効活用されてる、学会で盛り上がったネタなんて使われない。TF-IDFとかまだまだ現役。いかに高度な技術使うかより、いかにノイズを削減するかの工夫が必要
- すごい研究をしようとするのではなく、事業に役に立つ研究をしよう