ver.1.1

乱択アルゴリズムを使った『国語研日本語ウェブコーパス』からのUniDic新規語彙素候補の自動抽出

コーパス開発センター

特任助教　岡　照晃

本発表の概要

① 国語研日本語ウェブコーパス（NWJC）から
② UniDic DBにまだ登録されていない短単位の
③ “候補”を
④ 自動的に列挙する
手法の提案と結果

国語研日本語ウェブコーパス（NWJC）

- 『国語研日本語ウェブコーパス』はウェブを母集団として100 億語規模を目標として構築した日本語コーパスです。
- ウェブ（WWW）上の日本語テキストを利用して100億語を超える規模の現代日本語コーパスを構築することによって、稀言語現象の言語学的、心理学的および情報処理的視点からの究明の可能性を開くことを目的としています。
- 具体的な応用として、言語研究のための用例収集、日本語使用実態の定量的な把握などを想定しています。

実際のサイズ(2014-4Q)

URL数	83,992,556
延べ文数	3,885,889,575
異なり文数	1,463,142,939
総文字数（異なり文）	33,226,333,292

アノテーションは全自動

- 形態素解析
  - 形態素解析器 MeCab (0.996) + 解析用UniDic (2.1.2)
  - 総短単位数：25,836,947,421
- 係り受け解析
  - 係り受け解析器 CaboCha (0.69) + UniDic 主辞規則

- アノテーションの人手修正を行なっていない

- UniDic DBに載っていない未知語を含んでいる

UniDic DBと大納言を使った短単位アノテーション

UniDicデータベース

書字形(出現形)	発音形(出現形)	語形(出現形)	品詞	語彙素	語彙素読み	類	語種	…
すもも	スモモ	スモモ	名詞-普通名詞-一般	李	スモモ	体	和語	
もも	モモ	モモ	名詞-普通名詞-一般	桃	モモ	体	和語	
も	モ	モ	助詞-係助詞	も	モ	係助	和語	

コーパスデータベース
(文字列テーブル)

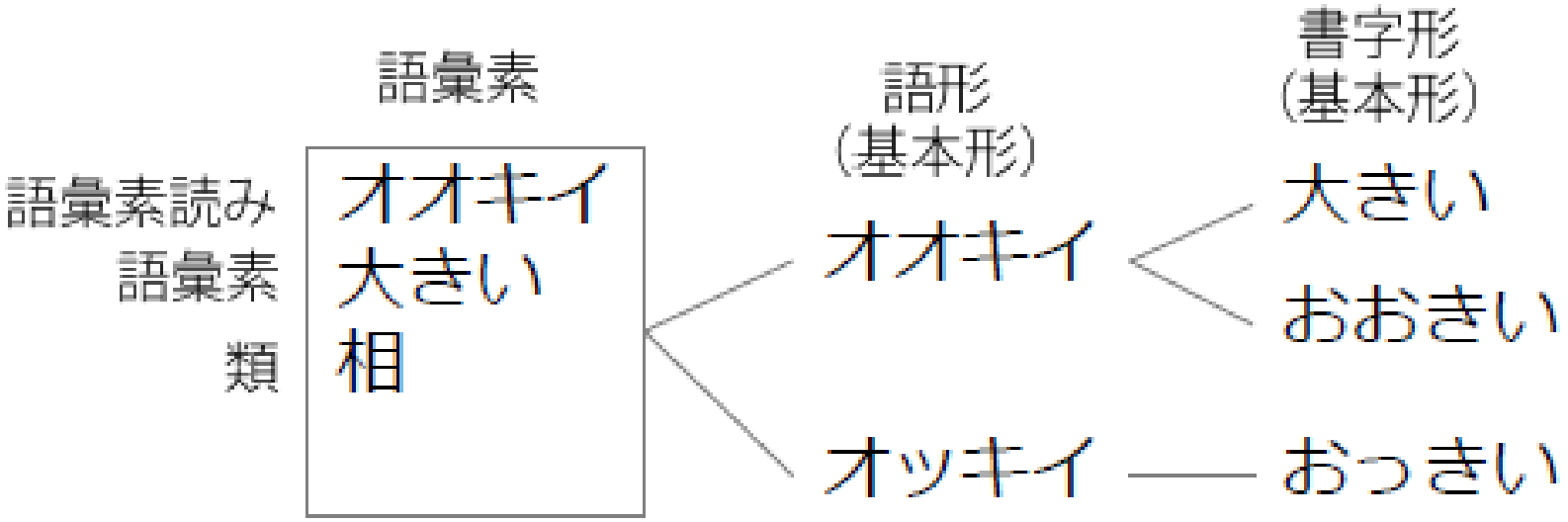
order	文字
10	す
20	も
30	も
40	も
50	も
60	も
70	も
80	も
90	も

- UniDic DBは手元の辞書的に利用し、辞書とコーパスを紐づけていく
- コーパス整備時にDBに未登録の短単位は、随時DBに追加する　←NWJCでは行なっていない

本発表の概要（再掲）

① 国語研日本語ウェブコーパス（NWJC）から
② UniDic DBにまだ登録されていない短単位の
③ “候補”を
④ 自動的に列挙する
手法の提案

UniDicの階層的見出し構造



未知語と新語

表 1　形態素解析における未知語の分類

未知語のタイプ		例
既知形態素からの派生	表記ゆれ	素晴しい（素晴らしい）
	連濁による濁音化	（堀り）ごたつ（こたつ）
	長音記号による置換	おはよー
	小書き文字による置換	ぁなた
	記号による置換	うれ∪い（うれしい）
	長音記号の挿入	冷たーーーい
	小書き文字の挿入	冷たぁぁぁい
	母音字の挿入	冷たあああい
	促音文字挿入	すっっっごく
既知形態素からの派生以外	反復型オノマトペ	ほいほい
	非反復型オノマトペ	ぺっちゃり
	感動詞	いやっほー
	新語・低頻度語	tsuda る，除染
	固有名詞	ツイッター

下線は本論文で扱う対象であることを示す.

先行研究：新語（未知語）の自動抽出

形態素解析で「未知語」と
判断された部分を取り出す

基本的に、字種に基づく未知語処理（MeCab, [村脇+, 10]）
例外：　[東+, 06]

生テキスト（文字列）から部分文字列の出現頻度に基づいて候補を列挙

当該部分文字列の左右の部分文字列の分布に基づく
（[長尾+, 93],[森+, 98], [萩原+, 11]）

今回は、下のアプローチをとります。

形態素解析器の誤解析

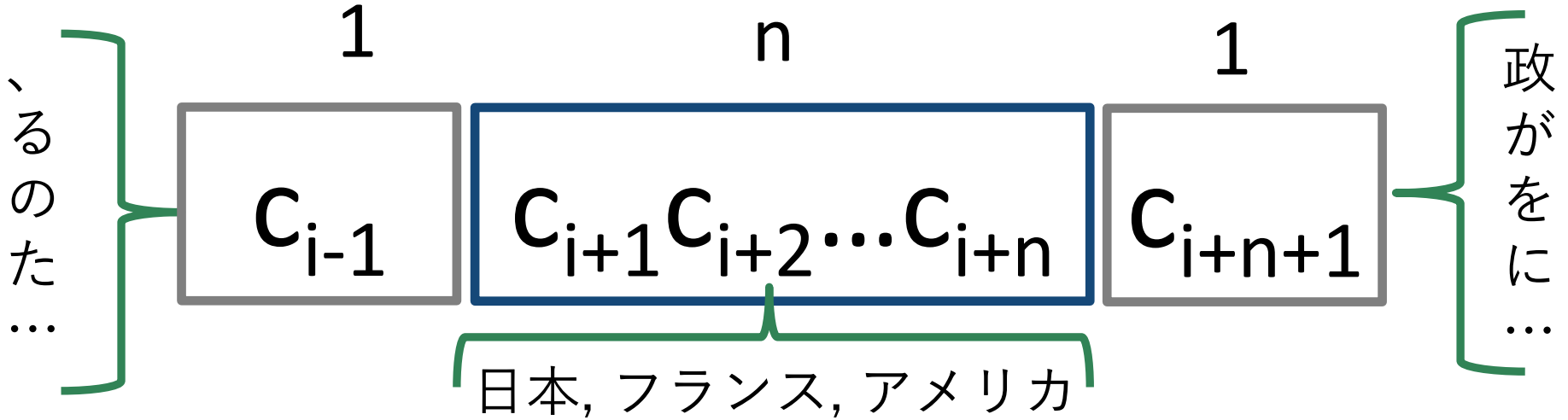
UniDicの場合は特に既知語に過分割し、未知語にならない。

書字形（=表層形）	語彙素	語彙素読み	品詞	大分類	中分類	小分類	活用型	活用形	発音形出現形	仮名形出現形	語種
Ｙｏｕ	ユー-you	ユー	名詞-普通名詞-一般	名詞	普通名詞	一般			ユー	ユー	外
ｔｕ	チュ-tu	チュ	名詞-普通名詞-一般	名詞	普通名詞	一般			チュ	チュ	外
ｂｅ	ビー-be	ビー	名詞-普通名詞-一般	名詞	普通名詞	一般			ビー	ビー	外

生テキスト（文字列）から部分文字列の出現頻度に基づいて候補を列挙

- [長尾+, 93]
  - 大規模日本語テキストのnグラム統計の作り方と語句の自動抽出
    - 自然言語処理, Vol. 96, No.1.
- [森+, 98]
  - nグラム統計によるコーパスからの未知語抽出
    - 情報処理学会論文誌, Vol.39, No 7.
- [萩原+, 11]
  - グラフカーネルを用いた非分かち書き文からの漸次的知識獲得
    - 人工知能学会論文誌, vol.26, No.3.

大規模日本語テキストのnグラム統計の作り方と語句の自動抽出 [長尾+, 93]



n	2	3	4	5	6	7	…
しきい値の組A	10	8	6	4	3	3	…
しきい値の組B	18	13	9	6	5	5	…
しきい値の組C	27	19	13	9	7	7	…

高頻度の部分文字列をその長さnに応じて左右の１文字のタイプ数で、しきい値をこえたものを取り出す。

基本的には、いずれも両サイドの部分文字列の統計情報を使う

- [長尾+, 93]
  - 先のとおり。
- [森+, 98]
  - 左右の文字のタイプ数でなく、左右各１文字が現れる条件付き確率の分布を使用。
  - その分布が特定の品詞の分布に近しければ、抽出する。
- [萩原+, 11]
  - ブートストラップ法。seedとして与えた部分文字列と出現の仕方が類似の部分文字列を集める。
  - 左右の部分文字数を1文字に限らない（n=1~8）。
  - 左右部分文字列の確からしさ（抽出対象部分文字列との自己相互情報量pmi）から抽出する部分文字列の単語らしさを計算する。

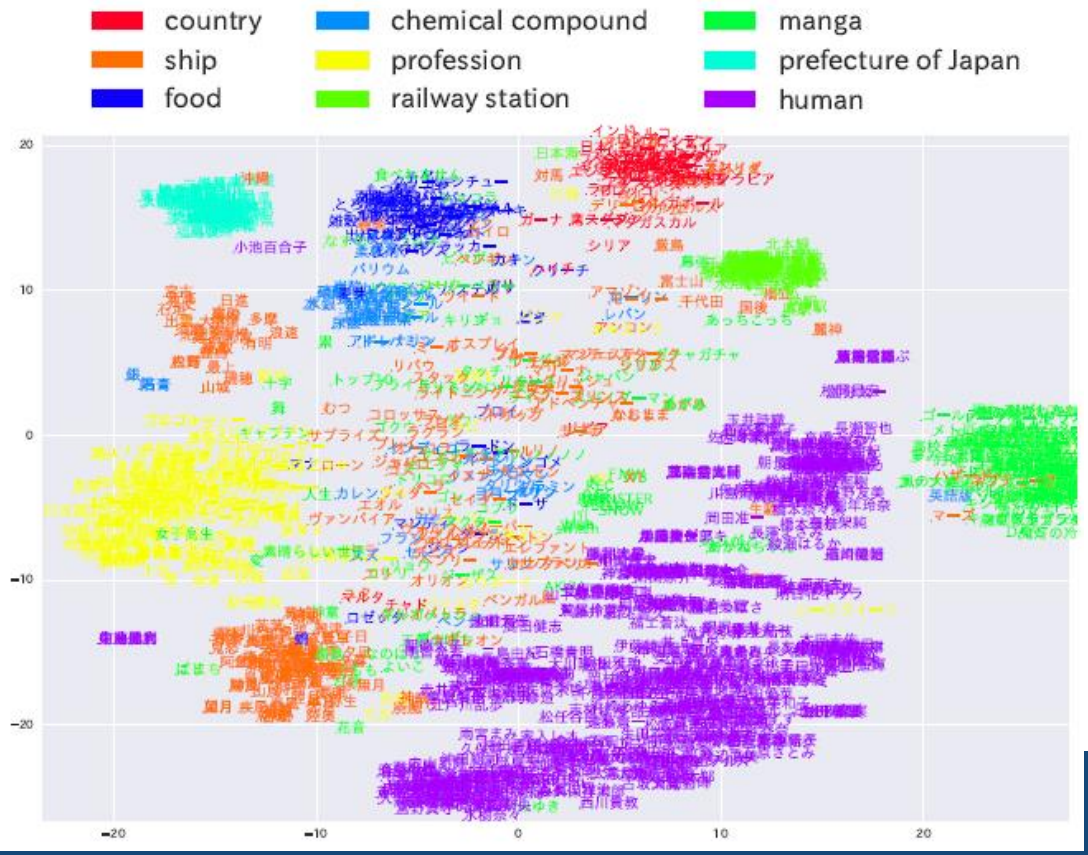
提案手法

- [押切+, 17] 単語分割を経由しない単語埋め込み
  - のアレンジ

① NWJCから長さ1~nすべての部分文字列を抽出
② 抽出した各部分文字列に対し、左連接・右連接する部分文字それぞれカウントし、左連接行列C_L・右連接行列C_Rを作る。
③ $C = [C_L, C_R]$を特異値分解
④ 左特異ベクトルを各部分文字列の分散表現とする。
⑤ UniDic DB内の登録済みの短単位と照合し、特定の品詞のみがついた部分文字列を正例、指定した以外の品詞がついた部分文字列を負例とし、分類器を構築。
⑥ UniDic DBに未登録の部分文字列を分類器（knn-classifier）にかける

単語分割を経由しない単語埋め込み（1）

- 分かち書きされていない文の集合から、分かち書きをしないまま、単語の分散表現を獲得する



分散表現

ある人がQiitaで公開していた「japan」という単語の分散表現

- [-0.17399372 0.138354 0.18780831 -0.09954771 -0.05048304 0.140431 -0.08839419 0.0392667 0.267914 -0.05268065 -0.04712765 0.09693304 -0.03826345 -0.11237499 -0.12375604 0.15184014 0.09791548 -0.0411933 -0.26620147 -0.14839527 -0.07404629 0.14330374 -0.15179957 0.00764518 0.01670248 0.15400286 0.03410995 -0.32461527 0.50180262 0.29173616 0.17549005 -0.13509558 -0.20063001 0.50294453 0.11713456 -0.1423867 -0.17336504 0.09798998 -0.22718145 -0.18548743 -0.08841871 -0.10192692 0.15840843 -0.12143259 0.14727007 0.2040498 0.30346033 -0.05397578 …]

単語分割を経由しない単語埋め込み（2）

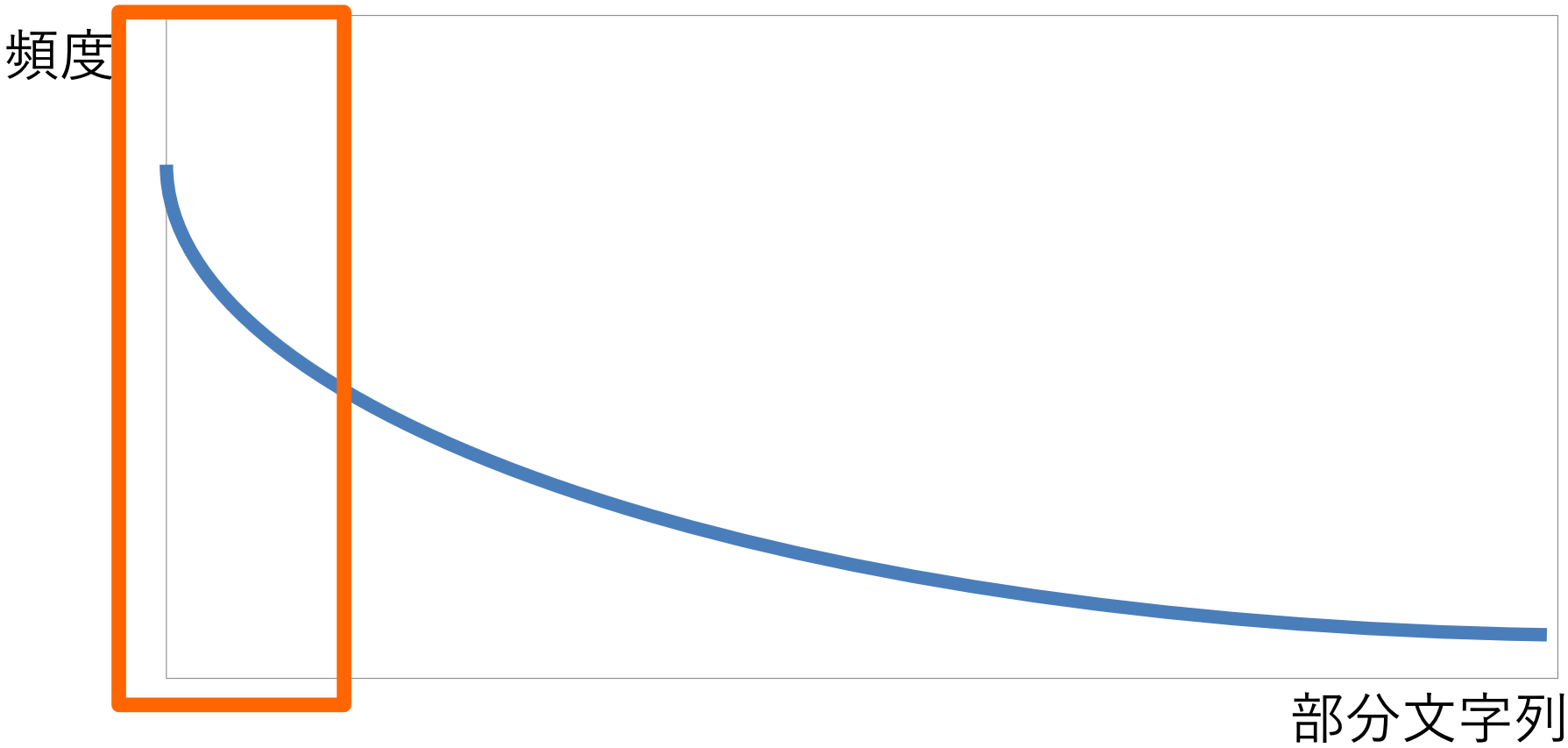
Query	Neighbors
💧	😅 😓 😥 💦 😂 💦💦 😀 😱 😟 😂😂
＼(^^)／	(≧∇≦) (*≧∀≦*) ٩(´Ｏ`*)۶ (o´∀`o) (*´∇｀*) (๑˃̵ᴗ˂̵) (❁´ω`❁) ヽ(*´∀｀)ノ ＼(^-^)／ o(^o^)o
(´;ω;`)	(;∀;) (´;ω;｀) (;_;) (;ω;) (T_T) (´;ω;`) (泣) (T＾T) (T-T) ヽ(;∇;)ノ
シン・ゴジラ	この世界の片隅に 四月は君の嘘 風夏 君の名は。 IQ246 ちはやふる 美女と野獣 北斗の拳 ねこあつめ サザエさん
サンキュー	さんきゅー さんくす あざす あざっす ありがとさん あざーす すごーい あざます おおきに サンキュー!
○ね	氏ね くたばれ 滅べ 散れ 消えろ 去れ 爆発しろ 首いてけ 連れてけ 死ねよ
でござる	でやんす でごわす でござるよ っス であります でしゅ でござる。 でござる… でございます ですぞ

NWJCの悩み：頻度が数えられない

- NWJCは総文字数３３億。そのなかに含まれる長さ1～８の部分文字列を列挙し、カウントするには、サーバのメモリ（512GB）が足りない。

- lossy count algorithm で、
  - 全体的に低頻度の部分文字列を排除
  - 頻度の近似値を計算
  - 全22,455,810個の部分文字列とその頻度の近似値を獲得
    - このうち、115万個の部分文字列を使用

115万件の選び方　[押切+, 17]



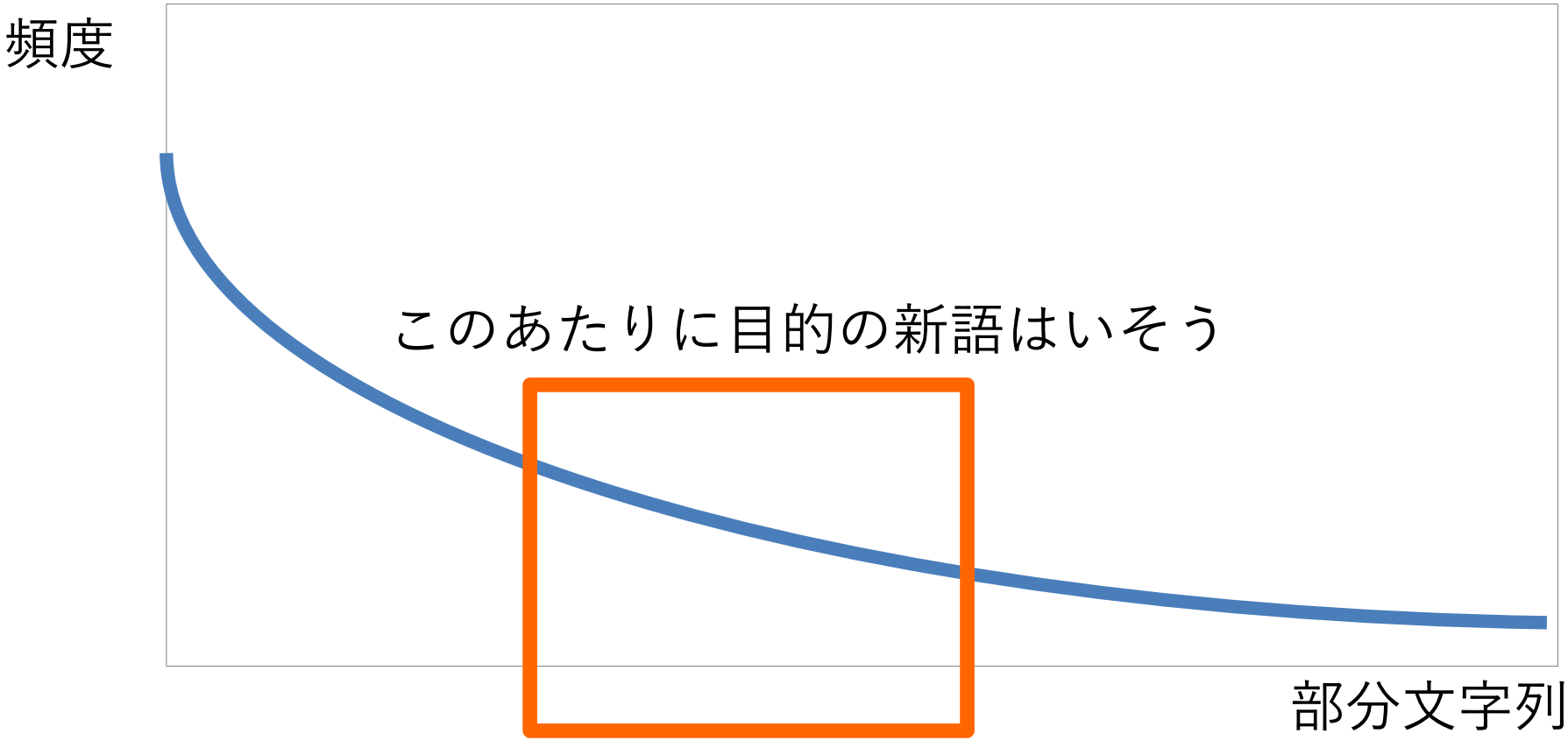
抽出された部分文字列の上位13件

- <<EOS>> 1384722213
- <<BOS>> 1384722213
- の 1133067869
- い 998155145
- 、 820991297
- で 793337889
- て 786002090
- し 753999738
- た 726467095
- に 700345283
- な 682454545
- と 630974447
- が 571650932

抽出された部分文字列の上位に現れた長さ２以上の部分文字列

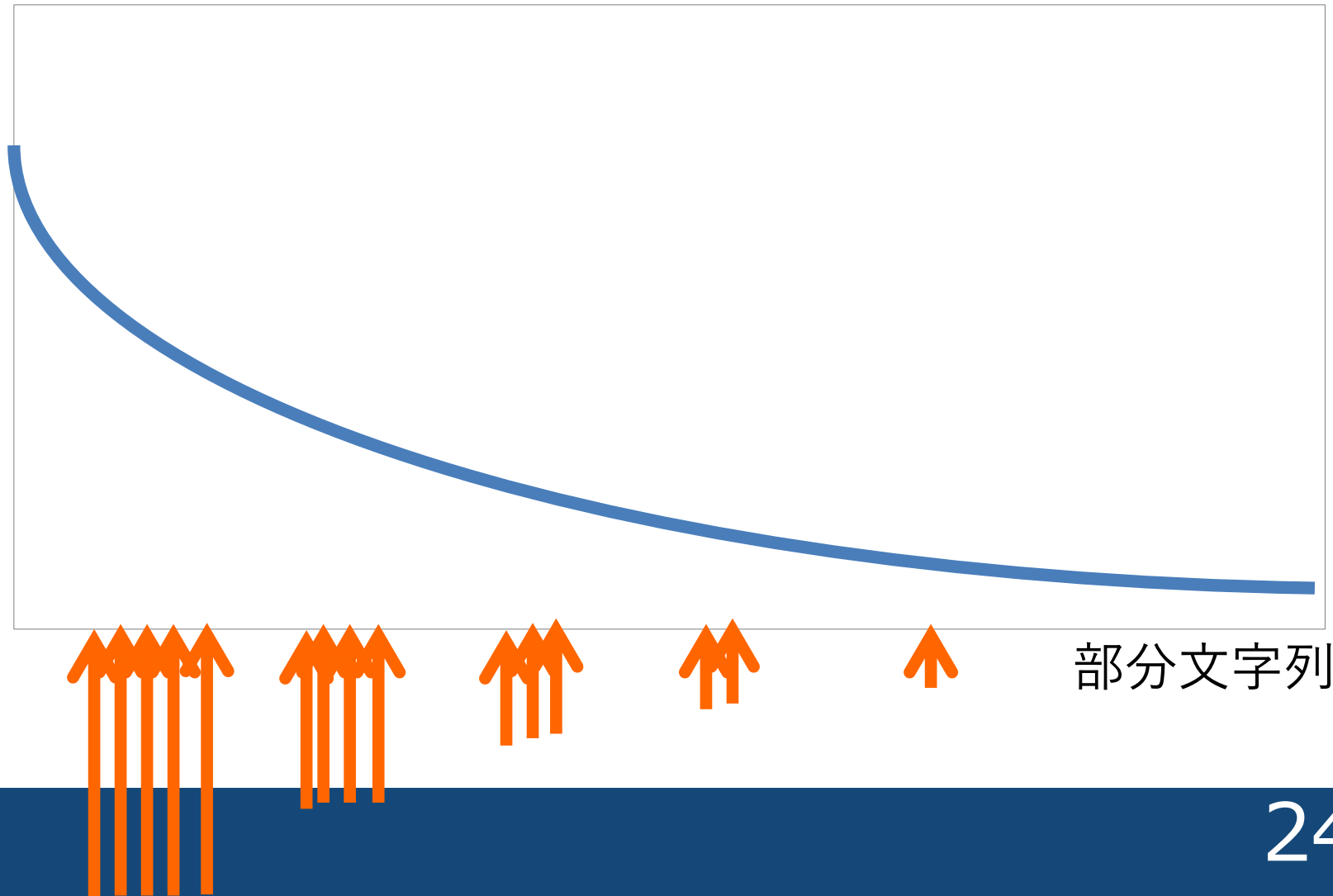
- って　254661654
- す<<EOS>>　214054411
- です　212560384
- した　210523707
- して　185924209
- った　181314981
- てい　177864597
- ます　173789792
- ない　154070495
- た<<EOS>>　141066785
- ・・　126182125
- から　119712392
- まし　118653135
- ます<<EOS>>　115676735

本当に獲得したい部分文字列は……



115万件の選び方（提案手法）

頻度の分布に従って、全体からランダムに115万件サンプリング



サンプリングのイメージ

- 22,455,810万の面を持ち、各面に抽出してきた部分文字列が書かれたサイコロがある。
- 各面は、書かれた部分文字列の頻度の大きさだけ出やすい。
- そのサイコロを115万個の異なる出目が現れるまで振り続ける。
- 出た115万個の出目の部分文字列を使用する。

提案手法

- [押切+, 17] 単語分割を経由しない単語埋め込み
  - のアレンジ

① NWJCから長さ1~nすべての部分文字列を抽出
② 抽出した各部分文字列に対し、左連接・右連接する部分文字それぞれカウントし、左連接行列C_L・右連接行列C_Rを作る。
③ $C = [C_L, C_R]$を特異値分解
④ 左特異ベクトルを各部分文字列の分散表現とする。
⑤ UniDic DB内の登録済みの短単位と照合し、特定の品詞のみがついた部分文字列を正例、指定した以外の品詞がついた部分文字列を負例とし、分類器を構築。
⑥ UniDic DBに未登録の部分文字列を分類器（knn-classifier）にかける

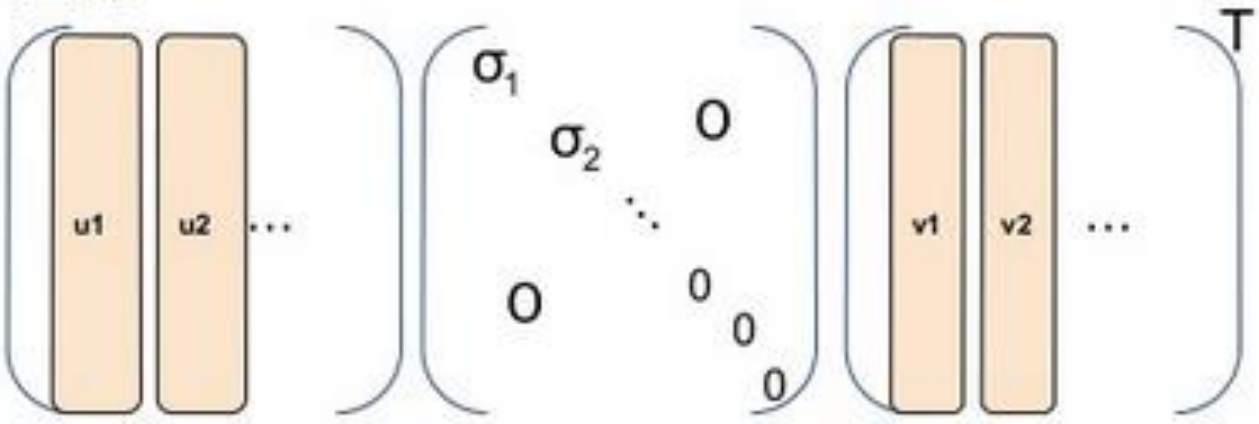
分散表現の作成

- 方法は様々だが、 [押切+, 17]に合わせて乱択化特異値分解を使用。
  - 大規模行列（今回の場合、115万×115万）の特異値分解を効率的に実行する。
  - [森+, 98]や[萩原+, 11]に対し、柔軟に素性が入れられる（e.g., 部分文字列長など）
  - w2v的な作り方よりも、調整するハイパーパラメータが少ない
  - 作成するベクトルの次元数は200次元に設定

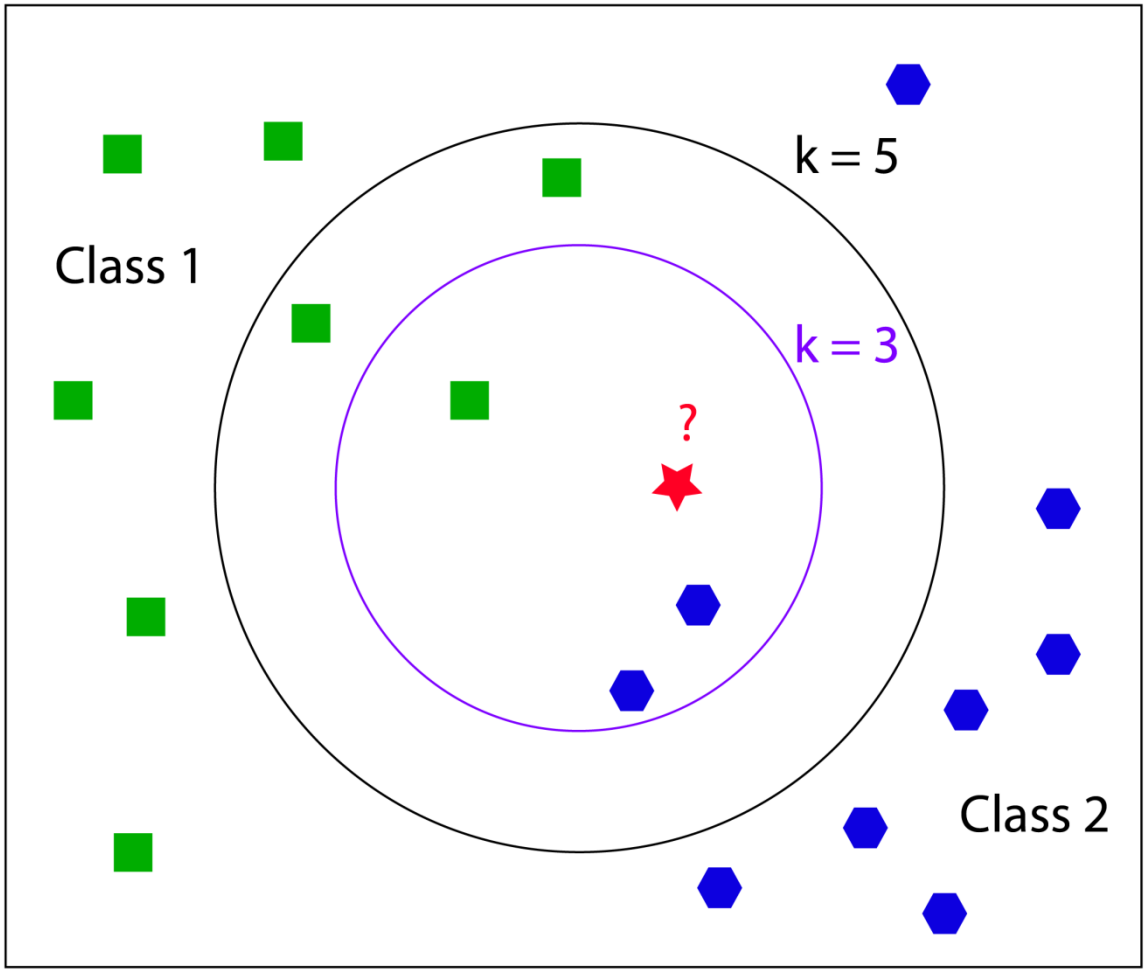
特異値分解

- 長方形行列の行列分解できるようにしたものが特異値分解

$$A = U\Sigma V^T =$$



分類器：weighted knn-classifier



近傍クラスの
多数決に
距離の重み付けを
して分類を行う

細々とした補足

- 連接行列の頻度を数える際もlossy countingを使用している。
- Classifierを作る際、「「」『』、。」のような記号が入った部分文字列はすべて負例にしている。
- サ変可能以外の可能性に基づく品詞を持つ部分文字列はすべて負例にしている。

抽出された候補の一部（名詞）

- ＦＣ２
- ＧＬ
- IPP
- ＭＭＯ
- MOD
- ＭＰ３
- ＮＡＮＡ
- Ｎ響
- ＳＩＭ
- Ｗｉ－Ｆｉ
- ＷｉＭＡＸ
- Ｗｉｋｉ
- ＹｏｕＴｕｂｅ
- ｉＭａｃ
- ｉＰａｄ
- ｉＰｈｏｎｅ
- ｉＰｏｄ
- ｉＴｕｎｅｓ
- ｉｅ
- ｐｉｘｉｖ
- ｔｗｉｔｔｅｒ
- アドセンス
- アラフォー
- コメ欄
- スパロボ
- ホルミル
- マツコ
- ラジコ
- 夏コミ
- 絵茶

懸案事項

- まだPrecisionが低い
  - 人手で見ないといけない箇所が多い
  - 長単位・文節のリストではじく？
  - 素性に入れるとRecallが落ちた
- 動詞・形容詞・副詞がうまく取れない
  - 部分文字列の内部の部分文字列を素性に入れてみたが……
  - そもそも数があるのか？
- このタスクにあった素性を組む必要がある
  - 現状、追加した素性は[東+, 06]のもの

まとめ・今後の課題

- NWJCからUniDic DBへ新規追加する短単位の候補を列挙する手法を提案
  - [押切+, 17]をアレンジした分散表現の作成
  - 頻度の分布に基づく部分文字列の絞り込み
  - Weighted knn-classifierによる候補の列挙とソート

- w2vライクな分散表現の作成手法も提案されたので、こちらも検討に入れる
  - [Oshikiri+, 17] Segmentation-Free Word Embedding for Unsegmented Language
    - In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.

謝辞

- 本研究は国立国語研究所コーパス開発センターの共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」(2016-2021 年度) の成果である。