単語分かち書き辞書『mecab-ipadic-NEologd』を公開して得た知見について

LINE株式会社 佐藤敏紀 (@overlast)

2017-03-06 語彙資源活用シンポジウム

佐藤 敏紀 (@overlast)

所属　LINE 株式会社 Data Labs エンジニア

業務　自然言語処理・機械学習・検索に関連する研究・開発

趣味　ボードゲーム

過去の開発経験

ブログ検索エンジンblogWatcher : 2の終盤～3の開発(6人)

近似文字列照合エンジン : 企画, 開発, 運用(2人)

検索エンジン用のクエリ訂正・提案 : 企画, 開発(6人)

NEologdによるWeb上からの語彙収集と資源生成 : 全て(1人)

研究としての概要

見出し語の
辞書を作ってます

Webサービス開発に必要な形態素解析

以下の条件を満たす必要があり、実現したかった

- **高速 && 簡単**

- **形態素解析 + チャンキング + 固有表現抽出**
  - “形態素より長い単語境界”の高精度な同定
  - 固有表現に対する”雑なカテゴリ付与”

- **新語や流行語、有名な固有表現の網羅**

言語資源の確保による解決が必要

「Web検索エンジンのクエリログとクリックログがあれば、概ね解決できる」という雑感があった

正道: コーパスとして言語資源を蓄える

選択: 辞書の見出し語として言語資源を蓄える

- 固有表現の採録速度と低い人的コストの両立

研究の目標

“更新され続ける”単語分かち書き用の辞書を実現

- 固有表現や複合名詞を単語として扱う
- システム運用により自動化・半自動化する
- 無料での商用利用と研究利用を可能にする

本研究の貢献：

新語・未知語・高頻度な固有表現・複合名詞等の語彙が不足していることが原因の解析誤りを改善

NEologd って?

システム名です

我々が取り組む「語彙獲得タスク」

以下の4要素の組(4つ組)を収集する

- ガイドライン(割愛)に沿い採録/非採録を決定

要素名	例
見出し語の表層	NL研
読み仮名	エヌエルケン
原型	情報処理学会自然言語処理研究会
品詞情報	名詞,固有名詞,一般

NEologd って?

NEologd: 語彙獲得と4つ組リスト生成するシステム

語彙獲得の処理

1. 新語・未知語の検出
2. Webサイトのクロール
3. 語彙不足なドメインに属する用語を網羅
4. テンプレートによる生成
5. ホワイトリスト、ブラックリストの管理

任意のタイミングで「4つ組リスト」を生成する

NEologd で獲得する4つ組の例

表層	読み仮名	原型	品詞情報
東京工業大学	トウキョウコウギョウダイガク	東京工業大学	名詞, 固有名詞, 一般,*,*,*
東京工業大学	トウキョウコウギョウダイガク	東京工業大学	名詞, 固有名詞, 組織,*,*,*
東工大	トウコウダイ	東京工業大学	名詞, 固有名詞, 一般,*,*,*
東工大	トウコウダイ	東京工業大学	名詞, 固有名詞, 組織,*,*,*
MacBook Pro	マックブックプロ	MacBook Pro	名詞, 固有名詞, 一般,*,*,*
東京都渋谷区渋谷	トウキョウトシブヤクシブヤ	東京都渋谷区渋谷	名詞, 固有名詞, 地域, 一般,*,*
東京都渋谷	トウキョウシブヤ	東京都渋谷区渋谷	名詞, 固有名詞, 地域, 一般,*,*
西川仁	ニシカワヒトシ	西川仁	名詞, 固有名詞, 人名, 一般,*,*
2016 年	ニセンジュウロクネン	2016 年	名詞, 固有名詞, 一般,*,*,*
生麦生米生卵	ナマムギナマゴメナマタマゴ	生麦生米生卵	名詞, 固有名詞, 一般,*,*,*

1. 新語・未知語の検出

インターネット内外のイベン急激に流行る単語
⇒ 新語や未知語の出現を監視し続ける

監視対象コンテンツの例

- ニュース記事
- Twitter のトレンド / ハッシュタグ
- 各種検索エンジンの人気キーワード
- 放映中のテレビ番組名、掲示板

1ヶ月程度の幅で集計し採録する語を抽出、で十分

2. Webサイトのクロール

Web上の網羅的かつ定期的に更新される言語資源

- なるべく差分だけをクロールしてる
- クロール時のマナーは守っている

単語とその読み仮名の組の正確性、に着目している

- 正確性の高いサイトが満たしている条件があった

自由に利用できる辞書に仕上げるため以下だけ使う

- 使用許諾が得られるコンテンツから得た4つ組
- 複数母体のサイトから共通で抽出できた4つ組

3. 語彙不足なドメインに属する用語を網羅

新語・未知語の監視をしていると、、
「特定ドメインの単語の網羅性が極めて低い!!!」と思う瞬間

⇒ 一般性が十分に高いと判断したら網羅的な収集を試みる

例1. 人名
- 災害時などに姓と名の網羅が、とくに必要である
  - クロール結果に基いて、問題が起きないように採録

例2. Unicode 絵文字や顔文字
- 実際に頻繁に使われているが、まとまっていない
  - 人手で4つ組リストを整備している

4. テンプレートによる生成

テンプレートで4つ組を生成できる場合

- 極めて正確性の高い4つ組の元データが使えた
- 高頻度な未知語を網羅できた

例1. 住所文字列

- もともと粒度毎にカラム分けされている
    - 実際に出現頻度が高い組み合わせを使う

例2. 時間表現と数値表現

- 網羅は目指さないが高頻度な範囲を生成する

NEologd: Webからの語彙収集システム

NEologdで解くタスク

- Webクロールの結果に基づく4つ組リスト作成

4つ組とは

- 語の表層、読み、語の表層の原型、品詞情報
  - 西川仁、ニシカワヒトシ、西川仁、名詞,固有名詞,人名,一般

4つ組リストの生成方法

- 自動、半自動、人手を使い分ける
- 採録の速さと読み仮名の正しさを追求する

mecab-ipadic-NEologdって

読めますか?

mecab-ipadic による解析例

ー ー ー

```
[overlast@]$mecab
任天堂は3月3日にNintendo Switchを発売した。
任天堂　　名詞,固有名詞,組織,*,*,*,任天堂,ニンテンドウ,ニンテンドー
は　助詞,係助詞,*,*,*,*,は,ハ,ワ
3　　名詞,数,*,*,*,*,*
月　名詞,一般,*,*,*,*,月,ツキ,ツキ
3　　名詞,数,*,*,*,*,*
日　名詞,接尾,助数詞,*,*,*,日,ニチ,ニチ
に　助詞,格助詞,一般,*,*,*,に,ニ,ニ
Nintendo 名詞,一般,*,*,*,*,*
Switch　　名詞,一般,*,*,*,*,*
を　助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
発売 名詞,サ変接続,*,*,*,*,発売,ハツバイ,ハツバイ
し　動詞,自立,*,*,サ変・スル,連用形,する,シ,シ
た　助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
。　記号,句点,*,*,*,*,。,。,。
EOS
```

形態素に分割される

固有表現が複数形態素に分割される理由

歴史的な経緯をあまり考慮せず、理由を挙げると

1. IPADIC が形態素解析のための辞書だから
2. 更新の頻度が低いから(または更新停止状態)
3. 固有表現抽出は別の研究トピックだから
4. 未知語を検出するには言語資源が必要だから

など様々な理由を考えることができる

固有表現バラバラ問題を解決するには

前提: MeCab による処理以外は想定できない

- 開発コストを下げる
- 既に普及している実装を使う

以下の条件を満たす言語資源を作る

1. 固有表現を1単語として分割する
2. 定期的に更新して、現実の状況を反映する
3. よく使われる固有名詞にあらかじめ対応する
4. 未知語は見つかり次第対応する

さらに: 既存の形態素解析の結果が実用上正しい時は尊重

単語分かち書き用の辞書 mecab-ipadic-NEologd

4つ組リストから生成される MeCab 用の辞書

- 単語の単位: 形態素と固有名詞・複合名詞
- 固有表現や複合名詞を形態素に分割しない
- IPADIC による分割結果を極力活かす

表 4　辞書による分かち書き結果の違い

解析器と辞書の名前	分かち書きの結果
MeCab & mecab-ipadic-NEologd	国土交通省 / は / 2001 年 / に / 設置 / さ / れ / まし / た / 。
MeCab & IPADIC	国土 / 交通省 / は / 2001 / 年 / に / 設置 /さ / れ / まし / た / 。
MeCab & UniDic	国土 / 交通 / 省 / は / 2 / 0 / 0 / 1 / 年 / に / 設置 / さ / れ / まし / た / 。
KyTea	国土 / 交通 / 省 / は / 2001 / 年 / に / 設置 / さ / れ / ま / し / た / 。
Juman++	国土 / 交通 / 省 / は / 2001 / 年 / に / 設置 /さ / れ / ました / 。

mecab-ipadic-NEologd による解析例

```
[overlast@]$mecab -d /usr/local/lib/mecab/dic/mecab-ipadic-neologd
任天堂は3月3日にNintendo Switchを発売した。
任天堂  名詞,固有名詞,組織,*,*,*,任天堂,ニンテンドウ,ニンテンドー
は      助詞,係助詞,*,*,*,*,は,ハ,ワ
3月3日  名詞,固有名詞,一般,*,*,*,3月3日,サンガツミッカ,サンガツミッカ
に      助詞,格助詞,一般,*,*,*,に,ニ,ニ
Nintendo Switch 名詞,固有名詞,一般,*,*,*,Nintendo Switch,ニンテンドースイッチ, ニンテンドースイッチ
を      助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
発売    名詞,サ変接続,*,*,*,*,発売,ハツバイ,ハツバイ
し      動詞,自立,*,*,サ変・スル,連用形,する,シ,シ
た      助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
。      記号,句点,*,*,*,*,。,。,。
EOS
```

固有名詞や固有表現が
1単語になっている

mecab-ipadic-NEologd: 単語分かち書き辞書

その他の特徴(2017年2月末時点)

- 週2回更新
- IPADICに530万語以上の語彙を追加
- IPADICが正しく分か書きできる語は尊重する
- 新語や流行した未知語の採録が速い
- WebニュースやSNS上の表記に対応
- 頻出な顔文字やUnicode絵文字を採録
- Twitterでのエラー報告を監視している

IPADIC と mecab-ipadic-NEologd の改善

NEologd で生成できる4つ組リストを使うだけでは不十分

IPADIC の改善 ⇒ パッチ作成

- 単純な誤りの訂正、分割誤りの原因となるコスト値の修正

見出し語を網羅的に追加

- 用言と副詞（次は動詞を追加予定）
- 高頻度な感動詞
- 一般名詞・固有表現・サ変接続名詞の表記揺れ
- SNSにありがちな崩れ表記語

追加した見出し語の解説	品詞情報	見出し語の例	見出し語数
細かく分けられない固有名詞	名詞,固有名詞,一般	Nintendo Switch, Chainer	1,070,047
日本の株式会社、法人、学校	名詞,固有名詞,組織	LINE株式会社	317,196
市区町村、観光地、駅	名詞,固有名詞,地域	JR新宿ミライナタワー	495,229
偉人や著名人の氏名	名詞,固有名詞,人名,一般	有吉弘行, BABYMETAL	751,264
知識があれば読める名前	名詞,固有名詞,人名,名	明日奈, 大也	95,232
知識があれば読める苗字	名詞,固有名詞,人名,姓	上垣内, 国ノ十	87,567
頻出なサ変接続名詞	名詞,サ変接続	勝越, 吹替	491
頻出な一般名詞, Unicode絵文字	名詞,一般	北北東, 北北東, よだれ, 🫩	1,747
頻出な形容詞とその表記ゆれ	形容詞,自立	うまぁ〜	507,812
非頻出な形容詞の表記ゆれ	形容詞,自立	ウマーーーーーーイ	1,051,146
形容動詞語幹の表記揺れ	名詞,形容動詞語幹	ユウイギ, 世間並	20,268
副詞とその表記ゆれ	副詞,一般	ぐわぁ〜っと, にこっ	139,792
IPADICの一般名詞の表記揺れ	名詞,一般	お好焼き, おこのみ焼	153,104
IPADICの固有名詞の表記揺れ	名詞,固有名詞,一般	アソサン,タカハタフドウ	138,379
頻出な時間に関する固有表現	名詞,固有名詞,一般	17代目, 99年間	17,653
非頻出な時間に関する固有表現	名詞,固有名詞,一般	十七代目, 九十九年間	16,533
頻出な数値に関する固有表現	名詞,固有名詞,一般	9999円, 100倍	158,831
非頻出な数値に関する固有表現	名詞,固有名詞,一般	九千九百九十九円	205,926
IPADICのサ変接続名詞の表記揺れ	名詞,サ変接続	夜あそび, 夜アソビ	26,058
SNSに見られる形容詞の崩れ表記	形容詞,自立	可愛いいい	60,616
SNSで頻出な感動詞	感動詞	無駄無駄無駄ァ	4,701
		合計	5,319,592

mecab-ipadic-NEologd はどんな時に有効か

想定するテキストの難易度

- Min: Web上のニュース記事
- Max: Twitter、インスタ、アプリのレビュー

実用して効果を確認した用途

- 文書分類、単語分散表現の精度向上 ⇐ これがきっかけ
- 大雑把な固有表現の抽出
- 全文検索の結果の改善
- 任意の文字列への読み仮名付与
- 近似文字列検索や、スペル誤り訂正

派生の開発物

語彙を集めると
こういうことが
できます

mecab-ipadic-NEologd 以外のOSSデータ(1/2)

for UniDIC

- **mecab-unidic-NEologd**
  - 実はこれも週2回更新している
  - IPADIC 版ほど手を入れていない

for mecab-ipadic-NEologd

- **ext-column-unidic-tokenized-surface**
- **ext-column-ipadic-tokenized-surface**
  - 解析結果の各語の末尾に IPADIC / UniDIC で形態素に分割した結果(mecab -Owakati 相当)を追加

mecab-ipadic-NEologd の代表的なオプション

```
./bin/install-mecab-ipadic-neologd \
 -n \  # 更新してインストール
 -y \  # インストール時の確認をしない
 -a \  # 全見出し語をインストール
 -u \  # ユーザ権限でインストール
 -p /home/overlast/local/dic/ipadic-neologd \
 --extend-column unidic-tokenized-surface
```

ちなみに--helpで全オプションを確認できる

--extend-column unidic-tokenized-surface の結果

ー ー ー

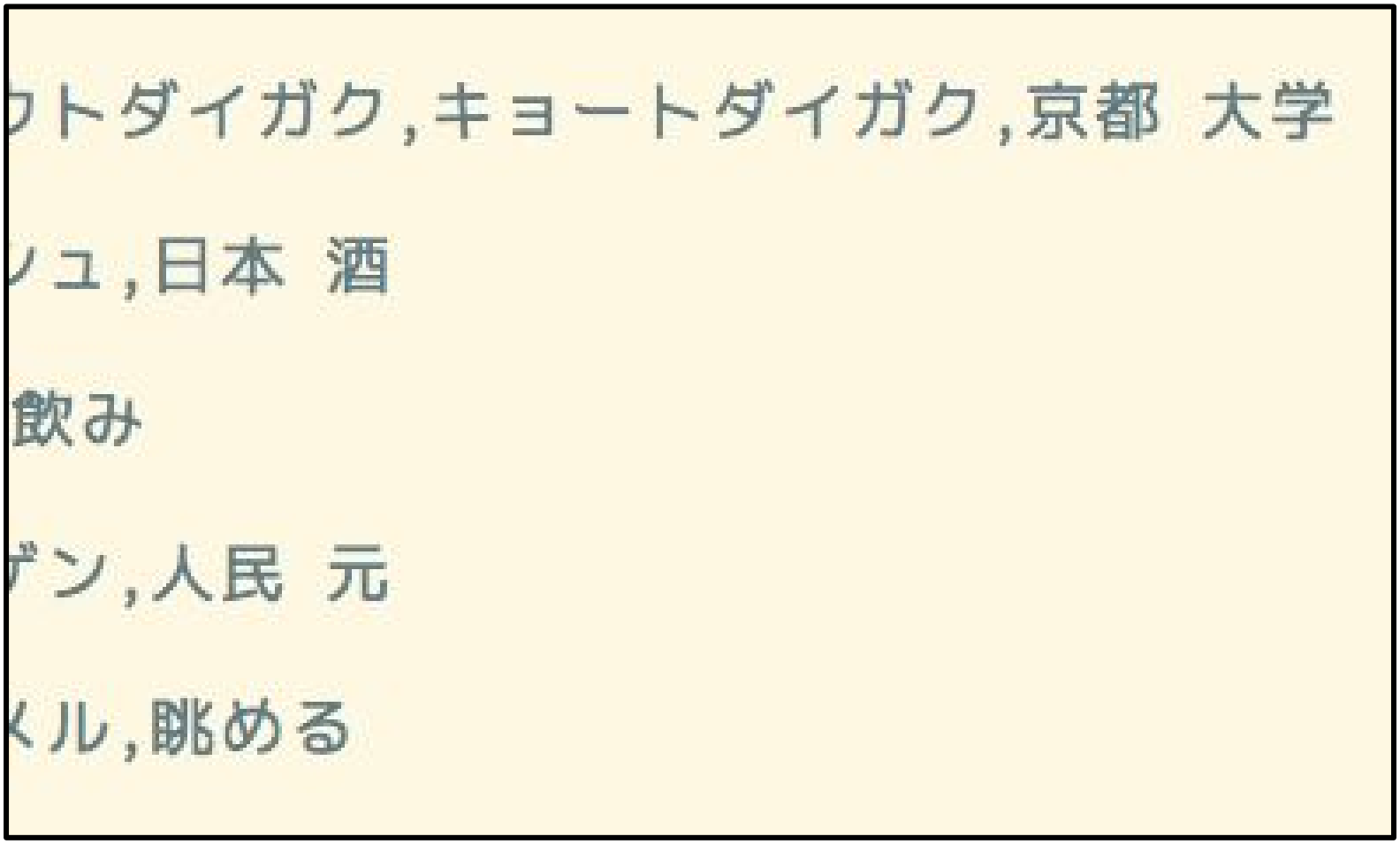
```
^C[overlast@]mecab -d /usr/local/Cellar/mecab/0.996/lib/mecab/dic/mecab-ipadic-neologd-ext-unidic
いつ京都大学で日本酒を飲みつつ人民元を眺める?
いつ	名詞,代名詞,一般,*,*,*,いつ,イツ,イツ,いつ
京都大学	名詞,固有名詞,組織,*,*,*,京都大学,キョウトダイガク,キョートダイガク,京都 大学
で	助詞,格助詞,一般,*,*,*,で,デ,デ,で
日本酒	名詞,一般,*,*,*,*,日本酒,ニッポンシュ,ニッポンシュ,日本 酒
を	助詞,格助詞,一般,*,*,*,を,ヲ,ヲ,を
飲み	動詞,自立,*,*,五段・マ行,連用形,飲む,ノミ,ノミ,飲み
つつ	助詞,接続助詞,*,*,*,*,つつ,ツツ,ツツ,つつ
人民元	名詞,一般,*,*,*,*,人民元,ジンミンゲン,ジンミンゲン,人民 元
を	助詞,格助詞,一般,*,*,*,を,ヲ,ヲ,を
眺める	動詞,自立,*,*,一段,基本形,眺める,ナガメル,ナガメル,眺める
?	記号,一般,*,*,*,*,*
EOS
```

--extend-column unidic-tokenized-surface の結果

ー ー ー

```
^C[overlast@]mecab -d /usr/local/Cellar/mecab/0.996/lib/mecab/dic/mecab-ipadic-neologd-ext-unidic
いつ京都大学で日本酒を飲みつつ人民元を眺める?
いつ	名詞,代名詞,一般,*,*,*,いつ,イツ,イツ,いつ
京都大学	名詞,固有名詞,組織,*,*,*,京都大学,キョウトダイガク,キョートダイガク,京都 大学
で	助詞,格助詞,一般,*,*,*,で,デ,デ,で
日本酒	名詞,一般,*,*,*,*,日本酒,ニッポンシュ,ニッポンシュ,日本 酒
を	助詞,格助詞,一般,*,*,*,を,ヲ,ヲ,を
飲み	動詞,自立,*,*,五段・マ行,連用形,飲む,ノミ,ノミ,飲み
つつ	助詞,接続助詞,*,*,*,*,つつ,ツツ,ツツ,つつ
人民元	名詞,一般,*,*,*,*,人民元,ジンミンゲン,ジンミンゲン,人民 元
を	助詞,格助詞,一般,*,*,*,を,ヲ,ヲ,を
眺める	動詞,自立,*,*,一段,基本形,眺める,ナガメル,ナガメル,眺める
?	記号,一般,*,*,*,*,*
EOS
```

UniDIC で分かち書きした結果が追加されている



Patch を当てたIPADICのみをインストールする

```
./bin/install-mecab-ipadic-neologd -n -y -u \
 -c \ # patchを当てて修正したIPADICのみinstall
 -p /home/overlast/local/dic/patched_ipadic
```

この機能は以前から個人的に便利に使っていた

IPADIC自体の改善できる点が見つかる

- おもに生起コストや読み仮名

修正点が見つかると”直すしか無い” 気分になる

- 10168 エントリ以上修正されている

```
[overlast@]$mecab
いつ京都大学で日本酒を飲みつつ人民元を眺める?
い      動詞,自立,*,*,一段,連用形,いる,イ,イ
つ      助動詞,*,*,*,下二・タ行,基本形,つ,ツ,ツ
京都大  名詞,固有名詞,組織,*,*,*,京都大,キョウトダイ,キョートダイ
学      名詞,接尾,一般,*,*,*,学,ガク,ガク
で      助詞,格助詞,一般,*,*,*,で,デ,デ
日本    名詞,固有名詞,地域,国,*,*,日本,ニッポン,ニッポン
酒      名詞,接尾,一般,*,*,*,酒,シュ,シュ
を      助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
飲み    動詞,自立,*,*,五段・マ行,連用形,飲む,ノミ,ノミ
つつ    助詞,接続助詞,*,*,*,*,つつ,ツツ,ツツ
人民元  名詞,一般,*,*,*,*,人民元,ジンミンゲ,ジンミンゲ
を      助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
眺める  動詞,自立,*,*,一段,基本形,眺める,ナガメル,ナガメル
?       名詞,サ変接続,*,*,*,*,*
EOS
^C
[overlast@]$mecab -d /usr/local/Cellar/mecab/0.996/lib/mecab/dic/patched_ipadic
いつ京都大学で日本酒を飲みつつ人民元を眺める?
いつ    名詞,代名詞,一般,*,*,*,いつ,イツ,イツ
京都大学        名詞,固有名詞,組織,*,*,*,京都大学,キョウトダイガク,キョートダイガク
で      助詞,格助詞,一般,*,*,*,で,デ,デ
日本酒  名詞,一般,*,*,*,*,日本酒,ニホンシュ,ニホンシュ
を      助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
飲み    動詞,自立,*,*,五段・マ行,連用形,飲む,ノミ,ノミ
つつ    助詞,接続助詞,*,*,*,*,つつ,ツツ,ツツ
人民元  名詞,一般,*,*,*,*,人民元,ジンミンゲン,ジンミンゲン
を      助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
眺める  動詞,自立,*,*,一段,基本形,眺める,ナガメル,ナガメル
?       記号,一般,*,*,*,*,*
EOS
```

いつ
京都大学
日本酒
人民元

が変化

mecab-ipadic-NEologd 以外のOSSデータ(2/2)

---

for Solr / Elasticsearch

- **Neologd-solr-elasticsearch-synonyms**
  - **シノニム辞書**
    - macab-ipadic-NEologdに採録した
      IPADIC の一般名詞・固有名詞の表記揺れを活用

- **例: 以下すべてを「お好み焼き」扱いできる**
  - お好み焼き, おこのみやき, おこのみ焼, おこのみ焼き,
    お好みやき, お好み焼, お好やき, お好焼, お好焼き,
    オコノミヤキ, オコノミ焼, オコノミ焼キ, オ好ミヤキ,
    オ好ミ焼, オ好ミ焼キ,オ好ヤキ, オ好焼

開発の動機

どんなものを作ることにしたのか

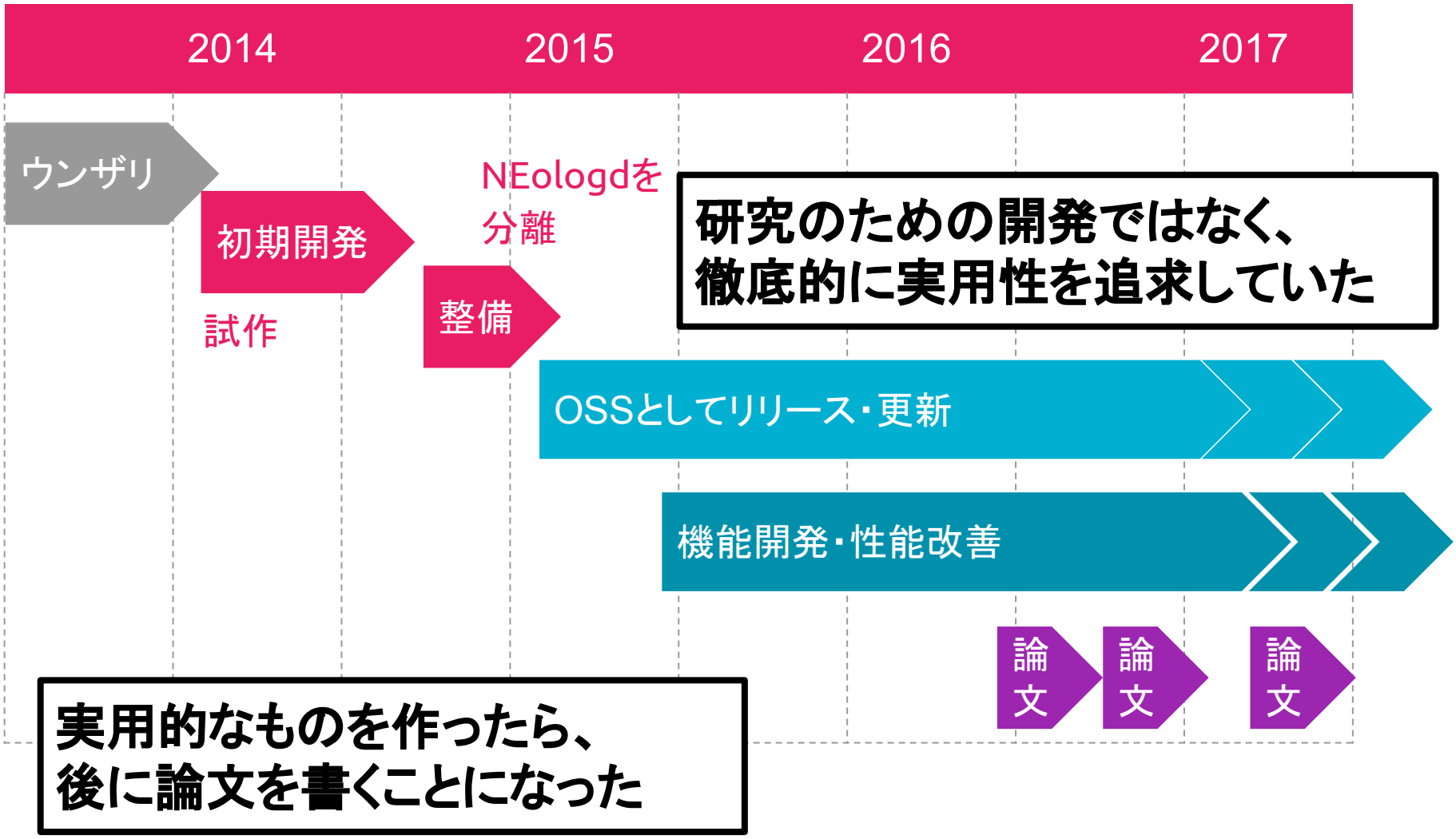
新語や固有表現を1単語として分割でき、
任意のタイミングで更新できるMeCab用辞書を作る

当時の僕の頭の中のイメージ

- 数年分の全文検索エンジンのクエリログ・クリックログを雑な集計をして得られるリストと同等以上の応用時の性能がある言語資源を作る

現状では追いついた部分と負けている部分がある

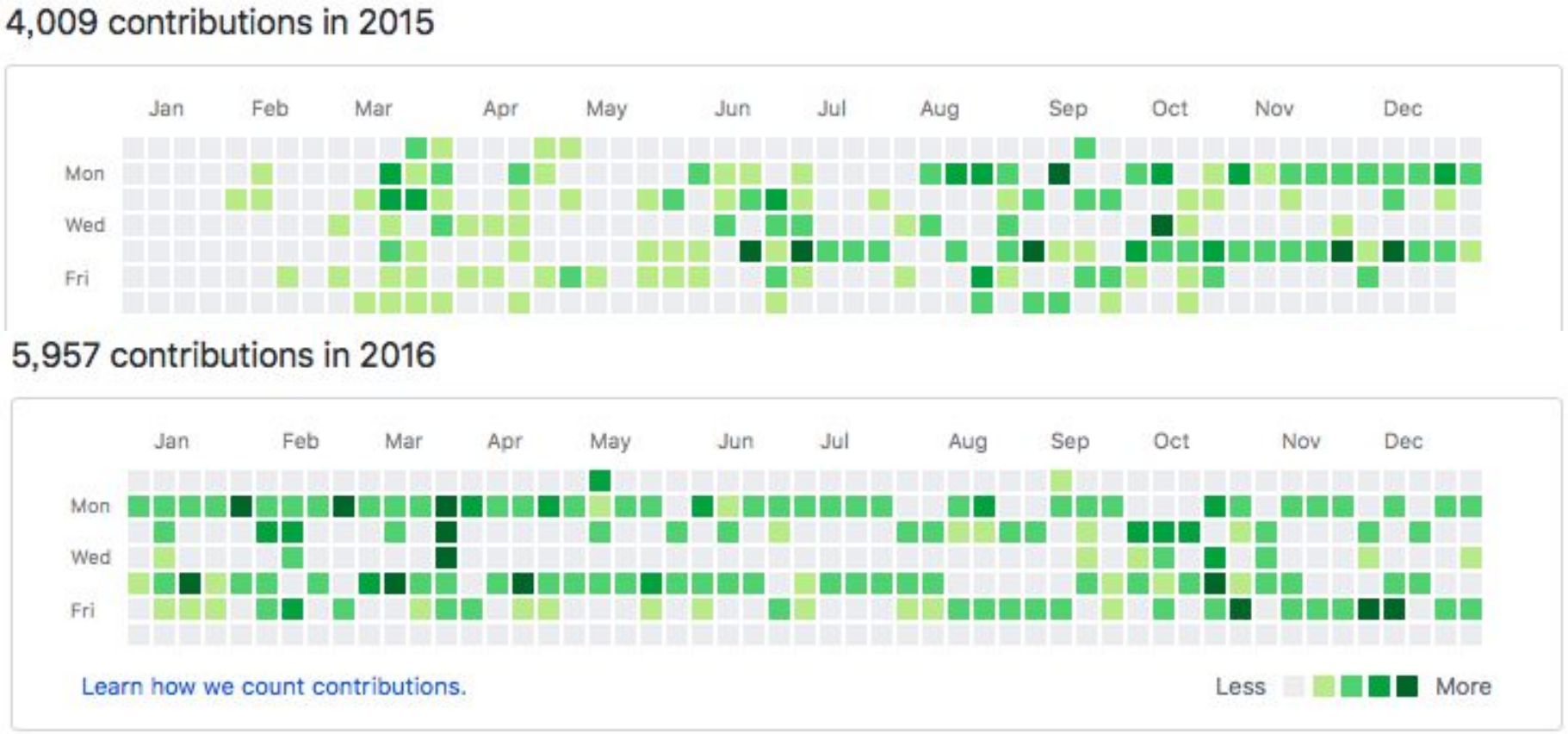
mecab-ipadic-NEologdの開発の流れ



語彙資源の更新を継続するメリット

GitHubのデータから考えます

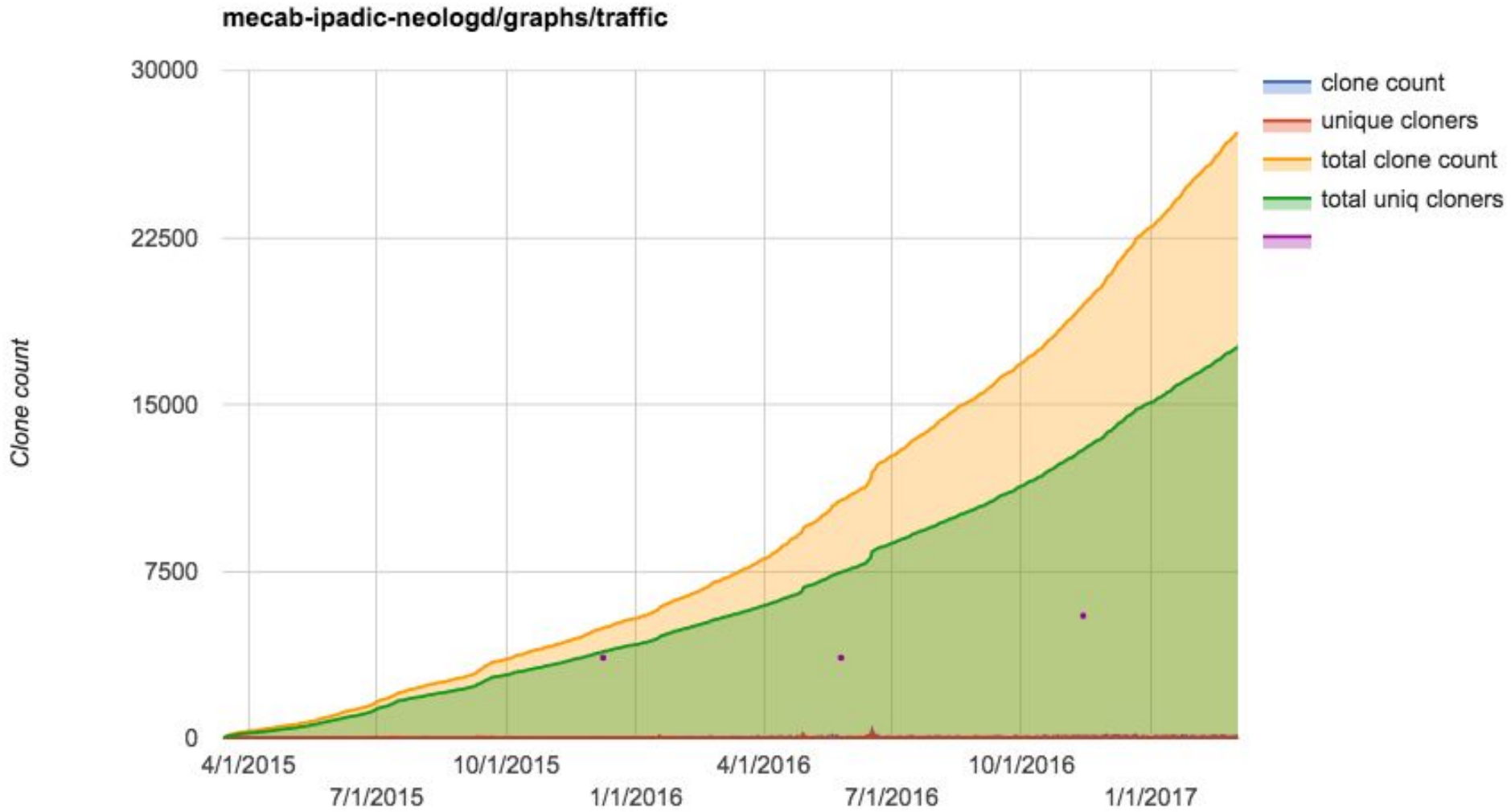
更新頻度がとても高い言語資源になった



- 2017年3月16日で公開から丸2年
  - 週2回の更新を始めてから1.5年強

2017年03月04日までの普及の様子

公開723日目　総clone数: 計27,227　異なりclone数: 計17,570



GitHub Stars: 826

世界一Starが多いGitHub上の辞書資源になった

GitHub Starの良いところ

- 押して下さった方が誰か分かるし、元気が出る
  - この人が!! と、ときどき気分が高まる
  - その方を調べるとお勤めの会社も分かる

更新することで感じたメリット(1/2)

更新の差分を得ると

- 新しい言語現象や流行に気がつく
  - 新たな開発物のアイディアになる

更新に差分が出る様にすると

- 差分を大きくしたくなり工夫が始まる
  - NEologdシステムの開発が捗る
    - 語彙の収集範囲が拡大する
    - 更新作業の効率が改善される

更新することで感じたメリット(2/2)

語彙を作ることが手法に影響することに気づく

- 単語の単位、採録基準を制御できるから

例: 何時の扱い

- 何/時
- 何/時代
- 何時/頃
- 何/曜日

品詞情報の系列を網羅的に処理
できそうなルールを書く

vs

何時、何時代、何時頃、何曜日など
何 + 時間単位を1単語として分かち書き

語彙資源の更新を継続するコスト

更新継続の計算機コストはそれ程でも

クローラー

- CPU、Disk I/O、HDD

4つ組リスト生成

- CPU、HDD

辞書の生成

- CPU、RAM、Disk I/O、HDD

一番困るのは収集・生成したデータを消せないこと

更新継続の人的コストはそれなりに高い

定期的に更新する語彙資源の価値は何か

- **新語・未知語問題が解決”され続ける”こと**、と感じた

新語・未知語問題の棚上げ状態の継続

- 更新が止まると徐々に影響が強くなる

対策: 2015年8月から休み・祝日も更新することに

- ほとんど自動化してるが、品質の担保は人手
- 世界的イベントが多発したら、最後は根性

更新継続コストは誰が負担するのか

OSSは開発者側から見ると無料ではない

- 人
- 計算機
- ストレージ
- 図書
- Webサイト・ドメインなど

個人的には企業が利益を得たら還元すべきと思う

- すべての企業がOSSから借りたら破綻する

実際に何処に、
どう普及してるのか

mecab-ipadic-NEologdを使ってますか?

使ってくださっている方!!

- 実はかなり悩ましい

GitHubではトラッキングできない

- Starを付けることと、使うことは妙にズレる

どこで使われているか確実に方法

- イベントを開催してアンケートを取る

好まれる公開方法
&
ライセンス

好まれる公開形態とライセンス

公開形態

- 無料ならGitHubで公開
- 有料でもダウンロード可能に(DVDは辛い)

ライセンス

- 無料ならApache License ver 2.0
  - CC 3.0などは面倒
- 有料でも契約して商用・研究ともに制限しない

mecab-ipadic-NEologdのライセンス

公開前にかなり議論

- 著作権違反は犯していないか
- 著作権違反じゃなくても訴える余地はないか
  - 訴えられると無実・無罪でも衰弱する

最終的に Apache License version 2.0 に

- そうなるような採録基準や判定をしている

使っていて不安にならない言語資源って

- 開発速度の速さは割と効く。すぐ直るから

解決できない課題

本研究が解決できない課題の例

低頻度な固有表現の抽出

- 原因: すべて登録するには限界がある(例: 金額や時間)

採録前の新語・未知語などに対する対処

- 原因: コーパスの蓄積が必要 or 後処理の工夫が必要

メモリ使用量の削減

- 原因: 用途が特定できず不要な見出し語を削減できない

MeCab 自体に手を加える必要がある

- 汎用な形態素解析ライブラリの開発は別のトピック

野望

長期的な野望

基礎的な語彙を整頓したい

- 一般名詞と固有名詞の境目がフワフワしてて面白い

日本語以外もやりたい

- NEologdで解決した課題は他の言語でも問題

NEologdの新語採録プロセスの前半部分を開放したい

- 確かな知り合いの手を借りてみたい
    - だけど、うまくいくいくか、信用されるか謎

僕が両手骨折しても大丈夫にしておきたい