文脈にもとづく未知語獲得における識別モデルの適用

研究者
鍛治伸裕　喜連川優
東京大学技術研究部

発表者　江口晃

はじめに

- 背景

  -未知語（辞書に登録されていない単語）の多いテキストは頑健に解析することができない

  -ウェブテキストには固有名詞や新造語などの未知語が頻繁に出現するため、高い精度で解析を行うことが難しい

  例：ググる、メタボる

- 研究の目的

  -未知語を獲得することで、ウェブテキストなどの未知語の多いテキストを高い精度で解析できるようにする

文脈に基づく未知語獲得

- 未知語獲得
  - -コーパスから未知語を抽出する
  - -抽出した未知語に適切な品詞を割り当てる

- 未知語に割り当てる品詞

  普通名詞、サ変名詞、母音動詞、子音動詞カ行、子音動詞サ行、子音動詞タ行、子音動詞ハ行、子音動詞マ行、子音動詞ラ行、子音動詞ワ行、子音動詞ザ変、イ形容詞、ナ形容詞

文脈情報の利用

- 弁別的文字列
  - 品詞tの直前または直後に出現しやすい文字列

- 弁別的文字列の例
  - 以下のテキストから普通名詞「XO醤」を抽出する

a.たくさんのXO醤をゲットしました。
b.XO醤などを軽くいためて香りを出す。
c.そのXO醤は、10年前に発明された。
弁別的先行文字列　→　「たくさんの」、「その」
弁別的後続文字列　→　「　を」、「　などを」、「は、」

識別モデルの適用

- 分類器
  - -品詞tの未知語候補である文字列を与える
  - -候補が品詞tを割り当て可能な単語であれば＋１、そうでなければ－１を出力する分類
  - -分類器は品詞の数だけ構築する

未知語候補の生成

- 未知語候補の生成

  -品詞tの弁別的先行文字列と弁別的後続文字列に囲まれて出現した全ての文字列を品詞tの未知語候補とする

  -素性数がσ以下のものは候補から外す

- 弁別的文字列の式

  $$coverage(p) = |\{w \in Wt \mid 0<f(pw)\}|$$

  p:文字列、 f(pw):文字列pwの頻度、 Wt:品詞tが割り当てられている辞書登録語の集合

  coverage(p):pが何種類の辞書登録語の直前・直後に出現したか

  -品詞tの弁別的文字列は、辞書を用いてコーパスから自動獲得する

  -coverage(p)にしきい値を設け、越えたpを品詞tの弁別的文字列とする

  -弁別的先行文字列と弁別的後続文字列それぞれについて行う

動的絞り込み

- 未知語候補の絞り込み

  -重複する候補が抽出された場合、不適格な弁別的文字列対の割合があるしきい値を上回ると候補から取り除く。この処理を未知語獲得の過程で動的に行う。

  -重複する候補のどちらかが既知語であるかを調べるさいは、辞書登録だけでなく、これまでの処理で判定された候補も含める

例：これまでは心配だったのですが。

「心配」（ナ形容詞語幹）＋「だったのですが」
「心配だっ」（母音動詞語幹）＋「たのですが」（後続文字列）
ナ形容詞語幹「心配」が既知の単語　→　後者が誤り

訓練事例

- 分類器の訓練事例の作成
  - -未知語候補cとそれに品詞tを割り当て可能であるかどうかを示す正解タグの組を作る
  - -正例は品詞tが割り当てられている辞書登録語を利用する
  - -負例はt以外の品詞が割り当てられている辞書登録語と、重複して抽出された候補を利用する

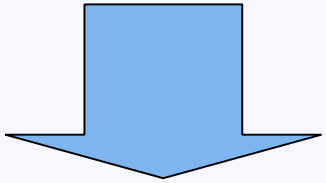
句判定

- 句判定

  -文脈に基づく未知語獲得の欠点の一つに、単語と句を明確に区別できないことがある

  例:「XO醤の味」

  誤って普通名詞であると判断されてしまう可能性がある



  -すべての未知語候補の処理が終わったのちに、分類器が＋1を出力したものに対して句判定を行う

  -句判定には単純な辞書引きを使う

実験の設定

- 実験の設定
  - 実験には1.7億文のウェブテキストとJUMAN辞書を用いた
  - 未知語候補の生成では弁別的文字列の長さを5とした
  - 素性数に対するしきい値σは1、16、32、64を試した
  - 分類器の性能を調べるため、評価用辞書を用いて適合率と再現率を調べた

実験結果（1）

表1：獲得した分別的文字列の例

品詞	先行文字列／後続文字列（上段／下段）
普通名詞	またとない、 など重要な、 どあらゆる があります、 はもちろん、 に対しても
子音動詞ラ行	あっけなく、 いい感じに、 から何かが るでしょう、 ろうとしました、 りましょう
イ形容詞	ヒジョーに、 たまらなく、 よりもっと かったので、 いのですが、 くなったの

先行文字列：299,574　後続文字列：153,583

表２：獲得した未知語

普通名詞	兄ィ、腐女子、音樂、特盛り
サ変名詞	逆ギレ、怪演、マターリ
子音動詞ラ行	ポシャる、帰える、ハショる
イ形容詞	甘っちょろい、ヤヴァい、ムズ痒い
ナ形容詞	ぴっかぴかだ、マッドだ、みょーだ

未知語：12,823

実験結果（3）

表3：適合率と再現率

σ	適合率	再現率	再現率の上限
1	80. 2	77. 1	89. 0
16	86. 7	75. 5	84. 9
32	89. 1	73. 2	80. 4
64	90. 2	69. 9	75. 5

おわりに

- 今後の課題
  -さらに高い精度の結果を得られるような手法の改良を進める
  -固有名詞や副詞など、今回の実験では対象としなかった品詞にも同様の手法が適用可能かを調査する