曖昧性解消 (WSD)

Word Sense Disambiguation

多義語の意味を同定する

ex) title

name of a book, music etc.
the right of legal ownership
an appellation of respect attached to a person 's name

他にも、そもそも品詞が違う場合もある

手法 (用いるリソース別)

教師あり学習
辞書ベース
教師なし学習

事前知識 [機会学習]

(この本では) 大別して 2 つ

- 教師あり学習
  - 入力データとその解のペアを用いて学習
  - 解を人力で与えるため面倒
  - 分類問題
- 教師なし学習
  - 解がわからない入力データで学習
  - クラスタリング問題
  - ex) 文書集合を似た文書の部分集合に分割

事前知識 [擬似語]

曖昧語を探すのは面倒

複数の語を一つの擬似語で置換する

> **文書中の** banana **と** door **を** banana-door **という単語で置換する**
> banana-door **は** 2 **つの意味を持った曖昧語となる**

事前知識 [上限下限]

アルゴリズムの良さの指標

WSD **では以下のように考える**

- **上限**
  - **人力で行った場合の正解率**
    大体 97~99%
- **下限**
  - **最もシンプルなアルゴリズム (すべてを同じ意味で捉える等) を用いた場合の正解率**
    大体 90%

勿論、この数値は単語の複雑さに依存する

表記ルール

w

曖昧語

$s_1, s_2, \ldots, s_k$

曖昧語 w の意味

$c_1, c_2, \ldots, c_i$

曖昧語 w の文脈

$v_1, v_2, \ldots, v_j$

文脈中の語句

表記ルール

以下のようなイメージ

I don ’t know the title of that book.

w = title
s_1 = name of a book, music etc
c_1 = {I don ’t know the of that book}
v_1 = know, v_2 = book, ...

教師あり学習

以下の 2 つを解説

- ベイズ分類器
  - ベイズ統計を用いる
- 情報理論による方法
  - 相互情報量を用いる

教師あり学習 [ベイズ分類器]

ある文脈 c のもとで、意味 s を取る可能性 P(s|c) が一番高い s を選択する。

すなわち、 $s = \arg\max_{s_k} P(s_k|c)$

$$
\begin{aligned}
s &= \arg\max_{s_k} P(s_k|c) \\
&= \arg\max_{s_k} \frac{P(c|s_k)P(s_k)}{P(c)} \\
&= \arg\max_{s_k} P(c|s_k)P(s_k) \\
&= \arg\max_{s_k} [\ln P(c|s_k) + \ln P(s_k)]
\end{aligned}
$$

観測の時点で P(c) は定数

教師あり学習 [ベイズ分類器]

単純ベイズ仮定

文脈に出現する各単語が出現する確率は他の単語と独立である。

$$P(c|s_k) = P(\{v_j|v_j \in c\}|s_k) = \Pi_{v_j \in c} P(v_j|s_k)$$

これを用いると前ページの s は以下のように変形できる

$$\begin{aligned} s &= \arg\max_{s_k}[\ln P(c|s_k) + \ln P(s_k)] \\ &= \arg\max_{s_k}[\Sigma_{v_j \in c} \ln P(v_j|s_k) + \ln P(s_k)] \end{aligned}$$

教師あり学習 [ベイズ分類器]

$$s = \arg\max_{s_k}[\Sigma_{v_j \in c} \ln P(v_j|s_k) + \ln P(s_k)]$$

$P(v_j|s_k) = \frac{C(v_j,s_k)}{C(s_k)}, P(s_k) = \frac{C(s_k)}{C(w)}$ **で求められる (最尤法)**

(C **は各要素の個数を表す**)

単純ベイズ分類器という

成功率 90%**くらい**
単純ベイズ仮定がそもそもおかしい

教師あり学習 [情報理論]

indicator

曖昧語の意味を示すような句
時制, **目的語**,etc...

例

フランス語 prendre

prendre une mesure

take a measure

prendre une dicision

make a dicison

目的語が indicator

教師あり学習 [情報理論]

Flip-Flop **アルゴリズム**

$t_1, .., t_m$(=T) **を翻訳語**,$x_1, .., x_n$(=X) **を** indicator **とする。**
ex) t_1 = take, t_2=make, x_1 = mesure, x_n = dicision

T **を適当に分割し、**$P = \{P1, P2\}$ **とする。**
while(**できるだけ**){
　　$I(P, Q)$ **を最大化する** X **の分割** $Q = \{Q1, Q2\}$ **を探す**
　　$I(P, Q)$ **を最大化する** T **の分割** $P = \{P1, P2\}$ **を探す**
}

2 **つ以上の意味解決をできるようにした拡張が存在する**

辞書ベース [定義利用]

文脈の単語と定義に登場する単語が似ていればその意味っぽい

$D_1, .., D_k$ **を** $s_1, .., s_k$ **の定義に登場する単語の集合とする**

ex) s_1= name of a book, music etc **の時、** D_1={name, of, a, book, music, etc}

E_{v_j} **を** v_j **の定義に登場する単語の集合とする**

$D_k \cap (\cup_{v_j \in c} E_{v_j})$ **の要素数が一番多い** s_k **を選ぶ**

$(\cup_{v_j \in c} E_{v_j})$ **は要するに、文脈に登場する単語の定義に使用される用語全体**

辞書ベース [定義利用]

正解率は 50%~70%

色々改良があるらしい

辞書ベース [シソーラス]

シソーラス

> **シソーラス** (Thesaurus) **とは、単語の上位 / 下位関係、部分 / 全体関係、同義関係、類義関係などによって単語を分類し、体系づけた辞書。**
> by Wikpedia(**シソーラス**)
> WordNet **に近い？**

文脈全体の意味的分類 $\Leftrightarrow$ 単語の意味的分類

> **単語の意味的分類がわかれば、単語の意味が分かる**

辞書ベース [シソーラス]

手順

$t(s_k)$ を s_k のシソーラスにおける subject コードとする。
$T(v_j)$ を v_j のシソーラスにおける subject コードの集合とする。

(各 v_j に対して意味が複数あるため集合)

$t(s_k) \in T(v_j)$ を満たす v_j の個数が最も多い s_k を採用する

精度は 50%程度

サンプルに用いた単語が難しすぎたらしい

問題点

例えば、コンピューターの本の中で mouse と出てきても、哺乳類というカテゴリを指す可能性は低い

辞書ベース [シソーラス 2]

目標

object **コードに対するナイーブベイズ分類器**

$$s = \arg\max_{s_k} [\Sigma_{v_j \in c} \ln P(v_j | t(s_k)) + \ln P(t(s_k))]$$

を構築する
$P(v_j|t(s_k), P(t(s_k))$ **をシソーラスを用いて計算したい**

辞書ベース [シソーラス 2]

手順 1

$$score(c_i, t_l) = \ln \frac{P(c_i|t_l)}{P(c_i)} P(t_l)$$
$$= \ln P(t_l) + \Sigma_{v \in c_i} \ln P(v|t_l) - \Sigma_{v \in c_i} \ln P(t_l)$$

お馴染みの単純ベイズ仮定

$$t(c_i) = \{t_l | score(c_i, t_l) > \alpha\} (\alpha \text{は何らかの定数})$$

文脈 c_i に妥当な分類を与えた

辞書ベース [シソーラス 2]

手順 2

$$V_j = \{c | v_j \in c\}$$
$$T_l = \{c | t_l \in t(c)\}$$

V_j は v_j を含む文脈
T_l は t_l に分類された文脈

$$P(v_j|t_l) = \frac{|V_j \cap T_l|}{\Sigma_j |V_j \cap T_l|}$$
$$P(t_l) = \frac{\Sigma_j |V_j \cap T_l|}{\Sigma_l \Sigma_j |V_j \cap T_l|}$$

辞書ベース [第二言語]

WSD **対象の言語を第一言語**

利用するリソースの言語を第二言語

手法

> **意味 s と文脈語 v の関係が、第二言語に翻訳しても成り立つときにその s を採用する**

辞書ベース [第二言語]

ex) interest

2 つの意味

legal share

ドイツ語では Beteiligung

attention

ドイツ語では Interesse

showed interest ではどちらの意味か

showed はドイツ語で zeigen
zeigen Beteiligung という言い回しは存在しない
zeigen Interesse という言い回しは存在する
なのでこの時の interest は attention

辞書ベース [語義の一貫性]

語義の一貫性という訳は去年の丸パクリ

対話レベル

　　一番採用回数が多い意味に合わせる

コロケーションレベル

　　意味不明

教師なし学習

教師なし学習の意義

- **専門的な文書の** WSD **を行うのに、一般的なリソースはほぼ使えない**
- **教師データが用意できない場合もある**

意味を正確に割り当てるのは不可能だが、意味を区別することはできる

単純ベイズ分類器と同じモデルを考える

$$s = \arg\max_{s_k} [\Sigma_{v_j \in c} \ln P(v_j|s_k) + \ln P(s_k)]$$

- **今回は教師データがないので各** P **が最尤法で求められない**
- EM **アルゴリズム** (14 **章でやるので解説は省略**)

正解率は教師あり学習より 5%~10%**程度落ちる**