Vol.2016-NL-228 No.6
2016/9/30

擬似用例を追加する能動学習を用いた一般単語の語義曖昧性解消

寺内賢志[†1]　佐々木稔[†2]　古宮嘉那子[†2]　新納浩幸[†2]

語義曖昧性解消において、教師あり学習に用いる訓練データすべてに人手でラベルを付けるためには膨大なコストがかかる。そのため、能動学習によってコストを削減する研究が行われている。先行研究では、分類器により負例と判定されたデータのみを訓練データとして追加する能動学習を用いて、固有名詞を対象とした語義曖昧性解消を行う。先行手法は固有名詞に対しては良い結果が得られるが、一般単語に対して識別精度が低いという問題点がある。そこで本論文では、一般単語に対する語義曖昧性解消のための能動学習手法を提案する。提案手法では、分類器により負例と判定されたデータのみでなく、正例と判定されたデータも訓練データとして追加する能動学習を用いて語義曖昧性解消を行う。提案手法での実験を行った結果、正例のみを対象とした場合には識別精度が向上した。しかし、正例と負例を対象とした場合には識別精度は向上しなかった。

Word Sense Disambiguation Based on Active Learning with Pseudo Examples for Popular Words

KATSUMUNE TERAUCHI[†1] MINORU SASAKI[†2] KANAKO KOMIYA[†2] HIROYUKI SHINNOU[†2]

1. はじめに

単語は文脈によって複数の意味を持つ。例えば「上げる」という単語において、「腕を上げる」と「根を上げる」では意味が異なる。このように文章によって異なる意味で使用されている単語に対し、どの意味で使用されているかを識別するための研究を語義曖昧性解消という。

語義曖昧性解消のための手法として、機械学習がある。語義曖昧性解消における機械学習では、対象単語が含まれる文とそこで使われている意味ラベルの組である訓練データを用いて分類器を学習させる。学習させた分類器を用いて、テストデータ中の対象単語がどの意味で使われているかを識別する。

機械学習の問題点として、訓練データとテストデータの分野があまりにも違う場合、識別精度が低下してしまう。そのため識別精度が高くなるようにデータを選んで訓練データ集合に追加する必要がある。そのような研究は能動学習と呼ばれる。

能動学習において、ラベルなしデータ全てに人手でラベルを割り振っていくと膨大なコストがかかってしまう。コスト削減を目的として、先行研究では擬似負例を追加する能動学習を行っている。擬似負例とは、ラベルなしデータにラベルを付けて訓練データ集合に追加する際、負例である可能性が高いと分類器によって判定されたデータである。このデータを追加することで、ラベルなしデータにラベルを付けるコストを削減している。先行研究では固有名詞を対象とした実験を行っているが、一般単語に対しては識別精度が低いという問題点がある。

本論文では一般単語に対しても効果的な能動学習を行うために、先行研究の改良を行う。提案手法では擬似負例だけでなく擬似正例を追加する能動学習を行い、一般単語に対しても語義曖昧性解消精度の向上を目指す。

2. 関連研究

本論文では、擬似用例を追加する能動学習を用いて語義曖昧性解消を行う。本節では、能動学習を用いて語義曖昧性解消を行う関連研究について記述する。

能動学習とは、機械学習において訓練データに追加するデータを選択する研究である。アプローチの一つに、不確実性サンプリング [1]がある。不確実性サンプリングとは、分類器によってどのラベルであるかが最も曖昧であると判定されたデータを選択する。Chan ら [2]は能動学習と領域適応を組み合わせた手法を提案した。この手法は、事前確率に基づく確信度スコアを用いて語義曖昧性解消に有用なデータを選択する。しかし、事前確率に基づく確信度スコアでは不十分であると考える。

Zhu ら [3]は、訓練データ集合中のラベルのバランスをとるため、オーバーサンプリングとアンダーサンプリングを用いる手法を提案した。オーバーサンプリングは、少ない

†1 茨城大学大学院理工学研究科情報工学専攻
Graduate School of Science and Engineering, Ibaraki University
†2 茨城大学工学部情報工学科
College of Engineering, Ibaraki University

ラベルに追加することでバランスをとる。アンダーサンプリングは、多いラベルからデータを削減することでバランスをとる。この手法では、正例と負例が最初から訓練データ集合に含まれているが、本研究では一部の正例のみで能動学習を始める。

3. 擬似負例を追加する能動学習

本節では、先行手法である擬似負例を追加する能動学習について記述する。

3.1 擬似負例を追加する能動学習の概要

擬似負例とは、ラベルなしデータにラベルを付けて訓練データ集合に追加する際、負例である可能性が高いと分類器によって判定されたデータである。このデータを追加することで、ラベルなしデータにラベルを付けるコストを削減している。分類器の訓練に用いる素性は以下の 5 つである。

(1)対象単語が含まれる文とその前後 1 文の内容語
(2)対象単語が含まれる文節内において対象単語より前に出現する内容語
(3)対象単語が含まれる文節内において対象単語より後に出現する内容語
(4)対象単語が含まれる文節に係る文節
(5)対象単語が含まれる文節を受ける文節

擬似負例であるかという判定には確信度スコアという値を用いる。確信度スコア c(d, s) は以下の式で計算する。

$$c(d,s) = \log p(s) \sum_{j=1}^{n} \log p(f_j \mid s)$$

c(d, s)は、データ d のラベルが s である確信度スコアであり、d が負例である確信度スコア c(d, neg)から正例である確信度スコア c(d, pos)を引いた値により、擬似負例として訓練データ集合に追加するデータであるかどうかを判定する。

3.2 擬似負例を追加する能動学習のアルゴリズム

擬似負例を追加する能動学習のアルゴリズムを以下に示す。

```
#初期化
Γ(P, N)   #P、N により訓練した分類器
T  :  #ラベルなしデータを含む訓練データ集合
P  :  #正の訓練データ集合
N  ←  ϕ  :  #負の訓練データ集合
PN  ←  ϕ  :  #擬似負例の集合
#能動学習
repeat
  foreach d (T - P - N)  :  #ラベルなしデータ
    c(d, pos)  :  #ラベルなしデータが正例である確信度スコア
    c(d, neg)  :  #ラベルなしデータが負例である確信度スコア
    if(c(d, neg) - c(d, pos))
        PN  ←  d
  end
  Γ  ←  Γ(P, PN + N)
  #人手でラベルを付ける
  cmin  ←  ∞
  foreach d  :  #ラベルなしデータ
    s(d)  :  #分類器 Γ によるデータ d の意味 s の予測
    c(d, s(d))  :  #予測した意味に対する確信度スコア
    if (cmin > c(d, s(d)))  then
      cmin  ←  c(d, s(d))
      dmin  ←  d
    end
  end
  if (s(dmin) = pos)
    P  ←  dmin
  else
    N  ←  dmin
until ラベル付きデータが所定の個数に到達
```

初期化として、正の訓練データ集合に正例の一部、負の訓練データ集合を空集合、そして擬似負例の集合を空集合とする。

次に擬似負例を追加する。ラベルなしデータ全てに対し、正例である確信度スコア c(d,pos)と負例である確信度スコア c(d,neg)を計算する。ここで c(d,neg) - c(d,pos)が τ 以上の場合、擬似負例として分類器を訓練する。閾値は τ =1 である。

訓練した分類器を用いて、ラベルなしデータ全ての意味を予測する。予測された意味の確信度スコアが最も低いデータに人手でラベルを付け、正または負の訓練データ集合に追加する。

4. 提案手法

本節では先行手法を改良し、擬似負例のみでなく擬似正例も含めた擬似用例を追加する能動学習の概要について記述する。

4.1　擬似用例を追加する能動学習

先行手法では能動学習において擬似負例の追加が行われている。しかし、擬似負例として訓練データ集合に追加されたデータの中には、実際には正例であるものが含まれている場合も多々ある。これによりテストデータの識別精度が低くなってしまうと考えたため、提案手法では擬似負例のみでなく擬似正例も追加する。分類器の訓練に用いる素性は変更せず、以下の 5 つである。

(1)対象単語が含まれる文とその前後 1 文の内容語

(2)対象単語が含まれる文節内において対象単語より前に出現する内容語

(3)対象単語が含まれる文節内において対象単語より後に出現する内容語

(4)対象単語が含まれる文節に係る文節

(5)対象単語が含まれる文節を受ける文節

分類器は naïve bayes 分類器を使用し、擬似用例の判定には確信度スコアを計算する。確信度スコアは以下のように定義する。

$$c(d,s)=\sum_{j=1}^{n}\log p(f_j \mid s)$$

この定義により、訓練データ集合のラベルごとの事前確率が全て統一される。今回の実験では一般単語を語義曖昧性解消の対象としているが、文章による素性にそれほど大きな違いがなかったため、識別において訓練データ集合のラベルごとの数の影響が大きいのではないかと考えたためである。

また、事前確率を統一したことにより、能動学習の初期段階ではラベルなしデータはほぼすべて正例と判定されてしまう。そのため、一回目の学習では正例である確信度スコアが最も低いデータを擬似負例として追加する。

4.2　擬似用例を追加する能動学習のアルゴリズム

擬似負例を追加する能動学習のアルゴリズムを以下に示す。

```
#初期化
Γ(P, N)  #P、N により訓練した分類器
T  :  #ラベルなしデータを含む訓練データ集合
P  :  #正の訓練データ集合
N ← ϕ  :  #負の訓練データ集合
PP ← ϕ  :  #擬似正例の集合
PN ← ϕ  :  #擬似負例の集合
 cmin1 ← ∞
 Dmin ← ϕ
#能動学習
#一回目に擬似負例を追加
foreach d (T - P - N)  :  #ラベルなしデータ
  c(d, pos)  :  #ラベルなしデータが正例である確信度スコア
  c(d, neg)  :  #ラベルなしデータが負例である確信度スコア
  if(c(d, pos) < min) then
    cmin ← c(d, pos)
    Dmin ← d
  elsif(c(d, pos) = min) then
    Dmin ← ϕ
    cmin ← c(d, pos)
    Dmin ← d
  end
  PN ← Dmin
  Γ ← Γ(P, PN + N)
end
#擬似正例含む擬似用例の追加
repeat
  foreach d (T - P - N)  :  #ラベルなしデータ
    c(d, pos)  :  #ラベルなしデータが正例である確信度スコア
    c(d, neg)  :  #ラベルなしデータが負例である確信度スコア
    if (c(d, neg) > c(d, pos))
      c(d, psdEx) ← c(d, neg) - c(d, pos)
      if(c(d, psdEx) >= τ)
        PP ← d
    else
      c(d, psdEx) ← c(d, pos) - c(d, neg)
      if(c(d, psdEx) >= τ)
        PN ← d
    end
  end
  Γ ← Γ(PP+P, PN + N)
  #人手でラベルを付ける
  cmin2 ← ∞
  foreach d (T - P - N)  :  #ラベルなしデータ
    s(d)  :  #分類器Γによるデータdの意味sの予測
    c(d, s(d))  :  #予測した意味に対する確信度スコア
    if (cmin2 > c(d, s(d)))
      cmin2 ← c(d, s(d))
      dmin ← d
    end
  end
  if (s(dmin) = pos)
    P ← dmin
  else
```

```
        N ← dmin                                    55
    end                                             56
until ラベル付きデータを 48 個追加                  57
```

初期化として、正の訓練データ集合に正例の一部、負の訓練データ集合を空集合、そして擬似正例、擬似負例の集合を空集合とする。

一回目の能動学習で、擬似負例を追加する。正例である確信度スコア c(d, pos)が最も低いもの全てを擬似負例の集合に追加する。

次に擬似正例含む擬似用例を追加する。ラベルなしデータ全てに対し、正例である確信度スコア c(d,pos)と負例である確信度スコア c(d,neg)を計算する。ここで c(d,Ex)が τ 以上の場合、擬似用例として分類器を訓練する。閾値は τ =6 である。

訓練した分類器を用いて、ラベルなしデータ全ての意味を予測する。予測された意味の確信度スコアが最も低いデータに人手でラベルを付け、正または負の訓練データ集合に追加する。

5. 実験

提案手法の有効性を評価するため、擬似正例の有無、分類器における各ラベルの事前確率の有無の組み合わせで合計 4 種類の実験を行った。

5.1 実験設定

実験に使用したデータ、対象単語、比較実験の条件の組み合わせを以下に示す。

・実験に使用したデータ

semeval2010　日本語タスク

訓練データ：

対象単語を含む 1 文とその前後の 1 文ずつを使用した。能動学習によって全ての訓練データを追加する。また、初期の訓練データにはできるだけ素性が多いものを選んだ。

テストデータ：

対象単語を含む 1 文であり、新語義としてラベルの付いていないデータが存在している場合はテストデータから除外した。

・対象単語

対象単語とその単語の semeval2010 の訓練データ、テストデータ、ラベルの数を表 1 に示す。「上げる・挙げる・揚げる」、「意味」のテストデータの数が 50 個に満たないのは semeval2010 の新語義として扱われているデータを除外したためである。また、初期の正の訓練データの数は 1 個または 2 個である。

表 1：各単語のデータの数とラベルの数

	訓練データの数	テストデータの数	ラベルの数
上げる・挙げる・揚げる	50	48	6
与える	50	50	3
意味	50	49	3
発つ・立つ・建つ	50	50	4
採る・取る・執る・捕る	50	50	6

・比較実験の条件の組み合わせ

提案手法をまとめると以下のようになる。

(1)能動学習の際、擬似負例のみでなく擬似正例も追加

(2)naïve bayes 分類器において各ラベルの事前確率を全て 1 に統一

これらの条件をそれぞれ適用した場合と適用しなかった場合の組み合わせは 4 つである。条件の組み合わせを表 2 に示す。

表 2：比較実験の組み合わせ

	擬似正例の追加	事前確率の統一
手法 1	○	○
手法 2	○	×
手法 3	×	○
手法 4	×	×

5.2 実験手順

実験の手順を以下に示す。

5.2.1 学習データの準備

訓練データを MeCab による形態素解析、そして CaboCha による係り受け解析をすることで抽出した素性の頻度を計算する。その中で初期の訓練データとするものに人手でラベルを付け、それ以外をラベルなしとする。

5.2.2 擬似用例の追加

訓練データにより分類器を学習させ、ラベルなしデータに対し確信度スコアを計算する。

確信度スコアの計算には以下の式を計算する。

$$c(d,s)=\sum_{j=1}^{n}\log p(f_j \mid s)$$

この式により、ラベルなしデータが正例である確信度スコア c(d, pos)、負例である確信度スコア c(d, neg)を計算する。ここで c(d, pos)の方が高い場合、c(d, pos) - c(d, neg)、c(d, neg)の方が高い場合、c(d, neg) - c(d, pos)を計算し、その値

が閾値τ以上の場合、確信度スコアが高い方のラベルを付け訓練データ集合に追加する。追加後、全てのラベルなしデータに対し確信度スコアを計算し、最も確信度スコアが低いデータに対し人手でラベルを付け訓練データに追加する。閾値はτ=6 と設定した。

5.2.3 テストデータの識別

擬似用例を追加した後で、語義を識別するモデルを作成する。このモデルに対して、テストデータを入力して対象単語の語義を識別する。

5.3 実験結果

比較実験は表 2 の 4 種類の条件で行った。テストデータ中の全てのデータの識別率、そして正例のみの識別率で評価を行う。テストデータには正例の数の方が少ないため、正例の識別率の重要度が大きいと考えた。

表 3：各単語と全ての単語の識別率(対象は正例負例)

	手法 1	手法 2	手法 3	手法 4
上げる・挙げる・揚げる	45.14%	48.61%	62.15%	46.88%
与える	43.33%	44.67%	58.67%	48.00%
意味	49.65%	48.30%	54.42%	51.02%
発つ・立つ・建つ	53.50%	47.00%	59.50%	50.50%
採る・取る・執る・捕る	43.00%	52.33%	55.00%	54.67%
ALL	46.45%	48.48%	58.16%	50.41%

表 4：各単語と全ての単語の識別率(対象は正例のみ)

	手法 1	手法 2	手法 3	手法 4
上げる・挙げる・揚げる	66.67%	47.92%	64.58%	43.75%
与える	44.00%	54.00%	46.00%	68.00%
意味	61.22%	59.18%	48.98%	51.02%
発つ・立つ・建つ	68.00%	38.00%	44.00%	48.00%
採る・取る・執る・捕る	63.82%	51.06%	61.70%	44.68%
ALL	60.66%	50.00%	54.92%	51.23%

6. 考察

提案手法では、対象を正例のみとした場合に識別精度が向上した。ここではそれぞれの手法についての考察を述べる。

6.1 手法 1(提案手法)

擬似正例を追加し、各ラベルの事前確率を統一する手法である。正例の識別率が大幅に向上した。正例である確信度スコアが高いデータを擬似正例として追加したことにより、分類器の訓練に用いる素性が増加したためであると考えられる。

6.2 手法 2

擬似負例だけでなく擬似正例も追加する手法である。先行手法よりわずかながら識別率が下がった。

6.3 手法 3

各ラベルの事前確率を統一し、さらに擬似正例を追加しない能動学習を行った手法である。全体の識別率、正例の識別率ともに高く、最も良い結果であった。擬似正例を追加する際、正例である確信度が高いものを追加した。そのため、正例と負例の境界が曖昧になってしまい、擬似正例を追加した場合より擬似負例のみを追加した場合に識別率が上がってしまったと考えられる。

6.4 手法 4(先行手法)

ラベルごとに事前確率を計算し、擬似負例のみを追加する手法である。全体の識別率は提案手法より高いが、正例の識別率が低かった。擬似正例を追加しない訓練データ中には負例が多く、各ラベルの事前確率を計算するため、正例の識別率が下がると考えられる。

7. おわりに

本論文では、擬似用例を追加する能動学習を用いて、一般単語に対する語義曖昧性解消を行った。単語の出現の偏りを抑えるため、各ラベルの事前確率を統一して確信度スコアを計算した。また確信度スコアにより、正例と判定されたデータも擬似用例として追加した。その結果、識別の対象を正例と負例とした場合には精度は低下した。しかし、対象を正例のみとした場合には識別精度が向上した。先行手法はデータを負例であると識別してしまう傾向があり、正例を識別することが重要であると考えたため、提案手法は有効であると考えられる。

しかし、比較実験において擬似正例を追加した場合、確実に精度が上がるという結果とはならなかった。正例である確信度スコアが高いデータを追加したため、正例と負例の境界が曖昧になってしまったと考えられる。今後、擬似正例の場合は確信度スコアが低いデータだけを追加するなど、効果的な能動学習手法を検討していく必要がある。

参考文献

[1] D. D. Lewis and W. A. Gale, “A sequential algorithm for training text classifiers,” in Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR ’94. New York, NY, USA: Springer-Verlag New York, Inc., pp. 3–12, 1994.

[2] Y. S. Chan and H. T. Ng, “Domain adaptation with active learning for word sense disambiguation,” in Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Prague, Czech Republic: In Proceedings of Association for Computational Linguistics, pp. 49–56, June 2007.

[3] J. Zhu, “Active learning for word sense disambiguation with methods for addressing the class imbalance problem,” in In Proceedings of Association for Computational Linguistics, pp. 783–790, 2007.

[4] 高山康博、今村誠、鍛冶伸裕、豊田正史、喜連川優 “web マイニングにおける語義曖昧性解消のための擬似負例を用いた能動学習” 情報処理学会論文誌：データベース、vol2、No2、pages 1-9(2009)