基礎

3.6 語義曖昧性解消

新納 浩幸（茨城大学工学部情報工学科）

語義曖昧性解消とは文中の多義語の語義を識別するタスクである．たとえば，「ドライバー」という単語には少なくとも，(a1) 自動車を運転する人，(a2) ねじを回すための工具，(a3) ゴルフのウッドクラブ 1 番の称，の 3 つの語義がある．そして入力文「行き先をドライバーに告げる」などが与えられたときに，入力文中の単語「ドライバー」が (a1)，(a2)，(a3) のどの語義として使われているかを識別するタスクが語義曖昧性解消である．

語義曖昧性解消は一般に教師あり機械学習手法を用いて解決される．上記の例で言えば，単語「ドライバー」を含む用例を適当な数（n 個）集め，それぞれの用例に対してその用例中の「ドライバー」の語義を付与しておく．これが訓練データとなる．特徴抽出により「ドライバー」を含む用例をベクトル $\boldsymbol{x}$ で表現する．訓練データ中の用例 $\boldsymbol{x}_i$ の「ドライバー」の語義を y_i とすると，訓練データ D は $\boldsymbol{x}_i$ と y_i のペアの集合 $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ となる．用例 $\boldsymbol{x}$ を入力，「ドライバー」の語義 y を出力と考えれば，「ドライバー」の語義曖昧性解消は $y=f(\boldsymbol{x})$ の関係を持つ関数 f を構築することで解決できる．D は関数 f の入出力の組が n 個集まったものであり，この入出力の例から f を構築するのが標準的な教師あり機械学習である（**図 -1** 参照）．

用例をベクトルに変換した後の処理は，一般のパターン認識と同様であり語義曖昧性解消に特化したものではない．そのため語義曖昧性解消の問題は，教師あり機械学習の枠組みで考えれば，「訓練データが不足している」および「特徴抽出時に十分な特徴を抽出できていない」の 2 つの問題に帰着される．現在の語義曖昧性解消システムのエラーの多くはこの 2 つの問題から生じている．この点を確認できたことが，エラー分析プロジェクトの語義曖昧性解消チームの成果といえる．ただしそれと同時にこれら問題の解決だけでは，高精度の語義曖昧性解消システムの実現が難しいことも確認できた [1]．

まず「訓練データが不足している」という問題を見てみる．語義曖昧性解消が困難な原因の 1 つは，自然言語が本来曖昧性を有していることである．たとえば入力文が「問題はドライバーです」の場合，文中の「ドライバー」の語義は曖昧であり，正解は文脈によって変化する．この入力文と正解を訓練データに追加したとしても，次に同じ入力文が現れた場合に正解するとは限らない．この問題は訓練データの領域と入力データの領域が異なる領域適応の問題とも関連している．曖昧性のある入力文であっても，そこそこシステムが正解できるのは，訓練データ中の最大頻度の語義をデフォルトとして出力するからである．たとえば訓練データの基になるコーパスがスポーツ記事であれば，「ドライバー」の語義は (a3) である確率が高い．そのため語義が判断できないときは (a3) を出力する．通常，入力文も同じ領域のコーパスから取られたものであるためデフォルトの解答は正解である確率が高い．しかし実際の入力文はスポーツ記事ではなく，別領域のコーパス（たとえば経済記事など）から取られている場合も多い．入力データが可変である限り，どんなに訓練データを増やしても領域適応の問題は生じる．

次に「特徴抽出時に十分な特徴を抽出できていない」という問題を見てみる．語義曖昧性解消が困難な原因として，システムの必要とする知識が膨大であることが挙げられる．たとえば「相手」には (b1) 物事をするとき，行為の対象となる人，(b2) 自分と対抗して物事を争う人，の 2 つの語義がある．今

「キャッチボールの相手」と「ボクシングの相手」という文中の「相手」の語義を考えてみる．「キャッチボール」も「ボクシング」も上位概念は「運動」である．そして「キャッチボール」は競技ではなく 2 人で行う単なる運動なので前者の正解は（b1）である．一方「ボクシング」は競技なので後者の正解は（b2）である可能性が高い☆1．このように語義を決めるには，単語の概念以上の知識が必要になる場面が多々ある．そして概念以上の知識をあらかじめ列挙しておくことは難しく，結果として，そのような知識を特徴抽出によって得ようとすることも困難である．

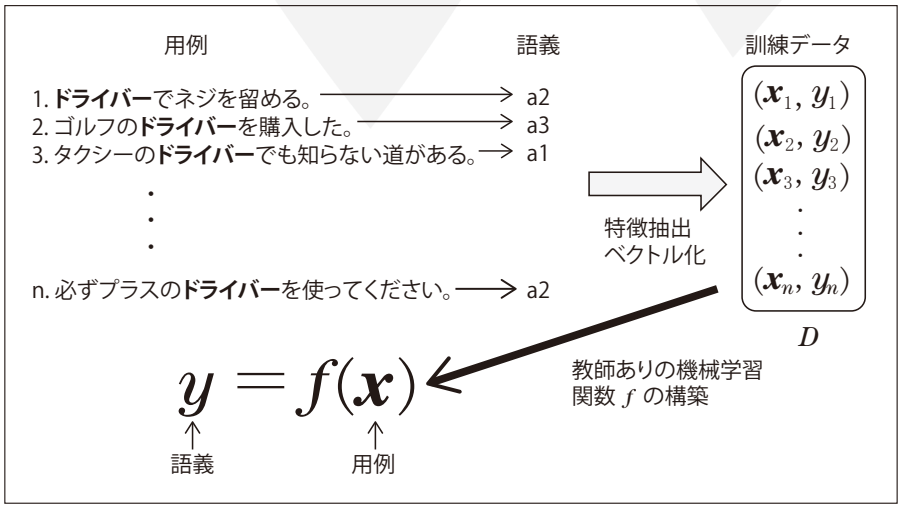

図 -1　教師あり機械学習による語義曖昧性解消

また上記 2 つの問題に収まらない問題として推論の問題がある．たとえば「開く」には（c1）あける，あいて広い状態になる，（c2）裂いて（または切れ目を入れて）処置する，の 2 つの語義がある．今「開くとカードが挟まっていた」と「開くとカードが入っていた」の「開く」の語義を考えてみる．これは開く対象を推理できないと識別できない．前者の開く対象はカードが挟まるようなたとえば本のようなものなので「開く」の語義は（c1），後者の開く対象はカードが入るようなたとえば封筒のようなものなので「開く」の語義は（c2）と推理できる．このような推論処理を要する語義曖昧性解消を，訓練データの追加や特徴抽出の改善によって解決するのは困難である．

最後に語義曖昧性解消の近い将来について述べる．いくつか語義曖昧性解消の困難な例を示したが，そのような問題を含んでいたとしても，実際は何らかの識別の手がかりが存在していることが多い．おそらく，現在の教師あり機械学習の枠組みで突き進めば 90％程度の正解率には達するだろう．しかしこれによって語義曖昧性解消システムが現実のさまざまなアプリケーションで利用されるとは思えない．先に挙げた領域適応の問題もあるが，より現実的な問題は，現在の教師あり機械学習の枠組みでは訓練データの作成コストが高く，対象単語が多くても数百程度に限定されてしまうことである．当然，現実のアプリケーションではすべての単語を対象にする必要がある．すべての単語に語義を付与する語義曖昧性解消は all-words WSD として研究されているが，そこでは教師あり機械学習によるアプローチは非現実的であり，主として，教師なし機械学習が用いられる☆2．今後，語義曖昧性解消の研究は高精度を追求する方向ではなく，どうしたら現実のアプリケーションで使える形になるかの研究にシフトしていくと予想している．

参考文献

1) 新納浩幸，村田真樹，白井清昭，福本文代，藤田早苗，佐々木稔，古宮嘉那子，乾　孝司：クラスタリングを利用した語義曖昧性解消の誤り原因のタイプ分け，自然言語処理，Vol.22, No.5：to appear (2015).

（2015 年 9 月 29 日受付）

☆1　正解はその「ボクシング」が試合として行われているのか，単なる運動として行われているかに依存する．

☆2　現状，正解率はおおむね 60 ～ 70％といったところである．

新納 浩幸（正会員）hiroyuki.shinnou.0828@vc.ibaraki.ac.jp
1985 年東京工業大学理学部情報科学科卒業．1987 年同大学院理工学研究科情報科学専攻修士課程修了．同年富士ゼロックス，翌年松下電器を経て，1993 年茨城大学工学部助手．2015 年同学部教授．現在に至る．