語義識別の誤り原因の調査とオンザフライの類似語判定

藤井丈明
茨城大学工学部
システム工学科

新納浩幸
茨城大学工学部
システム工学科

佐々木稔
茨城大学工学部
情報工学科

1 はじめに

本論文では語義識別の誤り原因の調査と on the fly の類似語判定を試みる．

一般に，曖昧性をもつ語義の識別では，人間が識別容易な問題は機械にとっても容易であり，人間が識別困難な問題は機械にとっても困難である．しかし Senseval2 の辞書タスクのいくつかの単語では，人間が識別容易でも機械にとっては識別困難であったことが報告されている [1]．その例として「開発」,「核」,「精神」,「乗る」,「生まれる」,「かかる」が挙げられている．本研究ではこの 6 単語を用いて実験を行なう．

ここでは，その原因の調査を遠方単語の影響と未知語の影響の 2 つの面から行なった．遠方単語を見ずに人間が識別した場合でも，見た場合と同様，正解率が機械よりも大きく高い．このことから遠方単語の影響は小さいと考えられる．また，トレーニングデータに未知語が減るように作為的に事例を数個増やした場合，正解率が大きく向上する．このことから，未知語の影響が大きいと考えられる．未知語の影響をなくすために，on the fly の類似語判定法を提案する．そこでは web を利用しての単語間の類似語を判定する．これを語義識別に応用する．

2 原因の特定

2.1 遠方単語

人間が正確に多義語の意味を識別できるのは，識別対象の多義語とその周辺単語だけではなく，文章全体を見ているからだとも考えられる．つまり識別対象の多義語から遠方にある単語（ここでは遠方単語と呼ぶ）をも識別の判断材料としているからだと考えられる．一方，多くの機械学習の手法では，遠方単語を考慮せずに，識別対象の多義語を中心とした周辺の数単語だけを見て語義識別を行なっている．つまり遠方単語が人間と機械の差を生じさせている原因として考えられる．

ここでは人間が機械学習の手法と同様に，識別対象の多義語を中心とした周辺単語だけを見て，つまり遠方単語を考慮しなくても多義語の識別が容易かどうかを調査する．このことにより，遠方単語の影響の有無を確認することができる．

また遠方単語の影響を調査するため，実際に機械学習により得られた規則を用いて，人間と機械の正解率を比較する必要がある．ここでは機械学習の手法として，Naive Bayes 法を用いる．その理由としては，Senseval2 の辞書タスクでは，Naive Bayes 法が最も良い成績を収めたためである．

3 人の人間が機械学習の手法と同様にして，遠方単語を見ずに語義識別を行なう．3 人の正解率の平均を，Naive Bayes 法によって得られた正解率と比較する．結果を表 1 に示す．

表 1: 人間と機械学習による正解率 (%) の比較

単語	人間	機械学習
開発	76.3	65.0
核	97.3	72.0
精神	84.0	64.0
乗る	82.3	66.0
生まれる	97.3	71.0
かかる	85.0	66.0

表 1 では人間の方が機械よりも正解率が，ある程度，高くなっている．これは人間が遠方単語を見て識別した場合と同様の傾向である．つまり人間が語義識別を行なう際には，遠方単語の影響は小さいと考えられる．

2.2 未知語

対象とした 6 単語の中で人間が正解し，機械が不正解したテストデータを調査した．その結果，人間が正解する場合でも単純に周辺単語にその語義を連想させる単語が出現するだけであることが多かった．機械がそのようなテスト文で不正解になるのは，その語義を連想させる単語が訓練データ中に存在しないためであった．つまりその単語が機械にとっては未知語であったためである．

このことを確認するために，訓練データに作為的に事例を 3 個増やして学習を行なってみる．その事例とは，先に述べた問題となる未知語を含んだ文である．具体的には以下の文である．

かかる 電話がかかった
お目にかかった
チームの命運がかかった重要な試合

開発 石油開発への貢献が関わってくる
油田開発への取り組み
ソフト開発を行なう

精神 立法の精神を活かす
日本的精神とその現在
精神的，経済的独立支援

のる 勢いにのる政党
脂がのった魚
電車にのって出かける

核 日本の核にいる人物
地域を核にした生活
マイネルト核から大脳皮質への投射系に障害がある

うまれる 芸術作品が生まれる時代
事態に動きが生まれる
演奏の時，生まれる音

訓練データ内に事例を増やす前と増やした後のテストデータに対する正解率を比較した．その結果が表 2 である．

表 2: データ追加前後の正解率 (%) の比較

単語	追加前	追加後
開発	65.0	70.0
核	72.0	75.0
精神	64.0	68.0
乗る	66.0	71.0
生まれる	71.0	75.0
かかる	66.0	70.0

事例を増やした後は，事例を増やす前よりもどの単語も正解率が 3〜5 % 向上した．このことから人間と機械の語義識別の正解率の差は，未知語の影響が大きいと考えられる．

3 on the fly の類似語判定

実験結果から，未知語の影響が大きいと考えられる．未知語は登録語数を増やすことで回避できそうだが，Zipf の法則によれば未知語は必ず出現するために，静的に単語を準備しておくアプローチには限界がある．

ここでは未知語の問題を解決するために，on the fly の類似語判定法を提案する．未知語が出現した段階で，与えられた単語群の中からその未知語と最も類似の単語を選択する手法である．

未知語が出現した段階で，未知語と単語群の各単語とが特徴ベクトルとして，表現することができれば，単純に距離を測るだけで問題の未知語と最も類似の単語を選択することができる．

問題はどのようにして，未知語が出現した段階で，その未知語の特徴ベクトルを得るかである．ここでは Web を利用する．

まず基底となる単語を 100 単語選出した $(v_1, v_2, \cdots, v_{100})$．これは論文 [3] の結果を考慮してアドホックに選んだものである．未知語 w の特徴ベクトルを得るのに，“w v_i” をクエリにして google から検索を行なう．そのヒット数を h_i とおく．そして未知語 w の特徴ベクトルを以下で表現する (図 1

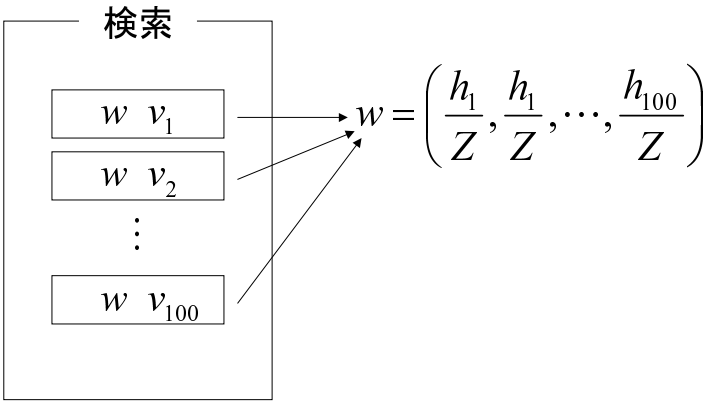

図 1: on the fly による特徴ベクトルの作成

参照).

$$w = (\frac{h_1}{Z}, \frac{h_2}{Z}, \cdots, \frac{h_i}{Z}, \cdots, \frac{h_{100}}{Z})$$

ここで Z は w を正規化する定数である.

$$Z = \sqrt{\sum_{i=1}^{100} h_i^2}$$

4 実験

ここでは on the fly の類似語判定を語義識別に応用した実験を行なう.

名詞の語義識別ではその名詞が複合語になっていれば，その名詞の直前あるいは直後の単語によってほぼ語義識別が可能であることが知られている．そこでテスト事例の問題の多義語が複合語になっていた場合には，その複合語を訓練データから探し，もし存在すれば，対応する語義を識別結果とする．そしてもしも存在しなかった場合に on the fly の類似語判定を行なう．この方法は論文 [2] の複合語を優先して識別を行なう手法と基本的に同じである.

例えばテスト事例の問題の多義語 q が複合語になっており問題の多義語の直前の単語が w だとする．今，“wq'' という複合語は訓練データ中に存在しない．しかし “x_1q''，“x_2q''，...，“x_nq'' という複合語は訓練データ中に存在するとする．ここで未知語を w，対象の単語群を $\{x_i\}$ として先に説明した on the fly の類似語判定を行なう．最も類似の単語が $\hat{x}$ であったとき，その “$\hat{x}q''$ を含む訓練データに対する語義を識別結果として返す.

ここでは「開発」に対して実験を行なった.「開発」のテストデータ (100 問) の中で「開発」が複合語の一部となっていたのは 71 問であった．このうち 21 問はその複合語が訓練データ中に存在した．この 21 問に対してはその複合語を含む訓練データの対応する語義が識別結果となる．正解数を以下に示す．参考のためにこの 21 問に対する Naive Bayes での正解数も示す.

表 3: 複合語を利用した識別

手法	正解数	不正解数
Naive Bayes	16	5
複合語の利用	17	4

「開発」のテストデータ (100 問) の中で「開発」が複合語の一部となっていたもので，その複合語を構成するもう一方の単語が未知語になっているものは，50 問である．この 50 問が on the fly の類似語判定の実験対象である．正解数を以下に示す．参考のためにこの 50 問に対する Naive Bayes での正解数も示す.

表 4: on the fly の類似語判定を利用した識別

手法	正解数	不正解数
Naive Bayes	19	31
on the fly	25	25

本方式を用いたことで「開発」の正解率は 65% から 72% に向上した．人間の正解率が 76% なので，その差は小さくなっている．表 4 では on the fly の類似語判定を利用した識別の正解率が 50% であり，この部分の精度を向上させることができれば，人間の正解率と同等になることも期待できる．

5 考察

今回の実験で，遠方単語の影響が小さいことがわかった．これは機械学習の手法に遠方単語の素性を取り入れなくても，十分な語義識別が可能であることを示している．

また，未知語の影響が大きいこともわかった．語義識別の精度向上には未知語の解決が重要だと考えられる．通常はシソーラスの作成により未知語の問題の解決が図られるが，未知語は必ず出現するために，on the fly による未知語の解決が望ましい．ここで提案した手法は Web を利用しているために未知語の解決が図れる可能性はあるが，まだいくつかの課題がある．

第 1 に，基底の単語の設定が挙げられる．今回は基底の単語として，アドホックに選定した 100 単語を用いたが，この 100 単語が基底として適しているかどうかは更に実験を重ねて確認する必要がある．

第 2 に，処理時間の問題が挙げられる．ここでは google を用いて検索を行なっているが，1 つの未知語に対して 100 回の検索を行なっている．このために処理時間がかかっている．次元数を増やすと更に処理時間がかかる．今後，処理時間を減らす方法も考える必要がある．

第 3 に，応用性の問題が挙げられる．今回は「開発」という一単語の名詞のみで実験を行なったが，識別対象の多義語が動詞であった場合は，今回のように複合語となる単語から語義識別を行なう方法を用いることはできない．そのため，複合語の代わりに識別対象の多義語の周辺単語を調べ，その共起頻度を得て on the fly の類似語判定を行なえば良いと考えている．これを確認し，他の品詞にも on the fly の類似語判定が有用であるかを確かめる必要がある．

6 おわりに

本論文では語義識別の誤り原因の調査と on the fly の類似語判定の提案を行なった．

語義識別の誤り原因として，遠方単語の影響および未知語の影響の 2 つの面から調査を行なった．結果，遠方単語の影響は小さく，未知語の影響は大きいことが確認できた．

またここでは未知語の問題を解決するために on the fly の類似語判定法を提案した．多義語の「開発」を用いた実験ではその効果を確認できた．

今後の課題としては，基底の単語の選定や処理時間の軽減の問題がある．また動詞への拡張や大規模な実験なども必要である．

参考文献

[1] 白井清昭： “SENSEVAL-2 日本語辞書タスク”, 自然言語処理, Vol.10, No.3, pp.3-24 (2003).

[2] 新納浩幸： “複合語からの証拠に重みをつけた決定リストによる同音異義語判別”, 情報処理学会論文誌, Vol.39, No.12, pp.3200-3206 (1998).

[3] 大城亜里沙，新納浩幸，佐々木稔： “検索エンジンを利用した単語クラスタリング”, 言語処理学会第 10 回年次大会, to appear, (2004).