シソーラス・文脈特徴空間の反復カーネル密度推定による教師なし語義曖昧性解消

谷垣 宏一　　柴　光輝　　高山 茂伸　　撫中 達司

三菱電機(株) 情報技術総合研究所

{Tanigaki.Koichi@ap, Shiba.Mitsuteru@ap, Takayama.Shigenobu@db, Munaka.Tatsuji@ct}.MitsubishiElectric.co.jp

1. はじめに

語義の曖昧性解消は，機械翻訳，情報検索など広範な分野に応用される基礎技術である[1]．本技術の応用先の一つとして，著者らは情報システムのデータベース・スキーマ照合[9]を検討している．大規模な企業・官公庁向け情報システムでは，データベース群の統合・整理がしばしば行われるが，その際，数万からの項目名（データベースのカラム名）を突き合わせて，統合する項目を調べ上げる作業が行われる．ところが項目名は，文字列長が制限された英数字で表記されることから任意の短縮表記が使用され，また日英語の混用やローマ字綴りによる曖昧性の増加と相まって，一見してそれが指し示す内容を想起しにくいものも少なくない．そうした項目を含むスキーマ照合作業の負荷軽減・自動化に，語義曖昧性解消を有効に適用可能と考えている．

実応用を考えるとき，語義曖昧性解消の学習データ確保はしばしば重要な問題となる[2]．特にスキーマ照合では，対象データを顧客情報システムから取得する必要があるため，1)入手可能なデータ量が限定される，2)内容や表現が業種・アプリケーション・設計者に依存する，3)アノテーションのような予備作業を実施する期間やコストを確保することが難しい，といった事情により，十分な量のラベル付き学習データを利用することは困難である．そこで本稿では，比較的少量のラベルなしデータと，シソーラスで定義された語義間の類似性を利用する教師なし学習により，語義の曖昧性を解消する方式を検討した．

語義曖昧性解消に関する先行研究のうち，本稿と同様教師なし，ないし半教師あり学習を用いる方法としては，bootstrapping法で学習事例を獲得していく方法（[7]など），語の出現文脈をクラスタリングする方法（[4] など），語の共起グラフにリンク解析を適用する方法（[8]など），noisy channel modelを用いる方法（[5] など）がある[2]．また，本稿と同様に辞書知識を利用する方法としては，語義の定義文を利用する方法や[2]，語義階層間のパス長を利用する方法（[6]など）などがある．

これに対し本稿では，疎なデータを平滑化し，頑健な語義推定を実現する方式として，文脈特徴空間と語義特徴空間に跨がるガウシアンカーネルを用いる方式を提案する．更に，カーネルに語義の割り当て確率で重み付けを行い，その重みをEMアルゴリズムを適用して推定することにより，全語の語義曖昧性を漸進的に解消する教師なし学習方式を提案する．また本稿では評価実験より得られた知見として，提案法による教師なし学習の信頼性が，曖昧性解消に用いた異なり語のパープレキシティと高い相関があることを述べる．

2. 語義曖昧性解消方式

本方式による語義曖昧性解消の基本的な考え方は，言い換え語が複数得られれば，それら言い換え語の間で語義の候補集合が異なることを利用することにより，教師なしでも正しい語義を見つけることができる，というものである．すなわち，正しい語義は全ての言い換え語において語義候補として出現するのに対し，その他の語義候補は語に依存してまばらに出現すると仮定し，その出現頻度の偏りを利用して正しい語義を推定する．ただし，実際にはスパースなデータセットから本来の言い換え語を得ることは困難であるから，より広い語を対象とした指針として，類似した文脈に出現する語の間では類似した語義に偏った確率分布が得られるほど尤もらしいと仮定する．

このような仮定に基づく語義曖昧性解消方式として，本稿では次の２つを提案する．1)ガウシアンカーネルによる文脈の類似性と語義階層の類似性の関係のモデル，2) EMアルゴリズムによる全語の同時・漸進的曖昧性解消．本方式の概要を図1に示す．各単語出現インスタンスの文脈は，文脈特徴空間上の位置（緑のライン）で表し，語義候補は，語義特徴空間上の複数の位置（緑のライン上にある赤矢印の先端）で表す．図ではそれぞれの特徴空間を模式的に1次元で示している．各語義の割り当て確率（初期値は一様に与える）を高さとしてガウシアンカーネルを配置し，それらカーネルの確率分布を重ね合わせることにより，類似した文脈に出現する類似した語義候補の尤度を高めることができる（図1左）．さらに，重ね合わせた山の高さの比に合わせて各語の語義割り当て確率を反復更新することにより，他の語と類似した語義候補を強化し，全語の曖昧性を漸進的に解消する（図1右）．このようにして，緩やかな換言語のクラスタの中で正しい語義を見つけることができる．

2.1. 語義と文脈の関係のモデル

データセットの単語出現インスタンスを$x \in X$ とし，出現文脈を無視したx の単語の種類を$w \in V$ （V はデー

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

タセットの語彙），wに対する語義の候補を$s \in S_w$とする．xの語義は次式により求めることができる．

$$s^* := \arg\max_{s \in S_w} p(s|x) \tag{1}$$

本稿では，タスク・ドメインを限定した場合，語wは一定の範疇の文脈のみで出現し，それゆえ$p(s|x)$はデータセット中の出現文脈に依らず一定であると仮定する．

$$p(s|x) \simeq p(s|w) \tag{2}$$

ところで本稿で対象とする語には，表記の短縮やローマ字綴りの揺れなどにより，文字列完全一致では辞書の見出し語と対応を取ることができない語が含まれる．そこで，$p(s|w)$を，辞書中で語義sを持ついずれかの見出し語とwが対応する確率λ_s^w $(0 \leq \lambda_s^w \leq 1)$，および，$w$が実際に語義$s$の意味で用いられている確率$\pi_s^w$ $(\sum_{s \in S_w} \pi_s^w = 1)$とに分離する．

$$p(s|w) := \lambda_s^w \pi_s^w \tag{3}$$

式(3)のλ_s^wは，辞書中で語義sを持つ見出し語$w_i \in V_s$のうち，wと最も近い語との文字列一致度で求める．文字列一致度にはギャップ重み付き部分列カーネル[3]を，文字列長の正規化を加えて用いた．

$$\lambda_s^w := \max_{w_i \in V_s} \frac{\mathcal{K}_{\mathrm{GWS}}(w, w_i)}{\sqrt{\mathcal{K}_{\mathrm{GWS}}(w,w)\,\mathcal{K}_{\mathrm{GWS}}(w_i,w_i)}} \tag{4}$$

一方，式(3)のπ_s^wは，データセット中の各単語出現インスタンスx_iの語義候補s_jに対する尤度$p(s_j|x_i)$を重みとするカーネル密度推定により，次式(5)で求める．ただし，$p(s_j|x_i)$には(3)式より別の語の$\pi_{s_j}^{w_i}$が含まれており，式(5)は循環定義となっている．これらπの計算方法については次節で述べる．

$$\pi_s^w \propto \sum_{x_i \in X} \sum_{s_j \in S_{w_i}} p(s_j|x_i)\,\mathcal{K}_{\mathrm{Gauss}}(x, s, x_i, s_j) \tag{5}$$

式(5)のガウシアンカーネル$\mathcal{K}_{\mathrm{Gauss}}(\cdot)$は，注目する2つの単語出現インスタンス$x, x_i$の文脈特徴ベクトル$\boldsymbol{\phi}_c(x)$, $\boldsymbol{\phi}_c(x_i)$，およびそれぞれの語義候補s, s_jの語義特徴ベクトルの距離$\boldsymbol{\phi}_s(s), \boldsymbol{\phi}_s(s_j)$により，次式(6)で定義する．式中の分散$\sigma_s^2$, σ_c^2は実験的に定めるパラメータである．これらのパラメータで文脈・語義それぞれに任意に平滑化した確率分布を構成し，語義を推定する．

$$\mathcal{K}_{\mathrm{Gauss}}(x, s, x_i, s_j) := \exp\left(-\frac{\|\boldsymbol{\phi}_c(x) - \boldsymbol{\phi}_c(x_i)\|^2}{\sigma_c^2} - \frac{\|\boldsymbol{\phi}_s(s) - \boldsymbol{\phi}_s(s_j)\|^2}{\sigma_s^2}\right) \tag{6}$$

後述の評価実験では，文脈特徴ベクトルとして，注目する語の直前・直後に出現する語，および，上位要素（テーブル名）・下位要素（カラム名）に出現する語を素性とするベクトルを用いた．語義特徴ベクトルとしては，注目する語義候補（WordNetシンセット[10]）とその全ての上位シンセットを素性とするベクトルを用いた．

2.2. 全語の漸進的曖昧性解消

語義割り当て確率π_s^wは、以下のようにEMアルゴリズムを適用することにより、全ての分類対象語について同時に推定することができる。

- Step 1（初期化）：全ての単語$w \in V$に対し、語義割り当て確率π_s^wに初期値$1/|S_w|$を設定する。$|S_w|$はwの語義候補集合S_wの要素数を表す。
- Step 2（Eステップ）：現在の語義割り当て確率${}^{(\mathrm{old})}\pi_s^w$による条件付き確率$p(s|x)$を、式(3)により全ての単語出現サンプル$x$とその語義候補$s$に対して求める。

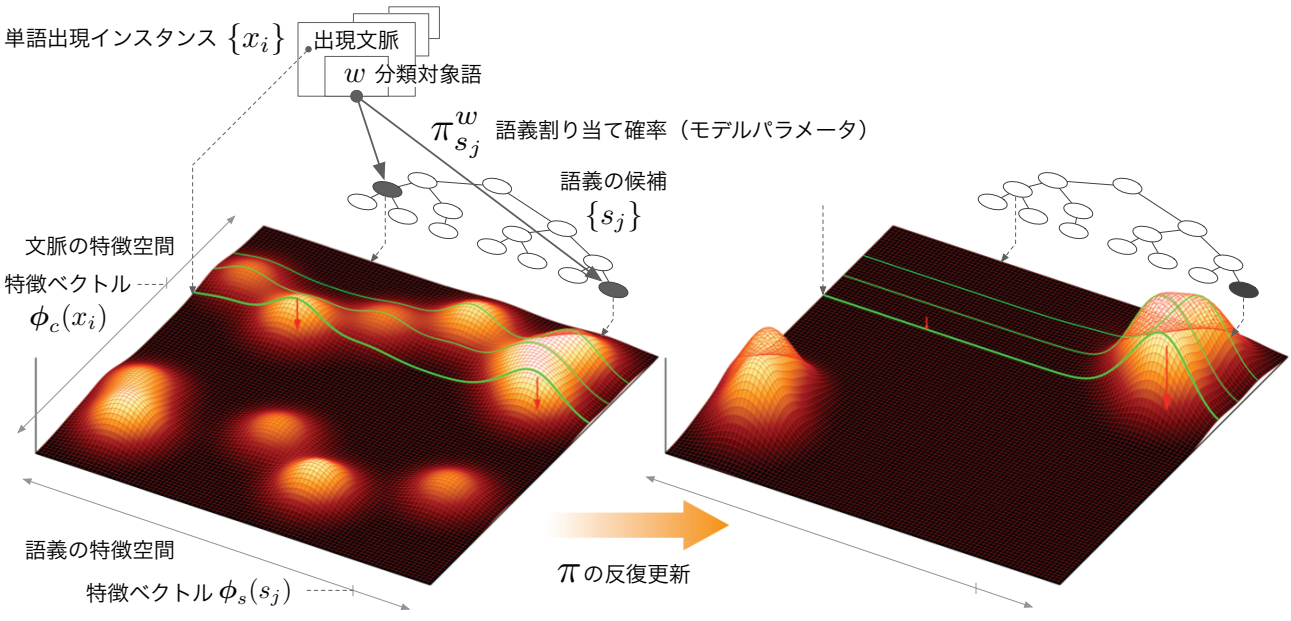

図1. 語義曖昧性解消モデル

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

- Step 3（Mステップ）：語義割り当て確率を次式により更新する。

$$ {}^{(\mathrm{new})}\pi_s^w := \frac{\sum_{x_i \in X_w} p(s|x_i)}{\sum_{x_i \in X_w} \sum_{s_j \in S_w} p(s_j|x_i)} \tag{7} $$

- Step 4（収束判定）：全ての分類対象語に対する尤度$\mathcal{L}$を次式で求め、$\mathcal{L}$の収束により学習を終了する．未収束ならStep 2へ戻り，π_s^wの更新を反復する。

$$ \mathcal{L} := \sum_{x_i \in X} \sum_{s_j \in S_{w_i}} \log p(s_j|x_i) \tag{8} $$

3. 評価実験

3.1. 実験条件

実験に使用したデータは，データベースのテーブルスキーマである．本データは，資材発注システムを想定したデータベースの模擬設計を社内で行い，情報システム技術者33名より収集したものである[1]．収集データの一部を表1に示す．これらテーブル名，カラム名を単語単位に分割し，各語の語義を推定する．単語分割は，人手により決定した正しい分割を与えて評価を行った．

語義としては，日本語WordNet[10]（1.1版）のシンセットIDを用い，正解を人手により付与して用いた．本実験で用いる語には，ローマ字綴り日本語や，短縮表記された語が含まれ，WordNetの見出し語と直接対応しないものがある．そこで，WordNetの日本語単語エントリにMeCab[11]で読みを付与し、読みをローマ字綴りに変換して見出し語の代わりに参照することで，語義の候補を抽出した．また，表記の短縮や活用により，出現形そのままではWordNetエントリと対応が取れない語については，本データセットから抽出した語で，WordNetエントリと完全一致する他の語の中から，先頭1字が一致する語を抽出し，式(4)の文字列一致度（長さの減衰係数は0.5）が0.5以上の語の語義を候補として使用した．

以上のようにして，語義の候補が得られた3,032語を分類対象として語義を推定し，そのうち，正解語義を候補に含む2,368語を精度評価対象とする．本データの詳細を表2に示す．なお，精度評価対象2,368語のうち，316語（異なり19語）は語義候補が1つであり，曖昧性を持たない．

分類に使う文脈の特徴ベクトルは，表2の文脈語に対しPorterのアルゴリズム[12]によるステミングを行って生成した．特徴ベクトルの次元数は文脈が616，語義が2,068である．頻度による選択や重み付け，次元圧縮は行っていない．

評価では以下i〜iiiの3つの手法を比較する．i)「提案法・EMあり」：3章で述べたカーネル密度推定とEMアルゴリズムによる語義割り当て確率の反復更新の双方を用いる．ii)「提案法・EMなし」：EMアルゴリズムによるπの反復更新を行わず，Step 1の初期化とStep 2のカーネル密度推定のみを行った状態で語義を推定する．iii)「ベースライン(k-NN)」：文脈特徴ベクトルの距離（内積）によりk近傍の単語出現インスタンスを抽出し，それらの語義候補となっているシンセットIDの頻度比を語義の尤度とする．各手法により候補語義の尤度$p(s|x)$を推定し，尤度に閾値を設定したときのF値を比較する．なお，提案法のパラメータσ_s^2 ,σ_c^2 は，それぞれの特徴空間における全仮説間距離の分散$\bar{\sigma_s^2}$,$\bar{\sigma_c^2}$ の 1×10^n倍， 5×10^n倍 (n∈{–4, –3, –2, –1, 0, 1}) を評価した．k近傍法のパラメータkは 1×10^n , 5×10^n (n ∈ {0, 1, 2, 3})にて評価を行った．

3.2. 実験結果・考察

提案法i, iiではパラメータを$0.05\,\bar{\sigma_s^2}$ ，$1.0\,\bar{\sigma_c^2}$ としたときにF値の最大値が得られ，k近傍法ではk=1,000にてF値の最大値が得られた．このときの再現率・適合率曲線を図2に示す．母数は表2の単語出現インスタンス2,368語である．

各手法の最大F値は，k近傍法で0.407，提案法i，iiでそれぞれ0.611，0.627であり，提案法i，iiによりk近傍法を上回る精度が得られた．k近傍法では，kを1,000まで広げた条件で最大精度が得られていることからデータのスパースネスに対応できておらず，これに対し，提案法i，iiでは特に語義側の平滑化が有効に機能した結果と考える．

一方，提案法iとiiを比較するとその差は僅かであり，本実験ではEMアルゴリズムの適用による曖昧性解消効果は限定的であった．この原因としては，以下に考察するように，一種の過学習が考えられる．2章冒頭で述べたように，提案法による教師なし学習は，類似した文脈で出現する言い換え語（異なり語）を利用して，正しい語義を見つける方式であるため，周囲の少数の異なり語の語義候

表1. 実験データの例

テーブル名	カラム名
HATCHUU	HATCHUU_ID, YOSAN_GAKU, NYUURYOKU_DATE, HENKOU_DATE, RIYUU_SHUBETSU_CD, SHAIN_ID, NOUNYUU_YOTEI_DATE, MITSUMORI_GAKU
HINMOKU_MASTER	HINMOKU_ID, HINMOKU_NAME
SHAIN_MASTER	SHAIN_ID, SHAIN_NAME_SEI, SHAIN_NAME_MEI

表2. 実験データ

	分類対象データセット			
	精度評価対象			
データセット	総数	異なり	総数	異なり
単語	2,368	196	3,032	251
正解語義	2,406	122	2,406	122
候補語義	23,316	1,012	28,381	1,104
単語出現インスタンス	平均	分散	平均	分散
正解語義	1.0	0.0158	0.79	0.189
候補語義	9.8	57.6	9.4	65.7
異なり文脈語	3.5	9.3	3.3	7.8

[1] 項目数は1,327であり，単純な比較はできないが，冒頭に述べた大規模情報システムの数十分の1程度のデータセットとなっている．

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

補に基づいて語義を推定した場合には，信頼性が低くなると思われる．語w　の曖昧性解消に使われた異なり語数の尺度としては，パープレキシティ $\mathrm{PP}_w = 2^{H_w}$　を用いることができる．ここでエントロピーH_w は，

$$H_w := \sum_{w_k \in V} p(w_k|w) \log_2 p(w_k|w) \quad (9)$$

であり，異なり語の条件付き確率 $p(w_k|w)$ は，

$$p(w_k|w) := \frac{1}{Z} \sum_{x \in X_w} \sum_{s \in S_w} \lambda_s^x \sum_{x_i \in X_{w_k}} \sum_{s_j \in S_{w_k}} \lambda_{s_j}^{x_i} \pi_{s_j}^{w_k} \mathcal{K}_{\mathrm{Gauss}}(x, s, x_i, s_j) \quad (10)$$

で定める．パープレキシティ PP_w と，EM適用による正解語義の尤度変化の関係を図3右に示す．また参考のため，EMを適用しない，初回カーネル密度推定のみの効果（提案法iiに相当）を図3左に示した．まず図3右より，上述の予想通り，パープレキシティが低い語では尤度が大きく下がことが多く，学習が不安定であったことがわかる．したがって，パープレキシティを教師なし学習時の信頼性の尺度として利用し，異なり語ごとに学習の強さを制御する

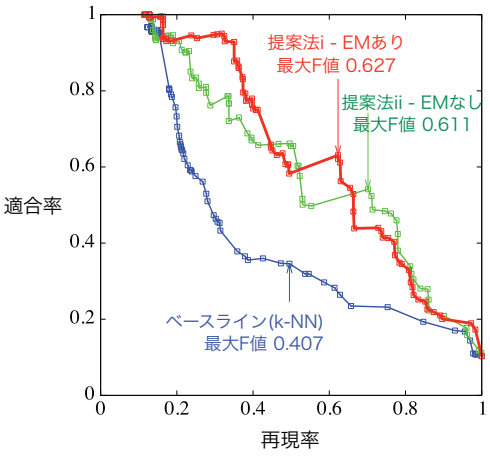

図2. 語義曖昧性解消精度の比較

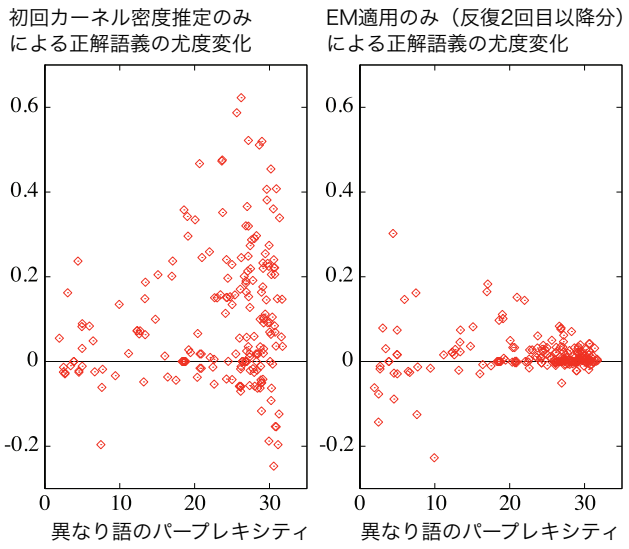

図3. 曖昧性解消に使われた異なり語の数と学習効果

方法で改善が期待できる．一方，パープレキシティが高い語では，図3右より，その幅は小さいもののほとんどの語で安定して正解語義の尤度が上がっていたことがわかる．したがって上述の方法で学習の強さを上げることでも一定の改善が期待できる．ただしこれらの語では，図3左で既に尤度が上がっていることから，少数に絞り込まれた語義候補の中で識別ができなかったものも含まれる．これに対しては，同様の信頼度をより細かく，語義の単位で考慮して学習の強さを制御することで改善の可能性がある．

4. おわりに

シソーラスと文脈特徴空間上の反復カーネル密度推定による教師なし語義曖昧性解消方式を提案し，本方式がスパースなデータにおいてk近傍法より頑健な語義曖昧性解消方式であることを示した．また，本方式による学習の信頼性が，曖昧性解消に使った異なり語の数と高い相関を持つことを示し，精度向上の可能性について述べた．今後，本知見に基づく改良を検討する．また，SemEvalなどの公開データセットを用いた評価を行いたい．

参考文献

[1] N. Ide and J. Veronis. “Introduction to the special issue on word sense disambiguation”. *Computational Linguistics*, Vol. 24, No. 1, pp. 1–40, 1998.

[2] R. Navigli. “Word Sense Disambiguation: a Survey”. *ACM Computing Surveys*, 41(2), ACM Press, pp. 1-69, 2009.

[3] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini and C. Watkins. “Text Classification using String Kernels”. *Journal of Machine Learning Research*, vol. 2, pp. 419–444, 2002.

[4] A. Purundare and T. Pedersen. “Word sense discrimination by clustering context in vector and similarity spaces”. In *Proc. of CoNLL*, pp. 41−48, 2004.

[5] D. Yuret and M. Yatbaz. “The Noisy Channel Model for Unsupervised Word Sense Disambiguation”. *Computational Linguistics*, Volume 36, Number 1, 2010.

[6] T. Pedersen and V. Kolhatkar. “WordNet::Sense-Relate::AllWords - a broad coverage word sense tagger that maximizes semantic relatedness”. In *Proc. of NAACL HLT*, pp. 17−20, 2009.

[7] D. Yarowsky. “Unsupervised word sense disambiguation rivaling supervised methods”. In *Proc. of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 189–196, 1995.

[8] E. Agirre, D. Martinez, O. Lacalle, and A. Soroa. “Two graph-based algorithms for state-of-the-art wsd”. In *Proc. of EMNLP HLP*, pp. 585–593, 2006.

[9] E. Rahm and P. Bernstein. “A survey of approaches to automatic schema matching”. *International Journal on Very Large Data Bases*, 10(4), pp. 334–350, 2001.

[10] 栗林, Bond, 黒田, 内元, 井佐原, 神崎, 鳥澤. “日本語ワードネット1.0”. 言語処理学会第16回年次大会, pp. 978−981, 2010.

[11] MeCab: Yet Another Part-of-Speech and Morphological Analyzer. http://mecab.sourceforge.net/

[12] Porter Stemming Algorithm. http://tartarus.org/~martin/PorterStemmer/

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved