ナイーブベイズの拡張による確率的概念辞書を用いた固有表現のクラス推定

白川 真澄[1]　Haixun Wang[2]　Yangqiu Song[2]　Zhongyuan Wang[2]
中山 浩太郎[3]　原 隆浩[1]　西尾 章治郎[1]

[1] 大阪大学 大学院情報科学研究科
[2] マイクロソフトリサーチアジア
[3] 東京大学 知の構造化センター

{shirakawa.masumi, hara, nishio}@ist.osaka-u.ac.jp
{Haixun.Wang, yangqiu.song, zhy.wang}@microsoft.com
nakayama@cks.u-tokyo.ac.jp

1　はじめに

自然言語で記述されたテキストの意味を計算機に理解させることは，情報科学において最も重要な課題の一つである．その中でも，テキスト中に出現する固有表現を検出・特定するタスクは，テキストの意味解析を行うにあたって根幹ともいえるタスクであり，これまで数多くの研究が行われてきた．検出した固有名詞の意味を特定する方法は大きく二種類に分けられ，クラス（Person，Organization，Location など）を特定する固有表現抽出（Named Entity Recognition）と，オントロジー辞書で定義されたエンティティにマッピングする曖昧性解消（Entity Disambiguation または Entity Linking）がある．

一方，人間が自然言語で記述されたテキストの意味を理解しようとしたとき，重要なプロセスとなるのは「概念化（Conceptualization）」である．心理学者 Gregory Murphy が著書 [5] で “*Concepts are the glue that holds our mental world together.*” と述べているように，概念は意味世界とテキスト世界を繋ぐカギとなるものである．このような知見に基づき，我々の研究ではこれまで，個々の語句に対して上位概念を確率的に定義した概念辞書を基盤知識として，ナイーブベイズを用いて複数の語句を概念化する手法を提案してきた [7]．この手法では，例えば，入力が “India” と “China” である場合，それらの概念（クラス）として “Asian country” や “developing country” などが推測される．さらに “Brazil” を入力に加えると，推測される概念は “BRIC” や “emerging market” に変化する．このように，入力系列に対し，コンテキストを考慮したクラス推定が可能である．

本研究では，上記のナイーブベイズによる概念化手法を用いて，テキスト中に出現する個々の固有表現の概念（クラス）を推定する．通常，一つの文書には複数の概念が存在するため，それらをうまく切り分けて概念化する必要がある．そこで，テキスト中に出現する個々の固有名詞について，1) 関連する（すなわち同じクラスに属すると思われる）語句を文書中から選択し，2) それらの語句を入力系列としてナイーブベイズを用いてクラス推定を行う．提案手法では，これらの二段階のプロセスをシームレスに実行することにより，閾値などのパラメータを必要としない安定したクラス推定を行う．

2　大規模確率的概念辞書 Probase

本研究で基盤知識として用いる Probase [9] は，16 億以上の Web ページを解析して構築された大規模な概念辞書であり，基本的には “... such as ...” や “... is a ...” などの Hearst パターン [2] を用いて上位下位関係を取得している．Probase の大きな特徴として，あるエンティティに対する上位概念（クラス）を確率的に定義している点が挙げられる．すなわち，エンティティ e のクラス c を確率 $P(c|e)$ として定義している．また，膨大な概念を定義している点も特徴として挙げられる．現時点では 800 万の概念を定義しており，Wikipedia をベースとした大規模オントロジー辞書である YAGO [8] の概念数が 35 万程度であることと比較しても，Probase の概念空間の大きさが確認できる．

エンティティ e がクラス c に属する確率 $P(c|e)$ は，Hearst パターンにおける出現頻度に比例して与え，$\sum_k P(c_k|e) = 1$ となるように正規化する．実際には，語句 t が与えられたときに，それがエンティティ $t^{(e)}$ だけでなくクラス $t^{(c)}$ そのものを意味する可能性も考慮し，以下の式によりクラス c_k を推定する．

$$P(c_k|t) = P(c_k|t^{(e)})P(t^{(e)}|t) + P(c_k|t^{(c)})P(t^{(c)}|t) \quad (1)$$

語句 t がエンティティ e を意味する確率 $P(t^{(e)}|t)$ およびクラス c を意味する確率 $P(t^{(c)}|t)$ は，それぞれの Hearst パターンにおける出現頻度から算出する．なお，$P(c_k|t^{(c)})$ は $c_k = t^{(c)}$ のときのみ 1 をとり，それ以外のときは 0 をとる．また，3 章の提案手法では事前確率 $P(c_k)$ も使用するが，これについても，Hearst パターンにおける出現頻度から算出する．

3 ナイーブベイズの拡張によるクラス推定

本研究が取り上げる問題は，ある文書から抽出された語句集合 $T = \{t_l, l \in 1, ..., L\}$ が与えられた時に，語句 $t \in T$ が属するクラス c_k を決定することである．また，$P(c_k|t)$ および $P(c_k)$ は与えられるものとする．

語句間は互いに独立であるという仮定を置くと，T に含まれる複数の語句がある一つのクラス c_k に属する可能性があるとき，それらの語句は，各語句が単独で出現した場合と比べて c_k に属する可能性が高くなる．つまり，ある語句 t のクラスを決定する際，同じ文書中にある t の関連語句（同じクラスに属していると思われる語句）を利用することで，t のクラス推定精度が向上すると考えられる．また，一つの文書は通常複数のクラスを含むため，t と関連のない語句も T に含まれている．t と関連のない語句を利用した場合，その情報は t のクラスを推定する上でノイズとなり，結果として t のクラス推定精度は低下する．

したがって，ある語句 $t \in T$ に注目しているとき，1) t に関連している語句を T から選出し，T のサブセット T_t を決定した後，2) そのサブセットを入力とし，ナイーブベイズ $P(c_k|T_t)$ によってクラス推定を行う．ここで，仮に T_t が簡単に決定できる場合，すなわち t と関連のある語句が明らかな場合，以下のナイーブベイズの式を用いてクラス推定が可能である [7].

$$P(c_k|T_t) \propto P(c_k) \prod_{t_l \in T_t} P(t_l|c_k) \propto \frac{\prod_{t_l \in T_t} P(c_k|t_l)}{P(c_k)^{|T_t|-1}} \quad (2)$$

なお，$|T_t|$ は T_t に含まれる語句数である．

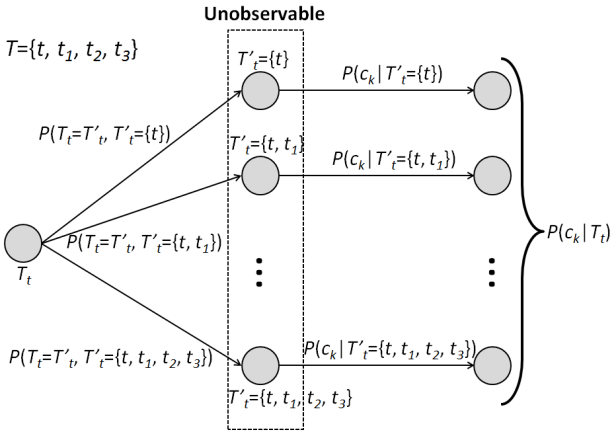

図 1: 要素が観測できない T_t に対するナイーブベイズの適用

ここで問題となるのは，T_t に含まれる語句をどのように決定するかということである．簡単な方法として，例えば，t と他の語句について，コサインや PMI などを用いてクラスの関連度を算出した後，閾値以上の語句のみを関連しているとみなす方法がある．4 章の評価実験より，適切な閾値が設定されている場合，高精度でクラス推定が可能であることが分かっている．しかし，適切な閾値は対象の文書ごとに異なるため，事前に決定することは難しい．実際，準備実験において，最も高い精度を達成した閾値では誤ったクラスを推定していた一方，別の閾値では正しいクラスを推定できていたケースも存在していた．

そこで提案手法では，T_t に含まれる語句の選択とナイーブベイズの適用 $P(c_k|T_t)$ をシームレスに行うことを考える．図 1 は提案手法を表している．中身の分からない語句集合 T_t が与えられたときに，関連度の閾値によって T_t の状態を決定する代わりに，全ての T_t の状態 T'_t を確率的に定義し，それぞれの状態 T'_t についてナイーブベイズを適用する[1]．まず，T_t が状態 T'_t である確率を以下の式により定義する．

$$\begin{aligned} P(T_t = T'_t) &= \prod_{t_l \in T'_t} P(t_l \in T_t) \prod_{t_l \notin T'_t} P(t_l \notin T_t) \\ &= \prod_{t_l \in T'_t} P(t_l \in T_t) \prod_{t_l \notin T'_t} (1 - P(t_l \in T_t)) \end{aligned} \quad (3)$$

t_l が T_t に含まれる確率 $P(t_l \in T_t)$ の算出方法については様々なものが考えられるが，4 章の評価実験では，精度の安定していたコサインによるスコアをそのまま確率として用いた．なお，$P(t_l \in T_t)$ の算出については今後，より理論的な方法を検討する必要がある．

[1]以下，T'_t とアポストロフィを付けた場合，中身が観測可能な語句集合を表すものとする．

図 1 に示す提案手法の拡張ナイーブベイズは，式 (2) と式 (3) より，以下のように表される．

$$P(c_k|T_t) \propto \sum_{T'_t}\left(P(T_t = T'_t)\frac{\prod_{t_l \in T'_t} P(c_k|t_l)}{P(c_k)^{|T'_t|-1}}\right) \quad (4)$$

上式を計算しようとすると，T の語句数に対して指数関数的に計算量が増加する．これは，すべての T'_t に対してナイーブベイズを適用しているためである．ここで，$t_l \in T'_t$ の場合と $t_l \notin T'_t$ の場合について整理すると，

$$\frac{\sum_{T'_t}\left(\prod_{t_l \in T'_t} P(t_l \in T_t)P(c_k|t_l)\prod_{t_l \notin T'_t}(1-P(t_l \in T_t))P(c_k)\right)}{P(c_k)^{L-1}} \quad (5)$$

上式の分子を t_l ごとに分解することにより，和集合部分を効率よく計算できる．その結果，以下の式が導かれる．

$$P(c_k|T_t) \propto \frac{\prod_{l=1}^{L}\left(P(t_l \in T_t)P(c_k|t_l) + \left(1 - P(t_l \in T_t)\right)P(c_k)\right)}{P(c_k)^{L-1}} \quad (6)$$

式 (6) は，通常のナイーブベイズの式 (2) における個々のクラス推定の確率 $P(c_k|t_l)$ を，$P(c_k|t_l)$ と事前確率 $P(c_k)$ の線形結合に置き換えたものであり，関連度 $P(t_l \in T_t)$ に応じてその比重が決定する（図 2）．結果的に，$P(t_l \in T_t)$ はスムージングの比重を決定するための係数の役割を果たしている．つまり，t_l が t に関連している場合は $P(c_k|t_l)$，関連していない場合は $P(c_k)$（何も与えられていない状態）であることを表している．関連度が 1，すなわち $P(c_k|t_l)$ と $P(c_k|t)$ が全ての c_k において同じ確率である場合，そのまま $P(c_k|t_l)$ が適用される．t_l と t の関連度が小さくなるほど $P(c_k|t_l)$ の影響が小さくなり，ナイーブベイズによるクラス推定の結果に及ぼす影響が小さくなる．全く関連がない語句（関連度が 0）の場合，分子は $P(c_k)$ となり，分母と相殺されるため，クラス推定の結果に全く影響を与えなくなる．また，t と関連のある他の語句が T に全く含まれていない場合，$P(c_k|t)$ そのものが式 (6) によるクラス推定の結果となり，直感的な結果と一致する．

4 評価実験

3 章で説明した提案手法について，エンティティの曖昧性解消タスクにおいて評価を行った．データセットには CoNLL2003 固有表現抽出タスク [6][2] で用いられたコーパス（Reuters Corpus RCV1 の一部，文書数 231）を利用した．正解データとして，このコーパスを曖昧性解消タスク用にタグ付けしたデータを用いた [3][3]．この正解データでは，データセットのコーパスから固有表現としてタグ付けされている語句を，Wikipedia 及び Freebase の該当するエンティティの ID にマッピングしている．ただし，該当するエンティティが存在しない場合は「該当なし」として評価の対象外としている．提案手法では推定したクラスが得られるが，このままでは比較できないため，得られたクラスと Freebase のカテゴリあるいは冒頭の説明文と照合を行い，該当するエンティティを決定した．推定したクラスだけでは複数の候補が残ってしまうことがあるが，その場合は Freebase での検索結果で最も上位にあるエンティティを選択した．

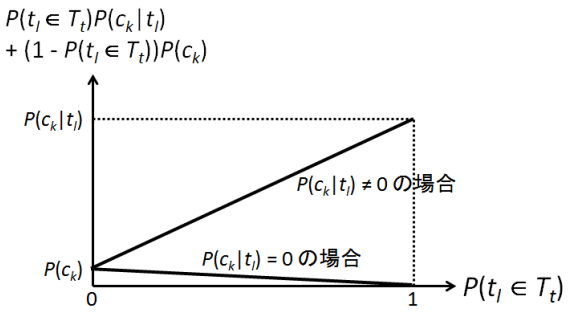

図 2: 確率 $P(t_l \in T_t)$ に応じたスムージング

評価尺度は文献 [3] と同様に，適合率のマクロ平均とマイクロ平均を用いた．提案手法である拡張ナイーブベイズ（関連度に応じたスムージング）の有効性を評価するため，適合率がほぼ最大となるようパラメータ調整した閾値ベースの手法（加算スムージングを併用）を，確率的概念辞書である Probase を用いて比較した．また，Probase を用いた曖昧性解消手法の有効性を検証するため，Hoffart らの手法 [3]，Kulkarni らの手法 [4]，Cucerzan の手法 [1] を比較対象とした[4]．

評価結果を表 1 に示す．適合率がほぼ最大となるよう関連度の閾値を調整した場合と比較して，閾値を用いずに提案手法のナイーブベイズを適用した場合の適合率がマクロ平均，マイクロ平均どちらの場合においても勝っている．このことから，入力の集合に含まれる要素が観測できない状況において，確率的に要素集

[2] http://www.cnts.ua.ac.be/conll2003/ner/
[3] http://www.mpi-inf.mpg.de/yago-naga/aida/
[4] これらの比較手法の結果は Hoffart らの実験結果を引用している．また，Kulkarni らの手法，Cucerzan の手法では，実装上の都合で全 231 文書のうち 229 文書をデータセットとして用いている．

表 1: CoNLL2003 データセットを用いた曖昧性解消タスクによる評価結果（値は%）

評価尺度	Probase		AIDA [3]	Kulkarni et al. [4]	Cucerzan [1]
	閾値 （調整済）	提案手法			
適合率のマクロ平均	83.14	83.57	82.02	76.74	43.74
適合率のマイクロ平均	80.26	81.40	82.29	72.87	51.03

合を決定してナイーブベイズを適用する手法の有効性が確認できる．また，閾値を調整した場合よりも適合率が高いことから，文書ごとに最適な閾値が異なっており，根本的に固定の閾値では問題に対処できないことが推測できる．

既存の曖昧性解消手法と比較しても，Probase を基盤知識とした概念化手法が十分な適合率を達成していることが分かる．また，CoNLL2003 データセットにおいて，正解データの固有表現の数が 18 以下の 136 文書（≒短い文書）のみを対象とした場合，AIDA ではマクロ平均 78.83%，マイクロ平均 82.93%であったのに対し，提案手法ではマクロ平均 86.30%，マイクロ平均 83.92%であった．このことから，Probase を用いた手法は，特に短文における曖昧性解消に有効であるといえる．Probase の is-a 関係のみを使ってこのような高精度を達成できたのは，Probase が豊富な概念空間を有していることに起因すると考えられる．Probase は他にも属性関係を定義しており，これらの情報を用いてさらに精度を向上できる可能性がある．

5 おわりに

本研究では，大規模な確率的概念辞書 Probase を基盤知識とし，ナイーブベイズを用いてテキスト中に出現する個々の固有表現の概念（クラス）を推定する手法を提案した．入力の集合に含まれる要素が観測できない状況において，集合の状態を確率的に定義してナイーブベイズを適用するという考え方を導入した．パラメータを設定することなく，安定したクラス推定が可能であることを，評価実験により確認した．提案手法の拡張ナイーブベイズは，固有表現のクラス推定以外にも，入力の集合に含まれる要素が確率的に定義できる場合に適用可能であり，汎用性が高い．

今後の課題として，入力の集合に含まれるか否かを決定する確率（関連度）をどのように決定するかについて検討する必要がある．現時点ではコサイン関連度としているが，本来は確率として定義するべきであるため，理論的な観点からみて正しい確率の算出方法について模索する予定である．

参考文献

[1] Silviu Cucerzan. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In *EMNLP-CoNLL*, pp. 708–716, 2007.

[2] Marti A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In *COLING*, pp. 539–545, 1992.

[3] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Furstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust Disambiguation of Named Entities in Text. In *EMNLP*, pp. 782–792, 2011.

[4] Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. Collective Annotation of Wikipedia Entities in Web Text. In *KDD*, pp. 457–465, 2009.

[5] Gregory L. Murphy. *The Big Book of Concepts*. The MIT Press, 2002.

[6] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *CoNLL*, 2003.

[7] Yangqiu Song, Haixun Wang, Zhongyuan Wang, Hongsong Li, and Weizhu Chen. Short Text Conceptualization Using a Probabilistic Knowledgebase. In *IJCAI*, pp. 2330–2336, 2011.

[8] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. YAGO: A Core of Semantic Knowledge. In *WWW*, pp. 697–706, 2007.

[9] Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q. Zhu. Towards a Probabilistic Taxonomy of Many Concepts. Technical Report MSR-TR-2011-25, Microsoft Research, 2011.