3I1-R-9-1

# 知識ベースに基づいた語義曖昧性解消における教師データの活用

Use of labeled training data for knowledge-based word sense disambiguation

松田 耕史*1　高村 大也*2　奥村 学*2
Koji Matsuda　Hiroya Takamura　Manabu Okumura

*1東京工業大学 総合理工学研究科
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology

*2東京工業大学 精密工学研究所
Precision and Inteligence Laboratory, Tokyo Institute of Technology

We propose a novel method for knowledge-based word sense disambiguation. Some pieces of prior work pointed out that context information can be captured by semantic relatedness of the context words between the target word. In the method proposed in this paper, the semantic relatedness is represented as a linear combination of several statistical measures calculated from the knowledge base. The weight of the linear combination are learned the weight of a classifier or a "Learning to Rank" method. The experimental result on a publicly-available dataset shows that the proposed method using a "Learning to Rank" learner achieves a higher accuracy than a typical SVM classifier in this task.

## 1. 緒論

### 1.1 研究の背景

語義曖昧性解消は自然言語処理の様々な応用タスクにおいて重要な役割を担う重要なタスクである．例えば，機械翻訳のタスクにおいてある英文に出現する，"bank" という語が "銀行" と認識されるか "土手" と認識されるかは，翻訳結果に大きな影響を及ぼす．こういった応用タスクにおける強い必要性にも関わらず，現在までに知られている，高い性能を実現する語義曖昧性解消アルゴリズムはタグ付きコーパスを用いた教師あり学習に基づくものが殆どである．しかしながら，語義タグ付きコーパスを作成することは非常にコストの大きい作業であることが知られており，曖昧性をもつすべての語に対して十分な質および量を備えたタグ付きコーパスを整備するのは困難である．この問題に対する一つの解決策として，知識ベースに基づく語義曖昧性解消法が研究されている．例えば，古典的なアルゴリズムの一つとして，Lesk アルゴリズム [Lesk 86] が知られている．このアルゴリズムは，語義曖昧性解消を行う対象語が出現する文脈と知識ベースに記述された語義定義文との比較を行い，最も文脈と類似度が高くなる語義を対象語の語義として割り当てるものである．最近の研究においては，知識ベースとして， WordNet*1, OntoNotes*2等，様々なものが利用されている．これらは単なる語の意味に関する情報を語釈文という形で記述した辞書ではなく，「語義」に着目し，語義と語義の間にどのような関係があるかをデータベース化したものである．そのため，本稿では，「辞書」ではなく，「知識ベース」という用語を用いる．知識ベースを語義曖昧性解消に有効に用いるためには，これらの知識ベースに記述された語釈文や用例などと共に，知識ベースに記述されている語義間の関係性に関する情報を利用することが不可欠であり，実際に知識ベースに記述された語義間の意味的関連性を用いることによって精度の高い語義曖昧性解消を行うことを目的とした研究が行われている [Pedersen 05]．しかしながら，知識ベースに基づいた語義曖昧性に関する既存の研究においては，前述した語義間の意味的関連性として単一の尺度を用いるものが殆どであり，複数の尺度を複合して用いることや，どういった尺度が有効であるかを教師データを用いて推定することは行われていない．

### 1.2 研究の目的と概要

本稿では，語義曖昧性解消に対する新たなアプローチとして，知識ベースに基づく語義曖昧性解消法において用いられる，語義概念間の意味的距離の尺度を既存の教師データから学習するという新しい語義曖昧性解消のフレームワークを提案する．教師データとして，過去に整備されてきた語義タグ付きコーパスを利用する．先に述べたように，これらのデータはすべての語に対して整備されたものではないため，特定の語に依存しない特性を教師データから獲得 (汎化) させる必要がある．そのために，これまでに提案されてきた複数の関連性尺度の線形結合に対するパラメータの最適化というアプローチを取る．これによって，過去に整備されてきた語義タグ付きコーパスを用いて，知識ベースの手法の利点である高い網羅性を維持しつつ，性能を向上させることを試みる．また，既存のデータセットに対して実験を行い，提案手法の優位性を確認する．

## 2. 関連研究

教師あり学習に基づく語義曖昧性解消手法は，多くの場合高い性能を発揮するが，曖昧性解消を行いたい対象語全てに対して，語義タグを付与したコーパスを整備する必要があり，コーパス作成に大きなコストがかかるという問題がある．しかしながら，過去の研究において，ある程度の量の語義タグ付きコーパスは既に整備されている*3．知識ベースに基づく語義曖昧性解消は，語義タグ付きコーパスを必要とせず，対象語が知識ベースに記述されていれば語義曖昧性解消を行うことがで

連絡先: 松田耕史，東京工業大学精密工学研究所，横浜市緑区長津田町 4259 R2-728，045-924-5295，matsuda@lr.pi.titech.ac.jp

*1 http://wordnet.princeton.edu/

*2 http://www.bbn.com/ontonotes/

*3 例えば，Senseval3 データセット
http://www.senseval.org/senseval3/data.html
や，SemEval2007 データセット
http://nlp.cs.swarthmore.edu/semeval/tasks/index.php
等が挙げられる

きる．つまり，高いカバレッジを実現することが可能である．しかし，語義曖昧性解消に知識ベースを有効に利用する手法については，いまだ十分な研究がなされているとは言えない．多くの手法では，語義と語義の間の関連性尺度を何らかの方法で用いるが，単一の尺度の提案や，知識ベース側の拡張による性能改善に関する研究が主であり，何らかの方法で尺度を適応的に決定するといった研究は行われていない．また，知識ベースに基づく手法では，最頻出語義を出力するというヒューリスティックスに基づくベースライン (Most Frequent Sense Baseline) を超える性能を出すのも難しいことが多くの研究により示されている．また，これに関連する問題としてバックオフ手法の選択が挙げられる．知識ベースの手法において，曖昧性解消を行う対象語と文脈との間で関連性を定義できなかった場合，バックオフとして，別の基準に基づいて選択した語義を出力することが広く行われている．多くの既存手法においては，知識ベースに付与された語義の頻度を用いて，上に述べた最頻出語義が出力されるが，知識ベースに付与された語義の頻度が，必ずしもすべてのドメインで有効ではない可能性がある．

### 2.1 WordNet

本稿では，知識ベースとして WordNet [*4] を用いる．WordNet はプリンストン大学が中心となって構築した英語の概念辞書である．WordNet においては，同義語は一つの synset という意味クラスとしてまとめられ，synset 同士の間に上位語，下位語，関連語等のリンク情報が付与されている．特に，動詞および名詞においては，すべての synset が is-a の階層構造を持った木構造に整理されているのがひとつの特徴である．同時に，それぞれの synset に対して，語釈文と用例が付与されている．このリンク情報や語釈文等を用いて語義曖昧性解消を行うアルゴリズムの一つとして，次節で述べる Maximum Relatedness Disambiguation がある．

## 3. Maximum Relatedness Disambiguation

本節では，本稿で提案する手法のベースとして用いる，Pedersen ら [Pedersen 05] が提案した，語義間に定義される任意の関連性尺度を用いて語義曖昧性解消を行うアルゴリズムについて述べる．このアルゴリズムは オリジナルの Lesk アルゴリズムと同様に，辞書の情報を利用して最も文脈との関連性が高い語義を出力するアルゴリズムである．彼らは Lesk アルゴリズムを一般化させ，単に語釈文と文脈の語彙の重なりを用いるのではなく，語義のペアに対して定義される任意の関連性尺度を曖昧性解消に用いることのできる枠組みを提案した．文脈内の語をそれぞれ，$w_1, w_2, ..., w_n$ とする．ただし，$1 \leq t \leq n$ であり，$w_t$ を語義を付与すべき対象語とする．それぞれの語 $w_i$ は $m_i$ 個の語義を持っており，それらを $s_{i1}, s_{i2}, ...s_{im_i}$ とする．語義曖昧性解消アルゴリズムの目的は，$w_t$ に対応する語義集合 $\{s_{t1}, s_{t2}, ..., s_{tm_t}\}$ の中から最適な語義を選択することである．Pedersen らは，以下の式で表されるスコアに基づいて語義曖昧性解消を行うアルゴリズムを提案した:

$$argmax_{i=1}^{m_t} \sum_{j=1, j\neq t}^{n} max_{k=1}^{m_j} rel(s_{ti}, s_{jk}). \quad (1)$$

ただし，$rel$ は任意の語義概念ペアに対してそれらの間の関連性を表す実数を返す関数である．つまり，$rel$ : $\{s_{t1}, s_{t2}, ..., s_{tm_t}\} \times \{s_{j1}, s_{j2}, ..., s_{jm_j}\} \rightarrow \mathcal{R}$ と定義することができる．例えば，Lesk アルゴリズムにおいては，二つの語義定義文の間での語彙の重なりが用いられる．

Pedersen らは，以下のような関連性尺度について調査を行った．括弧内に [Pedersen 05] 内での表記を示す．詳細な定義は元論文を参照いただきたい．

- WordNet 上の synset 間のパスに基づく尺度
  - is-a 階層における最短パスの長さの逆数 (path)
  - is-a 階層における least common subsumer の深さ (wup)
  - 最短パスの長さを階層の深さで正規化した値 (lch)
- 選択情報量に基づく尺度
  - least common subsumer の選択情報量 (res)
  - least common subsumer の選択情報量をそれぞれの synset が持つ選択情報量で正規化した値 (lin)
  - least common subsumer の選択情報量とそれぞれの synset が持つ選択情報量の差 (jcn)
- synset に付与された語釈文や用例に基づく尺度
  - それぞれの synset に付与された語釈文間の語義のオーバーラップ，ただし，近接する synset の語を含む (lesk)
  - 語釈文から二次の共起を用いて作成されたベクトル間の cosine 尺度 (vector)
  - 語釈文と用例それぞれから計算した vector 尺度の和 (vector pairs)

## 4. 提案手法

本稿では，Maximum Relatedness Disambiguation をベースとして語義概念間の意味的距離の尺度を教師あり学習を用いて最適化し，教師ありデータに対して最適な尺度を求めることで，語義曖昧性解消の性能を向上させることを目指す．つまり，語義曖昧性解消における知識ベースの有効な利用法を教師データから学習することを試みる．具体的には，以下の式で語義を決定する:

$$argmax_{i=1}^{m_t} \sum_{j=1, j\neq t}^{n} max_{k=1}^{m_j} \sum_{r=1}^{R} \lambda_r rel_r(s_{ti}, s_{jk}). \quad (2)$$

式 (2) に示すように，$R$ 個の関連性尺度の重み係数 $\lambda$ での線形結合を考え，その上で Maximum Relatedness Disambiguation を行う．この $\lambda$ は既存の語義タグつきデータから学習することができる．しかしながら，各関連性尺度はお互いに高い相関を持っていることが予備実験で明らかになったため，通常の線形分類器を用いた重みの学習は安定した推定が行えない．線形分類器においてはすべてのデータを同じ条件で扱うが，今回の問題は，語義の「選好関係」を学習する問題と捉えることもできる．データに対して順位付けを行い，選好関係を学習する手法は Learning to Rank という問題として，情報検索などの分野で盛んに研究が行われている [Trotman 05]．本稿においては，ある対象語と文脈が与えられたときに，正解語義のランクがそれ以外の語義のランクより高くなるように，語義のペアに対して制約を与える Pairwise アプローチをとる．できるだけ多くの制約を満たすようにパラメータを最適化することが，この場合の学習の目的となる．これにより，「選好関係」をより直接に取り入れたモデルの構築が可能になると推測できる．

### 4.1 学習

実際に式 (2) における各関連性尺度の重み $\lambda$ は次のようにして学習される．まず，訓練データ内のそれぞれの文に対して，

*4 http://wordnet.princeton.edu/

対象語 $w_t$ の語義がある $s_t$ であると仮定した場合の文脈内のそれぞれの語と語義 $s_t$ の Maximum Relatedness の値をそれぞれの関連性尺度 $rel_r \in R$ について求める:

$$x_{rel_r} = \sum_{j=1,j\neq t}^{n} max_{k=1}^{m_j} rel_r(s_t, s_{jk}). \quad (3)$$

ただし，$R$ は Pedersen らが用いた以下の 9 種類の尺度である．

$\{lesk, path, jcn, lch, lin, res, vector, vectorpairs, wup\}$

$x_{rel_r}$ を並べたものを教師データのベクトルとする．このようにして作られたベクトルについて，$s \in Senses(w_t)$ であるものを正例とし，そうでない場合に負例とする．ただし，$Senses(w_t)$ は訓練データにおいて対象語 $w_t$ に対してタグ付けされた語義の集合である．つまり，一つの教師データから文脈と特定の語義 $s_t$ との関連性を素性値として抽出し，$|Senses(w_t)|$ 個の正例と $m_{w_t} - |Senses(w_t)|$ 個の負例を生成する ($m_{w_t}$ は対象語 $w_t$ の語義数)．ランク学習においては，正例または負例のラベルを付与する代わりに，それぞれの事例を一つのランキング事例とみなし，$s \in Senses(w_t)$ である事例がそれ以外の事例より選好される，というラベルを与えている．このようにして作られた学習データを入力として学習を行い，各関連性尺度 $rel_r$ の重み $\lambda_r$ を推定する．

## 5. 実験

### 5.1 データ

ベンチマーク用の評価データとして SemEval-2007 Coarse-Grained English All Words Task [Navigli 07] で用いられたデータを用いる．教師データとしては，SemEval-2007 Lexcal sample task[Pradhan 07] において提供された訓練データを用いた．前処理として，原文に対して TreeTagger[Schmid 94] で品詞タグを付与し，WordNet に付属するユーティリティを用いて，複合語の検出を行った．

### 5.2 学習アルゴリズム

それぞれの関連性尺度の重みの学習には線形分類器としては Support Vector Machine (SVM), ランク学習器としては Stochastic Pairwise Descent (SPD)[Sculley 09] を使用した．各手法において，最適化手法には確率的勾配降下法に基づく手法の一つである Pegasos[Shalev-Shwartz 07] を用い，実装としては，オンラインで公開されている sofia-ml [*5]を用いた．SVM, SPD それぞれの正則化パラメータは 0.0001，繰り返し回数は $10^9$ 回の設定のもとで実験を行った．

### 5.3 Back-off 戦略

WordNet の synset 間には必ずリンクがあるとは限らないため，Maximum Relatedness Disambiguation においては，関連性尺度がゼロになる場合が考えられる．この場合にバックオフ戦略として，WordNet の各 synset に付与されている頻度の値を参照し，最も高頻度の語義を付与する手法と，候補となる語義集合からランダムに語義を選択する手法の二通りが考えられる．WordNet に付与されている語義頻度は，SemCor コーパス[*6]からとられたものであるが，これは主に Brown コーパスのサブセットに対してタグを付与したものである．Brown コーパスは様々なドメインから用例を収集した Balanced なコーパスであり，このコーパスから取られた頻度値は，英語の書き言葉としてはバランスの取れた統計値であると考えられる．しかしながら，語義の分布はドメインごとに大きく異なることが知られている [Koeling 05]．たとえば，“mouse” という単語は，コンピュータに関するドメインでは機器としてのマウスを表すことが多く，生命科学に関するドメインでは，動物としてのマウスが最頻出の語義であると考えることができる．そのことを考慮すると，バックオフとして SemCor から計算された語義頻度を用いることが常に最適な選択ではないといえる．そのため，WordNet の最頻出語義をバックオフとして付与した場合，手法の一般性を制限した条件のもとでの評価になることを付記しておく．そのため，以下の三つの条件で実験を行った．

- WordNet に付与された語義頻度に基づく最頻出語義 (MFS) でバックオフを行う
- ランダム選択でバックオフを行う
- バックオフを行わない (該当語義なし，と出力する)

表 2: 文脈幅 $n = 9$ における，バックオフ戦略が選択された事例数．各尺度の表記は [Pedersen 05] に合わせた．

| 尺度 | 事例数 | 割合 |
|---|---|---|
| path | 1977 | 0.871 |
| wup | 1977 | 0.871 |
| lch | 1977 | 0.871 |
| res | 2031 | 0.898 |
| lin | 2085 | 0.919 |
| jcn | 2021 | 0.890 |
| lesk | 57 | 0.025 |
| vector | 57 | 0.025 |
| vector pairs | 86 | 0.038 |
| 提案手法 | 57 | 0.025 |

### 5.4 実験結果

実験結果を表 1 に示す．線形分類器に基づく重みは，単一の尺度を用いた Pedersen らの手法に比べて高い性能を示してはいないことがわかる．それに対して，ランク学習器に基づく重みは，より高い性能を示している．バックオフに MFS を用いた場合は提案手法のいずれも Pedersen らの最高精度には達しなかったが，バックオフにランダム選択を用いた場合，バックオフを行わなかった場合は，ランク学習器に基づく重みを用いることで，Pedersen らが提案した手法の最高精度を上回る性能を示した．

## 6. 考察

### 6.1 バックオフ戦略

表 2 に，各尺度において，バックオフ戦略が選択された対象語の割合を示す．パスに基づく尺度においては，考慮する文脈内にパスを辿ることのできる語義ペアが存在しなければならないが，実際にはそうでないことが多いことが分かる．これは WordNet におけるグラフ構造に起因する問題である．まず，WordNet 内の synset 間に定義されているリンク情報は，殆どが同品詞間のものであり，文脈内に対象語と同じ品詞の語が含まれない場合には，殆どの場合パスに基づく関連性尺度は計算が行えない．また，WordNet には名詞，動詞，形容詞，副詞の 4 つの品詞に関する情報しか含まれていないため，文脈内にそれらの品詞に該当する語が含まれない場合にも同様の問題が発生する．今回は文脈幅を 9 に固定した設定のもとで実験を行ったが，より広い文脈幅を用いるなど，実験設定を見直すことでより性能が向上する可能性がある．

*5 http://code.google.com/p/sofia-ml/

*6 http://www.cse.unt.edu/~rada/downloads.html#semcor

表 1: 文脈の長さを 9 に固定した場合の実験結果 (F 値)．各尺度の表記は [Pedersen 05] に合わせた．

| 手法 | 尺度/最適化アルゴリズム | MFS backoff | Random backoff | backoff なし |
|---|---|---|---|---|
| Baselines | Most Frequent Sense | 0.788 | - | - |
| | Random | - | 0.524 | - |
| Pedersen ら [Pedersen 05] | path | 0.723 | 0.627 | 0.160 |
| | wup | 0.708 | 0.611 | 0.150 |
| | lch | 0.722 | 0.632 | 0.159 |
| | res | 0.699 | 0.596 | 0.115 |
| | lin | 0.715 | 0.595 | 0.095 |
| | jcn | 0.738 | 0.647 | 0.136 |
| | lesk | 0.699 | 0.699 | 0.697 |
| | vector | 0.699 | 0.700 | 0.696 |
| | vector pairs | 0.597 | 0.598 | 0.589 |
| | 9 尺度の平均 | 0.700 | 0.634 | 0.310 |
| 提案手法 (線形分類器) | SVM | 0.676 | 0.676 | 0.673 |
| **提案手法 (ランク学習器)** | **SPD** | 0.707 | **0.707** | **0.703** |

表 3: SemEval2007 Coarse-Grained English All-Words Task における他システムとの性能比較

| System | F 値 | MFS Backoff |
|---|---|---|
| UOR-SSI | 0.832 | ○ |
| NUS-PT | 0.825 | ○ |
| NUS-ML | 0.815 | ○ |
| LCC-WSD | 0.814 | ○ |
| GPLSI | 0.795 | ○ |
| UPV-WSD | 0.786 | ○ |
| **提案手法 (ランク学習器)** | **0.707** | - |
| TKB-UO | 0.702 | - |
| **提案手法 (線形分類器)** | **0.676** | - |
| PU-BCD | 0.661 | - |

### 6.2 他システムとの比較

表 3 に，SemEval 2007 のワークショップに提出された他のシステムとの比較結果を示す．ランク学習器に基づく提案手法は，最頻出語義に基づくバックオフを行わないシステムと比較して，高い性能を示していることが分かる．しかしながら，より精度の高いシステムはすべて何らかの形で最頻出語義に関する情報を用いている．知識ベースに基づく語義曖昧性解消においては，前節で述べたようにバックオフとして WordNet に付与されている語義頻度に基づく最頻出語義を用いることが広く行われているが，これは，語義頻度という多くの語に対しては容易には計算できない値を用いている．最頻出語義に関する情報を用いない手法は，他の言語や他の知識ベースへの適応を比較的容易に行うことが可能であると考える．しかしながら，本稿で述べた手法においても，バックオフを行う場合の条件について改良を行うことで，語義頻度の情報を用いることで性能の向上が見込めることは事実であり，今後の課題としたい．

## 7. 結論

本稿では，知識ベースに基づく語義曖昧性解消法において用いられる，語義概念間の意味的距離の尺度を既存の教師ありデータから学習するという新しい語義曖昧性解消のフレームワークを提案した．具体的には，複数の尺度の線形結合のパラメータを既存の語義タグ付きコーパスに対して最適化を行うために，ランク学習器を用いる手法を提案した．提案手法の性能を確認するため，SemEval 2007 において用いられたデータを用いて実験を行った．その結果，学習された距離尺度は，バックオフに MFS を用いないという条件の下では最も良い性能を示し，SemEval 2007 のワークショップに提出された他システムと比較しても良好な性能であった．今後の課題として，文脈幅への依存性の確認，品詞ごとのモデルの分割が挙げられる．

## 参考文献

[Koeling 05] Koeling, R., Mccarthy, D., and Carroll, J.: Domain-Specific Sense Distributions and Predominant Sense Acquisition, in *Proceedings of HLT and EMNLP* (2005)

[Lesk 86] Lesk, M.: Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone, in *Proceedings of the SIGDOC*, ACM (1986)

[Navigli 07] Navigli, R., Litkowski, K., and Hargraves, O.: Semeval-2007 task 07: Coarse-grained english all-words task, in *Proceedings of the SemEval* (2007)

[Pedersen 05] Pedersen, T., Banerjee, S., and Patwardhan, S.: Maximizing semantic relatedness to perform word sense disambiguation, in *Research Report UMSI* (2005)

[Pradhan 07] Pradhan, S. S., Loper, E., Dligach, D., and Palmer, M.: Semeval-2007 task-17: English lexical sample, srl and all words, in *Proceedings of the SemEval* (2007)

[Schmid 94] Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees, in *Proceedings of international conference on new methods in language processing* (1994)

[Sculley 09] Sculley, D.: Large scale learning to rank, in *NIPS 2009 Workshop on Advances in Ranking* (2009)

[Shalev-Shwartz 07] Shalev-Shwartz, S., Singer, Y., and Srebro, N.: Pegasos: Primal estimated sub-gradient solver for svm, in *Proceedings of the ICML*, ACM (2007)

[Trotman 05] Trotman, A.: Learning to rank, *Information Retrieval*, pp. 359–381 (2005)