語義曖昧性解消を対象とした領域固有のシソーラスの構築

新納 浩幸 (茨城大学 工学部 情報工学科) [1]
國井 慎也 (茨城大学大学院 情報工学専攻) [2]
佐々木 稔 (茨城大学 工学部 情報工学科) [3]

Construction of Domain Dependent Thesaurus for Word Sense Disambiguation

Hiroyuki Shinnou (Ibaraki University, Department of Computer and Information Sciences)
Shinya Kunii (Ibaraki University, Department of Computer and Information Sciences)
Minoru Sasaki (Ibaraki University, Department of Computer and Information Sciences)

1　はじめに

多くの自然言語処理システムでは，その適用領域を限定することで，システムのパフォーマンスを改善できる．語義曖昧性解消 (Word Sense Diaambiguation, WSD) においても，適用領域を限定することで，識別精度を改善できると考えられるが，その具体的方法は明らかではない (Navigli (2009))．一方 WSD では，一般に，シソーラスを利用することで，その識別精度を改善できる．そこで本論文では領域固有のシソーラスを構築することで，適用領域を限定した WSD の精度向上を図る．

WSD は文中の多義語である対象単語に対して，その単語がどの語義で使われているかを識別するタスクである．このタスクは自然言語処理の中心課題であり，従来より多数の手法が提案されているが，近年は教師付き学習を用いる手法が中心である．そこではまず少量の対象単語の用例を用意し，その用例中の対象単語に語義 $c \in C$ を付与する．次に用例中の対象単語の周辺の文脈を特徴ベクトル $\boldsymbol{x}$ で表し，先に付与した語義とのペア $(\boldsymbol{x}, c)$ を作る．このペアの集合が訓練データとなり，この訓練データから SVM 等の分類器を学習することで WSD を解決する．

このアプローチで問題となるのは，訓練データのスパース性 (Feature Sparseness) である．例えば素性として周辺の単語表記だけを利用する場合，訓練データから得られる素性の集合は，取り得る素性全体に対してスパースとなる．そのためテストデータでは訓練データには全く出現しない単語のみで特徴ベクトルが構成される場合も多く，その場合，正しく語義識別ができない可能性が高い．そこでスパース性を解消するためにシソーラスを利用することが一般に行われている．シソーラスとしては分類語彙表 (分類語彙表 (2004)) などの手作業で作られたものが一般に利用されるが，コーパス内の単語をクラスタリングすることで自動構築したシソーラスも利用できる．WSD の適用領域を限定した場合，その領域のコーパス S から自動構築されるシソーラス T_S を利用すれば，一般的なシソーラスを利用するよりもより高い精度が実現できると考えられる．つまり上記のシソーラス T_S が領域 S の固有のシソーラスになると考えられる．ただしシソーラスは単語の語義の上位下位関係を表したものであり，シソーラスが領域に依存するとは考えづらい．そのため WSD の対象領域のコーパスに様々なコーパスを加え合わせたコーパス G からシソーラス T_G を構築した場合，シソーラス T_G はシソーラス T_S を包含していると考えられ，T_S が T_G よりも有用であるとは単純には言えない．

本論文では，基本的に T_G よりも T_S を使う方が WSD の識別精度の観点では有利であるが，G が S を拡大するような形になっている場合に，T_G を使う効果が現れると考える．そこで T_G が T_S に比較して有効な度合い α を導入し，T_G と T_S の以下で示す線形結合により領域 S に固有のシソーラス $T_{S'}$ を作成することを提案する．

$$T_{S'} = (1-\alpha)T_S + \alpha T_G \tag{1}$$

[3] shinnou@mx.ibaraki.ac.jp
[3] 13nm707s@hcs.ibaraki.ac.jp
[3] msasaki@mx.ibaraki.ac.jp

問題は α の算出方法と (式 1) の実現方法である．α の算出のために，コーパスに対する代表単語群という単語集合を定義する．ここではコーパス内の単語を頻度順に並べ，累積頻度が全体のある割合になるまでの単語群を代表単語群と定義する．今，コーパス S と G の代表単語群をそれぞれ W_S，W_G とする．これらを用いて α を以下で定義する．

$$\alpha = \frac{|W_S \cap W_G|}{|W_S|}$$

また (式 1) の実現方法としては，S と G のそれぞれから単語 w の縮約ベクトル w_s と w_g を作成し，w_s と w_g に (式 1) に対応する重みを乗じて結合させることで，単語 w の特徴ベクトル w_{sg} を構築し，$\{w_{sg}\}_{w \in G}$ をクラスタリングすることで $T_{S'}$ を構築する．

実験では現代日本語書き言葉均衡コーパス（BCCWJ (Maekawa (2007))）における 3 つの領域 OC (Yahoo! 知恵袋)，OY (Yahoo! ブログ) 及び PB (書籍) を利用する．提案手法を用いて各領域の固有のシソーラスを構築した．また東京工業大学の奥村研究室では BCCWJ に語義タグをつけたデータを構築しており，本実験の WSD ではこのデータを利用させていただいた．ある程度の頻度を持つ多義語を選び，OC では 32 単語，OY では 23 単語，PB では 46 単語の WSD を行う．構築できた領域の固有のシソーラスの他，いくつかの自動構築したシソーラスを用いて WSD を行う．5 分割交差検定による WSD の平均正解率からシソーラスの評価を行う．

2　領域固有シソーラスの作成

2.1　コーパスに対する代表単語群の非被覆

本論文では T_G が T_S に比較して有効な度合い α を導入し，α を用いた T_G と T_S の線形結合 (式 1) により領域 S に固有のシソーラス $T_{S'}$ を作成する．

α の算出法を示す．本論文では基本的に T_S は T_G よりも有効であると考える．そして T_G が T_S よりも有効になるとしたら，G が S を拡大するような形になっている場合だと考える．この拡大の度合いが α を意味する．そして本論文ではこの拡大の度合いを以下のように考える．

まず G や S には WSD で利用される重要な単語がある．それをここでは代表単語群とよぶ．そして G の代表単語群を W_G，S の代表単語群を W_S とする．ここで注意として $S \subset G$ だが $W_S \subset W_G$ とは限らない．今，そこで $W_S - W_G = W_S - W_S \cap W_G$ の意味を考えると，これは S においてのみ重要である単語と考えられ，その度合いは以下で示させる．

$$\frac{|W_S - W_G|}{|W_S|} = 1 - \frac{|W_S \cap W_G|}{|W_S|}$$

ここから α は以下となる．

$$\alpha = 1 - \frac{|W_S - W_G|}{|W_S|} = \frac{|W_S \cap W_G|}{|W_S|} \tag{2}$$

次にコーパスに対する代表単語群の定義であるが，本論文ではコーパス内の単語を頻度順に並べ，コーパス S の場合は，累積頻度が全体の 8 割になるまでの単語群，またコーパス G の場合は，累積頻度が全体の 7 割になるまでの単語群を代表単語群とした．

2.2　シソーラスの線形結合

前述した T_G が T_S に比較して有効な度合い α を利用して，領域 S に固有のシソーラス $T_{S'}$ の構築方法を述べる．

S と G のそれぞれから単語 w の縮約ベクトル w_s と w_g を作成する．具体的にはトピックモデルである Latent Dirichlet Allocation (LDA) (Blei et al. (2003)) を用いて作成する．コーパス S に対して LDA により $p(w|z_i)$ が得られる．ここで z_i は i 番目の潜在的トピックである．またここではトピックの数を 100 に設定する．そして w_s を以下とする．

$$w_s = \frac{1}{Z}(p(w|z_1), p(w|z_2), \cdots, p(w|z_{100}))$$

ここで $Z = \sum_{i=1}^{100} p(w|z_i)$ である．同様にしてコーパス G に対して LDA を適用して w_g を得る．注意として $w \notin S$ の場合，w_s は 0 ベクトルとなる．

LDA から求まった w_s や w_g はクラスタ数を 100 とした場合の，単語 w のソフトクラスタリング結果に相当する．これらのクラスタリング結果に重み α をつけて統合することで，領域 S に固有のシソーラス $T_{S'}$ を構築する．具体的には重み付きアンサンブルクラスタリングを行う．処理的には w_s に $1-\alpha$ を乗じた 100 次元のベクトルと，w_g に α を乗じた 100 次元のベクトルを結合させ，200 次元のベクトル w_{sg} を作成する (図 1 参照)．この w_{sg} を単語 w の特徴ベクトルと考えて G 内の単語をクラスタリングする．このときのクラスタ数は 1000 に設定する．またクラスタリングアルゴリズムには k-means を用いた．

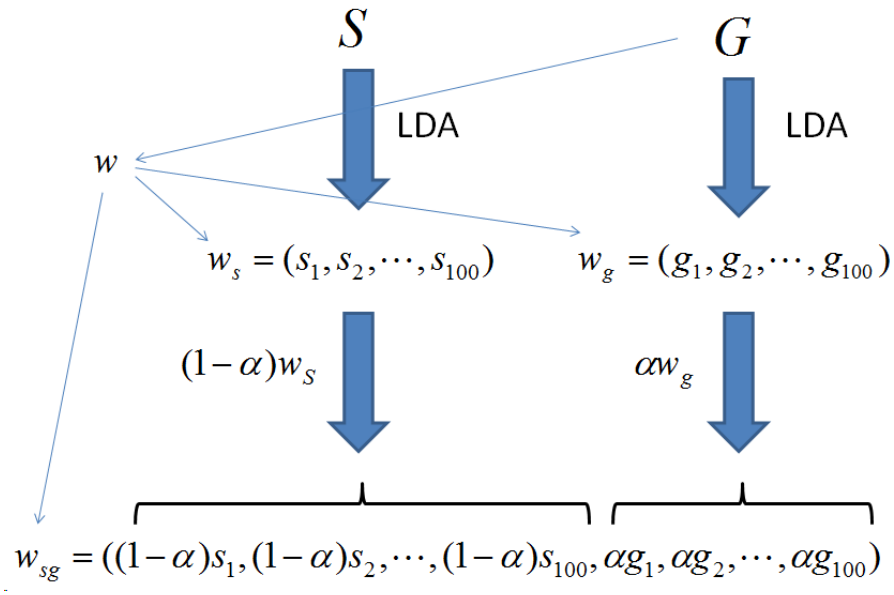

図 1: w_{sg} の作成

3　実験

シソーラスを評価するために，WSD の対象単語 w に対する訓練データ作成の際に，w の周辺の名詞に対して，シソーラスから得られるその名詞のクラスタ番号を素性として含めて識別を行う．また WSD の領域は固定する．ここでは BCCWJ の 3 つの領域 OC (Yahoo! 知恵袋)，OY (Yahoo! ブログ) 及び PB (書籍) を利用する．つまり提案手法により 3 つの領域固有のシソーラスを構築し，それぞれの領域の WSD において，領域固有のシソーラスの効果を確かめる．また対象領域を含むより大きなコーパス G としては，上記 3 つの領域のコーパスを合わせたコーパスとした．

表 1 にそれぞれの領域におけるコーパスのサイズ[4]，名詞の種類数，本実験における WSD の対象単語の数，その対象単語の平均語義数，及びそのた対象単語の平均用例数を示す．また表 2 に対象領域で用いた WSD の対象単語と語義数（括弧内の数）を示す．

[4]配布形式の（VARIABLE の）XML ファイルのサイズの合計であり，テキスト部分のサイズではない．

表 1: 利用コーパスのデータ

	OC	OY	PB	G (OC+OY+PB)
サイズ (KB)	125,347	129,384	255,528	510,260
名詞の種類数	83,130	141,950	181,356	278,629
WSD 対象単語数	32	23	46	-
WSD 平均語義数	2.81	2.91	3.06	-
WSD 平均用例数	158.4	139.5	128.4	-

表 2: 対象単語とその領域内の語義数

OC	ある (4)，いう (2)，意味付ける (3)，いれる (2)，教える (3)，買う (2)，書く (2)，聞く (2)，来る (2)，くれる (2)，子供騙し (2)，時間割り (2)，自分自身 (2)，ため (3)，できる (3)，出る (3)，とき (3)，ところ (3)，とる (7)，ない (2)，なか (3)，なる (2)，願う (2)，場合 (2)，入る (4)，前 (3)，みる (5)，持つ (4)，やる (3)，行く (2)，よい (3)，よる (3)
OY	買う (2)，来る (2)，くれる (2)，こと (3)，時 (2)，時間割り (4)，自分自身 (2)，する (8)，ため (3)，できる (3)，出る (3)，度 (3)，とき (2)，ない (2)，なか (2)，年 (3)，日 (3)，人 (3)，分 (5)，まあ (2)，前 (2)，もの (4)，やる (2)
PB	あう (6)，あげる (6)，歩く (2)，言う (2)，生きる (2)，意味付ける (3)，入れる (3)，書く (2)，考える (2)，関係付ける (3)，聞く (2)，技術屋 (2)，来る (2)，こどもだまし (2)，これ (2)，時間割り (2)，時代物 (5)，自分自身 (2)，社会問題 (3)，する (8)，それ (3)，高い (2)，出す (3)，立つ (4)，立てる (5)，強い (2)，手 (4)，出る (3)，とる (7)，人間業 (2)，場合 (2)，入る (4)，場所割り (2)，人 (2)，一つ (3)，ほとんど (2)，前 (2)，身 (2)，見える (4)，見る (6)，持つ (3)，モデルチェンジ (3)，山 (2)，やる (4)，行く (2)，呼ぶ (2)

また WSD で利用した素性は以下の 9 種類である．(e0) w の表記，(e1) w の品詞，(e2) w_{-1} の表記，(e3) w_{-1} の品詞，(e4) w_1 の表記，(e5) w_1 の品詞，(e6) w の前後 3 単語までの自立語の表記，(e7) e6 のクラスタ番号，(e8) w_{-1}，w_1，(e6) の各単語に対するクラスタ番号．なお対象単語の直前の単語を w_{-1}，直後の単語を w_1 としている．

シソーラス T を用いた各領域における WSD の実験を述べる．まず各対象単語の用例を素性リストに変換するが，その際，シソーラス T を利用して素性の一部に T におけるクラスタ番号が付与されている（上記の素性 (e8) に相当）．実験は 5 分割交差検定で行う．つまり対象単語 w の用例の素性リストの集合を 5 分割し，1 つをテストデータ，残り 4 つを訓練データとして SVM で学習し，w に対する正解率を測る．テストデータを変更することで 5 つの正解率が得られるが，それらの平均を w に対する正解率とする．領域 S における各対象単語ごとに上記の正解率を測り，それらを平均した値を領域 S におけるシソーラス T の平均正解率とする．

シソーラス T としては，シソーラスを使わないもの None とその領域 S のコーパスのみから作成したシソーラス T_S，コーパス G から作成したシソーラス T_G および提案手法に従って作成したシソーラス $T_{S'}$ を用いた．実験の結果を表 3 に示す．表 3 には $T_{S'}$ を作成する際に用いた α の値も示した．α の値は前章で説明した方法から算出した．

表 3: 利用したシソーラスに対する平均正解率 (%)

	OC	OY	PB	平均
None	**83.66**	84.58	82.58	83.51
T_S (対象コーパス)	83.54	84.92	83.18	83.88
T_G (全コーパス)	83.21	84.38	**83.21**	83.60
$T_{S'}$ (提案手法)	83.45	**85.49**	82.74	**83.89**
α	0.139	0.404	0.538	

総合的にみると，提案手法により作成したシソーラスによる正解率は，対象コーパスのみを用いて作成したシソーラスによる正解率や全体のコーパスを用いて作成したシソーラスによる正解率よりもわずかではあるが高かった．

4　考察

4.1　α の算出

本論文では T_S と T_G を以下の方法で結合している．

$$T_{S'} = (1-\alpha)T_S + \alpha T_G$$

この式において α は T_G の重みである．本論文では α の意味を T_G が T_S に比較して有効な度合い α と考えたが，それ自体は上記の式に沿っている．問題は α の算出方法であり，本論文で考案した (式 2) が妥当な値を示しているかどうかは不明である．

ここでは α の値を $0.2, 0.4, 0.6, 0.8$ と変化させた場合の平均正解率の変化を各領域に対して求めた．その結果を図 2，図 3，図 4 に示す．

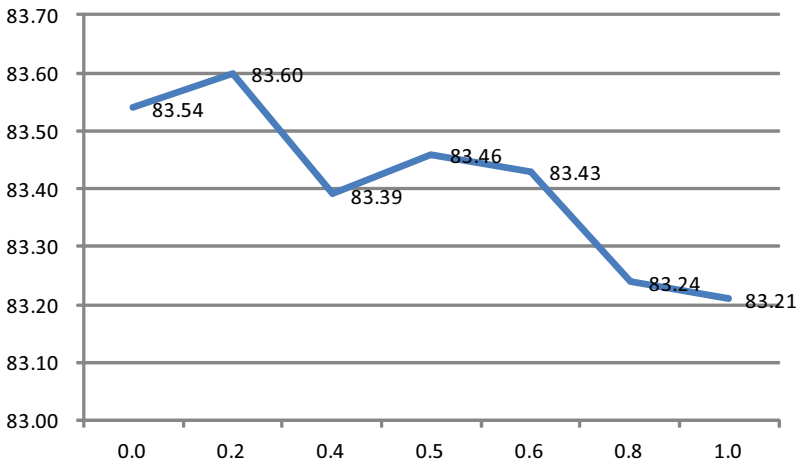

図 2: α に対する平均正解率（ＯＣ）

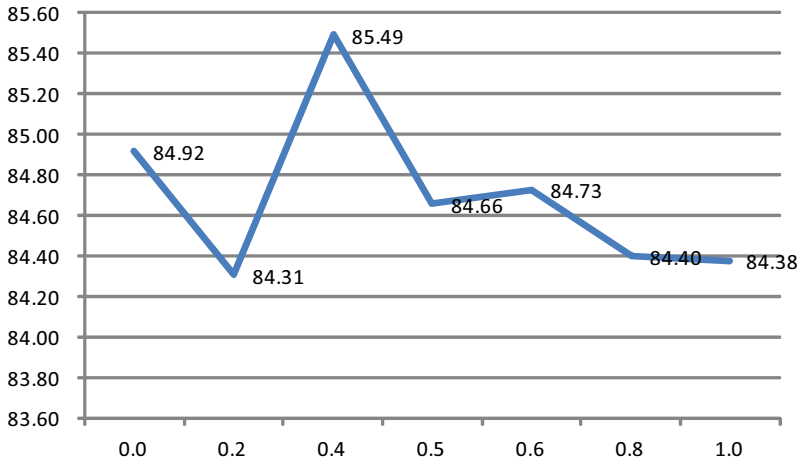

図 3: α に対する平均正解率（ＯＹ）

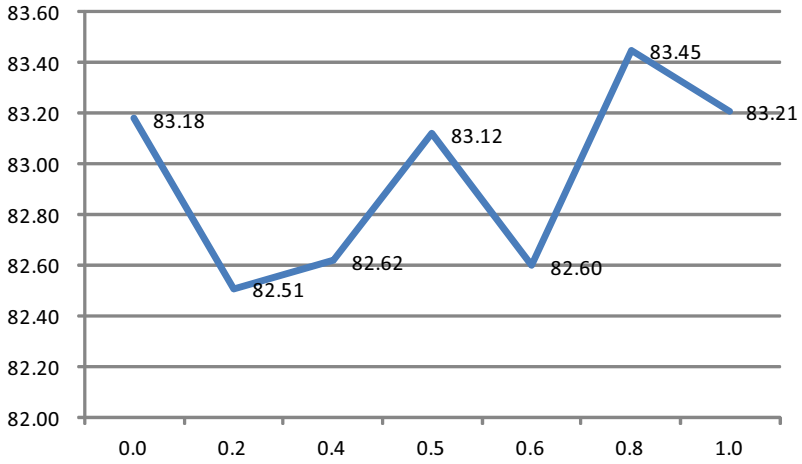

図 4: α に対する平均正解率（ＰＢ）

これらの図からわかるように適切な α の値を求めることができれば，T_S や T_G を単独で用いるよりも，それらを結合した $T_{s'}$ を用いた場合の方が WSD の精度が高まる．

しかしどの図も α と平均正解率が線形あるいは単峰性の関係になっていない．これは α の算出が困難であることを示すと共に，$T_{S'}$ を (式 1) でモデル化することが不適切であることも示唆している．対象単語毎に適切なシソーラスが異なることが原因だと予想している．

4.2　その他のシソーラスとの比較

T_S と T_G の結合方法として，本論文では LDA を用いて，それぞれのコーパスに対する単語の次元圧縮表現 (100 次元) を導き，それらを重み付きで結合して単語をベクトル化し，それをもとにクラスタリングする (クラスタ数 1000) という手順をとった．

ただし T_S と T_G の結合方法としては，コーパス S とコーパス G を結合させたコーパス $S+G$ を作り，$S+G$ からシソーラス T_{S+G} を作ることも考えられる．

またシソーラスとしては本論文で自動構築したもの以外に，既存の分類語彙表を用いることも可能である．さらに本論文で用いたコーパスとは別の一般に言語研究で用いられる新聞記事（毎日新聞 2005 年〜2008 年度版，4 年分）から作成したシソーラス T_{news} を用いることもできる．ここではこれらのシソーラスを用いた場合の正解率を求めた．表 4 に結果を示す．

表 4: その他のシソーラスによる平均正解率 (%)

	OC	OY	PB	平均
T_{S+G}	83.37	84.04	83.05	83.49
分類語彙表	83.04	84.53	**83.17**	83.58
T_{news}	**83.72**	84.45	82.59	83.59
提案手法 ($T_{S'}$)	83.45	**85.49**	82.74	**83.89**

提案手法のシソーラス $T_{S'}$ を用いる場合が最も良かった．その他のシソーラスには大きな差はなかった．

4.3　分野依存知識と一般知識の利用

本研究は WSD へ分野依存知識を利用する研究の一つである．従来，WSD へ分野依存知識を利用する研究は Domain-Driven Disambiguation (DDD) と呼ばれるアプローチの中で行われてきた (Strapparava et al. (2004))．これは基本的に対象単語の語義を内容語の領域と対象語義の領域との比較から選択する手法を取る (Gliozzo et al. (2004)，Gliozzo et al. (2005)，Magnini et al. (2002))．ただし利用している分野依存知識は語義分布である．本タスクのように領域をコーパスとして設定し，更に WSD の対象単語が一般的な単語である場合，語義分布が分野依存知識として活用できるケースは少ない．

また WSD へ分野依存知識を利用するアプローチは，WSD の領域適応の問題と関連が深い．領域適応とは学習元のデータが存在するソース領域と，実際に学習により得られた分類器を適用する先のターゲット領域が異なる問題である (Sogaard (2013))．ターゲット領域の知識をどのように取り込めるかが 1 つの課題である．我々は論文 (新納浩幸・佐々木稔 (2013)) において，WSD の領域適応にターゲット領域のトピックモデルを利用することを提案したが，そこではターゲット領域のシソーラス (本論文の T_S に相当) のみを利用している．一方，森は形態素解析の領域適応の問題を扱い，ターゲット領域の知識を含むより広範囲の領域の知識を利用することで，ターゲット領域の知識のみの利用よりも更に精度が上がったことを報告している (森信介 (2012))．しかし WSD などの他のタスクでもこの性質があるのかどうかは不明である．

ターゲット領域の知識は分野依存知識に相当し，ターゲット領域の知識を含むより大規模な知識は一般知識に相当する．WSD では分野依存知識（シソーラス T_s に相当）が概ね有効であるが，一般知識（シソーラス T_G に相当）を併用して利用することで精度向上が期待できる．論文 (Kunii and Shinnou (2013)) では T_s と T_G のそれぞれから得られる素性に重みをつけて，その素性を WSD の分類器の学習に利用している．本論文では T_s と T_G を融合した $T_{S'}$ を構築するというアプローチをとっている．どちらも分野依存知識と一般知識を融合して利用する研究である．

5　おわりに

本論文では適用領域を限定させた WSD の精度向上のために，WSD で利用されるシソーラスを，領域固有のシソーラスとして構築する方法を提案した．概略，適用領域のコーパス S と，S に別領域のコーパスを加えたコーパス G の各々からシソーラスを作成し，それらを線形結合した．線形結合する際の重みの算出には代表単語群の（非）被覆という概念を導入した．

BCCWJ の 3 つの領域 OC，OY，PB の各々で様々なシソーラスを用いて WSD を行うことで提案手法を評価した．提案手法により構築したシソーラスを用いた場合の正解率が，総合的に見れば，最も優れていたが，その差はわずかであり有意性には疑問もある．

ただし S のみから構築したシソーラス T_S や，G のみから構築したシソーラス T_G よりも何らかの形で融合させれば，より有効なシソーラスが構築できることは明らかである．今後はこの融合の方法を更に考えたい．

文献

David M. Blei, Andrew Y. Ng, and Michael I. Jordan (2003) “Latent dirichlet allocation,” *Machine Learning Reseach*, Vol. 3, pp. 993–1022.

Alfio Massimiliano Gliozzo, Bernardo Magnini, and Carlo Strapparava (2004) “Unsupervised Domain Relevance Estimation for Word Sense Disambiguation.,” in *EMNLP*, pp. 380–387.

Alfio Gliozzo, Claudio Giuliano, and Carlo Strapparava (2005) “Domain kernels for word sense disambiguation,” in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 403–410, Association for Computational Linguistics.

Shinya Kunii and Hiroyuki Shinnou (2013) “Combined Use of Topic Models on Unsupervised Domain Adaptation for Word Sense Disambiguation,” in *Proc. of PACLIC-2013*, pp. 415–422.

Kikuo Maekawa (2007) “Design of a Balanced Corpus of Contemporary Written Japanese,” in *Symposium on Large-Scale Knowledge Resources (LKR2007)*, pp. 55–58.

Bernardo Magnini, Carlo Strapparava, Giovanni Pezzulo, and Alfio Gliozzo (2002) “The role of domain information in word sense disambiguation,” *Natural Language Engineering*, Vol. 8, No. 04, pp. 359–373.

Roberto Navigli (2009) “Word sense disambiguation: A survey,” *ACM Computing Surveys (CSUR)*, Vol. 41, No. 2, p. 10.

Anders Sogaard (2013) *Semi-Supervised Learning and Domain Adaptation in Natural Language Processing*: Morgan & Claypool.

Carlo Strapparava, Alfio Gliozzo, and Claudio Giuliano (2004) “Pattern abstraction and term similarity for word sense disambiguation: Irst at senseval-3,” in *Proc. of SENSEVAL-3 Third International Workshop on Evaluation of Systems for the Semantic Analysis of Text*, pp. 229–234.

新納浩幸、佐々木稔 (2013)「k 近傍法とトピックモデルを利用した語義曖昧性解消の領域適応」，自然言語処理，第 20 巻，第 5 号，pp.707–726.

森信介 (2012)「自然言語処理における分野適応」，人工知能学会誌，第 27 巻，第 4 号，pp.365–372.

分類語彙表 (2004)「国立国語研究所」.