翻訳知識獲得のための多言語インターネットディレクトリィの自動統合（継続）

研究代表者　福本文代　山梨大学大学院医学工学総合研究部・教授
共同研究者　鈴木良弥　山梨大学大学院医学工学総合研究部・准教授

1 はじめに

本研究は，異なる言語で記述された複数のインターネットディレクトリィを統合することで得られる多言語関連文書から，機械翻訳システムに必要となる翻訳知識を自動的に獲得することを目的とする．機械翻訳は自然言語処理の成果の一つであり，コンピュータを介したユニバーサルコミュニケーションを実現するためのコア技術として注目されている．高品質な翻訳を生成するためには，対訳語に関する大量の語彙知識が必要である．近年，Web をはじめとする大規模データが手軽に入手できるようになったことを背景に，大規模データから互いに類似した内容を持つ多言語関連文書を自動的に抽出し，そこから対訳語を獲得する研究が盛んに行われている．この手法における対訳語の精度は，意味的に類似した多言語関連文書を高精度で抽出できるかに依存する．これまで統計手法や機械学習をはじめとする様々な手法が提案されているが，混沌としたWeb データが抽出対象であるために，いずれも質の高い対訳語を得るまでには至っていない．本研究はインターネットディレクトリィの階層構造が人手で構築されていることに注目し，これを利用することで質の高い大量の対訳語を獲得することを目指す．具体的には，日本語と英語で記述された2 つの階層構造を統合することで，互いに類似した内容を持つ文書対を抽出し，得られた文書対から対訳語を抽出する手法を提案する．

日本語インターネットディレクトリィの各分野と英語インターネットディレクトリィの各分野との対応付けを行うため，図 1 に示すように，各分野に属する英語文書（日本語文書）を辞書引きにより翻訳することで，日本語文書（英語文書）を作成し，これらの文書を日本語の階層（英語の階層）へ分類する．しかし，一般に，ディレクトリィが下位になるほど分野の粒度は細かくなるため，ノードに属する訓練文書数は少なくなる．そこで本研究では，(1) 分野名が付与されている少数のデータと分野名が付与されていない大量のデータを用いることで階層への分類を高精度で実現する手法を提案する．さらに，(2) 分野の対応付けを用いることで日英対応文書を抽出し，対応文書からの対訳語を抽出する手法を提案する．

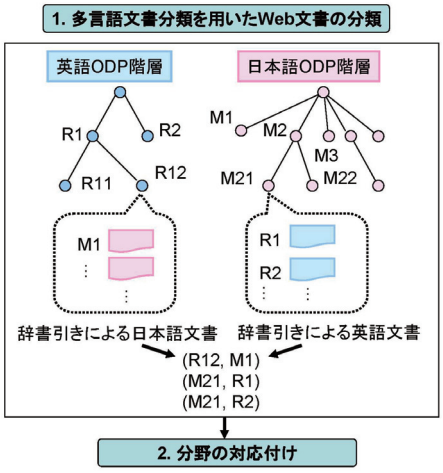

図1　インターネットディレクトリィの統合による翻訳知識の自動獲得

2 関連研究

少数から成る正例とラベルなしデータを用いた半教師付き学習法は，これまでにも数多く提案されている．Nigam は半教師付き学習の一つである EM 法を用いた文書分類手法を提案した(Nigam, 2000)．Joachims は，TSVM(Transductive Support Vector Machine)を用いることで少数の正例において高精度な分類が可能となることを示した(Joachims, 1999) しかしこれらの手法は，正例数が増えると必ずしも高い精度が得られないことが報告されている． Manevitz らは one-class SVMs を用いた分類法を提案した(Manevitz, 2001)．しか

し最適な識別境界面を設定するためのパラメータをどのように推定するかが課題として残されている．Yu らはこの問題を解決するため，正例に基づく学習手法 Positive Example Based Learning (PEBL) を提案した(Yu, 2002)．PEBL は，正例に対する負例の関連の度合いを考慮した手法である．予め用意された正例(集合)を用い，それとの類似度の値が最も小さいラベルなし文書を負例とし，機械学習 SVMs を用いて学習し,ラベルなし文書を分類する．分類の結果，負例と判定された文書を訓練データに加え学習し，残りのラベルなし文書をテストするという処理を負例と判断されるラベルなし文書がなくなるまで繰り返すという手法である．しかし，SVMs により負例と判断された事例が真に負例であるとは限らない．すなわち，本来正例である文書を誤って負例であると判定すると，正例である文書が負例として訓練データに加えられるため，結果的に正しく分類できない分類器が生成されてしまうという問題がある．本研究では，この問題を解決するため，順次抽出した負例に対し，誤りの検出と修正(Fukumoto, 2006)を行うことで，負例文書を正確に判定することを試みる(Fukumoto, 2013)．さらに SVMs の学習結果に対して Boosting を適用し弱い学習器を繰り返し学習することで分類器を作成し，テスト文書を高精度で分類する手法を提案する(Yamamoto,2012)．

一方，階層構造の統合における関連研究として，市瀬らは，データが持つオントロジー(分類階層) の対応を発見することで，階層構造を統合するHICAL システムを開発した(市瀬, 2003)．Web ディレクトリのWeb ページは人手で階層的に分類されているため，適切なカテゴリを選ぶことにより，情報の信頼度が高いWeb ページを容易に取得できる．市瀬らは，階層構造が保持する共有文書の数に基づいて異なる階層構造間におけるカテゴリ同士の類似性を推定した．しかし，この手法では，共有文書を持たないカテゴリの類似性推定が困難となる．濱崎らは異なる階層構造間における類似文書を見つけ，カテゴリ内に共有文書を仮想的に作り出すことで実際には共有文書を持たないカテゴリ間の類似性を求める手法を提案した(濱崎, 2004)．しかし，人手により分類した文書間の類似性と自然言語処理技術により求めた類似性は閾値を0.8 にした場合，正答率は，OpenDirectory で0.55 ，Lycosでは0.17 と精度面で課題が残されている． 本研究はこれらの問題に対し，文書分類を用いることで共有文書の数だけでなく分類処理の類似性を考慮しカテゴリ間の類似性を推定する統合手法を提案する．

類似文書抽出における研究として，内山らは大規模な日英対訳コーパスを作ることを目的に1989 年から2001 年までの読売新聞とThe Daily Yomiuri から日英記事対応と文対応を得た(内山, 2003)．日英対訳コーパスは機械翻訳や英語学，比較言語学あるいは英語教育や日本語教育などに非常に有用な言語資源である．しかし，一般に利用可能で大規模な日英対訳コーパス存在していなかった．そこで，大規模な日本語新聞記事集合とそれと内容的に一部対応している英語新聞記事集合から2 つの類似尺度AVSIM とSntScore を用いることで大規模な日英対訳コーパスの作成を試みた． 本研究における類似文書抽出も単語を要素とする類似度計算を用いることで抽出する．

3 少数の正例とラベルなし事例による文書の自動分類

3-1 正事例に基づく学習

Yu らは正例とラベルなし文書(集合)を用いた学習手法 (PEBL)を提案した(Yu, 2002)．PEBL は，正例に対する負例の関連の度合いを段階的に抽出することで学習・分類を行う手法であり，マッピング処理と収束という 2 つの処理から成る．マッピング処理では，ラベルが付与されていない文書集合 U と正例文書集合 P に対し，1-DNF (Disjunctive Normal Form) を用いて学習する．学習の結果得られた分類器を用いて U を分類し，負例と判定された文書(集合)を正例との類似度の値が最も低い負例 N_1 として保存する．また，$R_1 = U \setminus N_1$ を残りのラベルなし文書集合とする．収束処理では，正例 P とマッピング処理で得られた $N=N_1$ を負例として SVMs を用いて学習し，R_1 を分類する．負例に分類された文書を N_2 とし，$N = N_1 \cup N_2$ とする． P と N を用いて学習し，$R_2 = U \setminus (N_1 \cup N_2)$ を分類する．負例に分類された文書を N_3 とし，$N = N_1 \cup N_2 \cup N_3$ とする．以上の処理を負例に分類される文書がなくなるまで繰り返す．得られた負例文書集合を用いて学習し，最終的にテストデータを分類する．

3-2 誤り修正と Boosting を用いた学習・分類

Yu らが提案した PEBL 手法における収束処理では，ラベルなし文書集合中の本来正例である文書が誤って負例に分類されてしまうときがある．その場合，収束処理が収束しないばかりか結果的に分類精度を低下させてしまう． そこで本研究では， 収束処理において，SVMs により負例と判断された文書に対して我々が提

案した誤りの検出と修正手法(Fukumoto, 2006)を適用することで負例文書を正確に抽出することを試みる．さらに，誤り検出・修正を行った結果に対して Boosting (Schapire, 2000)を適用し，弱い分類器を繰り返し学習することで誤分類を減らすことを試みる．本手法の流れを図 2 に示す．

図 2 において，マッピング処理では，素性選択の一つである $\chi 2$ 統計量を用いて正例文書集合 P とラベルなし文書集合 U から正例を特徴づける単語を抽出する．収束処理では正例文書集合 P とマッピング処理で得られた負例文書集合 $N=N1$ を入力とし，誤り検出・修正処理を適用する．その結果に対して Boosting を適用することで負例集合 $N2$ を抽出し，$N=N1 \cup N2$ とする．P と N を入力し誤り検出･修正処理を適用し，その結果に対して Boosting を適用することで負例に分類された文書集合を $N3$ とし，$N=N1 \cup N2 \cup N3$ とする．以上の処理を負例に分類される文書がなくなるまで繰り返す．最終的に分類すべきテストデータは，Boosting の最終段階で得られた分類器の多数決により判定される．

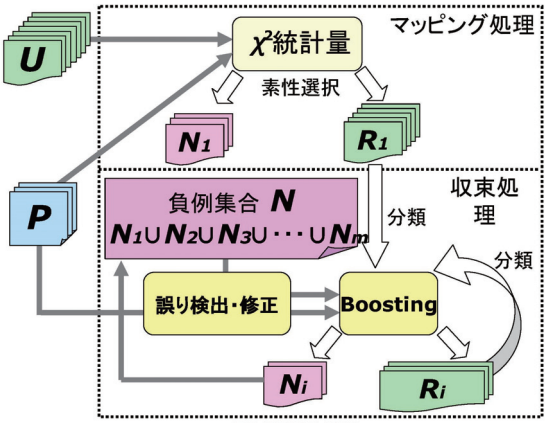

図2 提案手法の流れ

3-2-1 素性選択による $N1$ の決定

本手法では，マッピング処理における素性選択として $\chi 2$ 乗値を用いた．$\chi 2$ 値はその値が高いほど，単語 w が正例を特徴づける単語であるとみなせる．正例とラベルなし文書集合に対して $\chi 2$ 値を適用し，得られた値を降順にソートした結果，上位 X 語を抽出する．ラベルなし文書集合からこの X 語を含まない文書を抽出し，これを正例との類似度の値が最も低い負例(集合) $N1$ とする．

3-2-2 負例文書に対する誤り検出と修正

誤り検出と修正は，(1) 誤り候補文書の抽出，(2) 最小エラー率推定に基づく検出と修正からなる．処理の流れを図 3 にに示す．

3-2-2-1 誤り候補文書の抽出

分類に悪影響を与える文書候補を抽出するため，SVMs で得られるサポートベクトルを利用する．サポートベクトルはテストデータの分類に関与する文書であるため，分類に悪影響を与える文書があるとすれば，それはサポートベクトルの集合に含まれていると考えたためである．

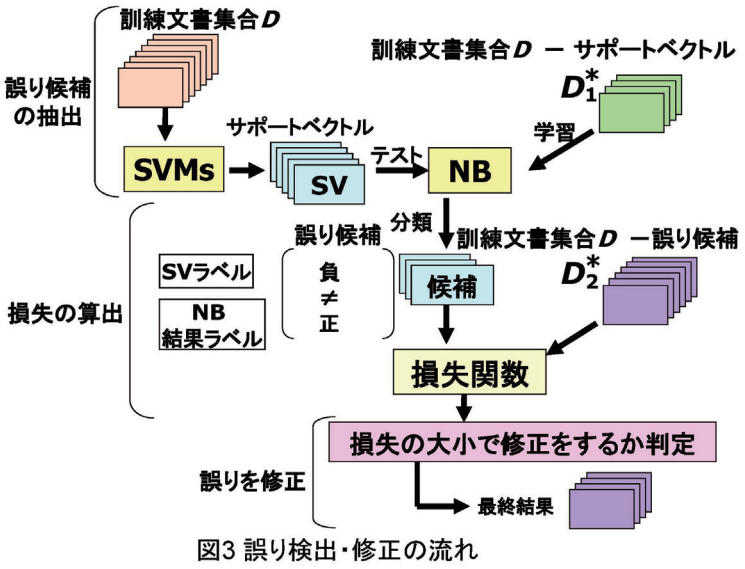

図3 誤り検出･修正の流れ

図 3 において訓練文書集合を D とし，SVMs により D を学習した結果得られる負のサポートベクトルを $\{x1, \ldots x1\}$ とする．$D1*$ は D から 1 個のサポートベクトルを除いた残りの文書集合とする．これをを用いて Naive Bayes(NB)で学習を行い，各サポートベクトルをテストデータとし，分類を行う．xk が NB により正例と判定された場合，その文書は誤りである可能性が高いと考え，これを誤り候補として抽出した．

3-2-2-2 最小エラー率推定に基づく検出と修正

誤り検出と修正は，最小エラー率を推定する損失関数に基づき判断される．ここで損失とはテスト文書の真の確率，すなわちテスト文書が与えられたときの分野への正しい事後確率と実際に機械学習を用いて求めた確率との差を示す．この差(損失値)が小さいほど両者のずれが少ないことを示し，学習により得られる確率が真の確率に近い，すなわち優れた学習法であると言える．損失関数を用いた誤り修正と検出は，誤り候補に対して損失関数を用い損失値を求めた結果，予め付与されたラベルによる損失値が大きい場合にそのラベルは誤りであると見なし，ラベルを正例に修正する．

3-2-3 Boosting 手法の適用

本手法は，順次抽出した負例に対し，負例文書を正確に判定するため誤りの検出と修正を行う．しかし，PEBL 手法で負例として抽出された文書の中に負例の特徴が弱い文書が含まれている場合がある．この場合，分類精度の低い分類器が作成され，次の収束処理において，正しい負例が抽出されにくくなるという問題が生じる．そこで本研究では，Boosting(Schapire, 2000)を適用し，特徴の弱い分類器を繰り返し学習することで誤分類を減らすことを試みる． Boosting は，分類が困難な特徴の弱い文書を集中的に学習することで分類精度の向上を図る手法である．本手法では，誤り検出・修正を行った結果である負例集合 Ni を Nに加えた結果と Pを訓練文書集合 Tdとし，Boosting を適用する．Boosting では，先ず全ての文書の重みを等しくしたデータを用いて既存の学習手法によりベース分類器を作成する．次に作成した分類器が誤分類した文書に対して高い重みを付与し，その文書を用いて再びベース分類器を作成する．これを T回繰り返すことで T個のベース分類器を作成する． 収束処理における誤り検出・修正処理の結果得られる負例集合 Niに対し，誤って分類された文書には Boosting により高い重みを付与することで Ni を修正していく．ラベルなし文書集合 Ri は T個の分類器の多数決を用いて分類される．最終的に分類すべきテストデータは，最終段階である Pと Nmを学習した結果得られる T個から成る分類器の多数決により判定される．

3-3 実験

3-3-1 実験データと評価尺度

本手法の有効性を検証するために UDC 国際十進分類法を用いた文書分類の実験を行った．実験では SVM-Light (Joachims, 1998)を使用し，SVMs のカーネル関数は線形カーネルを用いた．また全ての実験において Boosting の学習回数は 100 回とした．各実験における訓練,及びテストデータの各文書に対して「茶筌」(茶筌)を用いて形態素解析を行い，名詞及び動詞を抽出した．抽出した名詞，及び動詞を SVMs の素性として用い，素性値は単語の出現回数とした．マッピング処理では，名詞と動詞に対して$\chi 2$ 統計量を求め，値の大きい順に上位 100 語を抽出した．ラベルなし訓練文書のうち，この 100 語を含まない文書(集合)を負例 NI として用いた．細分化された分野が複数付与された文書に対する本手法の有効性を検証するため，UDC 国際十進分類法の分野が付与された UDC データを用いて文書分類を行った． 実験に用いた UDC データは，1994 年の毎日新聞記事 1 年分の 21,993 文書に対して人手により分野名を付与した文書集合である．1 文書当たりの平均分野数は 2.7 分野であり，最大分野数は 9 分野の文書を使用した．実験では,21,993 文書から 1 分野あたり訓練，及びテストにそれぞれ最低 50 文書，20 文書以上割り当てるという条件下で，無作為に 6,000 文書抽出し，テストデータとして用いた，

また，残り 15,993 文書を訓練データとして用いた．UDC の分野名は階層構造で表現されている．本研究では上位 1，2，及び 3 階層の分野名のうち，訓練データの文書が 50 文書以上存在する分野を抽出した．

本手法の有効性を検証するため，ラベルありの SVMs のみ用いた手法と PEBL のみ用いた手法，PEBL と誤り修正を行った手法(以下，誤り修正手法)，PEBL と Boosting を用いた手法(以下，Boosting 手法)の実験を行い，本手法(PEBL に誤り修正と Boosting を行った手法)との比較を行った．実験では，少数の正例とラベルなし文書集合を用い分類精度を求めた．各手法で用いたデータを以下に示す． また，分類精度は，適合率・再現率に基づく F 値を用いた．

1 SVMs 手法

各分野から 1 つの分野を正例の分野とし，無作為に 50 文書抽出する． この 50 文書を訓練データの正例として使用する．また，訓練データ 15,993 文書から正例の分野名が付与されていない文書を負例として用いた．負例文書数は，正例と同じ数である 50 文書と，精度向上のため数を増やした 250 文書で実験を行った．SVMs 手法の実験では，負例数が分類精度に影響するため，負例数を変え実験を行った結果，負例の数が 250 のときに最も高い精度を得られたため，実験では負例数を 250 とした．

2 PEBL 手法，誤り修正手法, Boosting 手法，及び提案手法

正例は，SVMs 手法と同じ 50 文書を用いた．ラベルなし文書集合は，訓練データ 15,993 文書から正例の 50 文書を除いた残りの文書とした．

3-3-2 文書分類実験

実験結果を表 1 に示す．表 1 は各手法における階層ごとの F 値のマクロ平均を示し，表中の"*"は提案手法の分類精度が有意水準 5%で有意差がみられたことを示す．SVMs_50，及び SVMs_250 は，50 文書の正例と 50 文書，及び 250 文書から成る負例文書集合を用い，SVMs により学習・分類を行った結果を示す．表中の括弧は，収束処理において負例と判定される文書数がゼロとなるまでの学習回数を示す．

表 1　分類精度

	SVM_50	PEBL	誤り修正	Boosting	提案手法	SVM_250
1 層	0.442*	0.653* (24)	0.671* (23)	0.674* (23)	0.694 (21)	0.688
2 層	0.334*	0.574* (28)	0.582* (27)	0.588 (26)	0.611 (24)	0.623
3 層	0.283*	0.525* (30)	0.537* (30)	0.538* (29)	0.567 (28)	0.567
平均	0.353*	0.584* (27)	0.597* (27)	0.600* (26)	0.624 (24)	0.626

＊ 有意水準 5%で有意差あり

表 1 によると，本手法のマクロ平均 F 値は 0.624 であり，PEBL，誤り修正，Boosting，及び SVMs_50 のいずれよりも高い精度が得られている．また各階層における本手法の分類精度はそれぞれ 0.694，0.611，0.567 であり，第 2 階層における Boosting のみ適用した結果を除き，精度向上に有意な差がみられている．負例と判定される文書数がゼロとなるまでの学習回数は，誤り修正，あるいは Boosting を行うことで PEBL よりも若干減少している．さらに，本手法の学習回数が僅かではあるが誤り修正や Boosting を単独で行った結果よりも減少していることから，誤り修正と Boosting の両者を適用することの効果が表れていると言える．また，本手法と SVMs_250 との比較においても分類精度はそれぞれ 0.624 と 0.626 であり，両者に有意な差はみられなかった．特に法律分野 第 1 階層(法律,法律学)では提案手法が SVMs_250 よりも，3%精度が向上している．

3-3-3 誤り修正の精度

文書分類の実験では，第 2 階層を除き，PEBL と Boosting を適用した結果に対し，誤り検出・修正を加えた本手法は有意水準 5%で有意差があった．そこで，収束処理における分類・テストの繰り返し過程で得られる負例文書に対して誤りをどれだけ検出・修正できたかを人手により確認した．実験では，第 3 階層における各分野を用い，その平均値を求めた．実験結果を図 4，及び図 5 に示す．図 5 における誤り修正の評価尺度は，適合率と再現率を用いた．

図 4 は第 3 階層を用いた場合の各学習における負例と判断された文書数，及びその中に含まれている誤りの文書数を示す．また図 5 は各学習における誤り修正の再現率と適合率を示す．

図 4 によると，学習回数が 10 回までは全負例数に対して誤り文書の占める割合は，約 10%である．11 回から負例と判断される文書数が少なくなるにつれ誤り文書の占める割合も高くなり，15 回では 15 文書中 7 文書が誤りであった．

学習回数が増加するにつれラベルなし文書は正例か負例かの判断が難しくなるため，誤り文書が含まれる割合も高くな

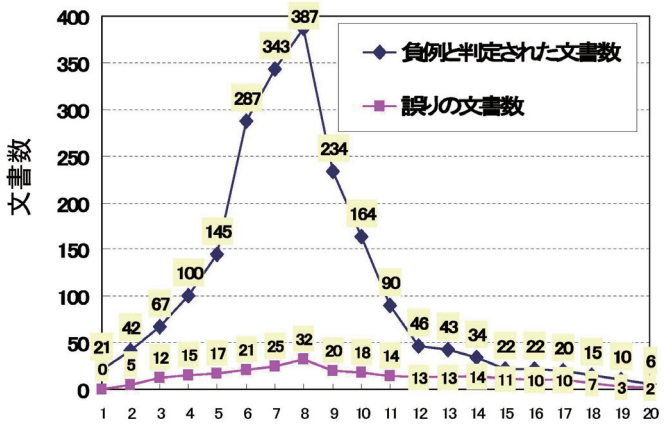

図4 収束処理における学習回数

ることは自然であると考えられる．図 5 によると，適合率は各学習において再現率よりも高く，平均適合率は 0.816 であった．このことから誤り検出・修正処理により，全ての誤りを検出できてはいないものの，検出された文書のうち 8 割程度の文書については正しく誤りを修正できていることがわかる．一方，再現率の平均は 0.502 であり適合率と比べると低い．特に学習回数が 12 回になると再現率は 0.385 であり 20 回で 0.300 まで低下する．これは，学習回数が多くなるにつれ負例と判断される文書数が少なくなると同時に，正例か負例かの判断が難しい文書が多く含まれてしまうためである．

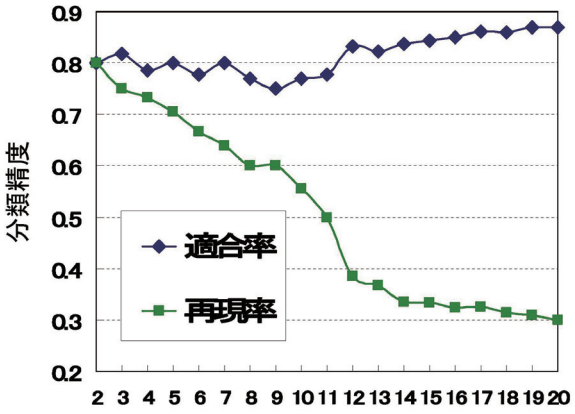

図5 収束処理における学習回数

本手法における誤り検出・抽出において誤りが正しく修正できない要因として(1) 誤り候補として正しく抽出できない，(2) 誤り候補として抽出できたものの誤りを正しく修正できないことがある．各学習における誤り文書を調査したところ，ほとんどの負例文書が(1)，すなわち誤りであるにもかかわらず，誤りの候補として抽出されなかった．本研究で用いた誤り検出・修正手法は，誤り候補を判定するための学習法として NB を用いていることから，少ない訓練文書数では誤り候補が正しく判定できず本手法の限界であると言える．今後は，誤り候補抽出に対して用いる学習法を検討する必要がある．

4. 対訳語の自動抽出

本研究では階層構造の各ノードに位置する文書の分類結果から類似カテゴリィの組を抽出した結果を用い，関連文書を抽出し，対訳語を自動的に抽出する．

4-1 類似カテゴリィ対の抽出

カテゴリィを統合するための一つ目の処理は，異なる階層間において類似したカテゴリィの組を生成することである．本研究では，類似した2つのカテゴリィ を類似カテゴリィ対と呼び，3つ以上のカテゴリィを類似カテゴリィ集合と呼ぶ．カテゴリィの類似性を判定するために，一方の階層構造に分類されている文書をテストデータとし，これらを他方の階層構造に分類する．我々は，Reuters'96とRWCPコーパス，及び Yahoo!JapanとOpen Directory Project(ODP)を用いて実験を行った．ここでは前者のデータを用いて説明する．Reuters'96を訓練データとし，日英機械翻訳ソフトウェアを用い，RWCPコーパスを英語に翻訳した結果をテストデータとして分類を行った．文書分類にはSVMsを用いた．次に文書分類結果を用いてカテゴリィ同士の類似性を求める．類似性の尺度として3章で述べたχ2値をここでも用いた．χ2値を正規化しその値をカテゴリィ同士の類似度とする．類似度が一定の閾値以上である場合，その対を類似カテゴリィ対として抽出した．

4-2 類似カテゴリィ集合の生成

カテゴリィを統合するための二つ目の処理は3つ以上から成る類似したカテゴリィの集合を生成する処理である．本研究では処理1で生成した類似カテゴリィ対に対してAgrawalらにより提案されたアプリオリアルゴリズムを適用することで，類似カテゴリィ集合を生成した(Agrawal, 1995)．アプリオリアルゴリズムは類似度が最小のサポート値を満たす要素数kの集合から末尾の要素のみが異なる集合を探し出し,その要素を結合することにより要素数k+1の集合を作成するというものである．例として図6を用い，要素数2のカテゴリィ集合から要素数3の類似カテゴリィ集合を生成する処理を説明する．要素数kを2，類似度の閾値を0.1とした場合，図6で示す要素数2のカテゴリィ集合の中で類似度の値が0.1以上のカテゴリィ集合の個数は4となる．これら4つの集合から要素数3の類似カテゴリィ集合の候補を生成する.要素数2の類似カテゴリィ集合の先頭に注目し，この集合に対して末尾のカテゴリィのみが異なる集合を探し出し，そのカテゴリィを注目しているカテゴリィ集合に結合することにより，要素数3の類似カテゴリィ集合の候補を生成する．以上の処理を4つの集合の各々に対して行い，候補を生成する．次にこの候補が類似カテゴリィ集合であるかどうかの判断

を行う．生成された候補の先頭に注目し，この候補から要素数が1だけ少ない部分集合の全てを作成し，候補の生成元である要素数2の類似カテゴリィ集合に存在するか否か判断する．全ての部分集合が存在した場合にのみ，その候補は類似カテゴリィ集合と判定される．この処理を繰り返すことにより要素数が3以上のカテゴリィ集合を生成する．得られた要素数k+1の類似カテゴリィ集合における類似度は，結合した2つの集合の平均とした．

類似したカテゴリィ集合の生成において，部分集合が候補生成元の類似カテゴリィ集合に存在するかを確認する際に，異なる階層構造のカテゴリィ要素を持つ類似カテゴリィ対だけでは，要素数3以上の類似カテゴリィ集合を生成することができない．そこで同じ階層内の親と子，及び親と孫の関係を持つカテゴリィ同士を無条件で類似カテゴリィ組とし，本手法により生成された類似カテゴリィ組に追加することにより，類似カテゴリィ集合の生成を行った．

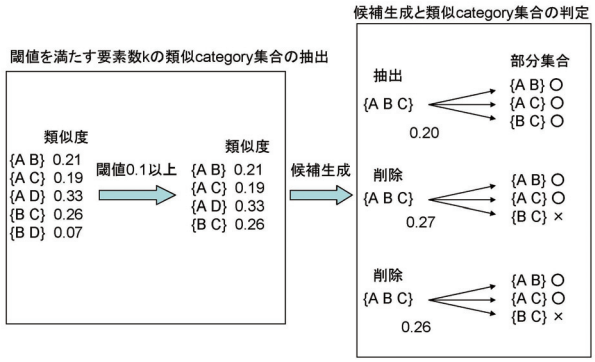

図6 候補生成と類似category集合の判定 (k=2)

4-3 関連文書抽出

類似カテゴリィ対，及び類似カテゴリィ集合の結果を用いることにより関連文書の抽出を行った．図7は統合結果である類似カテゴリィ対の例を示す． 図7においてReuters'96のカテゴリィの一つである"Science and Technology"と UDC 毎日新聞のカテゴリィの一つである"Space Navigation"が抽出されたとする．カテゴリィ "Science and Technology"に属する Reuters の文書，及び"Space Navigation"に属するUDC毎日新聞の翻訳結果である文書をそれぞれ単語の異なり数を次元とし単語数を要素とするベクトルで表現する．余弦尺度を用い2文書の類似度を求める．類似度が一定の閾値以上の2文書を類似した文書として抽出する．

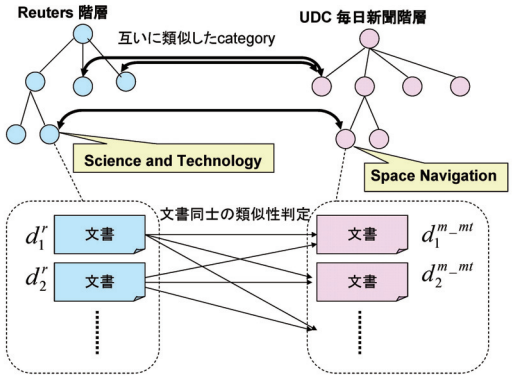

図7 関連文書の抽出

4-4 対訳語の抽出

4-3 節で得られた関連文書対の集合を用い，対訳語の抽出を行う．対訳語の抽出は以下で示す２つの処理から成る．

(1) 文書に基づく対訳語の抽出

式(1) を用い，Reuters'96 の動詞－名詞表現 *vnr*，及び UDC 毎日新聞の動詞－名詞表現 *vnm* を抽出する．

$$\{vn_r, vn_m\} s.t. vn_r \in \exists d_i^r, vnm \in \exists d_j^m, \qquad BM25(d_i^r, d_j^{m\_mt}) \geq L_\theta$$

$$BM25(d_i^r, d_j^{m\_mt}) = \sum_{w \in d_j^{m\_mt}} w^{(1)} \frac{(k_1+1)tf}{K+tf} \frac{(k_3+1)qtf}{k_3+qtf} \qquad (1)$$

ここで *w* はUDC毎日新聞記事の英語訳の文書 *dj* 内に出現する単語であり *w(1)* は重み付けされた単語である．*k1*，及び *k3* はそれぞれ 1，1,000 に設定されている。*tf*，及び *qtf* はそれぞれ文書 *di* 及び *dj* に出現する単語数を示す．式(1)により得られた{*vnr*，*vnm*}からそれらの $\chi 2$ 値を求める．

(2) 文に基づく対訳語の抽出

一般に，文書内の動詞－名詞組は多数存在するため(1)で求めた対訳語にはノイズが多く含まれてしまう．そこで，文レベルの組を抽出することによりそれらのノイズを除去する．文レベルの対訳語抽出の式を式(2)

に示す.

$$Ssim(vn_r, vn_m) = \max_{S\ vn_r \in Set_r, S\ vn_m \in Set_m} sim(S\ vn_r, S\ vn_m) \quad (2)$$

ここで, Set_rとSet_mはそれぞれReuters, 及びUDC毎日新聞の動詞－名詞対を含む文の集合を示す. 式(2)中の類似度sim は式(3)で示される.

$$sim(S\ vn_r, S\ vn_m) = co(S\ vn_r \cap S_{mt}\ vn_m) / (|S\ vn_r| + |S_{mt}\ vn_m| - 2co(S\ vn_r \cap S_{mt}\ vn_m) + 2) \quad (3)$$

|X| は文X内に含まれる内容語の個数を示し, 式(3)の分子は, Reuters, 及びUDC毎日新聞の動詞－名詞対の両者に含まれる内容語の個数を示す. 本研究では以下に示す式(4)を満たすReuters, 及びUDC毎日新聞の動詞－名詞対を対訳語として抽出する.

$$\{vn_r, vn_m\} = \underset{\{vn_{r\_} \in BP(vn_m),\ vn_m\}}{\operatorname{argmax}} Ssim(vn_{r\_}, vn_m) \quad (4)$$

4-5 実験

4-5-1 カテゴリィ の統合実験

実験では、ODPとYahoo!Japan 及び, Reuters'96とUDC毎日新聞記事を用いた. ODPは, 6,227文書が297カテゴリィ, Yahoo!Japanは28,691文書が305カテゴリィに分類されている. Reuters'96は1996年8月20日から1997年7月19日までの1年分, 806,791記事が4階層から成る126カテゴリィ(分野)に分類されているものを用いた. UDC毎日新聞記事は1994年の27,755記事が7階層から成る9,951カテゴリィ に分類されている. 本研究では, IBM機械翻訳ソフトウェアを用いてそれぞれの記事を翻訳した. 我々は, Reuters(ODP)及びUDC毎日新聞記事(Yahoo!Japan)を訓練データとテストデータに分割した. さらにテストデータをパラメータ推定用のテストデータと推定されたパラメータ値を用いてカテゴリィ対を抽出するためのデータとに分けた. パラメータ推定の結果, 式(1)の閾値を0.003 と設定した. 実験ではReuters(ODP), 及びUDC(Yahoo!Japan)のカテゴリィ のうち, 各カテゴリィで文書が5文書以上存在する109, 4,739のカテゴリィ を用いた. カテゴリィ統合の評価として再現率・適合率に基づくF値を用いた. Reuters, 及びUDCデータを用いた実験結果を表2に示す.

表2 カテゴリィの統合結果

	階層構造あり			階層構造なし		
	適合率	再現率	F値	適合率	再現率	F値
Reuters & UDC毎日	0.503	0.463	0.482	0.462	0.389	0.422
Reuters	0.342	0.329	0.335	0.240	0.296	0.265
UDC毎日	0.157	0.293	0.204	0.149	0.277	0.194

表2において"Reuters& UDC毎日"は本手法を示し, "Reuters", 及び"UDC毎日"はそれぞれ一方の階層構造を用いた結果を示す. 表2より階層構造ありの場合に本手法のF値は0.482でありReuters, 及びUDC毎日のF値はそれぞれ0.335, 0.204であることから, カテゴリィの統合結果は意味的に類似したカテゴリィ対の抽出に効果的であることがわかる. 抽出されたカテゴリィ対のうち, 類似度の最も高い上位7対を表3, またYahoo!JapanとODPの結果を表4に示す.

表3 カテゴリィ対の抽出例(Reuters & UDC毎日新聞)

Reuters	UDC毎日
Science and technology	運輸交通工学(Highway and transportation technology)
Environment	環境学(Environmental studies)
Annual results	調査結果(Annual results)
Unemployment	雇用者(Employer)
Domestic markets	商業(Commerce)
Labour	労働(Labour)
New products/services	通信業務(Telecommunications service)

表4 カテゴリィ対の抽出例(ODP & Yahoo!Japan)

ODP	Yahoo!Japan
Science/Math	自然科学と技術/数学
Games/Puzzles	趣味とスポーツ/ゲーム
Sports/Fantasy	趣味とスポーツ/ギャンブル
Reference/Dictionaries	各種資料と情報源/辞書
News/Weather	メディアとニュース/天気
Recreation/Motorcycles	趣味とスポーツ/自転車
Business	ビジネスと経済

4-5-2 関連文書抽出

4-5-2節で得られたカテゴリィ対を用い，関連文書を抽出する実験を行った．関連文書抽出で用いたデータはReuters，及び毎日新聞の1997年6月13日から21日までの記事を用いた．それぞれの記事の日付差は±3日とした．例えば毎日新聞記事が6月18日である場合，関連文書抽出の対象となるReutersの記事は6月15日から6月21日までとなる．日付差を用いた結果，抽出の対象となるReuters記事は15,482記事，毎日新聞記事は391となり，実際に関連する記事を人手により抽出した結果，513記事となった．実験結果を表5に示す．

表5 関連文書抽出(Reuters & UDC毎日新聞)

	適合率	再現率	F値	$L\theta$
階層構造なし	0.417	0.322	0.363	40
Reuters階層	0.356	0.544	0.430	20
階層構造の統合結果	0.839	0.585	0.689	20

表5は階層構造なし，Reutersの階層構造のみ使用，及び本手法で得られた統合結果を用いて類似文書を抽出した最高精度と，そのときの閾値$L\theta$の値を示す．表より明らかに，本手法のF値が0.689であることから統合することが有効であることを示している．同様にODPについては769，Yahoo!Japan3,048ページを用いた．実際に関連する記事を人手により抽出した結果，3,659記事となった．実験結果を表6に示す．

表6 関連文書抽出(ODP & Yahoo!Japan)

	適合率	再現率	F値	$L\theta$
階層構造なし	0.170	0.076	0.105	30
ODP階層	0.144	0.205	0.169	10
階層構造の統合結果	0.121	0.607	0.202	10

Reuters & UDC毎日新聞記事の精度と比較すると全般的に精度は低下しているが，表6によると階層構造の統合結果が一番優れていることから，統合が関連文書抽出に有効であるということがわかる．

4-5-3 対訳語抽出

関連記事抽出の結果を用いて対訳語を抽出する実験を行った．ここではReuters，及びUDC毎日新聞記事データを用いた結果を示す．対訳語の抽出実験ではReuters，及びUDC毎日新聞記事は同年月，すなわち1996年8月20日から1997年7月19日の１年分の記事である806,791，及び119,822記事を用いて実験を行った．記事の日付差は±3日とした．表7に4-5-2の結果得られた関連記事数を示す．

表7関連文書対の総数

手法 （$L\theta$）	対の数	英語文書	日本語文書

階層構造なし (40)	3,042,166	428,042	70,080
Reutersの階層構造 (20)	27,181,243	43,0181	99,452
本手法により得られたカテゴリィ対 (20)	81,904,243	45,965	654,787

我々は，表7のデータから対訳語の抽出を行った．実験結果を表6に示す．

表8 対訳語の抽出結果

手法	単言語の対		候補			上位1,000対		
			候補数		Rate(Doc&Sent/Doc)	対の数		Rate(Doc&Sent/Doc)
	日本語	英語	Doc & Sent	Doc		Doc & Sent	Doc	
階層構造なし	25,163	44,762	25,163	6,976,214	0.361	177	62	2.9
Reuters	10,576	37,022	10,576	1,273,102	0.831	268	64	4.2
本手法	8,347	21,524	8,347	5560,472	1.489	328	72	4.6

表8は対訳語の抽出結果を示す．"Doc&Sent"は文書と文ベースの抽出結果である本手法を示し"Doc"は文書から対訳語を抽出した結果を示す．表8より階層構造を統合した本手法は階層構造なしの場合と比較すると15.1%(32.8-17.7)，Reutersの階層構造のみ用いた場合と比較すると6.0%(32.8-26.8)の精度の向上がみられた．我々は328の対を人手により評価した結果，既存の辞書に含まれていない78対を抽出することができた．されに51.2%に相当する168対が機械翻訳ソフトウェアを用いても正しく翻訳されない対であることが明らかになった．このことから本手法は，既存の辞書にない対を知識源として抽出することができ，辞書を補完するための手法として有効であるということが言える.カテゴリィ名 スポーツから抽出された対訳語の一例を表9に示す．表9において(X,Y)のXはReuters，YはUDC毎日新聞のカテゴリィを示す．実験では，"earn medal"や"block shot"などのように154の対訳語のうち12の対訳語はReutersの階層構造のみを用いた場合にも抽出することができた.一方"get strikeout"や"make birdie"などスポーツ,野球,ゴルフに属するものは,Reutersの階層構造を用いた場合のみでは得られることができなかった．これらのことからも階層構造の統合は，対訳語を抽出する上で有効であることがわかる.

表9 対訳語の抽出例

手法 Reuters	カテゴリィ / カテゴリィ 対	対の数 Doc & Sent	対の数 Doc	対の数 (%) Doc & Sent	対訳語例
本手法	スポーツ	262	19,391	36(13.7)	Block shot Earn medal
	(スポーツ，野球)	110	8,838	24(21.8)	Get strikeout
	(スポーツ，ゴルフ)	177	3,418	28(21.4)	Make birdie
	(スポーツ,サッカー)	115	2,656	39(39.5)	Block shot Give free kick
	(スポーツ，スキージャンプ)	68	661	10(14.7)	Earn medal Postpone downhill

5. まとめ

本研究では，異なる言語で記述された複数のインターネットディレクトリィを統合することで得られる多言語関連文書から，機械翻訳システムに必要となる翻訳知識を自動的に獲得する手法を提案した．具体的には，(1) 分野名が付与されている少数のデータと分野名が付与されていない大量のデータを用いることで階層への分類を高精度で実現する手法を提案した．さらに，(2) 分野の対応付けを用いることで日英対応文書

を抽出し，対応文書からの対訳語を抽出する手法を提案した．文書分類の実験の結果，本手法のマクロ平均F値は0.624であり，従来手法であるPEBLが0.597，また人手で作成した50文書，及び250文書からなる負例文書を用いたSVMsによる結果がそれぞれ0.353,0.626であったことから負例収集に対する本手法の有効性が確認できた． また階層構造におけるカテゴリィの統合実験では，Reuters，及びUDC毎日新聞1種類の階層構造のみを用いた場合のF値はそれぞれ0.335，0.204に対して本手法では0.482であった．また，関連文書の抽出実験の結果，階層構造を用いない場合， Reuters の階層構造のみを用いた場合の F 値はそれぞれ0.363，0.430であるのに対して，本手法は0.689の精度が得られた．同様に，階層構造を用いない場合，ODPの階層構造を用いた場合のF値はそれぞれ0.105，0.169であるのに対して，本手法は0.202の精度が得られた．対訳語の抽出実験の結果，階層構造を統合した本手法は階層構造なしの場合と比較すると15.1%(32.8-17.7),Reutersの階層構造のみ用いた場合と比較すると6.0%(32.8-26.8)の精度の向上がみられた．さらに，ReutersとUDC毎日新聞記事データ，及びODPとYahoo!Japanいずれの場合も，既存の辞書に含まれていない対訳語も抽出することができたことから，辞書を補完するための手法として有効であるということが言える．今後の課題として，(1) 文書分類のさらなる精度の向上，(2) 動詞－名詞以外の共起に関する対訳語の抽出，(3) 既存知識への統合などを検討する予定である．

【参考文献】

[Nigam,2000] Nigam, K., "Text Classification from Labeled and Unlabeled Documents using EM", *Machine Learning*, Vol.32, No. 2, pp. 103-134, 2000.

[Joachims,1999] Joahims, T., "Transductive Inference for Text Classification using Support Vector Machines, Proc. of the ICML'09, pp. 200-209, 1999.

[Manevitz,2001] Manevitz, L. M. and Yousef, M. "One-class SVMs for Document Classification", Machine Learning, Vol. 2, No2. pp. 139-154, 2001.

[Yu,2000] Yu,H., Han, H. and Chang, K. C-C. "PEBL Positive Example based Learning for Web Page Classification using SVM, Proc. of the ACM Special Interest Group on Knowledge Discovery and Data Mining, pp. 239-248, 2002.

[Fukumoto,2004] "Correcting Category Errors in Text Classification", Prof. of the 20th International Conference on Computational Linguistics, pp. 868-875. 2004.

[Fukumoto,2013] " Text Classification from Positive and Unlabeled Data using Misclassified Data Correction", Proc. of the Annual Meeting of the Association for Computational Linguistics 2013, *To Appear*

[Yamamoto,2912] ″少数の正例とラベルなし事例による文書の自動分類″，山本剛士，福本文代，松吉俊，IEICE論文誌，Vol. J95-D, No. 9, pp. 1794-1801, 2012.

[Ichise,2003] Ichise,R. Takeda, H. and Honiden, S, "Integrating Multiple Internet Directories by Instance-based Learning, In Proc. of the 18th International Joint Conference on Artificial Intelligence, pp. 22-28, 2003.

[濱崎, 2004] 濱崎雅弘，″不均質な情報源間での情報共有支援″，博士論文，総合研究大学院大学，2004.

[内山,2003] 内山将夫，井佐原均，″日英の記事および文を対応付けるための高信頼性尺度″，言語処理学会，Vol. 10, No. 4, pp. 201-220, 2003.

[Schapire,2000] Schapire, R. E. and Singer, Y., "BoosTexter: A Boosting-based System for Text Categorization", Machine Learning, Vol. 39, No. 2, pp. 135-168, 2000.

[Joachims,1998] Joachims, T., "SVM Light Support Vector Machine", Dept. of Computer Science Cornell University, 1998.

[茶筌] 松本裕治他，″形態素解析システム「茶筌』，奈良先端科学技術大学院大学，松本研究室，2000.

[agrawal,2001] Agrawal,R. and Srikant, R., "On Integrating Catalogs", In Proc. of the 10th International World Wide Web Conference, pp. 603-612, 2001.

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
Text Classification with Relatively Small Positive Documents and Unlabeled Data	Proc. of the 21st ACM International Conference on Information and Knowledge Management, pp. 2315-2318, 2012	2012. 10
Classifying Hotel Reviews into Criteria for Review Summarization	Proc. of the 2nd Workshop on Sentiment Analysis where AI meets Psychology, pp. 65-72, 2012	2012.12
Text Classification from Positive and Unlabeled Data using Misclassified Data Correction	Proc. of the Annual Meeting of the Association for Computational Linguistics 2013, *To Appear*	2013. 8