3A4-1

言語横断テキストマイニング

Cross-lingual Text Mining

海野 裕也　　那須川 哲哉
Yuya Unno　　Tetsuya Nasukawa

日本アイ・ビー・エム株式会社 東京基礎研究所
IBM Research - Tokyo, IBM Japan Co., Ltd.

We propose a novel text-analytic approach, *cross-lingual text mining*, which analyzes documents in multiple languages from the perspective of a single language. This system automatically adjusts its parameters to obtain domain dependent translations only using a general bilingual dictionary, and analyzes foreign documents through them. The experimental results show that it successfully obtains optimal parameters and makes it possible to retrieve insights from foreign documents without knowing the foreign language.

1. はじめに

企業や組織が世界各国で活動を展開するにつれ，複数言語で書かれた文書を蓄えることが増えている．それに伴い，多言語の文書集合から知見を発見する需要が高まっている．テキストマイニングは大量の文書集合から有用な情報を引き出す技術の総称である．特に情報抽出技術に基づく重要表現の抽出と，その出現の偏りを分析することが知見の発見に有効であることが実証されてきた [Feldman 98, Feldman 07, 那須川 06]．これらの手法は解析対象の言語と専門知識を必要とし，解析者は一般に専門知識は有するものの他言語を理解することができない．従って，言語を超えた文書解析技術が必要であった．一方，言語の壁を越えるための機械翻訳技術は長年取り組まれているものの，未だに完璧な翻訳を得るには至っていない．本研究は，現状の機械翻訳のための基盤技術をテキストマイニング技術に応用し，多言語文書からの知見発見を目指すものである．

本研究のポイントは 2 点ある．1 つは分野依存性の高い文書集合から，対訳関係にある表現対を精度よく自動で獲得することである．特に解析対象の文書は，専門用語や業界用語を含むなど分野に強く依存するため，解析対象の文書集合から訳語を抽出する必要がある [那須川 09]．この際，多様で雑音の多い現実のデータからいかに高い精度で訳語を抽出できるかが課題となる．そこで，同一分野の文書集合から訳語を検出する手法をベースとし，汎用の訳語辞書を擬似正解とみなしてパラメタの最適化を図る．もう 1 点は，得られた訳語対を知見発見のためのテキストマイニングに組み合わせることである．完璧な訳語対が抽出されることを期待することは難しいが，関連語が検出されることは多い．表現の出現の偏りを検出する際，こうした関連語でも知見発見に結びつく可能性がある．

日米の政府組織が公開する自動車不具合データを用いて実験を行った結果，パラメタの最適化によって訳語検出の精度を高めることができたこと，および部分的ながら多言語文書内の特徴的な共起関係を検出できたことを報告する．

関連した技術として言語横断検索がある [Nie 99]．検索技術は文書集合から発見したい表現がはっきりしているときには役に立つ．一方で，テキストマイニングは明示的なクエリ表現を与えず，概念の集合から特徴的な表現の出現を見つけ出す．文書集合から新しい知見を発見するという本研究の目的には，非自明な情報を検出できない情報検索だけでは不十分である．

連絡先: 海野 裕也，日本アイ・ビー・エム株式会社，大和市下鶴間 1623-14，yunno@jp.ibm.com

2. テキストマイニング

一言で “テキストマイニング” といっても，その手法や目的は様々である [Hearst 99, Feldman 07]．本論文では大量の文書から有用な情報を検出することを目的とする．このために役立つのが，情報の出現の偏りや変化を捉えることである．ある 2 つの情報が特徴的に偏って共起している場合，何らかの意味を持っている可能性が高い．

2.1 自己相互情報量による偏りの分析

自己相互情報量（Point-wise mutual information, PMI）は，2 つの条件の関連を測る最も代表的な指標の 1 つである．条件 x, y に対する PMI は

$$\mathrm{PMI}(x, y) = \log \frac{p(x, y)}{p(x)p(y)}$$

で定義される．これは条件 x と y が独立な場合に比べて，どれくらい特徴的に共起しているかを示す指標となる．2 条件が独立であると値は 0 となり，関連があるほど大きな値をとる．

文書から概念の偏りを捉える上で重要なのは，適切な単位で表現を抽出することである [Feldman 98]．例えば，単に単語の共起を捉えると,「東京」と「都庁」のような自明な共起，特に連語関係が見つかるに過ぎないことが多い．このために重要なのが情報抽出技術である．例えば名詞の連なりや単語の係り受けを，あるいは概念の辞書を事前に用意して，抽出したい概念の表現を明確にする必要がある．

PMI を計算する際に，同時確率と周辺確率は各条件を満たす文書数から推定される．単純に相対頻度を使って推定すると，文書数が少ないときに偶然高い値を示すことがあり，これらを適切に排除する必要がある．そこで，相対頻度の代わりに区間推定の下側信頼限界の値を使う．この値は，サンプル数が少ないとき相対頻度に比べて小さく，無限大のサンプル数で相対頻度に収束する．本論文では 90%の信頼区間を使用した．

実際に分析を行うときの典型的なシナリオを紹介する．1) まず 2 つのカテゴリーを選択する．ここでカテゴリーとは，例えば「部品名」などのように解析対象の表現のまとまりである.「車種名」や「報告日」のような定型データである場合もある．2) 次に 2 つのカテゴリー中の全表現に対して，その組み合わせの PMI を計算し，高いものを探す．3) 検出された表

現対の両方を含む文書を読み，似たような文脈で言及されているか確認する．最後のステップは重要で，共起していたとしても実際にはまったく共通項がないことはよくある．また，以上の過程は文書集合から**気づき**を得ることが目的であり，このあとどういうアクションを取るかは個別に判断する．

3. 訳語対の自動抽出

我々の目的は企業のもつ現実の文書集合から言語横断的に解析を行うことである．こうしたデータは多数の専門用語や固有名を含み，分野への依存度が強い．そのため，一般の訳語辞書には含まれない語，あるいは特定分野では訳として不適切な語が存在する．また，分野ごとに訳語辞書を作るのはコストの点で現実的ではない．そのため解析対象の文書から訳語対を抽出する必要がある．ただし，データ形式や得られる量も様々なので，対訳コーパスなどの特殊なデータを仮定した手法を使うことはできない．精度の面でこれらの最新の手法には劣るが，同一分野の二言語コーパスのみを仮定した手法を採用する．

那須川らは，似たような意味内容の記述であるが，対訳関係にない文書集合のみから訳語対を抽出する手法を提案している [那須川 09]．こうした文書集合内では同じ意味表現は別言語でも似た文脈で現れやすいため，共起表現の頻度を元に訳語対を抽出する．まず元言語の表現に対して，元言語文書集合中で特徴的に共起する表現を取得し，これを**ピボット表現**と呼ぶ．ここでは前節で定義した PMI が，あらかじめ定めた閾値 θ_s より高い表現をピボットとする．ピボット表現は元の表現の特徴を表す．次に，元言語と対象言語間の汎用的な辞書を使って，ピボット表現を翻訳する．翻訳された各ピボット表現に対して，先と同様に特徴的に共起する表現を対象言語文書から探すことで，訳語候補を生成する．この閾値を θ_t とする．最後に，得られた各訳語候補に対して，先と同様 θ_t 以上の PMI をもつ表現を抽出し，元の表現のピボットとの一致数でランキングする．ランクに応じて上位 n 件を訳語候補として出力する．

この手法は特殊なデータが必要ない上，分野特化した訳語対を獲得できるため適用範囲が広い．一方で，高い精度を得るには適切なパラメタを設定する必要がある．

4. 言語横断テキストマイニング

解析対象の文書集合とは異なる言語を使って文書の解析を行うことを言語横断テキストマイニングと呼ぶことにする．本研究は以下のような状況を想定する．解析者が理解できない言語で記述された文書集合があり，ここから知見を得たい．解析者は母語による解析の経験はあるため，分野に関する知識，例えば「車の部品名」や「車種名」が分析に重要であることは知っている．そこで，解析したい概念を母語で提示すると，システムは対応する対象言語の表現，つまり “engine” に対する「エンジン」を自動で検知し，その出現の偏りを示す．こうした状況は多国籍企業では自然であり，このシステムにより対象分野の専門知識のみで多言語の解析ができるようになる．

分野依存の表現の多い別言語の文書に対して PMI を推定するにはどうすればよいか．まず，解析対象の各表現の訳語候補を獲得する．対象文書からこれを得ることで，分野特化した訳語辞書を作ることができる．獲得された辞書で元言語の表現を翻訳し，対象文書から訳語表現の出現の偏りを検出する．

4.1 擬似正解を利用したパラメタの自動調整

那須川らの訳語対抽出手法 [那須川 09] にはいくつかのパラメタが存在するが，最適なパラメタはデータの規模や分野によって異なることが予想される．そこで，小規模な訳語辞書からこれを自動調整する方法を提案する．提案する手法は 2 つの手順からなる．まず，汎用の訳語辞書から分野に特化した訳語対だけを選択し，これを擬似正解とする．次に，訳語抽出手法のパラメタを変化させ，擬似正解に対する正解率が最大になるパラメタを選択する．

汎用的な訳語辞書から擬似正解を作る場合，分野に特化した訳語対を選択する必要がある．たとえば，辞書には “door” の訳語として「玄関」が存在するが，これは自動車の文脈では適切とはいえない．分野から逸脱した訳語にパラメタが調整されないよう，こうした訳語対を正解から排除する．そこで，訳語関係にある表現の出現頻度は，言語に依らず同程度と仮定する．すると，例えば自動車分野において訳語関係にない “door” と「玄関」の出現頻度は大きく異なるはずなので，これを擬似正解から排除できる．1 以下の正のパラメタ σ を使い，元言語表現 e と対象言語表現 j，出現文書の相対頻度 $\mathrm{DF}(\cdot)$ に対して，

$$\sigma \mathrm{DF}(e) < \mathrm{DF}(j)$$

となる訳語 j のみを擬似正解とする．そして，得られた擬似正解に対して精度が最大になるパラメタを選択する．ただし，調整したい元言語の表現集合に対して，得られた訳語候補中に擬似正解がひとつ以上含まれているものの割合を精度とする．

4.2 自動獲得した訳語対を使ったテキストマイニング

自動獲得された訳語対をテキストマイニングに応用する．テキストマイニングは 2 つの表現集合 A，B 間で特徴的な共起の組み合わせを探す点にあるが，多言語環境ではこの表現が対象文書とは別言語で与えられる．文書と異なる言語の表現を探すため，ここでは単純に「A 中の表現 a を含む」という条件の代わりに，「a の訳語候補のいずれかを含む」で置き換える．例えば，自動車部品の表現集合の内，“engine” を含む文の代わりに，訳語候補「エンジン」か「発動機」を含む文書を探す．

訳語候補は前節で得られた候補を利用するが，完璧な訳語が得られることは期待しづらい．ただし，重要なのは元の元言語表現と関連がある文書を見つけられるかである．従って，例えば “engine” という元言語表現の代わりに，「シリンダー」や「ピストン」などの関連語との共起を発見できれば，“engine” の問題を発見するのに役立つかもしれない．

5. 実験

訳語対の自動抽出，パラメタの自動調整，および獲得した訳語を利用した言語横断テキストマイニングの評価を行った．

5.1 実験設定

実験用に政府機関の収集している自動車不具合情報データを用いる．日本語のデータは MLITJ の収集する自動車不具合情報 20,269 文書[*1]を，英語のデータとして NHTSA の収集したデータ 525,055 文書[*2]を用いる．文書数のみならず，前者は 1 文に平均 35 文字，後者は 51 単語と文長も大きく異なる．しかし，現実の企業のデータが多様であることは珍しくなく，いかなるデータにも解析を行えるようにしたい我々の目的からすれば妥当な設定である．

訳語辞書には機械翻訳のために作られた 159,749 単語対からなる訳語辞書を用いる．ただしこの辞書は公開されていない

以下の実験では，英語話者の解析者が日本語文書を解析することを想定する．すなわち，原言語を英語，対象言語を日本

∗1 http://www.mlit.go.jp/jidosha/carinf/rcl/defects.html
∗2 http://www-odi.nhtsa.dot.gov/downloads/index.cfm

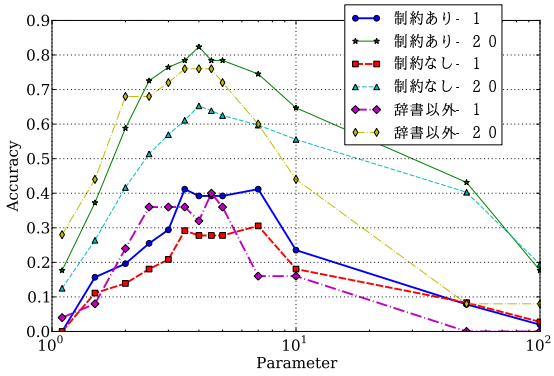

図 1: パラメタを変えたときの訳語正解セットに対する正解率

語とし，英語の表現集合に対して日本語の文書集合から知見を得る実験を行った．

5.2 単言語における解析の予備実験

最初に日本語文書を日本語で解析する．こうして得られた共起を，言語横断解析との比較対照とする．

まず名詞の全出現の半分をカバーする出現頻度上位 104 の名詞を選んだ．ここから解析対象として意味のありそうな 36 の名詞を人手で選別した．各名詞の出現と，およそ 5,000 の車種名との共起分析を行ったところ，PMI の値が log(2) 以上のペアは全部で 192 対あった．なお，車種名は文書の定型データとしてあらかじめ与えられている．

得られた名詞と車種名のペアに対して，例えば「車種 A」が「エンジン」の問題を抱えているかを判断するには車に対する調査が必要である．この調査は難しいので，両方を含む文書集合を人手で調べ，問題がありそうかどうかの主観評価を行った．主観評価には，一貫した問題に言及しているかで判断する．例えば，“door” に対して,「ドアで異音がする」「バックドアがへこんだ」「ドアの鍵が閉まらない」といった文書集合が得られたとする．これらは全て別々の言及をしており，関連がありそうとはいえない．逆に特徴的に同じことを言及しているときは関連があるとみなす．この主観評価の結果，192 対の内，142 対に問題がありそうと判断され，29 対は関連が無く，残りの 21 対は判断できなかった．このように，特徴的な共起関係の多くが製品の不具合を暗示する可能性が強い．評価指標として，いくつの共起関係を検出できたか，また検出された共起関係の内のいくつに関連がありそうか，また日本語で検出された 192 共起対のいくつをカバーできたかという点で評価する．

5.3 パラメタの自動調整実験の結果

次に訳語抽出の実験を行った．パラメタの調整のために，自動車部品を示す 100 の英語表現を選ぶ．このうち，訳語辞書，およびその訳語が対象の文書集合中に含まれたのは 65 表現 142 対であった．4.1 節で定義した擬似正解の脚きりパラメタ σ を 0.1 として制約をかけた．つまり，対象言語での相対頻度が，元言語での相対頻度の 10%未満の表現は擬似正解から排除した．この結果，51 表現と対応する 69 訳語対が擬似正解として得られた．得られた訳語対が適切か人手で判定した結果，前者は 59.9%の 85 対が，後者は 89.2%の 58 対が，自動車分野に適切な訳語であると判定された．この脚きり制約によって 30%近く，正解データとしての精度が上がったことになる．

訳語抽出パラメタをかえて，各訳語正解セットに対して正解率をプロットしたのが図 1 である．制約ありとなしの擬似正解，および訳語辞書に含まれなかった 35 の英語表現に対して，それぞれ訳語候補数を 1 と 20 にしたときの合計 6 通りで，対象言語のパラメタ θ_t に対する正解率をプロットした．まずわかるのは，正解率の程度は異なるものの，いずれのグラフも似たようなパラメタでピークを持っている点である．訳語候補数をさらに変えたり，もう一方のパラメタ θ_s に対するプロットでも同様であった．ここから 3 つのことがわかる．まず，汎用辞書から作った擬似正解に対して最適化したパラメタによって，解析対象の表現に対する訳語が適切に抽出できた．次に，抽出訳語の候補数に，最適なパラメタは依存しなかった．最後に，擬似正解のノイズを除去するために脚きり制約をしても，平均的な正解率は異なるものの，ピークを与えるパラメタは変わらなかった．脚きりの影響がない原因として，訳語の獲得手法が文書集合での出現頻度の高いものしか対象にしないからと考えられる．文書集合中に頻度の低い表現は 3. 節の手法で獲得されないため，パラメタをいくら変更しても相対的な精度の良し悪しは変わらない．結果として，ピークをとるパラメタは制約ありとなしであまり変わらなかった考えられる．またこのことは，擬似正解に対する制約は不要なことも示している．

表 1: 取得された訳語対の例

順位	gas	air bag	reverse	system
1	**燃料**	スライドドア	レンジ	エンジン
2	**ガソリン**	**エアバッグ**	駐車場	不良
3	ホース	衝突	D	センサ
4	燃料タンク	**エアバック**	壁	エンスト
5	タンク	助手席	車	制御
6	給油	壁	車庫	警告

※下線かつ太字: 正解　　下線: 関連語

実際に獲得された訳語の例を表 1 に示した．“gas” と “air bag” は全体的にうまくいった例である．特に後者では，表記ゆれの「エアバック」も訳語として取れている点に注目したい．“reverse” は正しい訳語は得られなかったが，いくつかの関連語がとれた．正しい訳語の「バック」が取れなかった原因として，エアバックやバックカメラなどの，複数の文脈で出現するので，適切に特徴を捉えられなかったためと推察される．さらに，“system” はまったく正しい訳語が取れなかったのは，より意味が曖昧で，色々な単語と連語を作っていたのが原因であろう．このように，精度は表現によって大きく異なり，特に文脈の曖昧な表現の訳語は検出が難しいことがわかる．

5.4 獲得した訳語対を使ったテキストマイニングの結果

獲得された訳語対を使い，英語を使って日本語文書の解析を行った．訳語候補の数と抽出される表現対の関係を図 2 に示した．PMI が log(2) 以上の共起対と，日本語表現で検出された 192 対との一致数を表す．最も多くの共起が得られたのは，訳語候補が上位 3 件のときで，283 対が抽出された．これは日本語で解析したときの 192 対に比べて極めて多い．一方，192 対との一致度では上位 1 件を使ったとき，およそ半分の 108 対を検出した．これらの原因は，複数の訳語を許すことで，いわば情報検索のクエリ拡張と同等の効果が得られて，多様な相関を見つけ出すことができたためと考えられる．訳語を増やすと，元の表現で期待される以上の共起を発見できた半面，徐々に元の結果から離れていく．訳語対を増やしすぎると検出数が下がるのは，無相関なキーワードも訳語として認識してしまい，PMI が下がりやすくなるためと考えられる．

得られた共起表現を含む文書集合中で，実際に問題が起こっ

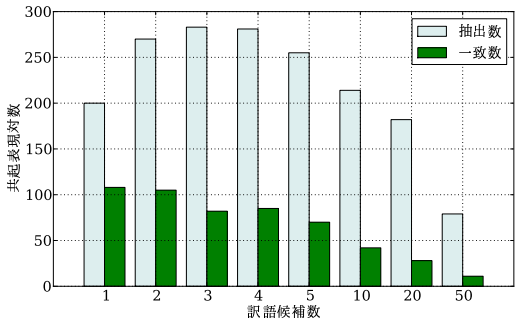

図 2: 訳語候補数と抽出された共起対の数

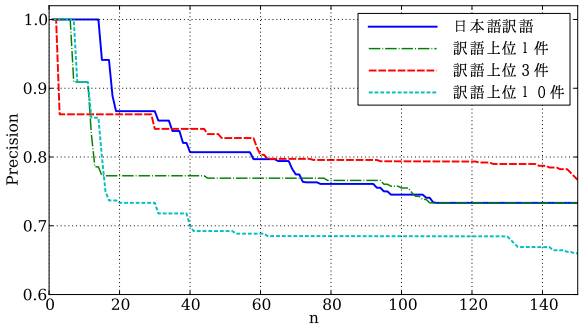

図 3: 上位 n 件の共起対に対する補完適合率

表 2: 同義語や関連語の訳語によって検出された文書の例

表現	訳語	本文（抜粋）
gas	ガソリン	燃料タンクから**ガソリン**が漏れる
	燃料	**燃料**が漏れて、エンストした
transmission	シフト	変速機の不良により、**シフト**ダウンした後に前進しなくなる
	ギヤ	走行中、**ギヤ**トラブルで突然減速した
key	キー	**キー**レスエントリー機構が働かない
	ドア	右側のスライド**ドア**から異音が生ずる

ていそうか主観評価を行い，その精度を示したのが図 3 である．PMI の高い上位 n 件の共起表現対中での精度を人手で測り，補完適合率をプロットした．訳語候補の上位 1，3，10 件を対象言語表現として与えたときを，日本語訳語を与えたときとで比較した．ばらつきはあるものの，上位 3 件のときが平均的にはよく，日本語で与えたときと同程度か若干高かった．これは，訳語候補を増やしたことで新たに得られた共起関係が，実際に意味のある共起であったことを表す．また，$n < 20$ での性能が異なるのは，訳語抽出に失敗した表現中に顕著な共起関係が含まれていたからである．

最後に，同義語や関連語によって得られた文書の具体例を表 2 に示した．それぞれ，表現として与えた用語と，その訳語候補，そして特定の車種名と共起した文書を抜粋した．最初の例は同義語である．自動車の文脈で “gas” は「燃料」や「ガソリン」を意味する．これらの同義語が適切に取れると，精度を落とさずに検出数が向上する．次の例は関連語である．“transmission” の正しい訳語である「変速機」は抽出できなかったが，関連語を抽出できたため変速機に関する問題を検出できた．特に最初の例には，抽出できなかった「変速機」自体を文中に含んでいた．以上の 2 例のように，例えば単純に分野特化した辞書を使う場合よりも幅広い関係を抽出できる可能性がある．一方，最後の例は関連語が悪影響を及ぼしている．「ドア」は “key” の関連語ではあるが，多くの文書で「ドア」への言及は “key” と関係なかった．これは「ギア」の問題が「変速機」と結びつくことと対照的である．また，「キー」と「ドア」の異なる問題を同一視したせいで，共起が得られにくくなる．特に使用する訳語候補数を増やすとこの傾向が顕著で，検出された文書集合内で一貫性が無くなる傾向にある．先の図 3 で訳語候補数をあげすぎると精度が悪化したのはこのためである．

6. まとめ

多言語の文書集合から知見を得るため，言語横断テキストマイニングという新しい文書解析手法を提案した．情報抽出とその出現の偏り検出からなる，従来のテキストマイニング技術に，訳語抽出手法を組み合わせた手法である．重要な点は 2 つある．1 つは汎用の訳語辞書を擬似正解とみなして訳語抽出に必要なパラメタを自動調整した点である．この結果，辞書に含まれない訳語に対しても高い精度で訳語を抽出することができた．もう 1 点は，こうして得た訳語で元言語とは異なる言語の文書を解析した点である．特に，得られた訳語は間違いを多く含むものの，同義語や関連語も含まれるため，単純に人手で用語を訳した場合よりもたくさんの共起関係を検出できた．また，翻訳の失敗により検出できなかった共起もあるものの，人手で訳したときと同程度の精度で問題を発見できる可能性があった．訳語精度の問題はあるものの，部分的にその有用性を実証できたといえる．これにより，分野の専門知識があれば，多様な言語のデータを活用することができるようになる．

謝辞

本研究を行うにあたって，東京大学の Daniel Andrade さん，長岡技術科学大学の村松祐希さんのご協力を仰ぎました．ここに感謝の意を表します．

参考文献

[Feldman 98] Feldman, R., Fresko, M., Kinar, Y., Lindell, Y., Liphstat, O., Rajman, M., Schler, Y., and Zamir, O.: Text mining at the term level, *Principles of Data Mining and Knowledge Discovery*, pp. 65–73 (1998)

[Feldman 07] Feldman, R. and Sanger, J.: *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press (2007)

[Hearst 99] Hearst, M. A.: Untangling Text Data Mining, in *Proceedings of ACL '99*, pp. 3–10 (1999)

[Nie 99] Nie, J.-Y., Simard, M., Isabelle, P., and Durand, R.: Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web, in *Proceedings of SIGIR '99*, pp. 74–81 (1999)

[那須川 06] 那須川 哲哉：テキストマイニングを使う技術/作る技術 - 基礎技術と適用事例から導く本質と活用法, 東京電機大学出版局 (2006)

[那須川 09] 那須川 哲哉, Andrade, D., 海野 裕也, 村松 祐希, 山本 和英：言語横断テキストマイニングのための翻訳対抽出, 言語処理学会第 15 回年次大会発表論文集, pp. 108–111 (2009)