言語解析の手法について

田代 佑妃

2004 年に株式会社 NTT データ数理システムより販売開始された完全自社開発テキストマイニングツール Text Mining Studio を用いた言語解析の手法につきいくつか紹介したい。

単語頻出解析

単語頻出解析とは、データ全体内でどんなキーワードが出現しているのかを確認する手段である。データ内での頻出頻度が高いキーワード順にグラフ化され、解析する際には、頻度の指定（〇回以上〇回以下）、文字数の指定（〇文字以上〇文字以下）、行中に現れる重複単語のカウント方法、上位何件を抽出するのかなど、様々な条件を指定することができる。

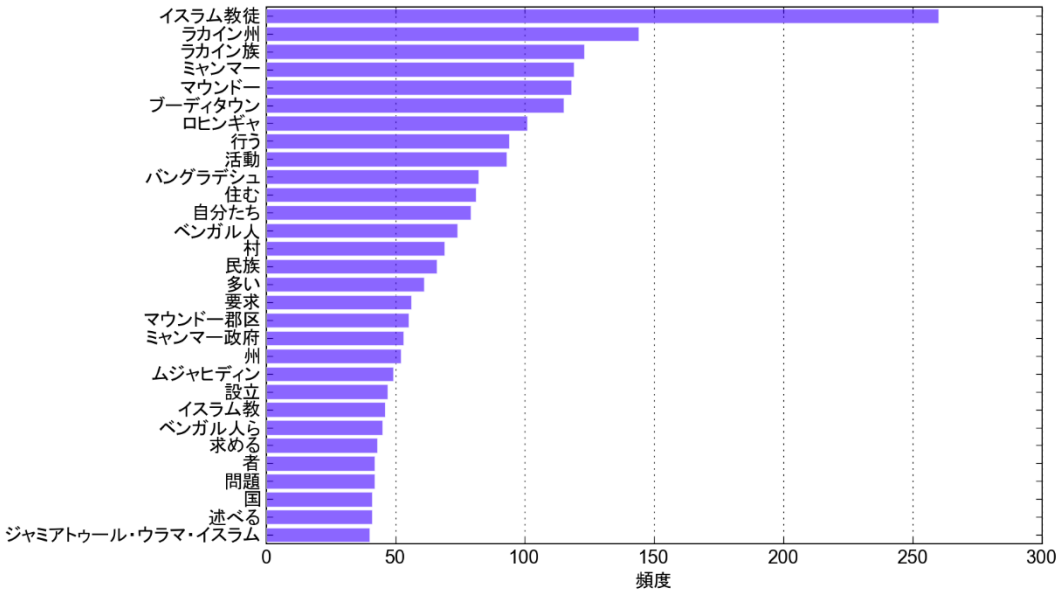

図 1　単語頻出解析「ミャンマー西門の難題」

条件：

頻度（1）回以上（ ）回以下

文字数（1）文字以上（ ）回以下

☑行中に現れる重複単語のカウントを 1 とする。

☑上記の条件を満たすもののうち上位（30）件を抽出する。

　☑上位が同じものは上位件数を超えても出力する。

□「その他」をカウント　　□「合計」をカウント

□「合計」からの割合も抽出する

※（ ）内の数字は変更可能

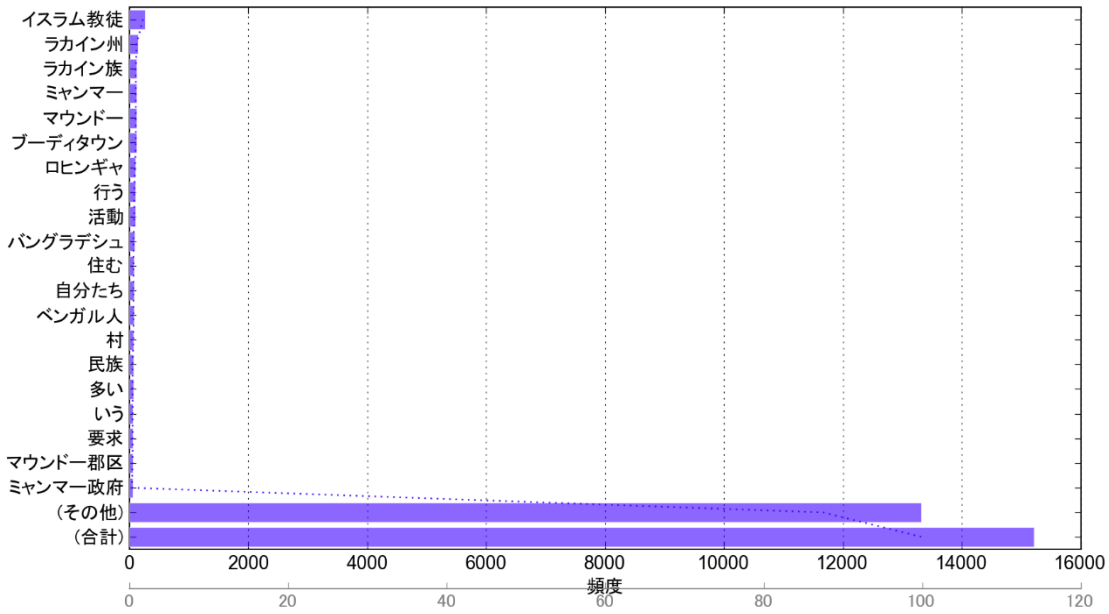

図２　単語頻出解析「ミャンマー西門の難題」

条件：

頻度（1）回以上（ ）回以下

文字数（1）文字以上（ ）回以下

☑行中に現れる重複単語のカウントを 1 とする。

☑上記の条件を満たすもののうち上位（20）件を抽出する。

　☑上位が同じものは上位件数を超えても出力する。

☑「その他」をカウント　　☑「合計」をカウント

☑「合計」からの割合も抽出する

図 2 のように、その他や合計をカウントすることで、それぞれのキーワードが全体（合計）の内、どの位の比率を占めるかもグラフで表すことができる。

また、品詞設定・単語フィルタや態度表現フィルタを活用することにより、キーワードを絞り込んで抽出することも可能である。

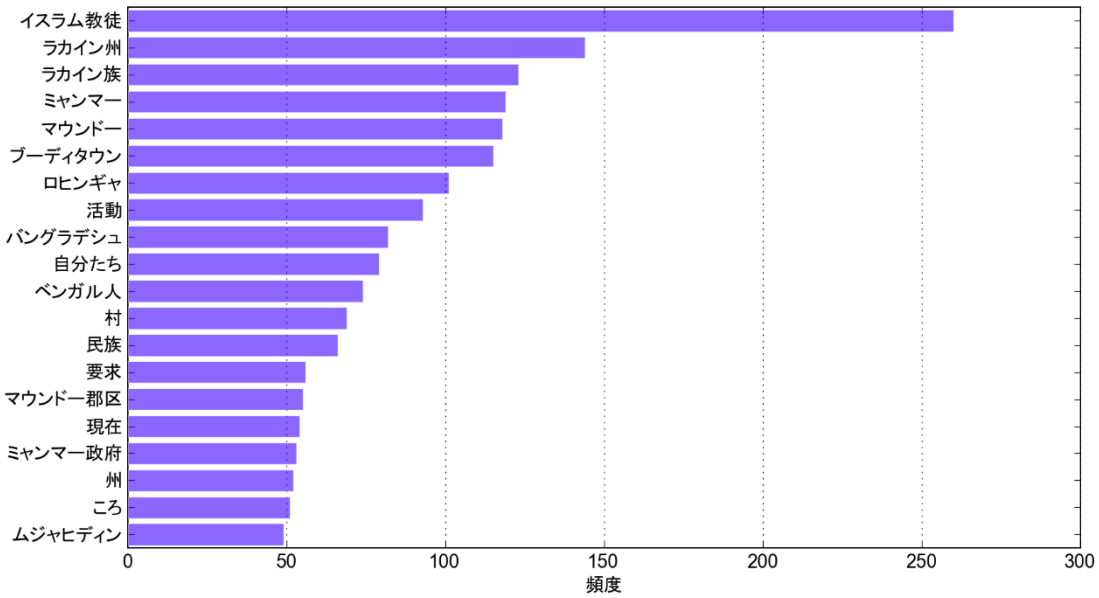

図３　単語頻出解析・「名詞」から始まる品詞のみ抽出　「ミャンマー西門の難題」

係り受け頻度解析

係り受け頻度解析とは、文章中に現れる係り受けの回数を数え、表やグラフに示すものである。解析する際には、品詞の設定・頻度・行中に現れる重複表現のカウント方法、上位何件を抽出するのかなど、様々な条件を指定することが可能である。

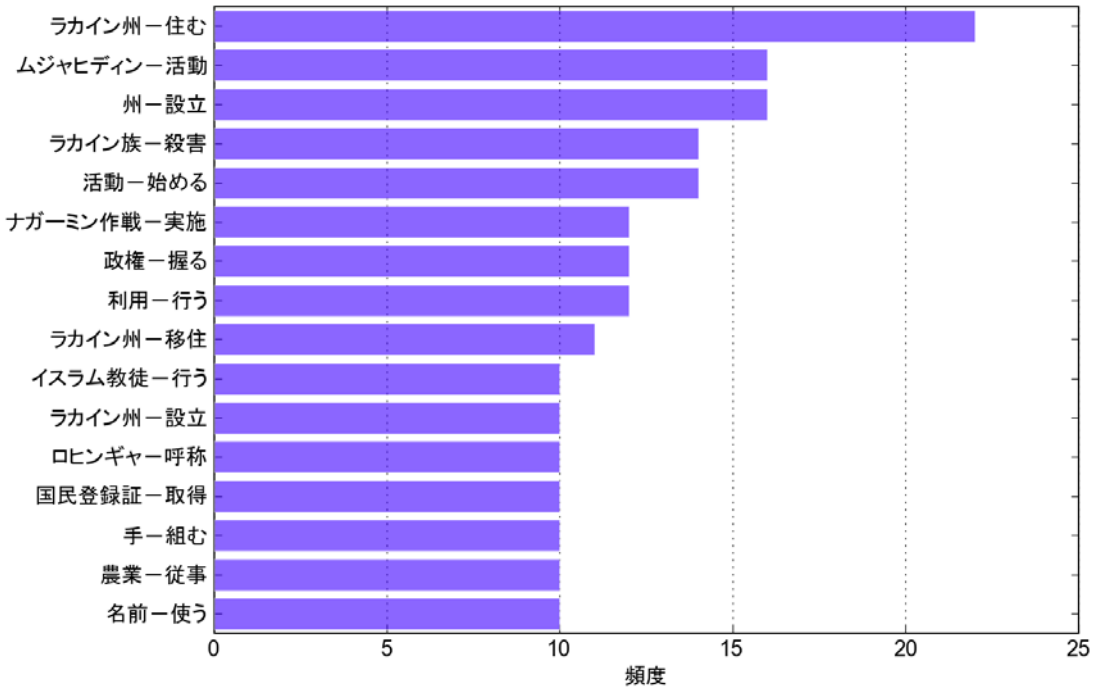

図 4 係り受け頻度解析　「ミャンマー西門の難題」

注目語情報

注目語情報とは、解析の際、気になるキーワードと同時に使用されている単語をみるための機能である。 文章または行（一つの意見）の中で同時に使用されることが多い単語の組み合わせが抽出され、注目した単語が、どのような表現で用いられているか、他のどのような単語・属性と同時に出現（共起）しているかを示す（Text Mining Studio バージョン 6.0 マニュアル p211）。

図 4 は「ラカイン族」をキーワードとし、共起抽出設定・最低信頼度 60、出現回数 5 回以上の共起ルールを最大 100 抽出するという条件で抽出された結果をネットワーク図で表したものである。

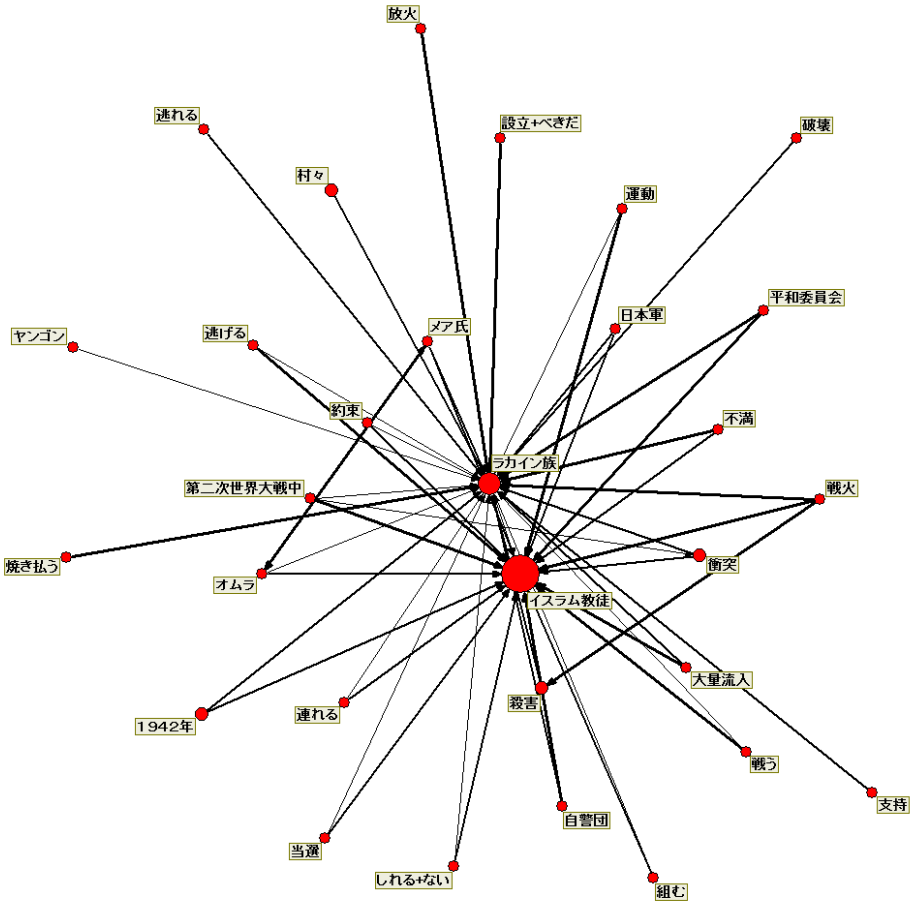

図 4　注目語情報「ラカイン族」　「ミャンマー西門の難題」より

注目語情報において、ある単語 A からある単語 B に矢印が引かれている場合、それは**ある行（または文）に単語 A が出現した場合に、同じ行（文）に単語 B も出現する確率が高い**という、その事実を表している。この場合の単語 A を「前提」、単語 B を「結論」と呼ん

でいる。

確率の値は、パラメータや結果の表にある「信頼度」の数値によって表される(Text Mining Studio バージョン 6.0 マニュアル p347)。例えば、「破壊」といった言葉が出現した場合、同じ行ないし文に「ラカイン族」という言葉が出現する確率が高いといえる。(図 4 参照)

また単語間をつなぐ矢印の太さは、「信頼度」という数値により決定され、ノードの大きさはこの頻度・出現数に対応している (Text Mining Studio バージョン 6.0 マニュアル p256・347)。前提単語 A と結論単語 B の間の「信頼度」とは、同一文章中もしくは同一行中で、単語 A が現れたときに単語 B が同時に出現する確率を表す。この値は 0～100 の間で与えられ、もしこの値が 100 ならば、単語 A が出現するときには必ず単語 B も同一文中に出現していることを示す。値が 50 ならば、単語 A が出現する文章もしくは行のうち、半数において単語 B も同時に出現していることになる (Text Mining Studio バージョン 6.0 マニュアル p214)。

特徴語抽出

特徴語抽出とは、データに付随する属性ごとに、特徴的に出現する単語及び係り受け表現（特徴語表現抽出）を抽出するものであり、解析の際に、対象としたい属性や抽出することばの品詞の選定が可能である。全体の頻度と属性ごとの頻度をもとに、抽出指標となる統計量を求めることで特徴語を抽出しているため、その際の抽出指標の算出方法を選択できる（Text Mining Studio バージョン 6.0 マニュアル p222～226)。

抽出指標の算定方法一覧

・**補完類似度**:単語頻度の大小を考慮した上で、その属性に偏って多く出現する言葉を抽出。
・**x 二乗値**:属性間でもし全く偏りがない場合には本来この言葉はこの程度出現するはずであるといった値(期待頻度)を基準として、実際にどの程度偏りがあるかを示す指標を計算。（頻度の小さい言葉が重視され易くなる傾向がある。）
・**Yates 補正 x 二乗値**:頻度が小さい場合に制度が向上するよう、x 二乗値に Yates の補正と呼ばれる操作を加えたもの。
・**相互情報量**:その単語が、実質的に属性についてどれだけの情報を持っているかに着目し、属性についてのエントロピーをより減少させる単語を抽出。
・**Dice 係数**:テキストの中から、ある言葉が出現している部分の集合とある属性を持つ部分の集合とを考え、その間の集合の類似度を与える尺度。
・**Cosine**:テキストの全体を 1 つのベクトルと考え、ある属性で特徴付けられるベクトルとある言葉で特徴付けられるベクトルとの類似を計算。
・**頻度**:言葉の頻度をそのまま指標とし母数を考慮しない。

・**頻度割合**：その言葉の、その属性での出現割合に基づいて判断。
・**Fisher の直接確率**：現状の状態以上の言葉の偏りが偶然発生するとしたら、その確率はどの程度かという値を直接求める。この値が小さいほど、偶然には起こりえない状況、すなわち偶然ではなく実際にその属性において本当に特徴的なケースであると考えられる。
（Text Mining Studio バージョン 6.0 マニュアル p226～227）

下記図 5～8 において、解析条件は変えず、抽出指標のみを変更して解析した結果を参考までに転記する。

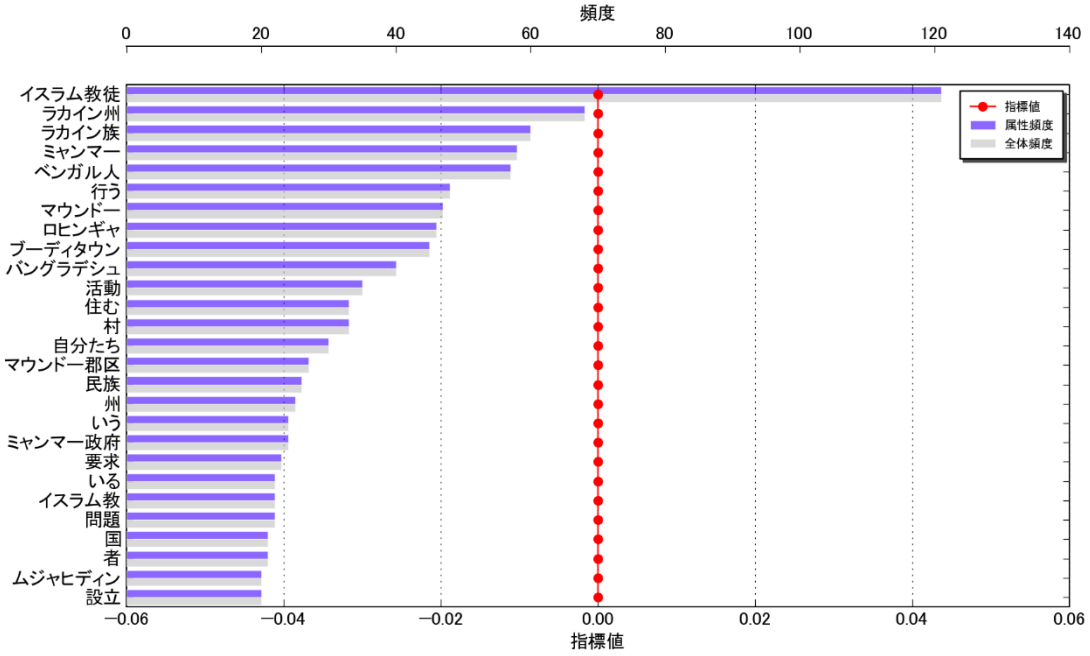

図 5　特徴語抽出「ミャンマー西門の難題」より　抽出指標：補完類似度の場合

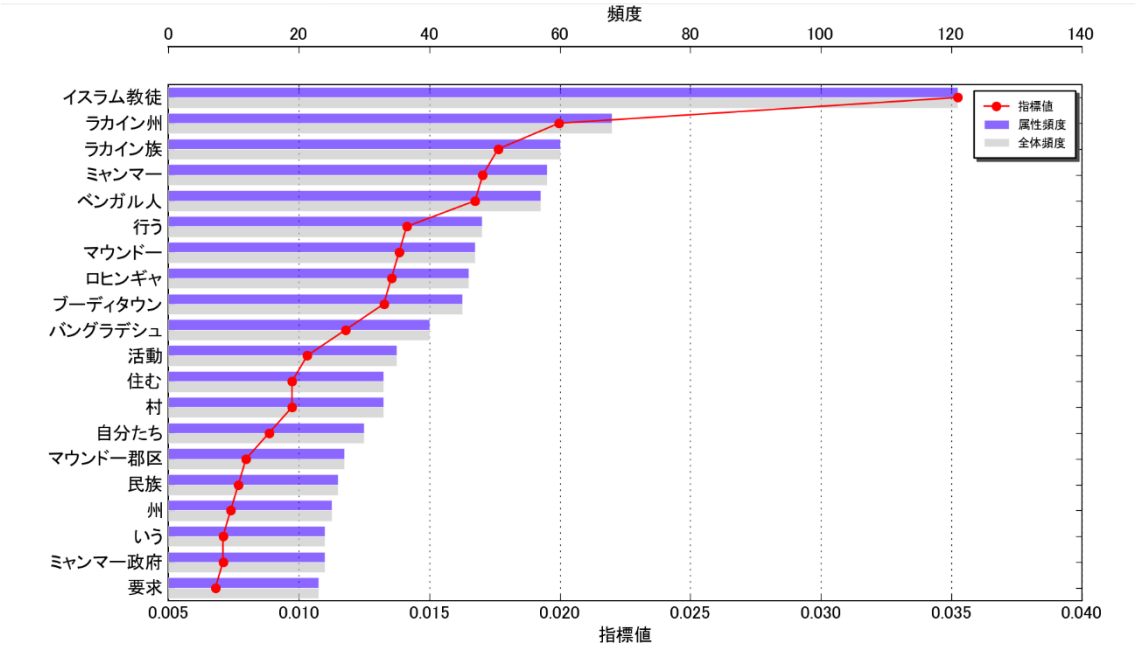

図 6　特徴語抽出「ミャンマー西門の難題」より　抽出指標：Dice 係数の場合

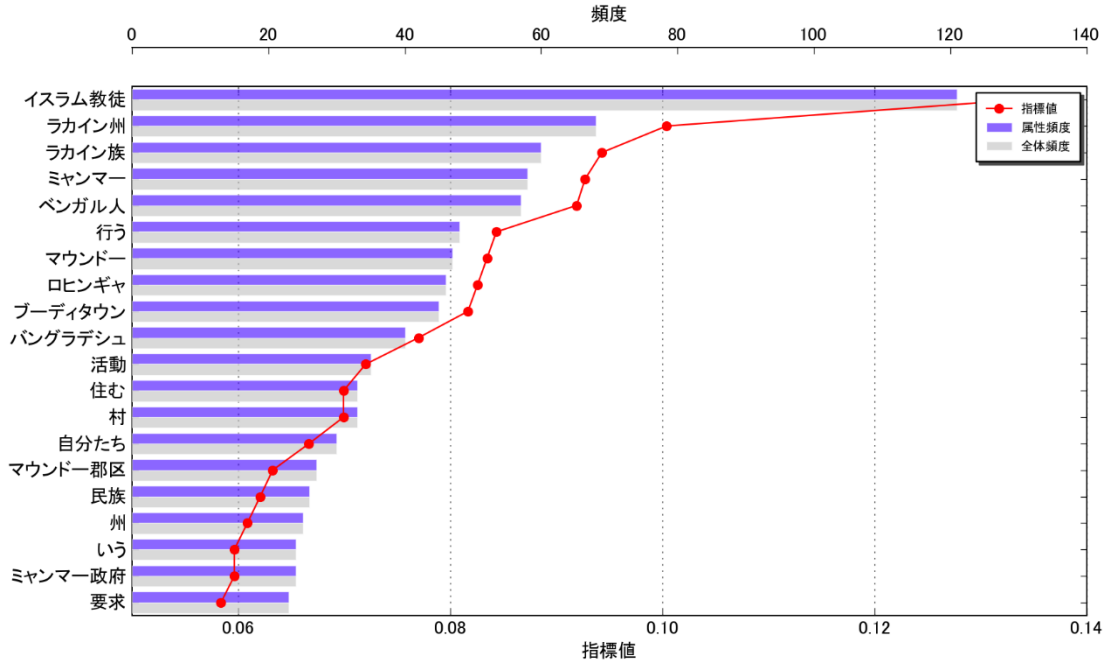

図 7　特徴語抽出「ミャンマー西門の難題」より　抽出指標：Cosine の場合

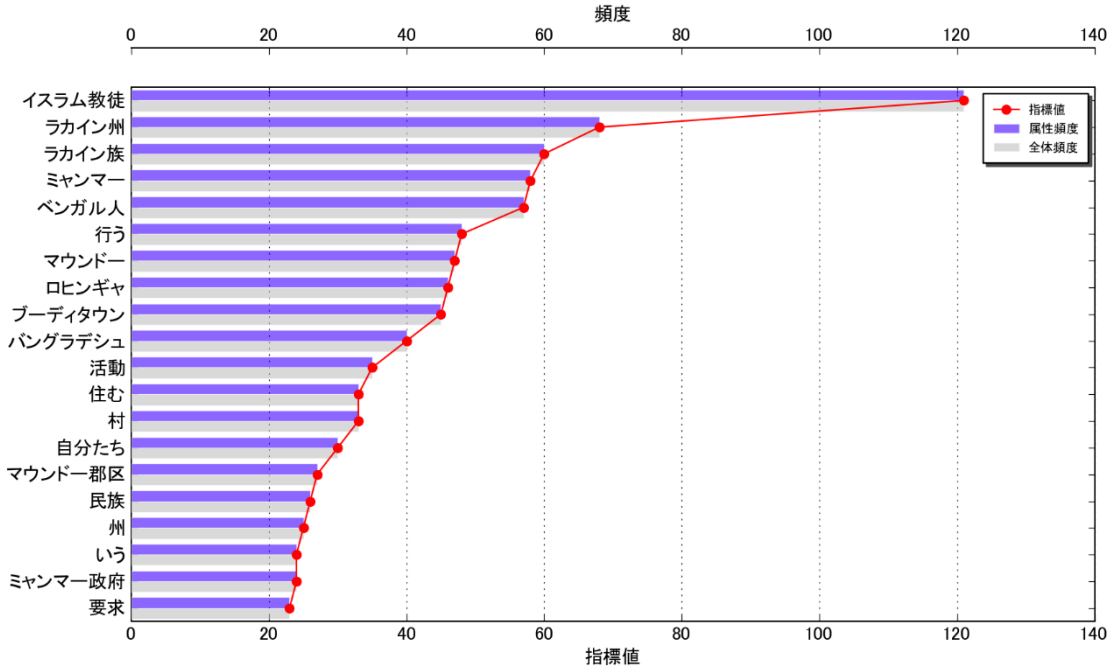

図８　特徴語抽出「ミャンマー西門の難題」より　抽出指標：頻度の場合

評判抽出

評判抽出とは、データ全体内で良いイメージ・悪いイメージで語られる言葉を抽出する手段である。単語に対して、好意的・非好意的表現のそれぞれ語られた回数をカウントし、それをもとに好評語・不評語のランキングを作成する（Text Mining Studio バージョン 6.0 マニュアル p235）。

Text Mining Studio では、好評・もしくは不評の評価を与える単語を「評価語」と呼び、3000 語強の評価語データベースを保持しており（別紙評価語一覧を参照のこと）、これらの評価語と各々の単語との係り受けに関係に着目して、好意的な評価語との係り受けが発生していることばには好評（Positive）の点数を、非好意的な評価語との係り受けが発生していることばについて不評（Negative）の点数を与えている（Text Mining Studio バージョン 6.0 マニュアル p235-236）。

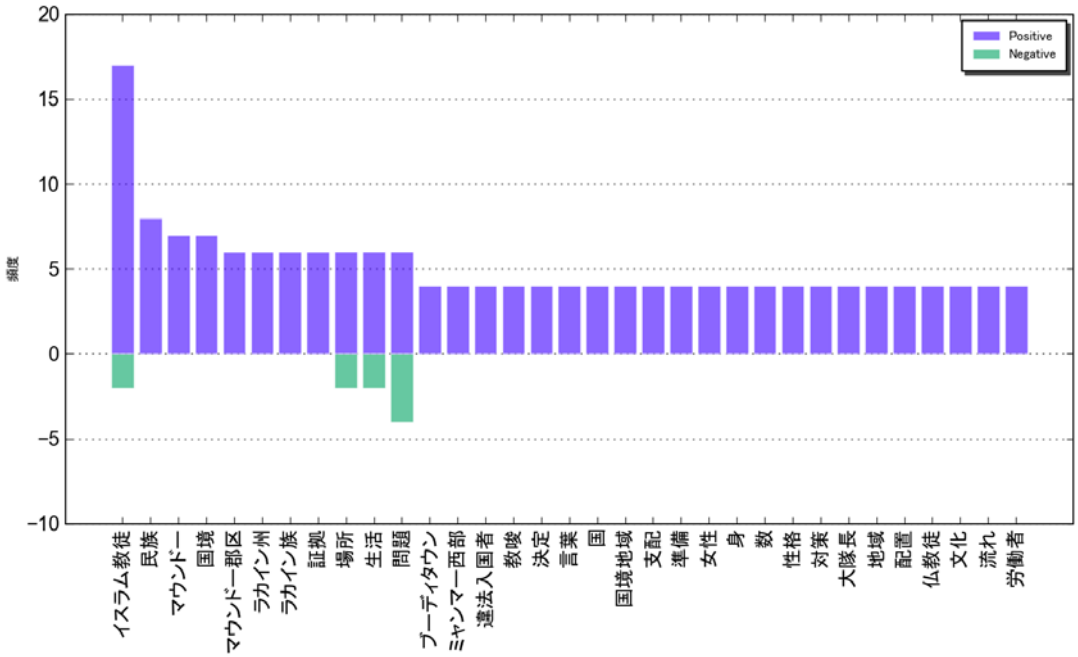

図 9　好評語ランキング　「ミャンマー西門の難題」

好評語ランキングでは、Positive（好意的）な表現で用いられている単語の一覧が回数の多い順に表示され、「Negative（非好意的）」は同一の単語が非好意的な表現で語られた回数を負数で示している。

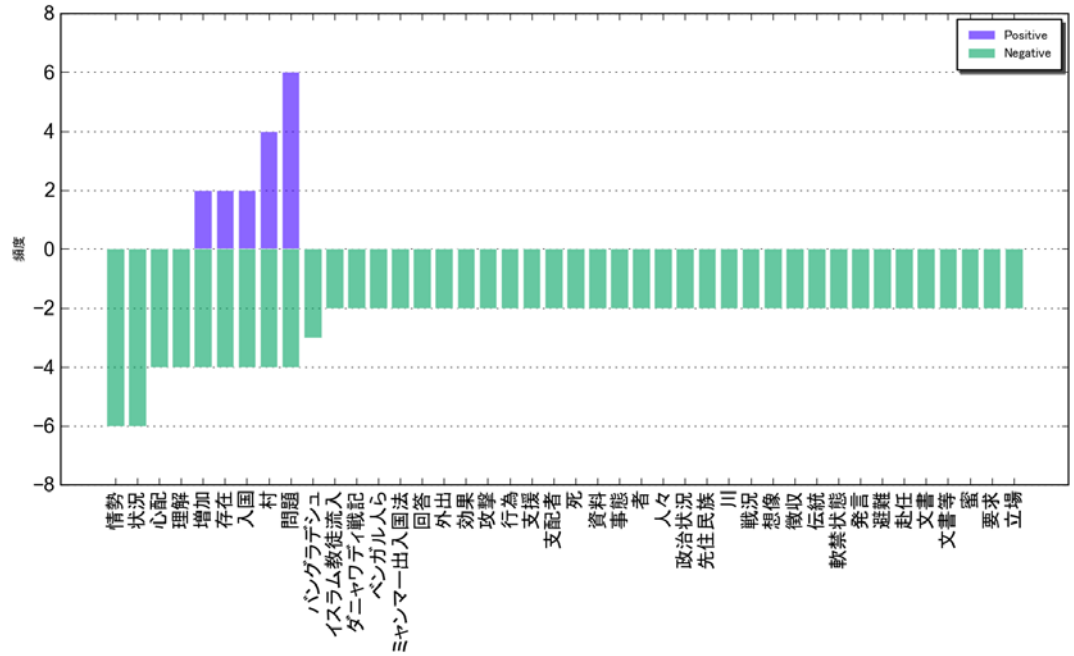

図 10　不評語ランキング　「ミャンマー西門の難題」

不評語ランキングでは、Negative（非好意的）な表現で用いられている単語の回数を負数で表しており、絶対値の大きい順に並んでいる。「Positive(好意的)」は同じ単語が好意的な表現で語られた回数を示している。

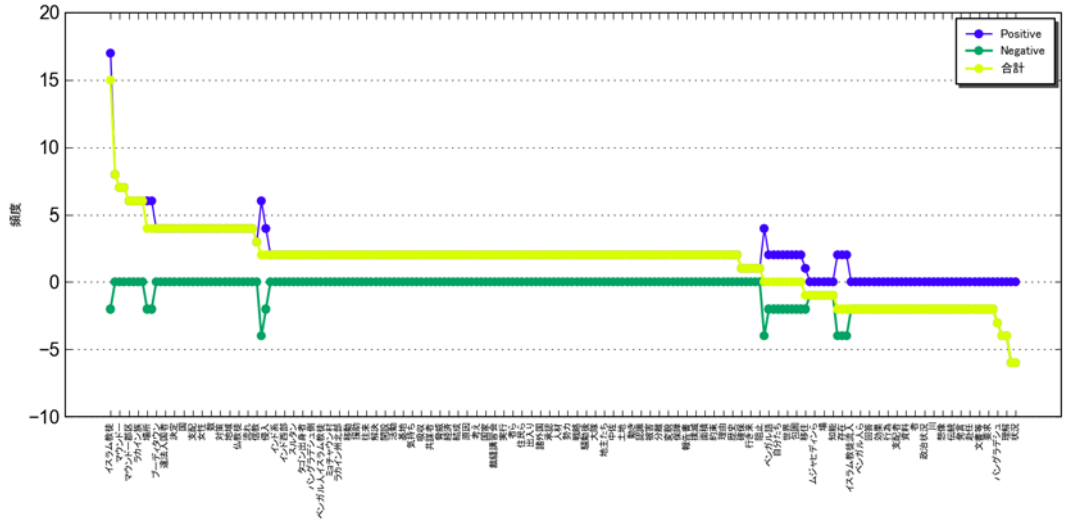

図 11　総合評価　「ミャンマー西門の難題」

総合評価では、評価の対象となった全ての単語が一覧化される。「Positive」「Negative」の示す内容は、「好評語ランキング」「不評語ランキング」と同一だが、「合計」ではこれらの正負を相殺した値を示しており、これの大きい順に並んでいる（Text Mining Studio バージョン 6.0 マニュアル p242)。

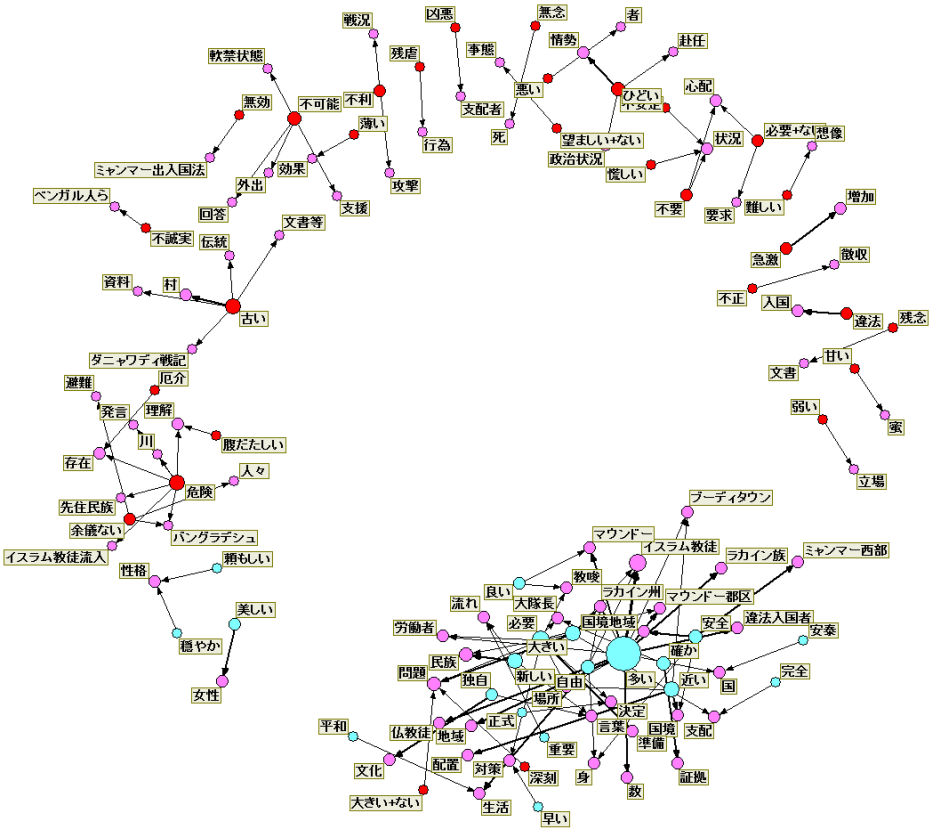

図 12　ネットワーク図　「ミャンマー西門の難題」

ネットワーク図では、Positive、Negative 表現抽出の結果を用いて、どのような言葉がどのような表現を用いて語られているのかを観察することができ、それぞれのノードの色は、

水色（Positive 評価を与える語）

赤（Negative 評価を与える語）

ピンク（評価を受ける語）

を意味している（Text Mining Studio バージョン 6.0 マニュアル p244）。矢印は評価を与える単語から評価を受ける単語の向きで作成され、ノードの大きさは、Positive、Negative 表現中での単語の出現数に応じ、リンクの太さは、それらが結合することば間の表現が出現した回数に対応している。（Text Mining Studio バージョン 6.0 マニュアル p244）。

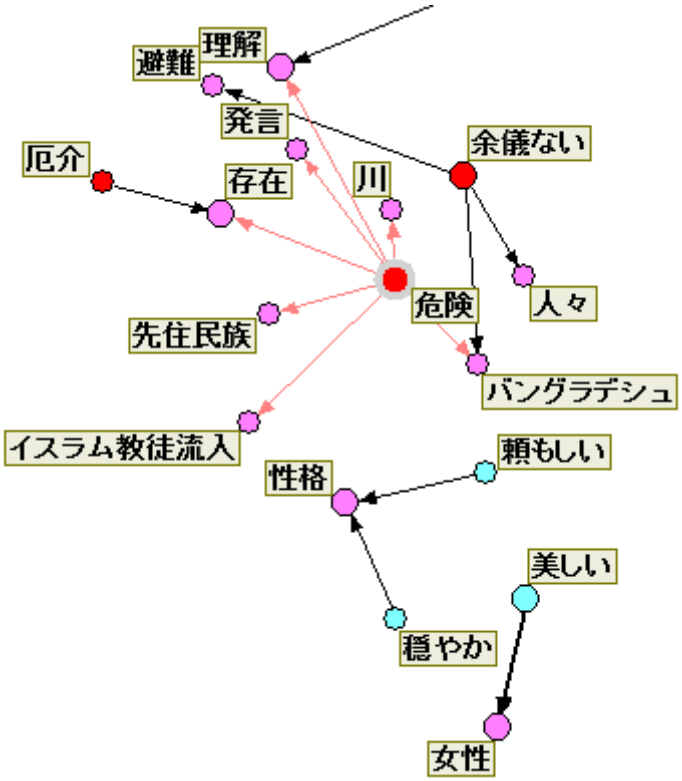

図 13　図 12 のネットワーク図一部抜粋・拡大

例えば図 13 のように「危険」という Negative 評価を与える語が、バングラデシュ・イスラム教徒流入・先住民族・存在・発言・理解・川といった言葉を評価し、「頼もしい」「穏やか」といった Positive 評価を与える語が、性格という言葉を評価しているということを読み取ることができる。

今回の分析では、「イスラム教徒」という言葉が Positive 評価を与える語として現れているが、図 14　原文参照を確認すると分かるように、「イスラム教徒」という言葉が必ずしも好意的な表現で用いられているわけではないということが読み取れる。

原文より一部抜粋

イスラム教徒を中央政府の大臣に任命すること。
イスラム教徒がラカイン州議会議員に就任できるようにすること。
イスラム教徒がミャンマー連邦政府議会議員に就任できるようにすること。
イスラム教徒を政府各局の重役に任命すること。
イスラム教徒にも信教の自由などの人権を与えること。
イスラム教徒のために新たな裁判所を開設し、宗教に関する法律の専門家のもと裁判を受

ける権利を与えること。
イスラム教徒のための小学校から大学までを開設することを許可し、ミャンマー政府が然るべき支援をすること。
イスラム教の文化と尊厳を破壊するような行為はしないこと。

上記のように、その当時イスラム教徒に与えられていなかったことを「できるようにする」・「任命する」・「与える」などといった「今後への要望事項」がプラスの表現として用いられることで、「イスラム教徒」が positive な評価を与える語として分析されてしまっている。したがって評判抽出を行う際には、原文を参照し、本来の意味と分析がかい離されていないかを（人間の脳により）判別することが必要となる。

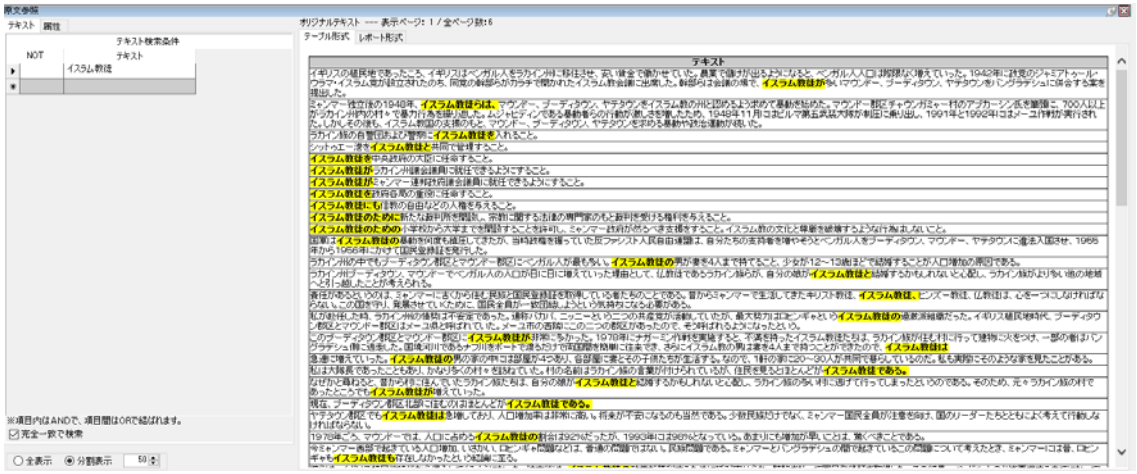

図 14　原文参照

また、図 15 が示すように、「イスラム教徒」と「イスラム教徒ら」や「ラカイン族」と「ラカイン族ら」はそれぞれ異なる単語として分析される。厳密に「イスラム教徒」のみとその他の人々を区別して分析するのか、それとも一つの括りとして分析するのかは、解析者の判断になるため、必要に応じて、Text Mining Studio の辞書機能を使い、類義語として扱うことも可能である。

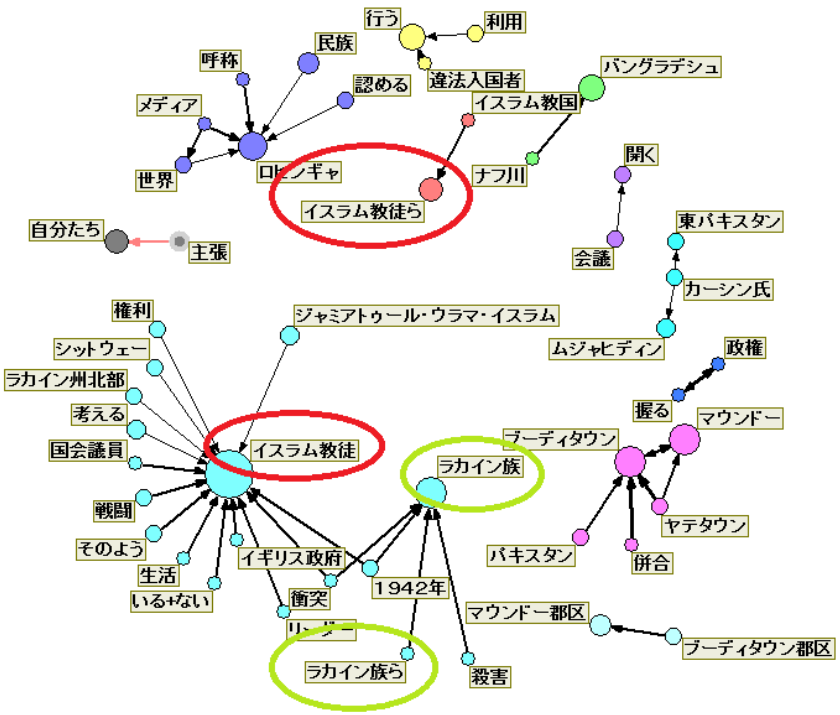

図 15　ネットワーク図　「ミャンマー西門の難題」より

本資料において記述した手法をさらに活用して、グローバルな関係性について詳細に分析することが、グローバル関係学を志向する科研費プロジェクトにおける新領域の取り組み課題である。ミャンマーにおけるいわゆるロヒンギャ問題および少数民族と政府との関係性の分析を端緒に、さらなる研究課題を取り扱うことを予定している。

参考文献

キンニュン著「ミャンマー西門の難題」(未定稿)。
株式会社 NTT データ数理システム (2016) Text Mining Studio バージョン 6.0 マニュアル