意味解析が開く自然語処理の世界

青山学院大学教授　原田実

これまでの自然言語処理は、形態素解析、係り受け解析までが実用域に達している。今回我々が開発した意味解析システムによって、自動要約、質問応答、内容検索、文分類、テキストマイングなどが意味を理解した高度なものに進化した現状を報告する。

１．

従来の自然語処理は、形態素解析・係り受け解析などが中心でした。形態素解析では、文を文節に、さらに文節を形態素（語）にわけ、各語の品詞や読みを辞書から調べ、係り受け解析では、各文節がどの文節に係っているかの依存関係を明らかにします。

例えば、「石油情報センターが 6 日発表したレギュラーガソリンの店頭価格（全国平均、消費税込み）は、前週比 4 円上がり 1 リットル 122 円となった。」を形態素解析ツール juman で解析すると、図１のような結果が得られます。ここで、文が、「石油」や「情報」など辞書にある語に分割され、その読みや品詞などが決定されている。さらに、係り受け解析ツール knp を用いると、図２のような結果が得られる。ここでは、「センター」と「６日」が「発表した」に係り、「発表した」が「レギュラー」に係っていることが分かる。

```
石油 せきゆ 石油 名詞 6 普通名詞 1 * 0 * 0 "代表表記:石油"
情報 じょうほう 情報 名詞 6 普通名詞 1 * 0 * 0 "代表表記:情報"
センター せんたー センター 名詞 6 普通名詞 1 * 0 * 0 "代表表記:センター"
が が が 助詞 9 格助詞 1 * 0 * 0 NIL
6 6 6 未定義語 15 その他 1 * 0 * 0 NIL
日 にち 日 名詞 6 普通名詞 1 * 0 * 0 "漢字読み:音 代表表記:日"
@ 日 ひ 日 名詞 6 時相名詞 10 * 0 * 0 "漢字読み:訓 代表表記:日"
発表 はっぴょう 発表 名詞 6 サ変名詞 2 * 0 * 0 "代表表記:発表"
した した する 動詞 2 * 0 サ変動詞 16 タ形 8 "代表表記:する"
・・・＜以下省略＞・・・
```

図１　例文を形態素解析した結果。１行が１つの形態素に関する情報で、文中での表記、読み、語幹、品詞などが並んでいる。

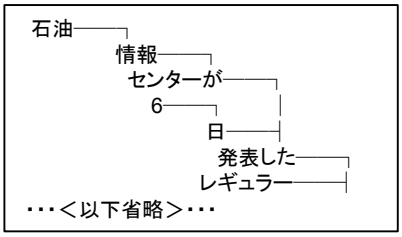

図２　例文を係り受け解析した結果。係り受け関係が線で表示されている。

今回われわれが実現した意味解析システム SAGE は、形態素解析・係り受け解析の後を受け、各語の意味を明らかにし、また係り受け関係にある（文節の代表）語間の役割的関係を明らかにします。SAGE は、語の意味を調べるのに EDR 電子辞書を用いる。EDR 電子辞書は、旧通産省の外郭団体基盤技術研究促進センターとコンピュータ関係の民間企業の共同出資の日本電子化辞書研究所が、１９８６年より９年かけて開発した電子辞書で、単

語辞書では２７万個の語に対して、その意味を１６進数６桁の数として表される概念で与えています。ただし、語の意味は語が使われる状況によって異なるので、図３に示すように語には可能な複数の概念が与えられています。一方、語意を表す概念は４１万個用意され、概念辞書では、概念の間の図４に示すような上位下位関係やその間にどんな役割的関係が存在するかを調べることができます。また、共起辞書では、新聞などに現れた係り受け関係にある２つの語の９０万個の共起事例に対して、図５に示すようにそれぞれの語意と語間の役割的関係（ガやハなどの助詞が表層格と言われるのに対して、深層格と言われる）を与えています。

```
1|0||  発表  ハッピョウ  ハッピョー  JSA  JLN3  JRN4  1faf73  発表する[ハッピョウ・スル]  作品を大ぜいの人に見せたり聞かせたりする
2|0||  発表  ハッピョウ  ハッピョー  JSA  JLN3  JRN4  3cf256  公表する[コウヒョウ・スル]  広く世間の人に知らせる
```

図３　発表するという語には２つに意味（1faf73[作品を大ぜいの人に見せたり聞かせたりする]と 3cf256[広く世間の人に知らせる]）がある

```
[0]:3cf256; {公表する[コウヒョウ・スル]} 広く世間の人に知らせる
  [1]:444578; 不特定の相手に伝える／
    [2]:30f835; 伝える／
      [3]:444c01; 情報の移動を伴う対人行為
        [4]:30f8dd; 対人行為
          [5]:444dd8; 対象行為
            [6]:30f83e; 行為
              [7]:30f7e4; 事象
                [8]:3aa966; 概念
```

図４　概念 3cf256[広く世間の人に知らせる]の上位概念構造を表すリスト

```
4|5||  発表  ハッピョウ  動詞  が  が  センター  センター  名詞
3cf256  公表する[コウヒョウ・スル]  広く世間の人に知らせる
agent  0
3cfc56  本拠地[ホンキョチ]  ある事柄の中心となる場所
006000010768-28-7/<センター>が…(発表)し
```

図５　「センター→（が）→発表」に関する共起事例、この場合、センターは語意 3cfc56[ある事柄の中心となる場所]、発表は語意 3cf2563cf256[広く世間の人に知らせる]、センターの発表に対する役割的関係は agent（行為の主体）であることが分かる。

SAGE による意味解析結果は一般には文節や形態素ごとにそれらの意味や品詞や他の文節との役割的関係（深層格という）などを保持したリストの集合として表現されますが、文節を頂点、文節間の深層格を弧と考えたグラフを意味グラフと読んでいます。先の例文の意味グラフを、ビジュアルツール vivi によって図示すると図６のようになります。このように、語の意味と語間の深層格を決定するには、各語の語意を単純に単語辞書を引くだけでは求められません、先に述べたように語には幾通りもの語意があるので、その中からこの文に最適なものを選ぶ必要があります。各語の語意は、文中の他の語の語意と関連しているので、結局共起辞書などを検索して、与えられた文中での語の使われ方と最も近い使われ方をしている共起事例を基に決定します。SAGE はこれらを高速（１文約１秒）かつ高精度（人手による正解に比べて語意９６％、深層格９３％）に決定します。

Vivitool の概念体系図

2 つの概念を比較できる→語の意味を比較→文の意味的な近さ（文類似度）を計算できる。」

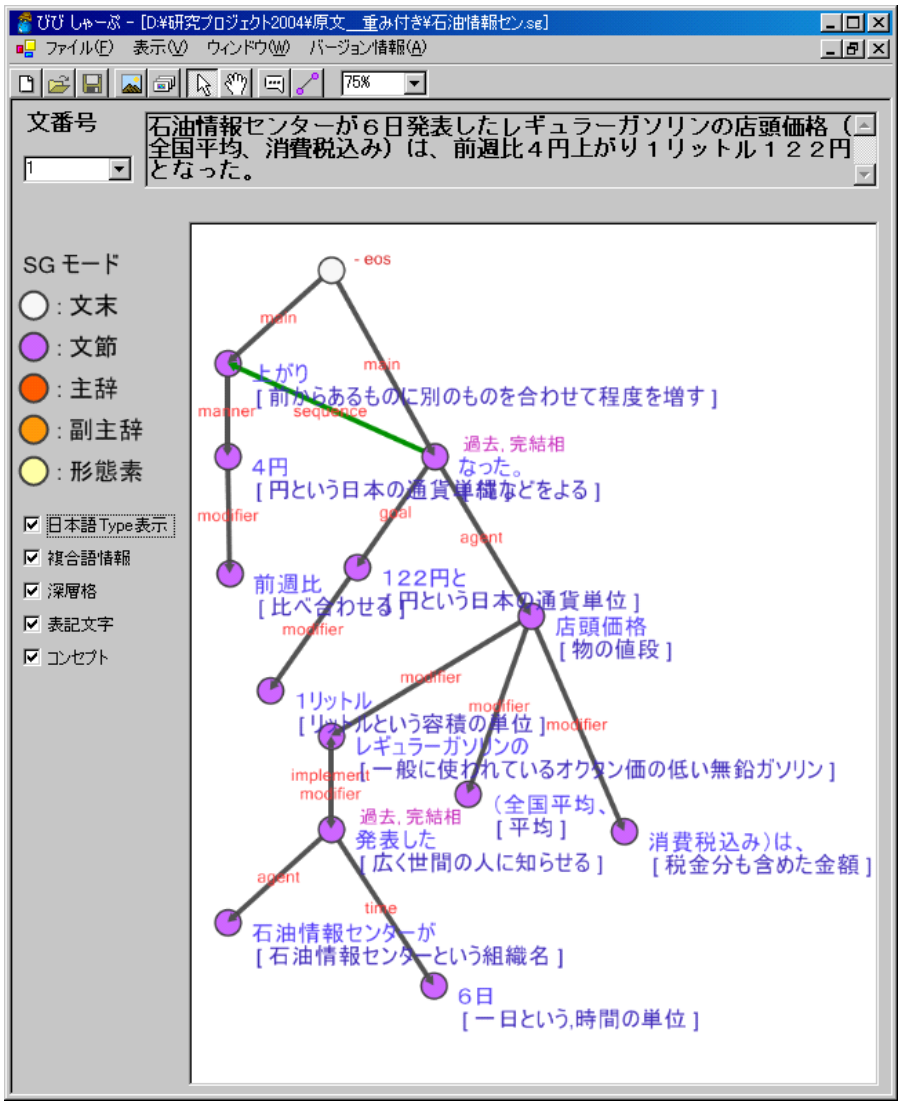

図６　上部窓枠内の文を意味解析した結果の意味グラフを図示した物。例えば、左下で文節「石油情報センター」は文節「発表した」の agent（行為の主体）的役割を演じていることが分かる。また、各文節には、文節内で中心となる形態素（主辞といわれる）の語意の説明が[]内に表記されている。

２．

このように文を構成する語の意味が分かると、例えば企業が持っている膨大な文書データから良質なセールス知識やマーケッティング知識などが抽出できるようになります。これはテキストマイニングといわれる技術で、文書データからの知識の発掘として注目を浴びています。しかし、従来のテキストマイニングでは、文書データに形態素解析と係り受け解析を行い、その結果の単語群をその表記や読みを伴って関係データベースなどに登録し、このデータベースに対して得たい情報をキーワードで指定して、表記の一致を条件にキーワードを含む文を抽出するなどをしていました。しかし、同一の概念を表す語はたくさんあり多様な表現で記述された実際の文書から思い通りの知識を検索することは大変です。これに対し、意味解析を行い、語に語意や係り先の役割を表す深層格を伴ってデータベー

スに登録すると、システムが文書中の各語と指定されたキーワードとの語意ベースの類似度を計算して、一定以上の類似度を持つ語を含む文を一括して抽出するので、キーワードの選択に迷うことなく求めている文を抽出することができます。

３．オブジェクト指向分析の自動化

意味解析技術を使えば、文章表現された内容を別の形式に変換するプログラムを簡単にすることができます。例えば、文章に記されたソフトウエアに対する要求仕様書をもとに、オブジェクト指向分析を行い、図６に示すような設計図を自動的に作成できます。ここでは、プログラム内にオブジェクト指向分析の知識として、「教師や医者や作家や会社員などといった人の役割はクラスとして設計する」などのルールが蓄えられています。このような知識をプログラミングする場合、従来の形態素解析をベースにしたシステムでは、入力した文の各語が人の役割を表すものであるかを検査するために、「教師」や「医者」や「作家」や「会社員」などの人の役割をすべて列挙する必要があり、結果としてルールの数が非常に多くなり、維持管理も難しくなるという問題がありました。一方、意味解析された結果を入力すると各語には語意が付いているので、これらが上位概念として“444df2; 職業, 肩書,役割で限定した人間”を持つかどうかを EDR の概念体系辞書で検査するだけで、その語が人の役割を表わす語かどうかを判断するこができます。言い換えれば、無数のルールが１つのルールと概念体系辞書に置き換わっているのです。これは、自然語文を入力して何らかの知的処理を行うプログラムの簡略化という点で非常に大きな意味を持ちます。

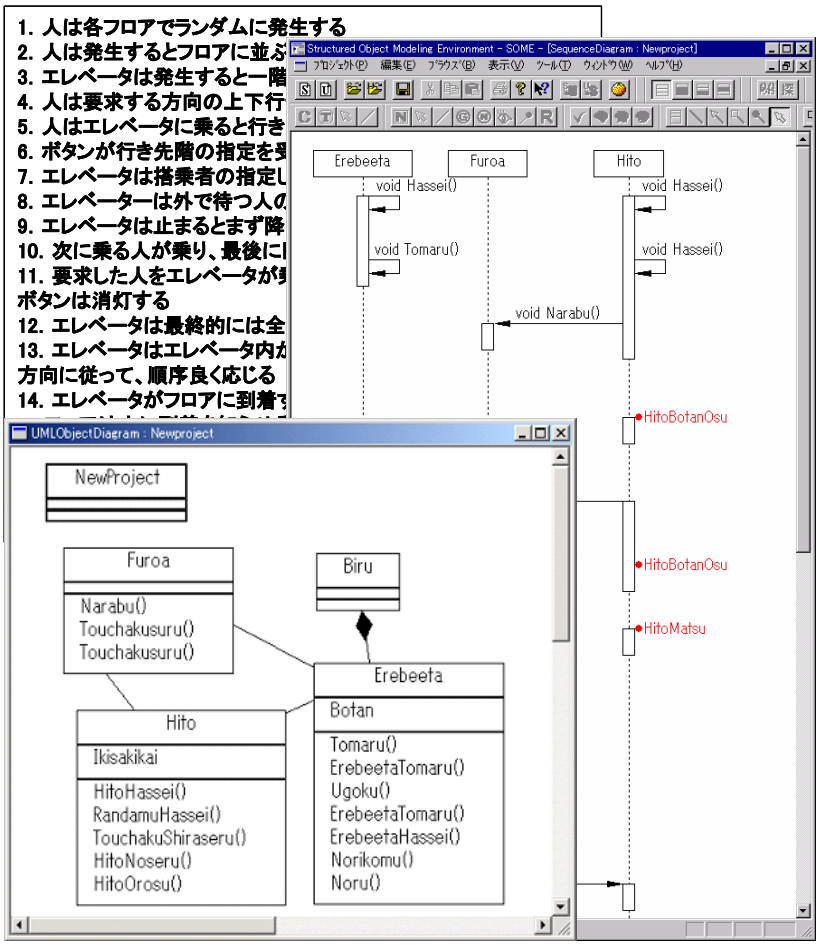

図６　左上の文章を意味解析後、オブジェクト指向分析ツール cameo が内蔵した分析知識に従って、右のシーケンス図や左下のクラス図を生成した。

４．

市販の Word などの文書作成ツールに付随している要約機能は、形態素解析や係り受け解析に基づいた重要文抽出を行う機能であり、文章を構成する文の中から重要な語を含む文を選んで色づけするだけです。文を選定する尺度が甘く取り出された文よりも取り出されなかった文の方が重要であることも多く、また要約文が原文と同じなので短くもならないという欠点があり、実際に利用されることはほとんどありませんでした。我々が開発した自動要約システム ABISYS は、図７に示すように意味解析結果を用いているので良質な要約文を作成できます。

自動要約においては、要約文のタネとなる重要語の選定が成否の鍵を握ります。この際従来は、語の出現頻度を計算して頻出語を重要語としていました。しかし、通常の文章では同じ概念を表すのに同じ語を用いずに別の表記で表現する場合が多いので同一表記の語を数えるだけでは見落としにつながります。また、「紅茶」、「茶」、「ウーロン茶」などが文章に出現すれば、これらをまとめて「飲み物」が話題の中心であることを判断しなければなりません。このような判断を行うためには、同じような概念を表すものは１まとまりとして頻度を計算しなければなりません。このような、語の意味的類似度の計算は、各語に与えられている語意をベースに概念体系辞書を検索することで実現できます。

また要約文の生成においては、重要語を種にできるだけ短い文を生成する必要があります。このためには、文中で各文節が係り先にとってどの程度必要な文節かを判断する必要があります。これは、意味解析によって係り先からみて各係り元に割り当てられた深層格を基に判断することができます。

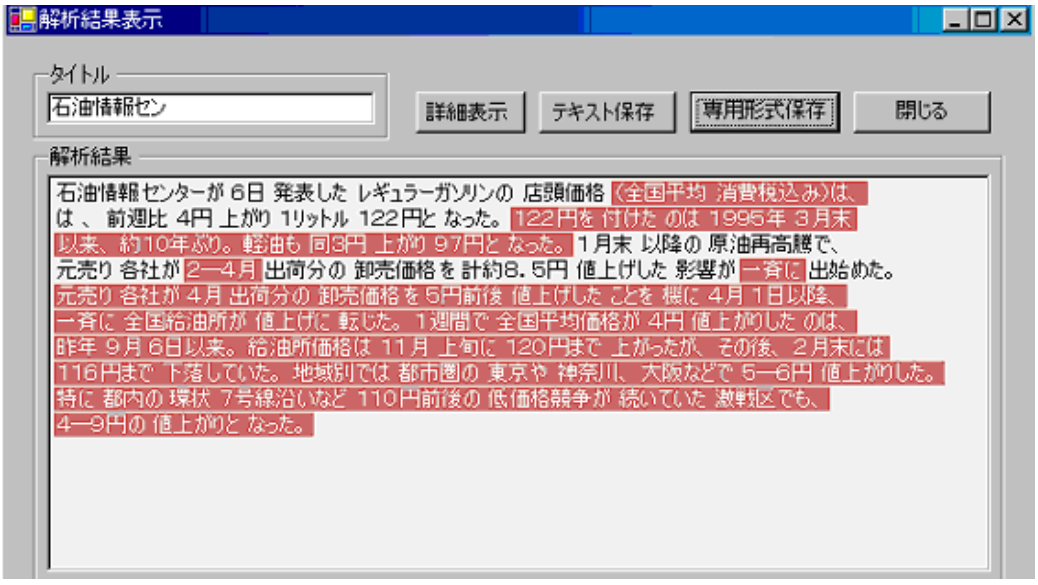

図７　画面内の全文を SAGE が意味解析し、その結果を基に ABISYS が要約として生成した要約文が白地に黒字で記されている。赤地に白字は要約としては不要とされた部分。

５．

文を意味解析すると、語の語意や語間の深層格をもとに、２つの文の類似度を計算することができます。具体的には、文を意味解析した結果の意味グラフ上の共通部分グラフの大きさをベースに類似度を計算します。共通部分グラフとは、２つのグラフ上の対応する頂点同士が類似した語意を持つ語を表し、さらにそれらの間の深層格も類似している頂点と深層格からなるグラフのことで、それが大きいほど２つのグラフはより広範囲に類似していると言えます。類似度を効率的に求めるアルゴリズムを考案し、質問文に対する意味グラフとインターネットから検索された知識文に対する意味グラフとの類似度を求めることによって、最大の類似度を持つ知識文における質問箇所に対応する頂点が表す語を質問に対する回答として生成する図８に示すよう質問応答システム METIS を作成しました。得られた知識文に解があれば 65%の精度で第１位として正解を返すことができます。誤回答率は 7%と非常に低く、システムの回答の信頼性が高いのも特徴です。

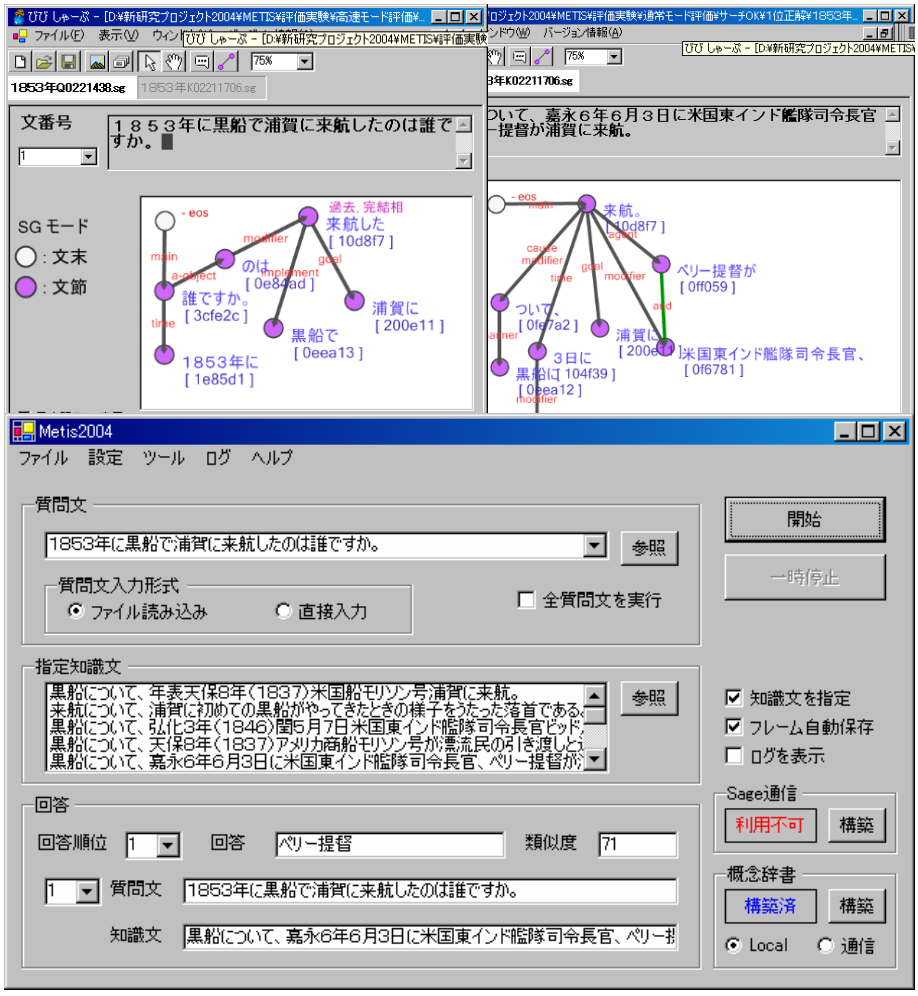

図８

図下の画面上部のように、質問文を入れると、主要なキーワードを取り出し、これを基にインターネットを検索し、その下の窓に表示さえたような複数の知識文を通り出し、質問文とこれらの知識文を図上のように意味解析し、その類似性（来航、浦賀、黒船、goal 格など）から左の質問文の誰に対応する右の知識文のペリー提督を回答として返す。

６．

質問応答と同様にして、文と文との類似度を計算することによって、沢山の文を似ている意味を持つ文同志が集まったクラスターに分けることができます。この技術を応用して、例えば図９に示すようなアンケート分類を行うシステム AQUA を製作しました。同じ技術を使えば、ある事件にもっとも類似した状況で起きた過去の事件に関する裁判所の判例を容易に検索することもできます。このように意味に基づく文の類似性の判定は、大量に存在する文章から最も望ましい文章を探し出す有力な手段で、今後広範囲に応用されることが期待されます。

文番号 3： 国連決議があっても憲法にてらせば、日本は協力できない。
文番号 8： 憲法上、軍事行動に対する協力はできないと思います。
文番号 10： 武力攻撃に協力すべきではない。
文番号 11： 憲法で武力による国際紛争解決を否定している限り、武力による解決には、協力すべきでない。
文番号 27： 極力攻撃は回避すべきだし、協力すべきでない。

文番号 22： 国連中心主義を掲げるわが国としては、応分の協力をすべき。
文番号 25： 国連の加盟国としての責務として、憲法で認められた範囲内で応分の協力は行うべき。
文番号 26： 協力をせざるをえないと思います。
文番号 31： 攻撃を承認する国連決議があれば国連決議に基づき国連に協力をすべき。
文番号 35： 仮に新たな決議がされても、支持はしても、資金、自衛隊派遣ではなく、復興に協力すべき。

文番号 24： 協力は、日本国憲法の枠内の平和的なものに限定される。
文番号 30： 攻撃容認決議がある場合でも、武力行使には関与せず人道的協力、復興支援に限るべき。
文番号 32： 人道的支援、特に復興支援に積極的に協力することに限定するべき。

文番号 1： アメリカへの協力は、集団的自衛権の行使であり、憲法に反します。
文番号 19： 大量殺りく兵器があるのなら問題であるが、まずは日本国憲法を考えるべきである。

文番号 16： 国連を中心に国際社会との関係を構築していくためには新決議があれば協力もやむをえない。
文番号 28： 国連決議の内容にもよるが、国際社会が一致するなら協力もやむをえない。

文番号 17： 日本は国連主軸外交に徹すべきであり、国連決議を精査し憲法の許容する範囲で行動するべきだ。
文番号 18： 日本国憲法の国際協調主義原則に則って行動するべき。

文番号 6： アメリカのイラク攻撃はあらゆる手立てをとって回避しなければなりません。
文番号 7： 日本がとるべき行動は、アメリカに対し、無法なイラク攻撃を止めるべきであるということを断固として働きかけることです。
文番号 15： イラクの大量破壊兵器の破壊は、査察によって達成されるべき。
文番号 34： できれば攻撃できるための決議を急がず、査察を強化継続すべき。

文番号 2： 平和的解決をめざすべきである。
文番号 13： 平和的手段で解決をするべき。

文番号 4： 国連による査察の継続強化が唯一の平和的解決の道である。
文番号 14： アメリカ･ブッシュ政権のイラク攻撃姿勢は、石油利権を主目的にした明白な侵略行為である。

文番号 9： 日本は良心的兵役拒否国家であることを、世界に向けて発信すべき時です。
文番号 29： 日本は軍隊も持たさず、一人前の独立国家として意見を述べよ。

文番号 23： 日米関係、北朝鮮問題などを考えた時、国連中心主義を貫くことが、国益上もっとも望ましい現実的判断と考える。
文番号 33： 査察が続いている限りは、新しい決議は必要ない。

その他の別意見
文番号 5： 憲法9条をもつ国として、戦争への協力はきっぱり拒否し、平和的解決に向けた外交努力に徹するべきと考えます。
文番号 12： 武力行使を容認する新たな国連決議に到らない努力こそ必要。
文番号 20： 攻撃をした場合、一般国民が被害を受けることを忘れてはならない。
文番号 21： 承認決議がある場合は、国際社会の総意、意思と理解せざるを得ない。

図９ 国会議員に対してイラク攻撃について実施したアンケート 35 文（インターネットに公表されている）を AQUA が自動分類した結果。「----」線で区切られているのが、似ている文を集めたクラスターの境界。