関連語抽出法の広範な適用可能性の検証

山本英子 [1, 2]　　井佐原均 [2, 1]

[1] 神戸大学　工学研究科　　[2] 独立行政法人　情報通信研究機構

1. はじめに

我々は、単語の意味は、それが使われる文脈によって規定されると考え、各単語の意味をその単語と共起する語を要素とするベクトルで表し、ベクトル間の包含関係を数値化することにより、コーパスから関連性のある単語セットを抽出する手法をこれまでに開発し、またその有効性を実証してきた[8, 9]。関連語抽出法はこれまでにも多く提案されているが、それぞれの目的に沿った関連語セットを抽出するために開発されているものが多い。

ある語に対して、なんらかの関係を持つ語を関連語というが、それらの語彙間の関係はさまざまである。言語処理技術の応用を念頭に置いたとき、注目したい、獲得したい関連語は、上位語・下位語や、同義語・反義語といった、個々の語彙を意味的に捉えるために役立ちうる関連語や、因果関係や含意関係、連想関係などといった、文または文書における展開の推移を推測するために役立ちうる関連語である。また、翻訳や検索といった利用目的の下で、表記の揺れの照合を目的とする場合、一般的な辞書のように詳細な語義の記述は必要ない場合もあるが、関連語は精度向上に役立つであろう。

語彙間の関係で分類された関連語は、有用な言語資料であるが、実世界でのタスクに適用可能であるかを検証されることは少ない。これはユーザビリティの評価が難しいためである。本稿では、関連語抽出結果の例を示し、考察するとともに、その関連語を応用できうる分野について議論する。

2. 関連語抽出法

ここで用いた関連語抽出法は単語の出現パターンを表すベクトル間の包含度を計り、それを関連度とする。包含度の測定には補完類似度を使う(詳細は文献[8]を参照)。出現パターンの包含度が高い場合、意味的にも上下関係にある関連語が対となると予測される。コーパス中の文を構文解析し、次の 3 種類の言語データを作成し、単語の出現パターンを得る。

- 一文中の名詞と名詞の共起関係に基づく NN データ
- 名詞と動詞の係り受け関係に基づく NV データ
  - 主語、目的語、間接目的語ごとに Ga データ、Wo データ、Ni データを作成。
- 主語と直接目的語の共起関係に基づく SO データ

これらは共起語となる名詞を制限する条件の厳しさが異なる。NN データは共起語が名詞であるという制限しかないが、NV データは係り受け関係を持つ動詞に限定している。さらに、SO データでは、主語となる名詞の共起語を直接目的語に限定しているため、もっとも自由度の少ない言語データとなっている。このように言語データの制限の厳しさを変えることで、さまざまな関係を持つ関連語セットを抽出できる。

この手法の特徴は、データ量が少なくてもそのデータに依存した関連語セットを抽出できることである。また、構文解析で得られる表面的な構文関係を利用しているので、構文解析器がある言語であれば、その言語の文書にも適用可能である。

3. 関連語抽出の事例と考察

コーパスから関連語を自動的に取り出すということは、そのコーパス(の分野)における語彙の関係を取り出すということであるため、得られる結果は、対象とするコーパスに依存する。これまでの言語処理研究の多くは比較的扱いやすい新聞記事などの一般的な文書を対象とし、既存のシソーラスや辞書を再現、もしくは補足・改善することを目標とした。しかし、実世界のニーズは、ある事柄(分野)に特化したデータ、つまり偏ったデータに現れる特有の情報を得て、活用することである。近年、医療やバイオ、マネージメントの分野など、蓄積された文書に関して言語処理の研究がされる傾向にある。たとえば、医学分野において、症状から病名を推測するための知識収集や Web 上の医療に関する情報獲得支援といった需要に応えるために、医学分野の様々な言語資源を用いて、医学用語間における関係抽出や用語辞書の構築などを行う研究が進められている。また、航空分野においても、事故やトラブルの防止を目指して収集されたパイロットや整備士による事例レポートなどを分析する研究が進められている[3, 4, 6]。

ここでは、事例として、我々の提案する関連語抽出法を用いて、医学分野と航空分野のデータについて関連語抽出を行った結果を示し、考察する。

3.1. 医学データ

医学分野のコーパスからの関連語抽出に関しては、文献[8]で報告し、その結果である関連語セットを分類的関係にあるものと主題的(少なくとも非分類的)

表 1: 医学データから得られた関連語セットの例

NN データから得た関連語セット
卵巣 - 脾臓 - 触診
データ - 原因 - うつ病 - 減少 - 血小板数 - 骨髄検査
NV データから得た関連語セット
アイスクリーム - チョコレート - ワイン
出血 - 発熱 - 血尿 - 意識障害 - めまい - 高血圧
SO データから得た関連語セット
潜伏期間 - 赤血球 - 肝細胞
雪 - 学校 - ガス

表 2: 関連語セットの分類結果の例

用語が 1 つのカテゴリに分布する関連語セットの例
手 - 口 - 耳 - 指
貧血 - 嘔吐 - 腰痛 - 尺骨神経麻痺 - 脳内出血 - 閉塞性黄疸
用語が複数のカテゴリに分布する関連語セットの例
卵巣 - 脾臓 - 触診
出血 - 発熱 - 血尿 - 意識障害 - めまい - 高血圧

関係にあるものとに分け、それぞれの検索キーワード群としての特徴を文献[9]で分析した。ある大学の医学部のホームページから得られた関連語セットの例を表 1 に示す。

得られた関連語セットには、様々な関係が含まれる。これを医学分野のシソーラスである MeSH シソーラスのカテゴリ分類と比較する(表 2)。用語が単一のカテゴリに含まれるセットは分類的関係を持つ関連語セットであり、複数のカテゴリに分布するセットは主題的関係などで結ばれたものと考えられる。

たとえば、「卵巣 - 脾臓 - 触診」を用いて、Web 検索すると「卵巣や脾臓の病気は触診によって診断される」という情報が得られ、この単語セットは医学的知識に対応していると見ることができる。同様に、「データ - 原因 - うつ病 - 減少 - 血小板数 - 骨髄検査」は「骨髄の疾患はうつ病や血小板の減少を引き起こす原因となるので、骨髄検査は必要である。」を表す医学的知識と見ることができる。MeSH シソーラスにおいて、この関連語セットを構成する用語は 3 つ以上の異なるカテゴリに分類されており、このような医学的知識は明示的には表現されていない。

3.2. 航空データ

実験コーパスとして用いた文書集合は 1992 年から 2003 年までの航空安全レポートをまとめ、回答が付与されたもの(6,427 件)である。個人情報保護の観点から、事前に名前等の個人情報は削除し、個人を特定できないような処理を行った。このレポートの内容には、出発地と到着地などを示す定型情報と、フリーテキストで記述された表題と本文が含まれている。本実験では、フリーテキストで記述された表題と本文を対象として関連語対を抽出した。

表 3 に Wo データから抽出された関連語対の上位 10 件を、その関係を人手により判定した結果とともに示す。各行の最初の数値は関連度である。取り出された用語対はすべてなんらかの関係を持つ関連語対であるとし、分類的関係にあると判断した用語対は同義・類義・反義・上位・下位などに細分し、分類的関係にないと判断した対は非分類的関係を持つとした。実験結果を見ると、関連度が高い対には同義・類義(略語を含む)関係の対が多く見られた。特に、目的語(Wo データ)による結果は、他の格の結果に比べ、この傾向がはっきり見られた。たとえば、表 3 には現れていないが「GTB (Ground Turn Back)」と「ATB (Air Turn Back)」が関連用語対として抽出されており、これらは何らかのトラブルのための「引き返し」という事象を表すので、類義関係に分類した。また、表 3 の「作業」と「ENG-Start」の関係は、整備作業のなかで、ENG-Start をする場合は「ENG-Start」が「作業」の下位(部分)と解釈できるが、一般的な解釈として、「作業」は整備作業、「ENG-Start」は通常運航でのエンジン始動を表すので、非分類的関係とした。

Wo データから得た包含度の上位 100 件の用語対について、人手により分類した結果を表 4 に示す。この表から、Wo データから抽出した用語対の多くは分類的関係にあることがわかる。同義・類義関係と階層(上位・下位)関係の判別は、文字列の包含関係の利用(「準備」は「出発準備」の部分文字列なので、「準備」が「出発準備」の上位概念)や、別途得られた類義語の利用(「処置」と「作業」が類義であることが分かれば、「整備処置」と「整備作業」が類義だと判断できる)や、Wo データからの結果だけではなく、Ga データや Ni データからの結果も合わせて

表 3: Wo データから得た関連語対の上位 10 件

1.000000	準備	出発準備	下位
0.963333	FLT	飛行	同義
0.942069	作業	整備	類義
0.919837	作業	出発準備	下位
0.877443	整備処置	整備作業	類義
0.876059	着陸	ATB	下位
0.874417	運航	FLT	類義
0.868159	作業	ENG-Start	非分類
0.835848	改善	善処	類義
0.835078	調査	検討	類義

表 4: 上位 100 件の関連語対が持つ関係

同義・類義関係	22
反義関係	2
階層関係	24
その他	35
分類的関係	計 83
非分類的関係	17

表 5: NN データから得た関連語対の上位 10 件

1.000000	感謝	意
0.782266	今回	ケース
0.660146	救急車	手配
0.641839	医師	診察
0.623064	医師	診断
0.619127	今回	事例
0.560951	お客様	ご迷惑
0.533606	原因	究明
0.489483	旅客	疾病
0.485799	発生	急病人

用いることにより、可能になると考えている。また、関連語対には、同義語、類義語、階層関係等の分類的関係だけでなく、非分類的関係を持つものがあり、それらの自動判別も必要である。

表 5 に NN データから抽出された用語対の上位 10 件を示す。NN データについては、主題的関係を見つけ出せそうな用語対が多かった。NN データは共起動詞に着目しての抽出ではなく、文中の名詞の共起に注目しての抽出であり、同一文中によく共起する用語対が抽出されるため、それらの用語対は同義関係や階層関係といった分類的関係ではなく、非分類的な、意味のある関係を持つことが多い。これらの用語対は、主に「の」で繋げられる関係にあり、ある場面を思い起こさせる、もしくは概念同士を結合させる主題的関係を持ちうる。たとえば、「救急車」と「手配」の対は、「救急車を手配した＝機内で急病人が発生した」といった場面を思い起こさせる。これらの対は「旅客」と「迷惑行為」の対よりも上位に位置した。これは、機長が報告すべきと考える事象の順位の表れかもしれない。また、「Delay」と「40 分」の対から Delay が 40 分であることが多いと推測される。一方でこれは、データ量の少なさによる影響の可能性もある。実験で「子供」と「7 歳」の関係が抽出され、データを確認したところ、子供の年齢を明記したレポートは非常に少なく、「子供といえば 7 歳である」といった推論に意味はない。

また、NN データから得られた用語対を用いて用語を連結して、関連語セットを構築した。NN データについては、主題的関係を持つ用語対が多く得られているため、用語セットも全体として主題的関係のある用語からなるセットが見られる。たとえば、「時間 - Delay - 40 分 - GND-Facility」といった関連語セットが得られ、主に「40 分の Delay は地上整備にかかる時間」であるとわかる。

4. 関連語の応用

関連語は言語資料として、さまざまな分野に応用できる。ここでは、我々の得た主題的、少なくとも非分類的関係を持つ関連語の適用可能性を検討する。

4.1. 創造的情報検索とその支援

関連語は言語理解や言語生成といった言語処理だけに有効なのではなく、情報検索においても有用なものであることはよく知られている。情報検索において、以前は、ユーザが目標とする、知りたい情報に導くために、ユーザが入力した検索キーワードに対して、それらを言い換える類義語や同義語などを提示・追加する検索支援が主流であった。一方、現在は、MSN Live Search や Google's keyword tool に見られる、他のユーザの履歴や Web データを活用し、ユーザの情報ニーズをより特定することに役立つ関連語を提示する検索支援もある[2]。前者は Query Expansion、後者は Query Suggestion による検索支援と呼ばれる。こうした移行の背景には、インターネットが普及し、知りたい情報を得るために検索するのではなく、Web 上にある情報を楽しむニーズが増えたことも影響しているのだろう。この視点において、我々は、情報検索結果から、情報量の多い、意図していなかった情報や知識を得る方向へ導く検索支援もユーザのニーズに応えることになると考える。本稿では、このニーズに応える情報検索を創造的情報検索と呼ぶ。創造的情報検索を支援する技術はユーザの現在のニーズを特定するものではなく、関連語を提示し、それを用いた検索によって創造的情報にユーザを導く、新しい Query Suggestion と捉えることができる。

創造的情報検索は、図書館や本屋に通ったり、店を覗いたりなど、我々が日常生活で意識することなく外的刺激から知識を得ている行為を、インターネットを通じて行うことに相当するだろう。それは、自分が興味を持った情報から関連する情報を得るために行われるネットサーフィンにも通じる。

このようなニーズに応える創造的検索支援を目的とし、我々はこれまでにユーザが入力したキーワードと追加キーワードとして提示する関連語との間にはどういった関係があると有効なのかを実際に検索実験を通して調査し、分類的関係より主題的関係、少なくとも非分類的関係が、目標とする情報を見つける情報検索ではなく、創造的情報検索の支援に役立ちうることを示した[9]。

4.2. 創造設計と創造的思考プロセス

ものづくりの分野では、多くの知識を持つユーザの関心を引く、付加価値の高い製品の開発が必要とされている。そういった製品の開発には、高い創造力が必要である。こういった背景から、この分野では、「いかに発想力・創造力を広げられるか」が重要な問題の一つとされる。創造力は設計者の持つ知識と思考に依存する。しかし、新しい製品を継続的に創り出すためには、知識や思考を工学として体系化することが必要である。昨今、設計者の創造的思考プ

ロセスのメカニズムを分析・解明する研究が多くの視点・観点から行われており[1]、そして、創造設計を支援する技術が待ち望まれている。

これまでに、創造的思考プロセスのメカニズムに関する研究において、物を理解するためには、それを表す語彙とその関連語との関係を認識し、理解することが重要であると報告されている[7]。また、二つの既存物(概念)から創造し、一つの人工物を設計するタスクにおいて、種となる既存物間の関係について、主題的関係にある概念を用いることにより、分類的関係にある概念を用いるよりも、創造力豊かな物が設計された、つまり、設計者の創造力を広げたことが報告されている[5]。

創造力に影響を与える因子の一つが語彙間の主題的関係であるならば、語彙間の分類的関係がシソーラスとして体系化されているように、主題的関係をも体系化することで、人間の創造的思考を模倣でき、さらには人間の創造的思考プロセスを支援することが可能になると考える。具体的には、語彙間の関係を一般的知識、専門的な知識、雑学などさまざまな知識を含むコーパスから推定し、語彙間の関連度を得る。この関連度と、被験者実験によって得られる係数によって、語彙間の方向性のある連想度を求める。この連想度の値がついた体系(図 1)を用いることにより、計算機は人間の思考プロセスを模倣でき、人間の創造的思考プロセスの拡大を支援できる。

人間がある概念から別の概念を連想するとき、分類的関係で概念を体系化したシソーラスに沿って連想していくとは限らない。たとえば、「ペンギン」から「鯨」を連想する場合に、人間は「ペンギン→鳥類→動物→哺乳類→鯨」という分類的関係を辿るのではなく、「ペンギン→泳ぐ→鯨」という主題的関係を辿ることもある。人間は、ある程度の選択の自由を持ちつつ、連想度の高いパスを辿って連想を進めると考えられる。見方を変えれば、連想度の高いパスによって結ばれた語彙は思いつきやすく、実際の創造的思考プロセスにおいても想起されやすいと考えられる。創造設計を支援するシステムの構築を目指し、このような人間の創造的思考プロセスを計算機に模倣させることが直近の目標となる。

5. おわりに

検索や設計に限らず、広くユーザインターフェースの分野でも、関連語についての言語資料が必要であると議論されているが、実際に、言語資料を活用し、ユーザビリティを満たすインターフェースを開発するのではなく、ユーザビリティが満たされる事例の解釈方法について多く議論されている。

しかし、関連語を提示することによって、ヒューマンコンピュータインタラクション(HCI)を活発にさせ、その過程で、人は知識を得ていくことができる。その点においてもコーパスからの関連語セットの自動獲得技術の確立は重要な意味を持つと考える。

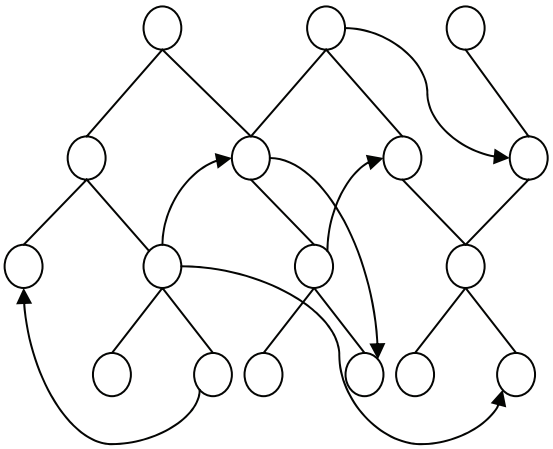

図 1: 語彙間の分類的／主題的関係の体系化

謝辞

本研究を行うにあたって、航空関連文書のテキストを提供していただきました㈱日本航空インターナショナルの寺田昭様と阿部泰典様、医学分野のコーパスのテキストを提供していただきました沖電気工業㈱の池野篤司様に感謝いたします。

参考文献

[1] Dong, A.: The latent semantic approach to studying design team communication. Design Studies, 26(5): pp. 445-461, 2005.

[2] Gao, W., Niu, C., Nie, J.Y., Zhou, M., Hu, J., Wong, K.F., Hon, H.W.: Cross-Lingual Query Suggestion Using Query Logs of Different Languages, SIGIR, pp. 463-470, 2007

[3] 清田, 中川: 航空安全情報分析ツール—human factor に着目したレポート分析手法の提案—. 第 44 回飛行機シンポジウム, 2D9, 2006.

[4] 齊藤, 薬師寺, 渡部, 松井, 佐々木, 寺田, 齋藤: 航空安全情報分析ツール—因果関係に着目したレポート分析手法の提案—. 第 44 回飛行機シンポジウム, 2D8, 2006.

[5] Taura, T. and Nagai, Y.: Primitives and principles of synthetic process for creative design, Computational and Cognitive Models of Creative Design VI, Gero, S. J., and Maher, M. L. (Eds.), pp.177-194, 2005.

[6] 寺田, 吉田, 中川: 文脈情報による同義語辞書作成支援ツール, 情報処理学会研究報告, NL-176, 2006.

[7] Wisniewski, E. J. and Bassok, M.: What makes a man similar to a tie? Cognitive Psychology, 39, pp.208-238, 1999.

[8] 山本, 井佐原: 共起語の包含関係を用いた分野固有知識の獲得, 言語処理学会第 12 回年次大会, C3-1, 2006.

[9] 山本, 井佐原: 文書集合から得た関連語集合のキーワード群としての特徴分析, 言語処理学会第 13 回年次大会, PC3-4, 2007.