文書集合から得た関連語集合の検索キーワード群としての特徴分析

山本英子　井佐原均

独立行政法人　情報通信研究機構

{eiko, isahara}@nict.go.jp

1. はじめに

なんらかの関係を持つ語をコーパスから集めた関連語集合は言語理解や言語生成、情報検索などに有効であると期待される。昨今、コーパスから語彙間のさまざまな関係を獲得するために、多くの手法が開発されるとともに[1, 2, 3, 4, 7]、関係を抽出するためのパターンを学習する手法も提案されている[5, 6]。

関連語集合は情報検索において、有益な情報にユーザを導くための手がかりとなる。Google の検索支援機能のようにユーザが入力したキーワードに関連する語を提示することが考えられるが、入力された語とどのような関係でつながる語を提示すれば、ユーザが適切な情報に到達することを支援できるだろうか。

本研究では、検索に用いるキーワードとしての有効性の観点から、文書集合から抽出した関連語集合の特徴を分析する。まず、主語と述語、目的語と述語などの係り受け関係を用いて文書集合から関連語集合を抽出し、得られた関連語を、既存のシソーラスと一致する(分類的関係にある)ものと、一致しない(非分類的関係にある)ものとに分類する。分類的関係にあるキーワード群を用いて検索した結果を元に、さらに分類的関係にある語を追加した場合と、分類的関係にない語を追加した場合との影響を、得られたページ数やページの内容を分析することによって調べ、検索支援に用いるために関連語が持つべき関係を求める。

2. 語彙間にある関係

語彙の間の関係には、少なくとも「分類的関係(taxonomical relation)」と「主題的関係 (thematic relation)」の 2 つがある。これらの関係は語彙間の関係を認識し、理解するために重要であると報告されている[9]。また、デザインの分野では、主題的関係はより創造を広げると報告されている[8]。

ここで「分類的関係」とは、概念の持つ属性の類似性を表す関係のことで、たとえば、「馬」、「牛」、「動物」といった単語の間にある関係である。同義関係、反義関係、階層関係などの意味的関係はこの分類的関係に含まれる。一方、「主題的関係」とは、主題的な場面を通して概念を結びつける関係のことで、たとえば、「牛 (cow)」と「ミルク (milk)」は「牛の乳を搾る(milking a cow)」、「赤ん坊 (baby)」と「ミルク (milk)」は「赤ん坊にミルクをあげる (giving baby milk)」といった場面を思い出させる、あるいはそのような場面で概念同士を結合する関係である。連想関係、因果関係、含意関係などはこの主題的関係に含まれる。

「分類的関係」は既存の辞書やシソーラスにも直接記述されており、これらから比較的容易に獲得し、利用できる。しかしながら、「主題的関係」を蓄積した言語資源は稀であり、既存のものから得ることは困難である。

このような背景から、本研究では、主題的関係に焦点を当て、主題的関係を持つと思われる関連語集合を抽出し、その関連語集合を構成する用語の検索支援における有効性を調査した。

3. 抽出手法

主題的関係を持つ関連語集合を抽出することを目的として、1)文書集合から係り受け関係を収集し、実験データを作成、2)自動階層構築方法[10]を用いて関連語集合を抽出、3)シソーラスを用いて非分類的関係を持つ関連語集合を選別する。

3.1. 共起関係の収集

まず、KNP によって文書集合を構文解析し、各文から「A<の>B」、「P<を>V」、「Q<が>V」、「R<に>V」、「S<は>V」のパターンにあてはまる係り受け関係を収集する. ここで、<X>は格助詞、A, B, P, Q, R, S は名詞、V は動詞を表す. 収集した関係集合から次の 3 種類のデータを作成した。

- **NN データ:** 各文について、共起する名詞を集めたデータである。ただし、対象となる名詞は上記の A, B, P, Q, R, S である。データ数は文書集合にある文の数に相当する。
- **NV データ**: 関係集合にある各動詞 V について、係り受け関係にある名詞 P、Q、R、S をそれに続く格助詞ごとに集めたデータである。それぞれのデータ数は格助詞ごとの係り受け関係に現れる動詞の種類数に相当する。
- **SO データ:** 関係集合にある目的語 P について、P と共起し、同じ動詞 V に係る主語 Q を集めた

データである。データ数は同じ動詞に係る主語と共起する目的語の種類数に相当する。

3.2. 関連語集合の抽出

次に、これまでに提案した自動階層構築方法[10]を拡張し、関連語集合の抽出を行う。この方法は、与えられた二語について、それぞれの共起語との出現パターンの包含関係から語彙間の関係を推定する。このとき、文献[10]では階層構造の抽出を目的としているため、用いる共起語をそれぞれの語の下位語に限定しているが、本研究では、広く 3.1 節に示す関係で得られる共起語を用いる。これによって、階層構造だけではなく、他の関係を持つ関連語集合も得られる。

3.3. 主題的関係を持つ関連語集合の選別

最後に、抽出された関連語集合から、分類的関係を持つ関連語集合をシソーラスを使って取り除き、主題的関係を持つ関連語集合を得る。

一般にシソーラスに含まれる語彙は分類的関係を表現するように配置されているので、分類的関係を持つ関連語集合は、シソーラス中で同じカテゴリに分類される。つまり、関連語集合がシソーラスに一致するなら、その関連語集合を構成する語彙は分類的に関連していると解釈できる。この考えに沿って、まず関連語集合を構成する語彙がシソーラスにおいて、どのようにカテゴリ分布しているかを調べた。次にシソーラスに一致する関連語集合を取り除くことによって得た非分類的関係を持つ関連語集合を、主題的関係を持つ関連語集合として抽出する。

4. 実験

実験では、医学部ドメインに限定して収集した文書集合(10,144 ページ、225,402 文)を使った。実験には医学用語辞書や専門用語辞書などは用いなかった。

この文書集合から収集した関係集合から作成されたデータの数は、NN データが 225,402、ヲ格データが 20,234、ガ格データが 15,924、ニ格データが 14,215、未格データが 15,896、SO データが 4,437 であった。

関連語集合を構成する語彙として、2005 Medical Subject Headings (MeSH®) シソーラスの見出し語とそれらのクロスリファレンスとして付随している類似語を和訳した医学用語を対象とした。実験データに現れた語はそのうち 2,557 個で、これらに関して、各データから関連語集合を抽出した。

主題的関係を持つ関連語集合の選別には、この MeSH シソーラスを使った。これは 15 個のカテゴリに見出し語を分類したものである。複数のカテゴリに分布される見出し語もあるが、本実験では、たとえば、「木―森―オラウータン」という関連語集合があり、「木」が二つのカテゴリに分布し、一方が「森」と同じカテゴリ、他方が「オラウータン」と同じカテゴリである場合、この関連語集合は「木」を介して、同じカテゴリに属し、したがって分類的関係にあるとして扱った。

5. 実験結果

表 1 に文書集合から抽出された関連語集合の数と、そのうちシソーラスに一致した関連語集合の数、一致しない関連語集合の数を示す。この表より、ヲ格データとガ格データが他のデータよりシソーラスに一致する関連語集合の割合が比較的高いことがわかる。これは、目的語や主語は共起する動詞によって、他のものよりも強く制約されるためと見られる。

また、NN データと NV データについて比較すると、NV データから得た関連語集合のほうが NN データからの関連語集合よりもシソーラスに一致する割合が高いことがわかる。すなわち、NV データのほうが NN データより分類的関係を持つ関連語集合を多く得られることを示している。また、SO データから得た関連語集合 37 個についても調査したが、シソーラスと一致する関連語集合はなかった。これは、通常主語と目的語の関係を抽出する場合、それらが係る動詞を主語と目的語のどちらかで制限するが、作成した SO データではこの制限を与えなかったため、NV データの特徴が現れなかったと見られる。

表 1. シソーラスと一致する／一致しない関連語集合の数と割合

データの種類	NN	NV			
		ヲ	ガ	ニ	未
関連語集合の数	594	199	62	37	85
一致数 (%)	45 (7.5)	58 (29.1)	14 (22.6)	6 (16.2)	7 (8.2)
不一致数	549	141	48	31	78

6. 分析

抽出した関連語集合が検索に有効であること、言い換えると、有益な Web ページに検索結果を限定できることを、Google を用いた検索によって、調査した。調査の対象は、構成する用語が二つのカテゴリに分布し、そのうちの一つの用語だけが残りの用語と異なるカテゴリに分布する関連語集合とした。そのような関連語集合は、表 1 におけるシソーラスと一致しない―非分類的関係を持つ―関連語集合の総和 847 個のうち、294 個あった。

これらの関連語集合を構成する用語をキーワード群として検索エンジンに入力し、Web 検索を行った。本研究では、各関連語集合から三種類の検索キーワードを作成した。ここで、調査に用いる関連語集合を

{X_1, X_2, ……, X_n, Y}と表すとする。X_iは同じカテゴリに分類される用語、Y は X_i と異なるカテゴリ分類される用語である。一つ目の検索キーワードは異なるカテゴリに分類される Y を除いた{X_1, X_2, ……, X_n}，二つ目は同じカテゴリに分類される用語のうち一つの用語 X_k と Y を除いた{X_1, X_2, …, X_{k-1}, X_{k+1}, …, X_n}である．実験では、X_i の中で最も高いまたは低い頻度を持つ用語 X_k を取り除いた。三つ目は同じカテゴリに分類される用語のうち一つの用語 X_k を除いた{X_1, X_2, …, X_{k-1}, X_{k+1}, …, X_n, Y}である。それぞれの検索キーワードを Type 1、Type 2、Type 3 と呼ぶ。

- Type 1: {X_1, X_2, ……, X_n}
- Type 2: {X_1, X_2, …, X_{k-1}, X_{k+1}, …, X_n}
- Type 3: {X_1, X_2, …, X_{k-1}, X_{k+1}, …, X_n, Y}

この三種類は、Type 2 を元となるキーワード、つまり初めに入力されたキーワードとしたとき、Type 1 は Type 2 に同じカテゴリに分類される高いまたは低い頻度を持つ用語を追加したキーワード群である。追加用語は本研究で使った文書集合において頻度に関する特徴を持ち、Type 2 にある用語と分類的に関連する用語である。一方、Type 3 は Type 2 に異なるカテゴリに分類される用語を追加したキーワード群で、この追加用語は Type 2 にある用語と主題的に関連すると思われる、非分類的に関連する用語である。

まず、Google の検索エンジンが推定し、提示するヒットページ数を使って、量的に検索結果を比較する。具体的には、Type 2 を用いて得たヒットページ数を基準に、Type 2 に一用語追加した Type 1 と Type 3 をそれぞれ用いた場合のヒットページ数を比較し、どちらのほうが検索において量的に有効かを考察する。

図 1 と 2 にそれぞれ高頻度と低頻度に関するヒットページ数による比較結果を示す。これらの図において、横軸は元となるキーワード（Type 2）を用いた場合のヒットページ数、縦軸は元となるキーワードに用語を一つ追加した場合(Type1 または Type 3)のヒットページ数である。図中の「○」は同じカテゴリにある用語を追加した場合(Type 1)のヒットページ数、「×」は異なるカテゴリにある用語を追加した場合(Type 3)のヒットページ数を表す。対角線は Type 2 に用語を一つ追加してもヒットページ数に影響がない場合を示す。

図 1 において、多くの「×」が対角線のかなり下にあることがわかる。これは、異なるカテゴリにある非分類的に関連する用語を追加するほうが、同じカテゴリにある分類的に関連する、高頻度の用語を追加するよりもヒットページ数を減少させる傾向にあることを示している。このことから、非分類的に関連する用語を追加することは有益なページを検索するために量的に有効であり、その用語は有益な用語であると考察できる。

表 2 は異なるカテゴリにある用語を追加した場合(Type 3)のヒットページ数が同じカテゴリにある高頻度の用語を追加した場合(Type 1)より減少する関連語集合の数を示す。この表から、非分類的に関連する追加用語のほうが高頻度の追加用語より、ヒットページ数の減少に貢献する場合が多いことがわかる。これは、高頻度の用語が検索に関してあまり有益な用語ではない場合が多いことを示している。

図 2 において、図 1 とは対称的に、多くの「○」が対角線のかなり下にあることがわかる。これは、同じカテゴリにある分類的に関連する、低頻度の用語を追加したほうが高頻度の用語を追加するよりもヒットページ数を減少させる傾向にあることを示している。実際に、低頻度の用語はインターネット上でも稀な用語である場合があり、それを含む Web ページ自体が少ないと予測できる。したがって、低頻度の用語は分類的に関連する用語であっても、有益な用語であると推測できる。

表 3 は異なるカテゴリにある用語を追加した場合(Type 3)のヒットページ数が同じカテゴリにある低頻度の用語を追加した場合(Type 1)より減少する関連語集合の数を示す。この結果から、低頻度の追加用語は関係の種類にかかわらず、ヒットページ数を減少させることに役立つことがわかる。このように、低頻度の用語を追加することは検索結果に対して量的に有効である。しかし、Type 1 を用いた結果と Type 3 を用いた結果との内容を考察すると、そこには大きな違いがある。

たとえば、SO データから得た関連語集合「潜伏期間－赤血球－肝細胞」について考察する。これは、「潜伏期間」が MeSH シソーラスにおいて他の用語と異なるカテゴリに分類される用語で、「肝細胞」が残りの「赤血球」と同じカテゴリに分類される低頻度の用語である。この関連語集合を構成する用語すべてをキーワードとして用いると、検索結果の一位に「マラリアとは?」というタイトルの日本語ページが位置する。「潜伏期間」と「赤血球」を用いた場合（Type 3）も同じページが一位に位置する結果を得る。しかし、「赤血球」と「肝細胞」を用いた場合(Type 1)は、このページは上位 10 ページ以内には入っていたが、一位ではなかった。

他の例として、NN データから得た関連語集合「卵巣－脾臓－触診」について考察する。これは、「触診」が MeSH シソーラスにおいて他の用語と異なるカテゴリに分類される用語である。この関連語集合を構成する用語すべてをキーワードとして用いると、「卵巣と脾臓の疾患は触診で診断できる。」という情報を

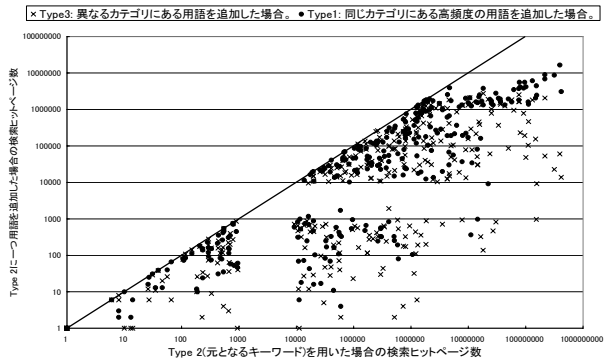

図 1. 高頻度の用語と異なるカテゴリにある用語をそれぞれ追加した場合のヒットページ数の変動

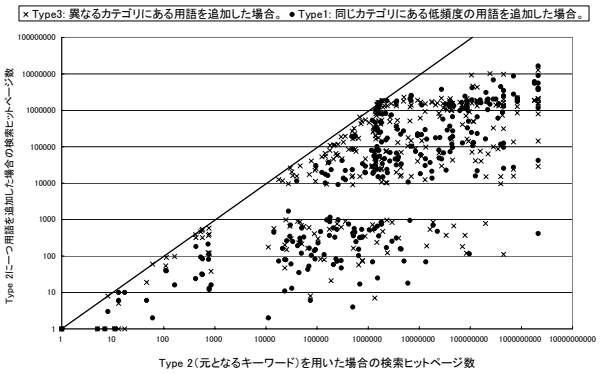

図 2. 低頻度の用語と異なるカテゴリにある用語をそれぞれ追加した場合のヒットページ数の変動

表 2. 異なるカテゴリにある追加用語が高頻度の追加用語よりヒットページ数を減少させる関連語集合の数

データの種類	NN	NV			
		ヲ	ガ	ニ	未
調査した関連語集合	175	43	23	13	26
ヒットページ数が Type 3 < Type 1 である関連語集合	108	37	15	12	18

表 3. 異なるカテゴリにある追加用語が低頻度の追加用語よりヒットページ数を減少させる関連語集合の数

データの種類	NN	NV			
		ヲ	ガ	ニ	未
調査した関連語集合	175	43	23	13	26
ヒットページ数が Type 3 < Type 1 である関連語集合	61	18	7	6	13

含むページが検索される。この結果から、この関連語集合は因果関係を持つと解釈できる。したがって、この関連語集合がユーザの意図を正確に定義し、関連のある Web ページを検索できることを示唆している。

実験において、他の用語と非分類的関係を持つ用語は有益なページに検索結果を限定することに有効であった。これに対して、検索支援に用いた場合、高頻度の用語は量的に有効ではなく、低頻度の用語は非分類的関係を持つ用語と比べ、質的に有意な傾向が見られなかった。今回は最初の試みとして、一つのドメインに限って実験を行い、考察したが、より正確に主題的関係を持つ関連語集合を抽出するために研究を発展させ、より量的かつ質的にその関連語集合の有用性を検証することが今後の課題である。

7. まとめ

本研究では、検索に用いるキーワードとしての有効性の観点から、文書集合から抽出した関連語集合の特徴を分析した。まず、文書集合から構文解析に基づき格助詞を利用し、関連語集合を抽出し、そのうち非分類的関係を持つ関連語集合を主題的関係を持つ関連語集合として得た。そして、このように抽出した関連語集合が検索に有効であること、言い換えると、有益な Web ページに検索結果を限定できることを、Google を用いた検索によって、調査した。その結果、非分類的関係を持つ用語は検索結果を有益なページに限定することに役立つと考察した。より正確に主題的関係にある関連語集合を抽出するために研究を発展させることが今後の課題である。

参考文献

[1] M. Geffet and I. Dagan. The distribution inclusion hypotheses and lexical entailment. In *Proceedings of ACL 2005*, 107–114, 2005.

[2] R. Girju. Automatic detection of causal relations for question answering. In *Proceedings of ACL Workshop on Multilingual summarization and question answering*, 76–114, 2003.

[3] R. Girju, A. Badulescu, and D. Moldovan. Automatic discovery of part-whole relations. *Computational Linguistics*, 32(1): 83–135, 2006.

[4] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora, In *Proceedings of Coling 92*, 539–545, 1992.

[5] P. Pantel and M. Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations In *Proceedings of ACL 2006*, 113–1200, 2006.

[6] D. Ravichanfran and E. H. Hovy. Learning surface text patterns for a question answering system. In *Proceedings of ACL 2002*, 41–47, 2002.

[7] I. Szpektor, H. Tanev, I. Dagan, and B. Coppola. Scaling web-based acquisition of entailment relations. In *Proceedings of EMNLP 2004*, 2004.

[8] T. Taura and Y. Nagai. Primitives and principles of synthetic process for creative design — Taxonomical relation and thematic relation. *Computational and Cognitive Models of Creative Design VI*, Gero, S. J., and Maher, M. L. (Eds.), 177–194, 2005.

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

[10] E. Yamamoto, K. Kanzaki, and H. Isahara. Extraction of hierarchies based on inclusion of co-occurring words with frequency information. In *IJCAI 2005*, 1166–1172, 2005.