

特定用語に関する Web 上の関連文書群からの用語説明情報の抽出

松本 航[†] 椎野 努[†]

愛知工業大学大学院工学研究科電気電子工学専攻[†]

1. はじめに

我々が用語の意味を知ろうとするときや用語の説明を求めるとき、Web 検索をしばしば用いる。しかし、Web 検索では関連文書が膨大な数になり、そこから目的の情報を求めるのは少なからず労力を要する。本稿では、特定用語を Web 検索したときの文書群から用語を説明する情報を抽出する方法を提案する。本稿で扱う特定用語を説明する情報とは特定用語そのものの内容を説明している情報とし、時間情報や金額情報などは含まないものとする。

2. 関連文書群

説明文を作成したい単語を特定用語と呼ぶことにする。Google を用いて特定用語を Web 検索し、検索結果上位 20 文書のコピーを保存して、これを関連文書群とした。図 2-1 に示す 5 つの単語について関連文書群を収集し、実験を試みた。形態素解析器は ChaSen を使用した。

特定用語	総文数	特定用語含有文の数
光電子増倍管	950	134
ポリモーフィズム	2401	141
クイックソート	1557	123
グライダー	1094	250
正規分布	1163	164

図 2-1 調査した特定用語

3. 説明的言い回し部分の抽出

特定用語の説明文を作成する目的で関連文書群を読むとき、「～とは」や「・・を～と言う」など、説明的な言い回し部分が含まれる文に注目して読む事が多い。そこで、関連文書集合から説明的な言い回し部分を抽出した。

説明したい単語を A と置いたとき、以下の(1)～(4)が含まれる文を抽出した。

- (1) A は～
- (2) A とは～
- (3) を A という | を A という
- (4) を A と言います | を A といいます

上の条件で抽出した結果、抽出文は金額情報や場所情報など不要情報まで含まれてしまっていた。

対策としてこの抽出文のうち不要と考えられる文を次の条件で除去した。

＜不要情報除去条件＞

- 1) 時間・地域・組織・人名・金額・章番号などの不要情報が含まれる場合
- 2) 特定用語に修飾句が付いている場合・特定用語を限定している場合
- 3) 数詞、記号、未知語だけで構成される場合
- 4) 疑問文

抽出文数を図 3-1 に示す。特定用語がポリモーフィズム、クイックソートの場合は文頭が「～とは」となる文がほとんどを占めており、同じような表現ばかりの文章になってしまった。また意味的に同じ文が多いため、抽出される情報量は少ない。そこで情報量を増やすために 4 章、5 章で述べる手法を試みた。

特定用語	単純な抽出	不要情報除去後
光電子増倍管	17	7
ポリモーフィズム	20	18
クイックソート	21	19
グライダー	67	43
正規分布	14	10

図 3-1 説明的言い回し部分の抽出文数(単位：文)

4. 特定用語が含まれる文中の必要関連用語

関連文書群において特定用語が含まれる文に存在する特定用語以外の名詞は特定用語と関連性が高い単語なのではないかと考え、関連文書群から次の条件で単語を抽出した。

＜抽出条件＞

- ・特定用語が含まれる文中の特定用語以外の複合名詞、単名詞、未知語で、かつ複数の文書に存在する単語
- ・3 章で述べた不要情報除去条件(1)～(3)に該当する単語は除く

ただし、単名詞同士あるいは未知語同士、単名詞と未知語が連続した単語を 1 単語として扱い、複合名詞とした。これらの単語を必要関連用語と呼ぶことにする。

抽出された単語の例を図 4-1 に示す。必要関連用語の品詞ごとの数は図 4-2 のようになった。

品詞	単語
複合名詞	光電面, ニュートリノ, 宇宙線, 光センサー, 電気信号, 検出器, 荷電粒子, 世界最大
未知語	スーパーカミオカンデ, カミオカンデ, PMT
単名詞	光, 電子, 使用, 開発, 観測, 事故, 当時, 電極, 発生, 水中, 装置, 検出, 信号, 水, 破損, 実験, 再開, タンク, 微弱

図 4-1 抽出された単語（光電子増倍管）

品詞	平均	最大	最小
複合名詞	5.8	8	3
未知語	3.0	6	0
単名詞	26.4	35	19

図 4-2 抽出された単語数（単位：個）

5. 必要関連用語をもとにした文抽出

必要関連用語は特定用語の説明に必要な性の高い単語と考え、これをもとにした文抽出を考える。

我々が特定用語の説明文を作成するとき、用語が含まれる文だけでなくその周辺まで参照することが多い。どの程度の領域を参照すれば十分な情報を含めることができるかを調査するため、特定用語が含まれる文から必要関連用語を含む文までの距離を

$$P_{Diff(i,j)} = P_{M(i)} - P_{X(j)}$$

(P_M : 特定用語が含まれる文の位置、

P_X : 必要関連用語を含む文の位置)

と定義し、

$$|P_{Diff(i,j)}| \leq \text{距離条件}$$

となる距離条件を満たす文を抽出する。

距離条件を設定するために特定用語を含む文の間隔を調査した。特定用語が含まれる文から直後の特定用語が含まれる文までの文の数を距離 D とする。特定用語が含まれる文の頻度を D ごとにとり、関連文書群中の特定用語が含まれる文の数で正規化した。ヒストグラムを図 5-1 に示す。特定用語が含まれる文はその 85%以上が 10 文以内の間隔、71%が 5 文以内の間隔で存在することが分かった。よって、特定用語が含まれている文の周辺 10 文以内を取れば十分な情報を得られると言える。

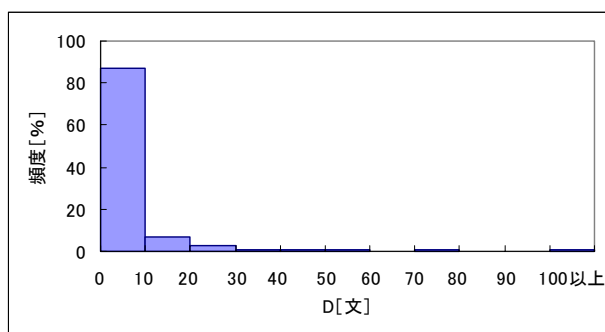


図 5-1 特定用語が含まれる文の間隔のヒストグラム

上記により、特定用語が含まれる文の近傍のみに存在する必要関連用語を文抽出に使用すること

とし、距離条件は 10 文までの範囲で変化させて以下の 5 つの品詞条件で文抽出を行った。

1. 複合名詞+単名詞+未知語
2. 単名詞+未知語
3. 複合名詞
4. 複合名詞+未知語
5. 単名詞

それぞれの品詞条件における距離条件ごとの平均文数のグラフを図 5-2 に示す。文は重複して抽出していない。

文の抽出に用いる距離条件は 3 以上にしても抽出文数はほとんど変化が無い。よって、特定用語が含まれる文に隣接する 3 文までの範囲に存在する必要関連用語を用いて文抽出を行えば説明情報が充分得られると言える。

また、品詞条件が複合名詞の場合と複合名詞+未知語の場合は抽出文の全文数に占める割合が多くとも 10%未満となっており、単純に特定用語が含まれる文を抽出した場合よりも抽出文数を少なく抑えることができた。

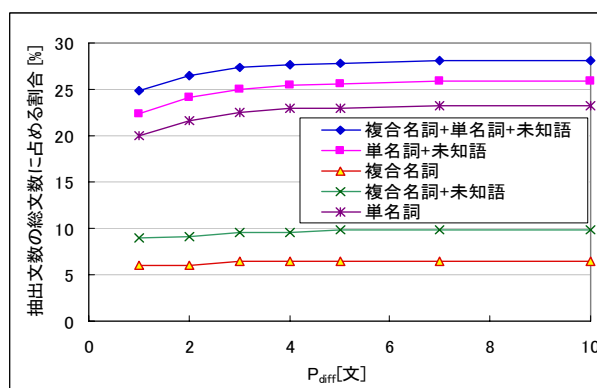


図 5-2 抽出文数の総文数に占める割合

6. 実験結果

必要関連用語の複合名詞および未知語のうち、特定用語が含まれる文に隣接する 3 文内に在る単語を用いて文を抽出したときの結果を図 6-1 に示す。単純に抽出を行った場合は不要情報まで含まれるので、3 章で述べた除去条件に従って不要情報を除去した。あわせて図 6-1 に示す。

これにより抽出した文は特定用語の説明に十分な情報を持つことが確認できた。しかし抽出した文は意味的に重複する文も多いため、今後の課題として、文を抽出した後での意味的に重複している文の統合が挙げられる。

特定用語	抽出文数	不要情報除去後
光電子増倍管	159	108
ポリモーフイズム	164	149
クイックソート	195	155
グライダー	40	34
正規分布	95	82

図 6-1 必要関連用語を用いて抽出した文数