E_032

LSAに基づいた辞書語義文の用語自動生成
Automatic Generation of Dictionary Definition Words Based on Latent Semantic Analysis

福田 ムフタル† 小川 泰弘‡ 外山勝彦‡
Muhtar Fukuda Yasuhiro Ogawa Katsuhiko Toyama

1．はじめに

語彙辞書を作成する場合，その見出し語の語義文に使用される語彙集である語義定義語(Dictionary Definition words)を予め決めておき，全ての語義文には基本的にその語義定義語しか使用しないという原則のもとで辞書開発を行うことは，一般的な言語生活だけでなく，自然言語処理における意味処理，言い換え，キーワードの自動付与，意味に基づいた情報検索，文書の意味抽出や自動要約などの知能情報処理を行う上で重要である．しかし，語義定義語の選定は，人手によるものが一般的であり，機械的に行われているとしても，テキストコーパスや新聞データベースにおける語の出現頻度に基づいて選定する傾向がある．語義定義語を人手によって選定する場合，見出し語の表す概念の上位概念をいくつかの側面から制限したり，概念コード化したりすることによって行う場合が多く，選定基準やルールが決まっているとしても，選定作業に携わる人の専門的背景やポリシーによって，その選定結果が違ってくる．従って，語義定義語の選定をできるだけ主観によらず，自動的に行う手法が望まれる．一方，語義定義語の選定で，その選定基準の他に，その数をいくつにするかも考える必要があり，数多くある上位概念語の中からどれとどれをその制限語数内に入れるべきかが重要かつ難しい問題であり，このような問題にも柔軟に対応できる手法が望ましい．

本研究では，語彙概念を数学的，及び工学的方法で数量化できるＬＳＡ法に基づいて，語義定義語を予め決めずに作成された一般辞(例えば，英英辞書や日本語国語辞書)を利用しての語義定義語の集合の自動生成を考え，それがどこまで可能かについて考察を行う．

ここでは，辞書の語義文や例文とそれに出現する単語から LSA に基づいた意味空間を生成し，その意味空間のベクトルの位置関係や近さに基づいて，語義定義語の生成を行う．まず，語義定義語の一般的な選定基準について述べる．それから LSA に基づいた意味空間の生成について述べる．その次に，語義定義語の自動生成の方法を示し，それに関して簡単な評価実験を行う．

2．語義定義語の選定基準

語義文に現れる語義定義語を明確に意識して開発された辞書にはロングマンの英英辞書[1]があり，その選定基準が定かではないが，語義定義語として 2000 語が列挙されている．一般的には，語義定義語の選定基準として次のいずれかが考えられるが，それぞれに一長一短がある．

(1) 使用頻度の高い語を語義定義語とする．最近，テキストコーパスや電子化された様々なテキストコンテンツが増えてきており，使用頻度の高い語を抽出するのは容易である．しかし，語義定義語は辞書見出し語の持つ各意味の明確な説明に使用できるものでなければならないし，高頻度語がそういった要求を満たすとは限らない．従って，高頻度語の中から，語義定義語になりうるものを選定する必要があるが，言語リソースによって出現頻度に偏りがあり，バランスよく選定を行うのが困難である．

(2) 上位概念に近い語を選定する．上位概念に近い語は語義定義語になりうる可能性が高いし，シソーラスや語彙概念階層体系などからの選定が考えられる．しかし，この場合，どの上位概念階層までですればよいか，どれを制限語数内に入れるべきか，頻度に基づくのか，選定作業チーム内の合意に基づくのか，それとも類義語の多い語を選ぶのかなど，考慮すべき点がいくつかある．

(3) 基本語彙の中から選定する．基本語彙は，日常の社会生活で必要不可欠とされる語彙集であり，基本語彙 2000 語，3000 語，...と言った具合に，重要度に応じてその数がいろいろあるが，ここでも何を基準にし，どれを選ぶのかということがある．

以上述べたように，語義定義語の選定に関して考慮すべき点が多く，いずれも重要であるが，それぞれの臨界点を意識せずにその選定を自動的に行えるのが本研究で示す手法の特長である．

3. LSAに基づいた意味空間の生成

LSA(Latent Semantic Analysis, 隠れ意味解析)[2]は，違った単語，あるいは違った文であっても，意味上の類似度，あるいは関連度を人手による前処理や背景知識なしで測ることのできる手法として提案され，多くの研究分野に応用されている．その基本アイデアは，単語が各行に，文書が各列に対応し，成分値が，その単語のその文書における関連度を表すような行列(それを単語・文書行列と呼ぶ)を作り，その行列に対して特異値分解を行い，分解された行列の行ベクトル，あるいは列ベクトルをその単語や文書を代表する意味ベクトルとし，それらのベクトル間の近さでもって単語や文書間の近さ，類似度，あるいは関連度を測る．例えば，次の d1 から d6 までの文書中の各単語(一部省略)の出現頻度からなる単語・文書行列 A を 図 1 のようにを作り，式(1)のように特異値分解する．

$$A = U \times S \times V^T \quad (1)$$

ここで，U, V は直交行列，S は対角行列，U×S，及び V×S の各行ベクトルが，それぞれに対応する各単語，各文書の意味ベクトルになる．

d1: the reaction occurs in the liquid phase of the system.
d2: the second phase of the system is implemented as an automatic mark-up system.

†名古屋大学大学院国際開発研究科
‡名古屋大学大学院情報科学研究科

d3: the system consists of two servers and a small computer.
d4: the phase of the moon changes over the course of one month in the solar system.
d5: we are in a transitional phase.
d6: the system generates hydrogen peroxide.

	d1	d2	d3	d4	d5	d6
computer	0	0	1	0	0	0
hydrogen	0	0	0	0	0	1
liquid	1	0	0	0	0	0
month	0	0	0	1	0	0
moon	0	0	0	1	0	0
peroxide	0	0	0	0	0	1
phase	1	1	0	1	1	0
reaction	1	0	0	0	0	0
server	0	0	1	0	0	0
solar	0	0	0	1	0	0
system	1	2	1	1	0	1

図 1: 単語・文書行列 A

本研究での意味空間の生成も基本的に従来の LSA 法に基づいているか，次の四つの点で従来と異なる．

（1）意味空間を生成するための単語・文書行列は，一般辞書の語彙見出しの語義文と，それに関する例文だけから作成する．

（2）各単語のそれぞれの語義を別々の単語と見なして単語・文書行列(正確には語義・文書行列)を作成する．

（3）単語・文書行列の成分は，語義とその語義文に対応する成分を 1，その語義と，その語義が現れる例文に関する成分を 1/(文の長さ)とする．

（4）単語・文書行列に関して，特異値分解後の語義の意味ベクトルとその語義の語義文意味ベクトル間の近さ(関連度)が，その語義の意味ベクトルとその語義の例文の意味ベクトル間の近さより大きくなるようにチューニングを行う[1]．

4. 語義定義語の自動生成

語義定義語の自動生成は次の手順で行う．

（1）前節で述べたように，辞書見出し語の語義と語義定義文や語義の例文から単語・文書行列 A を作成する．

（2）LSA 法に基づいて意味空間を生成する．即ち，単語・文書行列 A を式 (1)のような形に分解し，U×S の行ベクトル（語義の意味ベクトル）$T_1,\ldots, T_m$，及び V×S の行ベクトル（語義文や例文の意味ベクトル）$D_1,\ldots, D_n$ を得る．（$D_1,\ldots, D_n$ は「単語・文書行列のチューニング」で利用する．）

（3）必要に応じて単語・文書行列のチューニングを行う．

$$\cos\theta = \frac{T_i \cdot T_j}{|T_i||T_j|} \qquad (2)$$

（4）式(2)の値をベクトル T_i とベクトル T_j に対応する語義間の近さとし，クラスタ数を生成したい語義定義語の数に等しくし，$T_1,\ldots, T_m$ に関してクラスタリンクを行う．

（5）各クラスタの中心ベクトルに最も近い語義ベクトルに対応する語義の見出し語の集合を語義定義語とする．

5. 評価実験

辞書の語義定義語の自動生成に関して前節で述べた手順に従って簡単な評価実験を行った．まず，ロングマンの英英辞書[1]から名詞 100 単語をランダムに選び，Wordnet[3]からその 100 語の語義文及び例文をあわせて 538 文を取り出した．そして，その中に含まれる重複しない単語約 1020 語を語義別に列挙し，語義 1706，語義文・例文 538 から 1706×538 の語義・文書行列を作成した．単語 1020 語の約半分である 492 語が，ロングマンの英英辞書[1]に列挙されている語義定義語 2000 語に含まれたため，今回の実験で語義ベクトルのクラスタ数を同様の 492 に設定して実験を行った．その結果，自動生成された語の約 63%に相当する 308 語がロングマンの語義定義語に含まれた．これは非常によい結果とは言えないが，数万語も数十万語もある見出し語に対してここで示した手法を利用することにより，語義定義語の候補を絞り込むことができることを考えると，本研究で示した手法が有効であるといえる．

今回の実験では，ロングマンの語義定義語に含まれる 492 語の約 37%の 182 語を回収できなかったが，100 語の語義文及び例文に含まれる単語の語義文と例文を追加してから語義・文書行列を作成し，さらにその行列のチューニングを行えば，上記以上のよい結果が得られると考えている．

6. おわりに

本研究では，LSA モデルに基づいた語義定義の自動生成手法を示した．従来の方法とは違って，単語・文書行列をテキストコーパスや新聞データベースではなく，辞書の語義文とその語義の例文から作成した．また，単語ではなく，語義を単語の代わりに使用した．これは，一つの単語は全く違った語義を持っていても一つの意味ベクトルで表現することよりも正確な意味空間を生成できると考えたからである．今回の実験は小規模なものであるが，今後，単語数や語義文・例文をさらに増やして実験を行い，ここで示した手法の有効性を大規模なデータに関しても確認し，実用の語義定義語生成システムの実現を目指す．

参考文献

[1] Longman. (2000). Longman Advanced American Dictionary. Pearson Education Limited, Harlow, England.

[2] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman Indexing by Latent Semantic Analysis. Journal of the American Society of Information Science, 41(6), 391-407, 1990.

[3] http://wordnet.princeton.edu/

[1] 今回は行列のチューニングを行っていない．