2B1-4

自然言語の完備束表現の提案

A Data Structure on Complete Lattice Representation of Natural Language

由井 卓哉[*1]
Yui Takuya

石崎 俊[*2]
Ishizaki Shun

[*1] 慶應義塾大学大学院政策メディア研究科
Graduate School of Media and Governance, Keio University
[*2] 慶應義塾大学環境情報学部
Faculty of Environment and Information Studies, Keio University

In this paper, we propose a data structure to add the bag-of-words to the word ordering information. Our data structure is complete lattice representation, which is derived from mathematical defined document.

1. はじめに

インターネットの普及と発展，さらにモバイル機器の一般化によって，インターネット空間上には嘗て無いほどの情報が蓄積されている．その情報の中でも，伝達に用いられるものとしてテキストがある．また，世界経済のグローバル化と新興国の著しい発展によって，インターネット空間上で求められる様々なテキストの多言語化の要請が増している．しかしながら，莫大なテキストのすべてを人手で処理するのは不可能であり，また，多くの人にとって数多くの言語を習得し，駆使する事は難しい． そのため，あらかじめコンピュータに多くのテキストを読み込ませて置き目的に合ったテキストを取り出す情報検索（information retrieval）や，ある自然言語で書かれたテキストを他の言語へと変換する機械翻訳（machine translation）の重要性は増すばかりである．

英語や日本語などの自然言語を，人間が翻訳や要約，索引付け，訂正を行う過程を模倣して，そのままの形で処理を計算機に行わせる事が難しい．その一因として，人間が持っている常識や感覚，単語や文の意味，そして推測を計算機で再現することの困難さが挙げられる．この様な問題に対して日本語など研究が活発な言語においては，形態素解析（morphological analysis）や構文解析（parsing），意味解析（semantic analysis），談話解析（discourse analysis）などの主目的の処理とは別に高度な前処理と計算機が利用可能な辞書を用いる事で，その精度の向上を図る事が行われている．しかしながら，個別の言語で研究が進んでいないものや，計算機が利用できる形の辞書が整備されていない言語では，上記の様な前処理の利用が厳しく言語障壁を築きやすい．

そのような中で，情報検索技術や機械翻訳技術の中には，自然言語で書かれたテキストの表層だけを用いて，bag-of-words モデルと呼ばれるデータ構造を構築して処理を行うものがある．この手法では，テキストに現れた単語だけを利用するために，高度な前処理が要らず様々な自然言語に応用可能である．

本稿では，この bag-of-words モデルの拡張を数学的に定義した文章（document）から考え，完備束表現(complete lattice representation)の提案を行う．

連絡先：由井卓哉，慶應義塾大学大学院政策メディア研究科，
メール：yui@sfc.wide.ad.jp

2. 関連研究

Bag-of-words モデルは，テキスト分類や情報検索，機械翻訳などに用いられている文章の特徴付け手法である．Bag-of-words の概念は [Zelling, 1954]には見つけることができ，自然言語処理では幅広く使われている．

その基本的なアイディアは，文章中の単語を単語毎に数え挙げたモノを文章の特徴量とするモノで，その特徴量をベクトルとして扱う事で計算機にも扱いやすくなる．

自然言語処理以外の分野でも，bag-of-words model は利用されており，動画のシーン検索などでも使われて他の手法と比べて成功している [Josef Sivic and Andrew Zisserman, 2003]．代表的なモノとして [Csurka, ほか, 2004]が提案した画像の認識手法の bags-of-keypoints モデルがあり，bag-of-features モデルとしても知られている．

この様に，bag-of-words モデルは幅広い対象や分野でも有用なモデルであり，このモデルの発展は大いに価値がある．

3. 文章,bag-of-words の定義

bag-of-words の拡張に際して，bag-of-words の生成元を文章と定義する．文章と従来の bag-of-words と完備束表現の数学的な定義の提案を行う．

数学では，要素をまとめて扱う時に集合という概念を用いる．集合は，その要素の扱い方について異なる2種類のまとまりを定義することができ，本論文でも [新井, 2011]にならって下記の集合を定義し，それらを用いて corpus, bag-of-words, 完備束表現を定義する．

(1) 集合

その元の現れる順番は関係なく，また同じ元は何度現れても一つある事と同じ．本論文では，{}を用いて表す．
a≠b とした場合

$$\{a,b\} = \{b,a\} = \{a,a,b\} \quad ①$$

集合を走る変数 x,y,…,について次が成り立つ：

$$\forall u(u \in x \leftrightarrow u \in y) \rightarrow x = y \quad ②$$

(2) 順序組

<x,y>を順序対とする．
順序対は下記の条件を満たすものとする．

$$\langle x, y\rangle = \langle u, v\rangle \Rightarrow x = u \ \wedge y = v \quad ③$$

三重対(triple)を

$$\langle x,y,z\rangle := \langle x,\langle y,z\rangle\rangle \quad ④$$

と定義，n-重対(n-tuple)も，帰納的に定義される．

全ての n-重対を指して，順序組とする．

a≠b とした場合

<a,a,b>≠<a,b,a,>≠<a,b> ⑤

3.1 集合論的定義

E を集合とし，またそのサイズ|E|は n としたとき，

文章，bag-of-words を以下の集合と定義する．

1. 文章

文章 D は，$s_i \in E$ を満たす有限順序組

D= $<s_1, s_2, \cdots, s_l>$ ⑥

2. Bag-of-words

Bag-of-words B は，s_i∈E と文章 D 中のs_iの重複度m_iからなる順序対$<s_i,m_i>$を元とする集合

B={$<s_1,m_1>$,$<s_2,m_2>$,$\cdots$,$<s_n,m_n>$} ⑦

4. 完備束表現の定義

$s_a, s_b \in E$と，文章 D 中の$s_i, s_j \in S$ (1≦i≦j≦l) が

$s_a = s_i \wedge s_b = s_j$ ⑧

を満たした回数を重複度m_{ab}とした時，

完備束表現 cLR は，順序対$<<s_a,s_b>,m_{ab}>$ を元とする集合

cLR={$<<s_1,s_1>,m_{1\,1}>$, $<<s_1,s_2>,m_{1\,2}>$, $\cdots$,

$<<s_n,s_n>,m_{nn}>$} ⑨

4.1 完備束表現の特徴

同じ文章 D から生成された bag-of-words B と完備束表現 cLR の関係は B におけるs_i∈E の重複度m_iと，cLR における重複度$m_{i\,i}$は，定義より下記の条件を常に満たす．

$m_{i\,i} = m_i!$ ⑩

また、完備束表現の定義より、常に要素間の前後関係が保存される。

このため，完備束表現は bag-of-words の上位互換のデータ構造として定義でき，bag-of-words を文章の各要素の重複度からなるベクトルとして表記する様に，完備束表現は文章の各要素から各要素への重複度からなる正方行列として表記する事ができる．

例えば，E={a,b,c}からなる文章 D=<a,b,a,b,c,a,c>は，bag-of-words では，

B = (3,2,2) ⑪

完備束表現では，

$$\text{cLR} = \begin{pmatrix} 6 & 3 & 5 \\ 3 & 3 & 4 \\ 1 & 0 & 3 \end{pmatrix} \quad ⑫$$

5. おわりに

本論文では，bag-of-words とその生成元の文章の集合論的定義を行い，完備束表現の提案を行った．完備束表現は，bag-of-words の持つ情報を持ち，また要素間の前後関係を保存する上位互換のデータ構造になる事を示した．

今後の展望としては，機械翻訳や自動要約などの処理技術への応用がある．自然言語では，文字や単語，句などと言った小さい単位への分解と解析，またそれらの合成で様々な課題に対処するが，その時に要素間の並び替えが問題になる事が多い．そのような問題に対して，要素間の前後情報を保存できる完備束表現が有効に働く事が予想される．

参考文献

CsurkaGabriella, ほか. 2004. Visual categorization with bags of keypoints. : Proc. of ECCV Workshop on Statical Learning in Computer Vision, 2004. ページ: 1-22.

Josef Sivic and Andrew Zisserman. 2003. Video Google:A Text Retrieval Approach to Object Matching in Videos. : International conference on computer vision, 2003. ページ: 1470-1477.

ZellingHarris. 1954. Distributional Structure. : Word, 1954. ページ: 146-62. 第 10 巻.

新井敏康. 2011. 数学基礎論. : 岩波書店, 2011.