平成 28 年度　学士論文

ハンガリアン法による語の相似関係抽出

Analogous word association and its detection by the Hungarian method.

平田　泰樹

Taiki Hirata

北海道大学 工学部 情報エレクトロニクス学科 コンピュータサイエンスコース 知識ベース研究室

Department of Electronics and Information Engineering.
Hokkaido University

2017 年 2 月 10 日

目次

図目次

表目次

第 1 章

序論

1.1 はじめに

SNS などのサービスが普及し、以前にも増して世界中の人々がインターネット上にテキストデータを投稿するようになった。こうして集められるデータは、その人の趣向、動向、人と為りや、世界の動向を知る手掛かりとなるだろう。しかし、インターネット上にあるテキストデータは膨大なものになっており、ユーザーが本当に欲しい情報は見えにくく、扱いづらいものとなってしまっている。膨大なデータを有効活用していくためには、効率的な分類、加工をしていかなければならないが、データのすべてを人手で解析し、まとめることは非常に困難である。そこで、計算機を利用してこのテキストデータを処理することがさまざまな分野で考えられている。
テキストデータを計算機で処理するにあたって、テキストを何らかの方法で数値化する必要がある。そもそもテキストを構成する文字自体は単なる記号であり、また、計算機が文字データを識別するために、文字一つ一つに事前に割り当てられている数値には、記号識別以上の意味がない。
そこで、テキストを計算機で処理するために、いかにして意味を持つ数値で表現し、どのように利用するか、が課題となっている。

1.2 研究背景

テキストを、意味を持つ数値で表現する手法について、現在まででいくつかの方法が研究されている。中でも 2013 年に発表された word2vec は、ニューラルネットワークによって学習した単語の意味表現ベクトルで、単語の和、差の計算ができるようになり話題となった。

このベクトルの加算減算が可能になったこと、形態素ベクトルの位置関係に相似性が見られることは、主成分分析、t-SNE などで次元を削減し、可視化したデータから見て取れる。

1.3　本研究の目的

本研究では、word2vec により得たベクトルデータを可視化することなしに、語の相似関係を抽出することを目的としている。

1.4　本論文の構成

はじめに本章は序論であり、研究背景に関して述べた。
本稿２章では、本研究で用いる単語の意味表現についての説明を述べた。３章では、本研究で用いるベクトルデータを出力する word2vec について述べた。
４章では、word2vec により得られたデータの解析に用いる、ハンガリアン法について述べた。
５章で、本研究で行った実験の流れについて説明し、結果を検証した。
最後に６章では、本研究における結論と、今後の展望について考慮していることを述べて最後のまとめとした。

第 2 章

意味表現の学習

本研究では、word2vec で学習した分散表現ベクトルから語の相似関係を抽出することを目的としている。本章では、単語の意味表現の学習についていくつかの表現方法と、word2vec で可能な分散表現ベクトルについてを説明する。

2.1 形態素と意味表現

2.1.1 形態素

自然言語とは、人と人とのコミュニケーションに用いられる道具であり、文字、あるいは音の並びによって構成されている。しかし、一般に文字、或いは音そのものは意味を持たないため、自然言語の最小単位として**形態素**を用いることが多い。
形態素とは、文字列が意味、或いは役割を持つ最小のまとまりである。本研究ではこの形態素に着目していく。

2.1.2 形態素解析

多くの自然言語処理に関わるタスクは、形態素を最小単位としているが、日本語などの言語は特に、形態素ごとに区切られた文章ではなく、また、計算機は、形態素のまとまりを認識できない。そこで、自然言語処理をする際に、第一の処理として**形態素解析**を行うのが一般的である。
形態素解析の目的は、入力された文を、形態素毎に区切り、品詞などを決定することである。

本研究では、京都大学情報学研究科と、日本電信電話株式会社コミュニケーション科学基礎研究所が、共同研究ユニットプロジェクトで開発した **MeCab**[*1] という、オープンソースの形態素解析エンジンを実験データの前処理に用いた。

[*1] http://taku910.github.io/mecab/

2.1.3　意味表現

自然言語を計算機で処理するにあたって、計算機が形態素の意味や役割を獲得すれば、文や文章など、より複雑な構造の意味を理解し、それに応じた処理ができるようになることが期待されている。この、計算機で認識するために形態素の意味や役割を何らかの数値で表現したもの、或いは表現することを意味表現という。
では、意味表現をどのように与えるかということが問題になる。

2.2　辞書

我々は、未知語の意味を知ろうとする時に検索エンジンでの意味を調べるなど、何らかの方法で意味が記述されたものを参照する。その参照対象をここでは一括して辞書と呼ぶこととする。辞書を計算機に与えることで、意味を教えることはできるかもしれない。日本語 **WordNet**[*2]は、計算機で使うことを想定して開発された辞書である。

2.2.1　WordNet

日本語版 WordNet は国立研究開発法人情報通信研究機構（NICT）で 2006 年から 2012 年まで開発されていた大規模な日本語意味辞書である。プリンストン大学で開発された Princeton WordNet と、ヨーロッパの EuroWordNet 協会が推進する Global WordNet Grid に着想を得て開発された。
WordNet では、単語の意味を **synset** と呼ばれるグループで表現しており、各 synset は単語の意味に関する記述と、上位関係や下位関係など、他の synset との関係などが保持されている。なお、今回使用しているこの日本語 WordNet の情報規模は以下の通りである。[2]

- synset 数 57,238
- 単語数 93,834
- 語義数：synset と単語のペア 158,058
- 定義文数 135,692
- 例文数 48,276

人手で整備された辞書は単語の意味に関してある程度精密な情報を備えているが、新語や、多くの複合語には対応しきれておらず、たとえば、MeCab を用いて日本語版 Wikipedia 全文から抽出された形態素数は 63〜84 万程度であった。（数字の幅があることに関しては後ほど説明するが、形態素の基本形で数えた場合と、活用された形そのままの場合の 2 パターンで実験をしたからである。）こうした点から、単語・複合語のすべてとすべての意味を収録し、あらゆる言語処理の応用に耐えうる辞書を作成することは、非常に困難であることがわかる。

[*2] http://nlpwww.nict.go.jp/wn-ja/

2.3　統計的手法

辞書などを作成することで、計算機に単語の意味を明示的に与えることには限界があることがわかった。そこで、現在数多くある、人間が生み出したテキストデータから自動的に単語の意味を学習・獲得させることを考える。

2.3.1　分布的意味表現

まず一つに**分布的意味表現**というものがある。
これは、**分布仮説**

> The Distributional Hypothesis is that words that occur in the same contexts tend to have similar meanings(Harris, 1954). (ACL wiki)

”同じ文脈に出現する単語は、似た意味を持つ。”という考え方によって構築されるものである。[1]

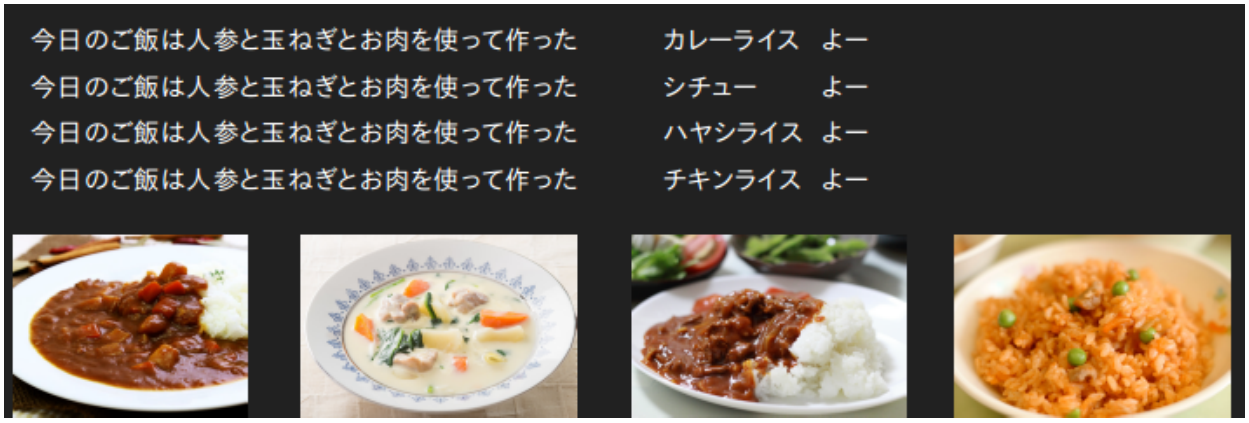

図 2.1: 分布仮設による意味の類推例

上述の例は全く同じ文脈で、似たような料理名が出てくる例文を並べたものである。この時、もしどれかひとつの料理名が全く無知なものであったとしても、それが料理であること、どのようなものであるか、ある程度類推することができるだろう。それは、まさしく同じ文脈に出現する単語であるからに違いない。

この考えに則ってベクトルを作成してみる。例えば、

- 今日の晩御飯は、ジャガイモの入ったカレーライスだ。
- 今日の晩御飯はハヤシライスだ。

という二本の文章を例にして、”カレーライス”と”ハヤシライス”の意味表現を考える。

まず、それぞれの文章を形態素解析すると、

- 今日 \ の \ 晩御飯 \ は \、\ ジャガイモ \ の \ 入っ \ た \ カレーライス \ だ \。
- 今日 \ の \ 晩御飯 \ は \ ハヤシライス \ だ \。

となる（上では \ によって形態素を区切っている）。

これをもとに、”カレーライス”と”ハヤシライス”に関して、同じ文脈に着目した意味表現ベクトルを作成する。

同じ文脈に出現する形態素（これ以降文脈語と呼ぶ）をベクトルの各次元とし、注目する単語と何個の同じ文脈でその次元の文脈語が出現しているかを値としてとると、

- カレーライス $=$ [(今日：1), (の：1), (晩御飯：1), (は：1), (、：1), (ジャガイモ：1), (入っ：1), (た：1), (だ：1), (。：1)]
- ハヤシライス $=$ [(今日：1), (の：1), (晩御飯：1), (は：1), (、：0), (ジャガイモ：0), (入っ：0), (た：0), (だ：1), (。：1)]

と表すことができる。

ここで、ある単語 x_i と別な単語 x_j が何らかの文脈で同時に出現していることを**共起**していると言い、何度共起しているかの回数を**共起頻度**と言う。つまり、上記の例で作成したベクトルは、共起頻度を値として持つ**共起ベクトル**である。[3] この方法でベクトルを作成する場合は、学習に用いられるテキストの集まり**コーパス**すべてから学習するため、一般に高次元で疎なベクトルが生成される。これは、コーパスに含まれる単語数が数万程度であったり、単語数に対し、語が共起することは多くないためである。

こうして作成される意味表現ベクトルを用いて別な自然言語処理のタスクを行う際、元々のベクトルが高密度で疎なものであるために、学習事例を正しく表現できなくなってしまうなどの問題が発生してしまう。

2.3.2　分散的意味表現

分布的意味表現ベクトルが高次元で疎なものであるという点を避けた学習方法として提案されたのが、**分散的意味表現**である。

分散的意味表現の学習手法の基本は以下のようになっている。

1. まず、すべての単語に、任意の固定次元のベクトルを割当て、ランダムな実数値で初期化する。
2. 与えられたベクトルを用いて、何らかの予測タスクを解く。
3. 予測タスクを解いた結果に応じてベクトルの値を更新する。

4. 2 へ戻る。

分散的意味表現を表すベクトルの各次元は、何らかの実数値変数が持つ値と解釈することができる。この実数値変数が持つ特徴はどのような予測タスクを解くか、固定次元数をどのくらいの値にするかなどによって変わると予想されるが、単純な共起頻度が値となるわけではなく、また、他の様々な形態素と次元を共有することから、コーパス中で共起することがない形態素同士の関係についても表現できることが期待される。

次章では、本研究で用いた形態素の分散的意味表現学習ツール word2vec で実装されている 2 種類の予測タスクについて述べる。

第 3 章

word2vec

word2vec[*1] とは、2013 年に Google の Mikolov らが発表した、単語の分散的意味表現学習ツールである。意味表現学習の過程で解くのは、ある文脈内で共起する形態素を予測するタスクである。文脈が与えられ、次に出てくる形態素を予測するモデルを**言語モデル**という。与えられた形態素列がどれほどその言語らしいかを評価するモデルである。

言語モデルでは着目形態素の前方にある形態素のみから次に来る形態素を予測しなければならないが、意味表現獲得を目的としているため、後方の形態素も予測に使える。よって、$x_1,...,x_{i-1}$ と、$x_{i+1},...,x_n$ から x_i を予測する。

この章では word2vec に実装されている 2 つの手法を紹介していく。

3.1 連続単語袋詰モデル

(Continuous Bag-of-Words model, CBoW model) **連続単語袋詰モデル**では、**文脈語** (着目形態素と共起している形態素) から着目形態素の出現確率を予測している。

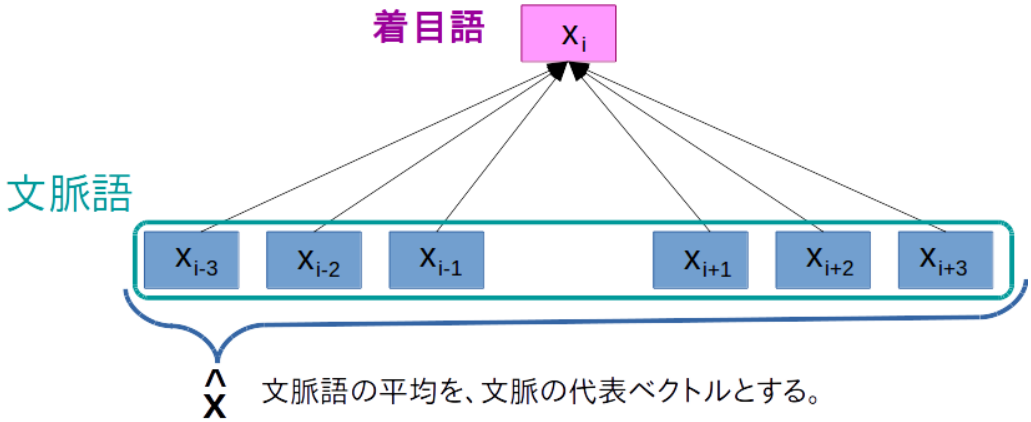

図 3.1: CBoW によるベクトル構築イメージ

[*1] https://github.com/svn2github/word2vec

3.2　連続スキップグラムモデル

(continuous Skip-gram model, Sg model)

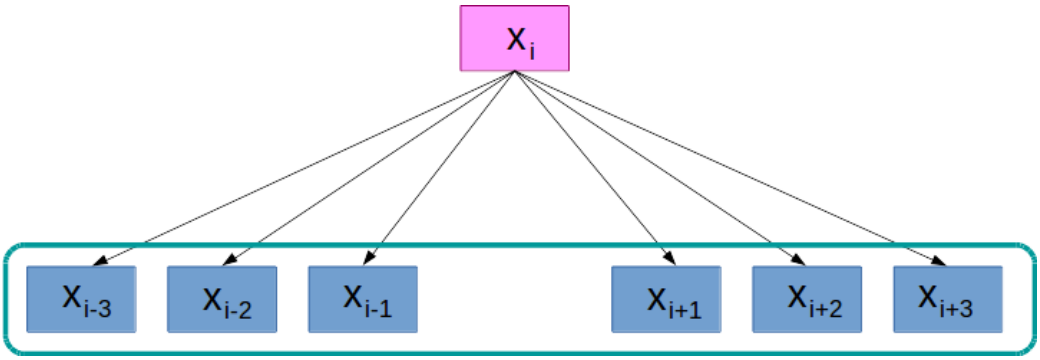

図 3.2: Sg によるベクトル構築イメージ

参考文献

[1] J.R.Firth. A synopsis of linguistic theory 1930-55. Studies in Linguistic Analysis, pp.1-32, 1957.

[2] 日本語 WordNet
http://nlpwww.nict.go.jp/wn-ja/

[3] ダヌシカ ボレガラ, 岡直観, 前原貴憲 機械学習プロフェッショナルシリーズ ウェブデータの機械学習. 株式会社 講談社, 2016.

[4] 黒橋禎夫, 河原大輔. 日本語形態素解析システム『juman』version 5.1 使用説明書. Technical report, 2005.

[5] 四ツ谷雅輝. 共起語を介した文間の相互依存関係に基づく重要文抽出法の提案. Master's thesis, 北海道大学大学院工学研究科, 2003.

[6] 松本祐治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸. 日本語形態素解析システム『茶筅』version 2.2.3 使用説明書. Technical report, 2003.

[7] 石崎俊. 自然言語処理. 照晃堂, 1999.

[8] 大島敦史. 判例文の論理展開と文を特徴付ける語の意味的類似性に基づく法律文要約手法の検討. Master's thesis, 北海道大学大学院工学研究科, 2005.