

トピックモデルから Word2Vec へ

From Topic Model to Word2Vec / 20171215 John Lau

自然言語処理領域のタスク

- 音声認識 / Speech Recognition
- 機械翻訳 / Machine Translation
- 感情分析 / Sentiment Analysis
- 質問応答 / Question Answering
- etc.

自然言語処理領域のタスク

- 音声認識
- 機械翻訳
- 感情分析
- 質問応答



works with the
Google Assistant

タスクを細かく分解すると

- 形態素解析 / Text Segmentation
- 類似度計算 / Semantic Similarity
- トピック判定 / Topic Detection
- etc.

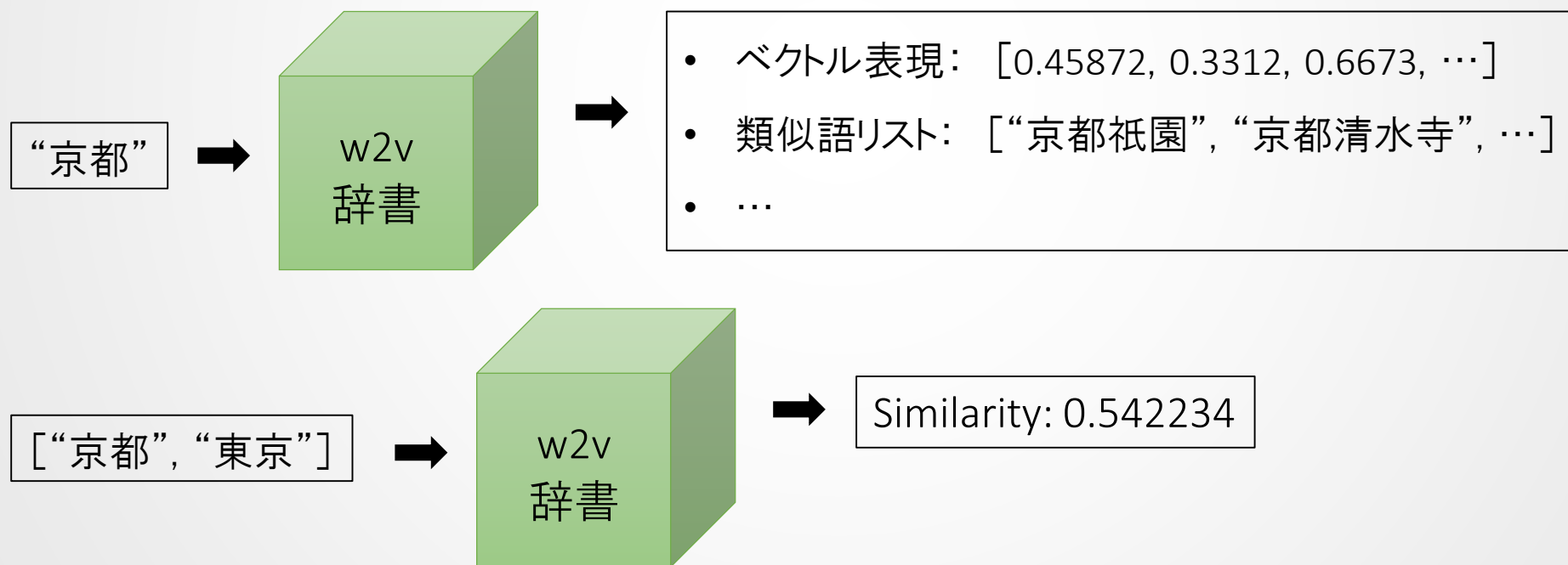
タスクを細かく分解すると

- 類似度計算 / Semantic Similarity



タスクを細かく分解すると

- 類似度計算 / Semantic Similarity



自然言語処理って何が難しい

● 言語表現

ドキュメント

1:07 P.M. EST

THE PRESIDENT: Thank you. When I came into office, I promised to look at the world's challenges with open eyes and very fresh thinking. We cannot solve our problems by making the same failed assumptions and repeating the same failed strategies of the past. Old challenges demand new approaches.

My announcement today marks the beginning of a new approach to conflict between Israel and the Palestinians.

In 1995, Congress adopted the Jerusalem Embassy Act, urging the federal government to relocate the American embassy to Jerusalem and to recognize that that city -- and so importantly -- is Israel's capital. This act passed Congress by an overwhelming bipartisan majority and was reaffirmed by a unanimous vote of the Senate only six months ago.

● データの取得

Congress adopted the Jerusalem Embassy Act

文章

- congress
- adopt
- ...

単語

単語をベクトルで表現する：I

- シンプルな考え方：One-hot encoding

ドキュメントの中のユニークな単語をリストアップし、各単語に番号を付ける

congress	0	0	0	1	...
adopt	0	1	0	0	...
Jerusalem	0	0	...	1	...
...

単語をベクトルで表現する：I

- シンプルな考え方：One-hot encoding

ドキュメントの中のユニークな単語をリストアップし、各単語に番号を付ける

congress	0	0	0	1	...
adopt	0	1	0	0	...
Jerusalem	0	0	...	1	...
...



※次元数は問題になる

単語をベクトルで表現する：Ⅱ

● 次元削減：トピックモデル

ドキュメント

1:07 P.M. EST

THE PRESIDENT: Thank you. When I came into office, I promised to look at the world's challenges with open eyes and very fresh thinking. We cannot solve our problems by making the same failed assumptions and repeating the same failed strategies of the past. Old challenges demand new approaches.

My announcement today marks the beginning of a new approach to conflict between Israel and the Palestinians.

In 1995, Congress adopted the Jerusalem Embassy Act, urging the federal government to relocate the American embassy to Jerusalem and to recognize that that city -- and so importantly -- is Israel's capital. This act passed Congress by an overwhelming bipartisan majority and was reaffirmed by a unanimous vote of the Senate only six months ago.



潜在意味
latent semantics



単語

- congress
- adopt
- government
- president
- fresh
- ...

単語をベクトルで表現する：Ⅱ

● 次元削減：トピックモデル

ドキュメント

1:07 P.M. EST

THE PRESIDENT: Thank you. When I came into office, I promised to look at the world's challenges with open eyes and very fresh thinking. We cannot solve our problems by making the same failed assumptions and repeating the same failed strategies of the past. Old challenges demand new approaches.

My announcement today marks the beginning of a new approach to conflict between Israel and the Palestinians.

In 1995, Congress adopted the Jerusalem Embassy Act, urging the federal government to relocate the American embassy to Jerusalem and to recognize that that city -- and so importantly -- is Israel's capital. This act passed Congress by an overwhelming bipartisan majority and was reaffirmed by a unanimous vote of the Senate only six months ago.



単語

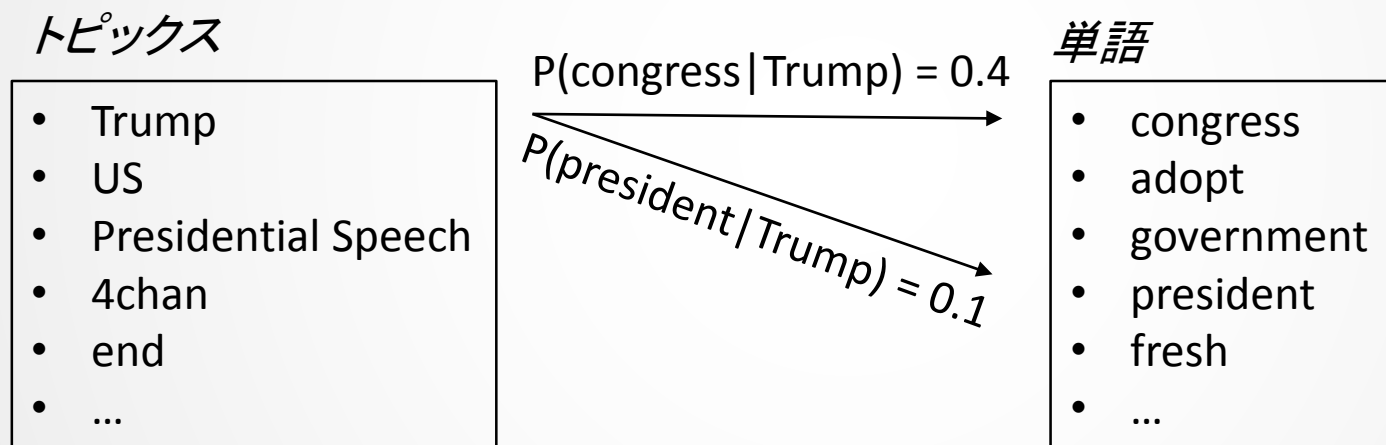
- congress
- adopt
- government
- president
- fresh
- ...



トピックス	Trump	US	Presidential speech	4chan	...
-------	-------	----	---------------------	-------	-----

単語をベクトルで表現する：Ⅱ

● 次元削減：トピックモデル*



単語をベクトルで表現する：Ⅱ

● 次元削減：トピックモデル*

ドキュメント

1:07 P.M. EST

THE PRESIDENT: Thank you. When I came into office, I promised to look at the world's challenges with open eyes and very fresh thinking. We cannot solve our problems by making the same failed assumptions and repeating the same failed strategies of the past. Old challenges demand new approaches.

My announcement today marks the beginning of a new approach to conflict between Israel and the Palestinians.

In 1995, Congress adopted the Jerusalem Embassy Act, urging the federal government to relocate the American embassy to Jerusalem and to recognize that that city -- and so importantly -- is Israel's capital. This act passed Congress by an overwhelming bipartisan majority and was reaffirmed by a unanimous vote of the Senate only six months ago.



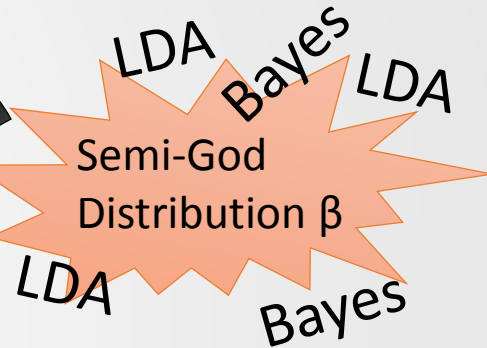
潜在要素
latent semantics



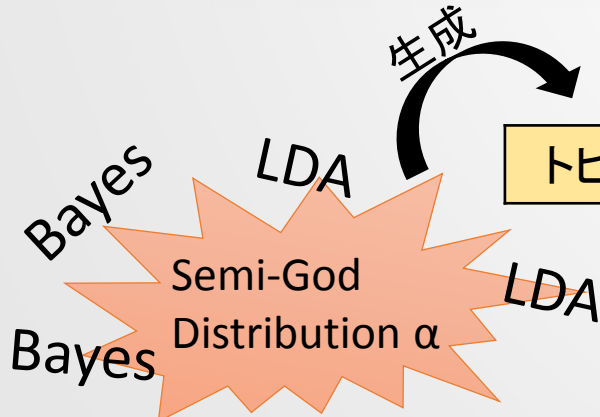
単語

- congress
- adopt
- government
- president
- fresh
- ...

生成



生成



トピックス

Trump

US

Presidential speech

4chan

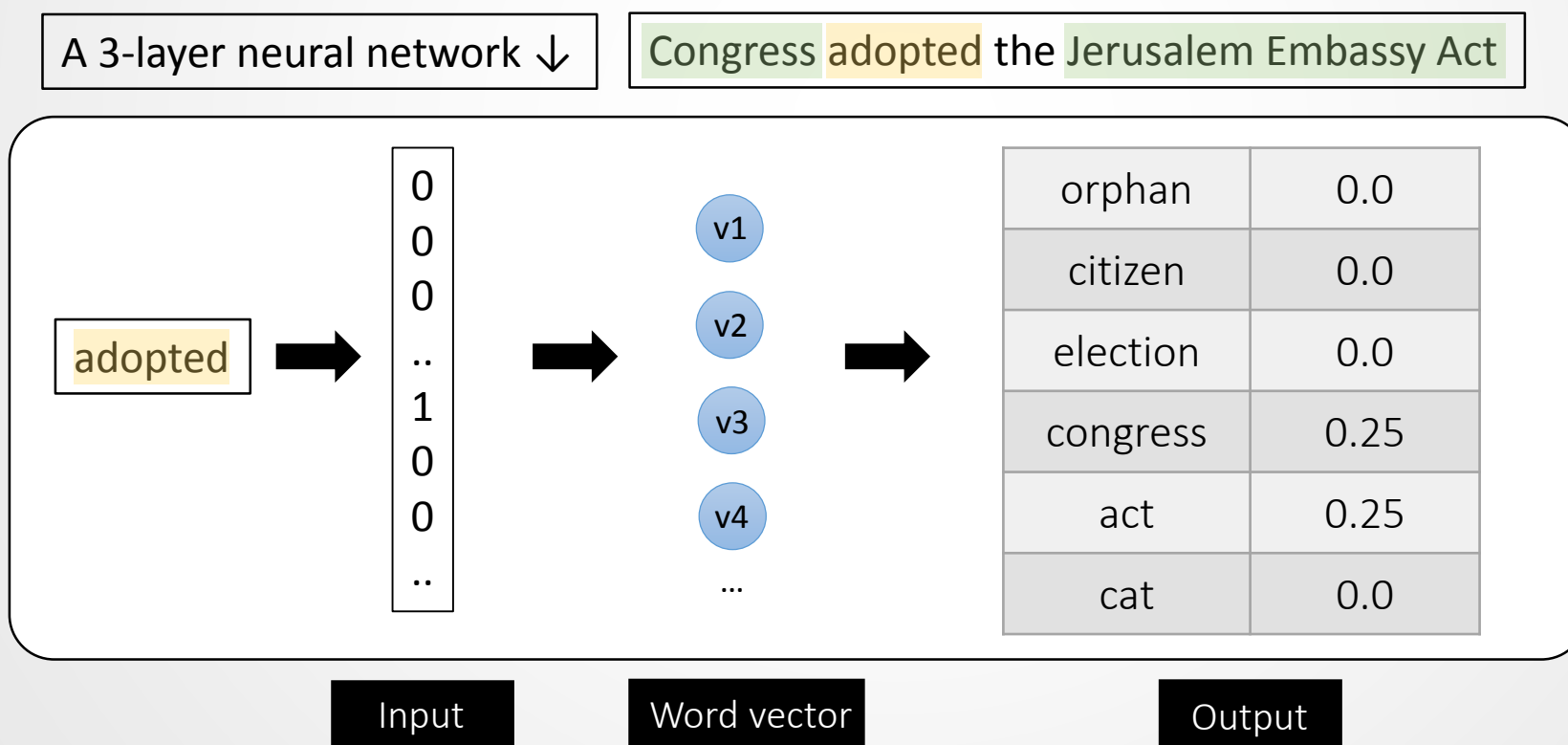
...

* Latent Dirichlet Allocation / 潜在的ディリクレ配分法

I と II のような、単語の前後文を考慮しないモデルは
Bag-of-words model / 詞袋模型 という。

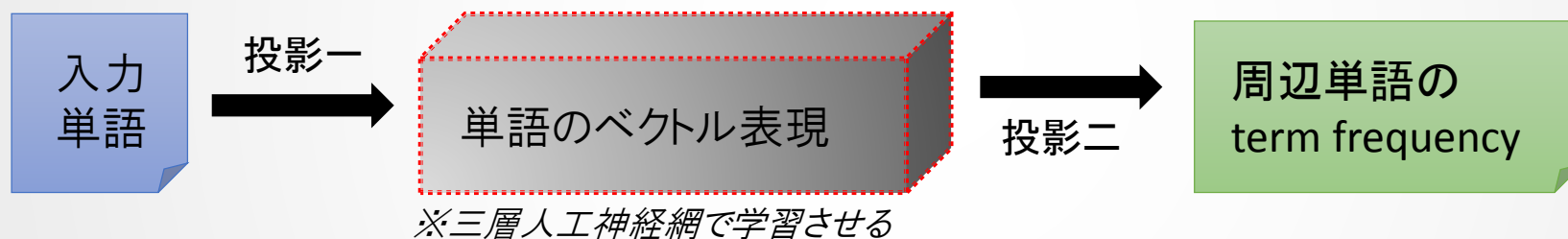
単語をベクトルで表現する：Ⅲ

- 単語のコンテキストを考えましょう



単語をベクトルで表現する：Ⅲ

- 単語のコンテキストを考えましょう



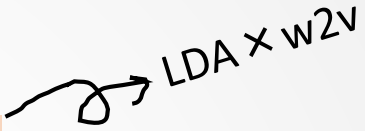
単語をベクトルで表現する：Ⅲ

- ベクトル化した単語はどこに使うか
 - ① ツイッターデータにおける共起単語調査
 - ② 文章類似度の算出、文章のクラスタリングなどの教師なし NLP タスク
 - ③ 文章分類器、感情分析などの教師ありの NLP タスク
 - ④ 短文エンベディング ⇔ 再帰型人工神経網との互換可能
 - ⑤ 文章のメタ特徴学習、深層神経網に入力

単語をベクトルで表現する：Ⅲ

- ベクトル化した単語はどこに使うか（応用）
 - ① EC サイト* のログデータ分析（ワンセッションを短文として扱う）
 - ② ユーザーの行動分析（①とは違ってアイテムをエンベディングする）
 - ③ 基本短文として使えるデータは対応できる（と思いたい）

Appendix: 亜種モデルとミックスモデル

- Sentence2Vec | Doc2Vec | LDA2Vec 
- Twitter-LDA \Rightarrow 短文用 LDA
- illus2vec \Rightarrow バナー分析 by テックチーム
- Bayes \times Word2Vec ?