畳み込みニューラルネットワークを用いた複単語表現の解析

進藤 裕之　　松本 裕治

奈良先端科学技術大学院大学

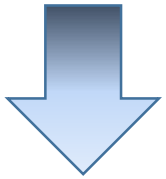
2015/09/27
自然言語処理研究会（NL）

背景

複単語表現の同定と品詞タグ付け

入力(文)

... in getting their money back ...



トークナイズ ＋ 品詞タグ付与

出力

VBG

... in getting their money back ...

IN PRP$ NN

複単語表現(MWE)

get ～ back: 取り戻す

これまでの複単語表現に関する研究

個別の複単語表現のみを対象としたコーパス構築と解析

- 複合名詞 [Kim and Baldwin '08]
- 軽動詞構文 [Tu and Roth '11]
- 句動詞の一部 [Cook et al. '08]

など

- データが小規模
- 解析手法の良し悪しを互いに比較することが困難

近年の複単語表現に関する研究

複単語表現の網羅的な注釈付けと解析

・Schneiderら（2014）:
　Web TreebankのReview部（3812文）に複単語表現の注釈付け

・Shigetoら（2013）:
　Penn TreebankのWSJ部に機能語相当の複単語表現の注釈付け
　ex. "according to", "as well as", "because of"

・駒井ら（2014）:
　OntoNotesコーパスのWSJ部に，句動詞の注釈付け
　ex. "get back", "pick up", "look forward to"

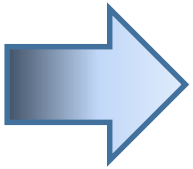
本研究の対象とする複単語表現

・Schneiderら（2014）:
Web TreebankのReview部（3812文）に複単語表現の注釈付け

・Shigetoら（2013）:
Penn TreebankのWSJ部に機能語相当の複単語表現の注釈付け
ex. "according to", "as well as", "because of"

・駒井ら（2014）:
OntoNotesコーパスのWSJ部に，句動詞の注釈付け
ex. "get back", "pick up", "look forward to"

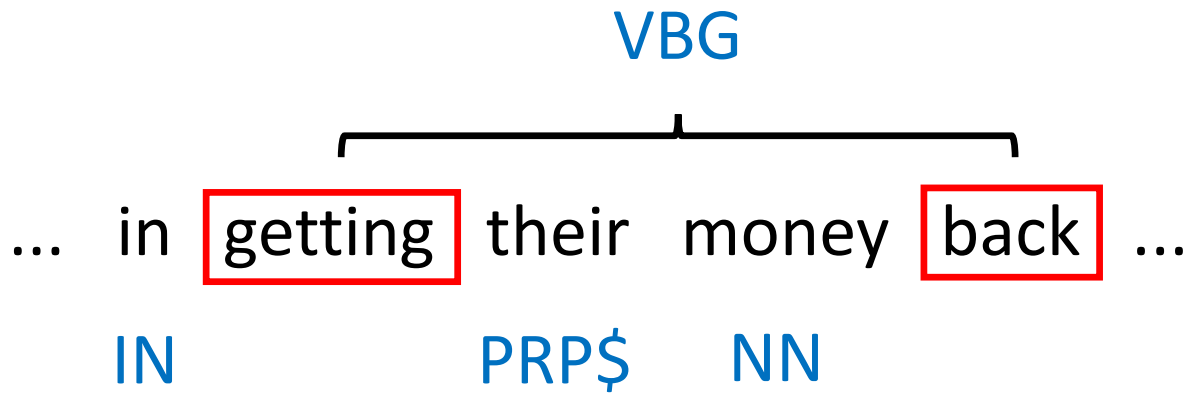
併合



OntoNotes： 約37000文
複単語表現の注釈： 約1400種類，12000事例

日本語形態素解析との違い

<u>複単語表現は，文中に連続して出現するとは限らない</u>



非連続パターンを扱える解析手法が必要

従来研究の問題点

非連続パターンを扱える系列ラベリング手法 [Schneider et al. '14]
（品詞を事前に与えて，複単語表現の同定のみを行う）

<u>素性を人手で設計する必要がある</u>

1. capital
2. word shape
3. prefix
4. suffix
5. has digit
6. has non-alphanumeric
7. context word
8. context bigram
9. lemma
10. context POS
11. context bigram POS

etc...

１．様々なレベル（文字，単語，複単語）の素性が必要

２．連続／非連続パターンの取り扱い

有効な特徴量を発見するのはコスト大

本研究

畳み込みニューラルネットワークを用いて，複単語表現の解析に有効な特徴量を自動的に学習する

・素性テンプレートが不要
・既存手法と同等以上の解析性能

関連研究：
畳み込みニューラルネットワークによる単語の品詞タグ付け
[Collobert et al. '11, Santos and Zadrozny '14]

→ 複単語表現レベルまで拡張したものが本研究

提案手法

解析の流れ

トークナイズと品詞推定のパイプライン処理（点推定）

（1） 入力文の単語列から，複単語表現の候補を検索する

（2） 各複単語表現の候補に対して，

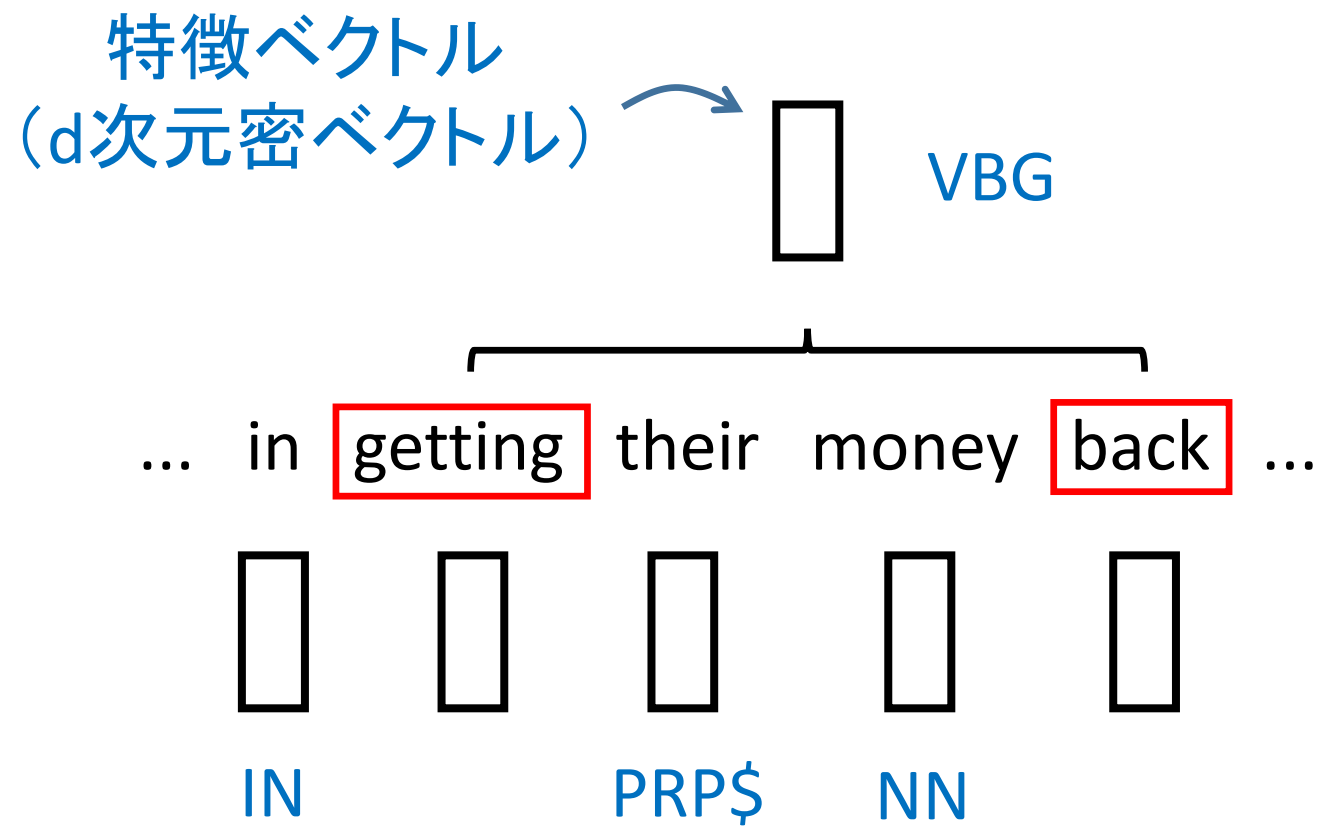
トークナイズ

a) 畳み込みニューラルネットワークで特徴ベクトルを計算
b) 複単語表現かどうかを判別

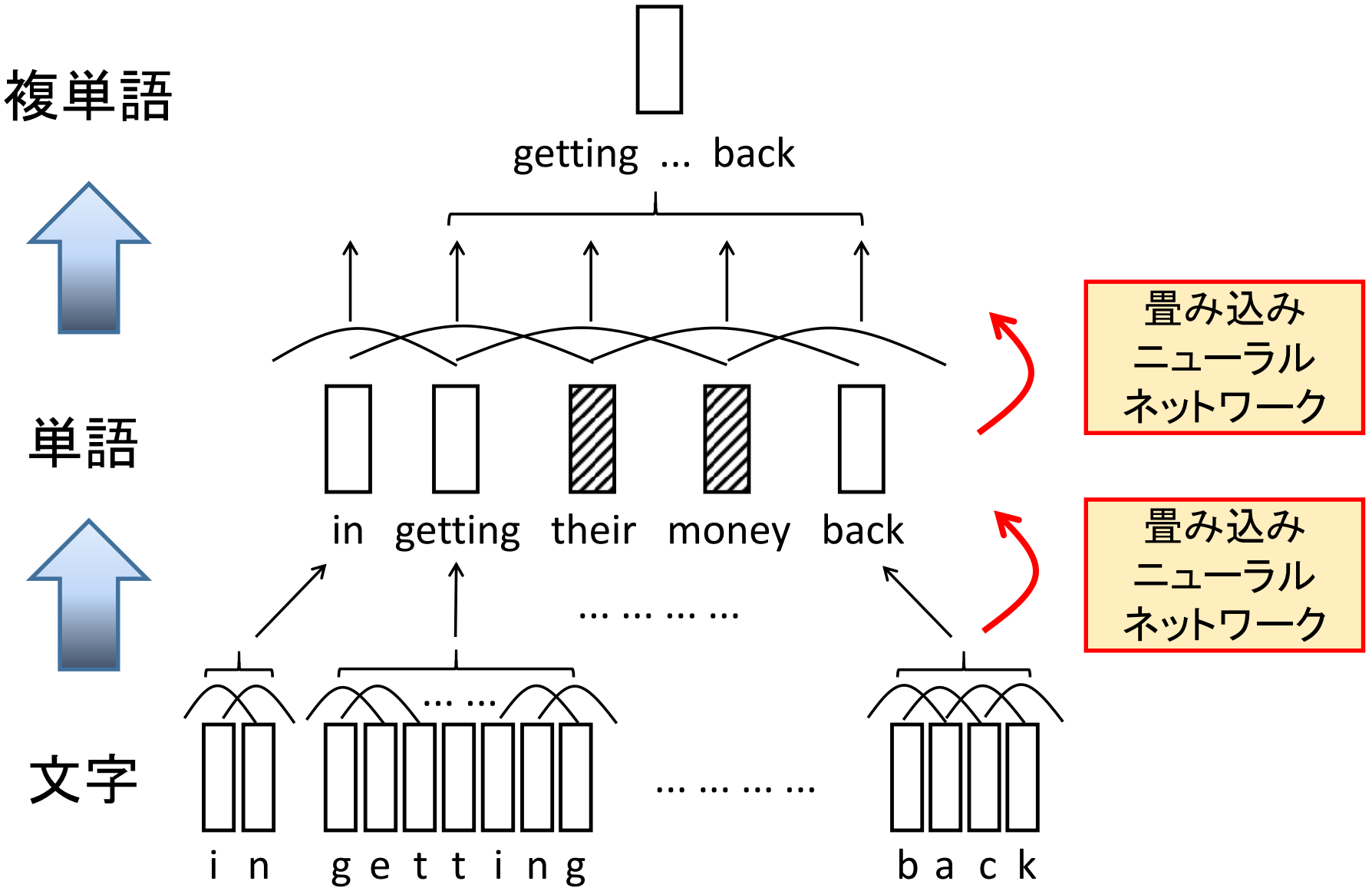
品詞タグ付け

（3） 各トークンに対して，（2）の特徴ベクトルから品詞を判別

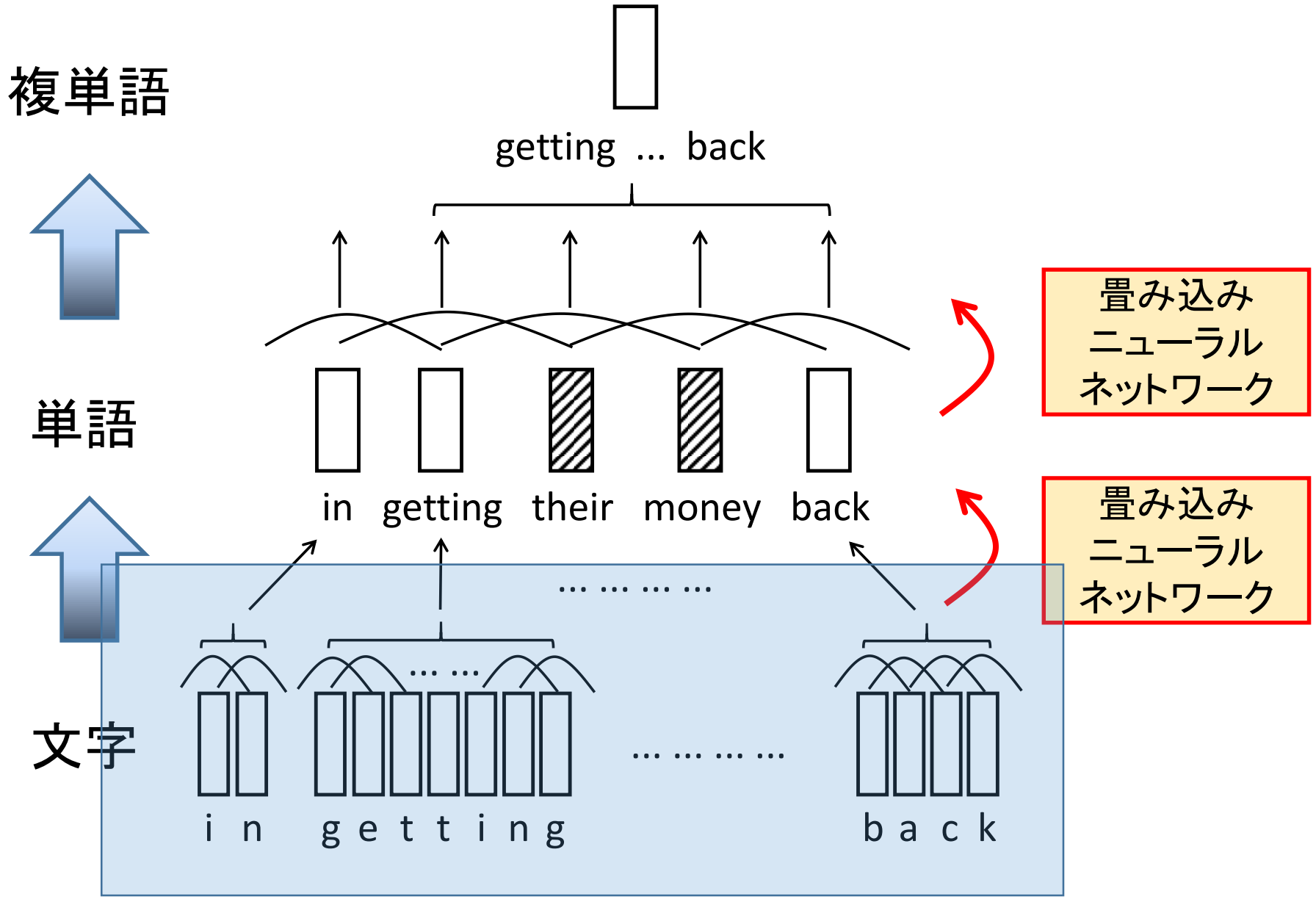
解析の流れ



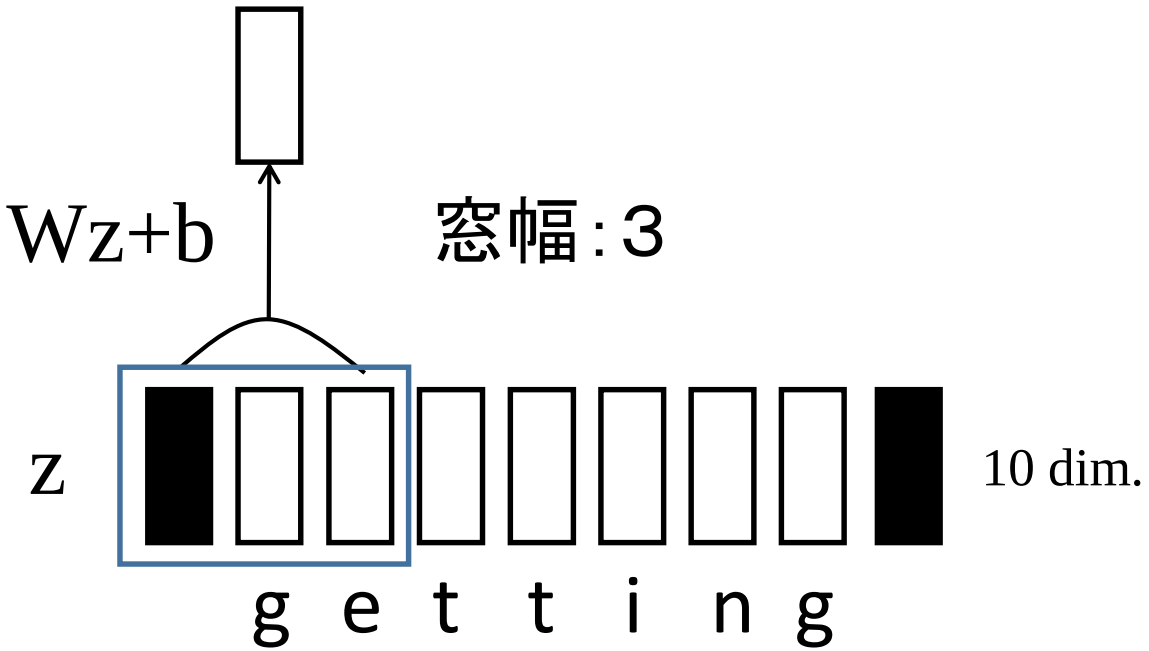
特徴ベクトルの計算


複単語
getting ... back
単語
in getting their money back
文字
i n g e t t i n g
b a c k
畳み込み
ニューラル
ネットワーク
畳み込み
ニューラル
ネットワーク

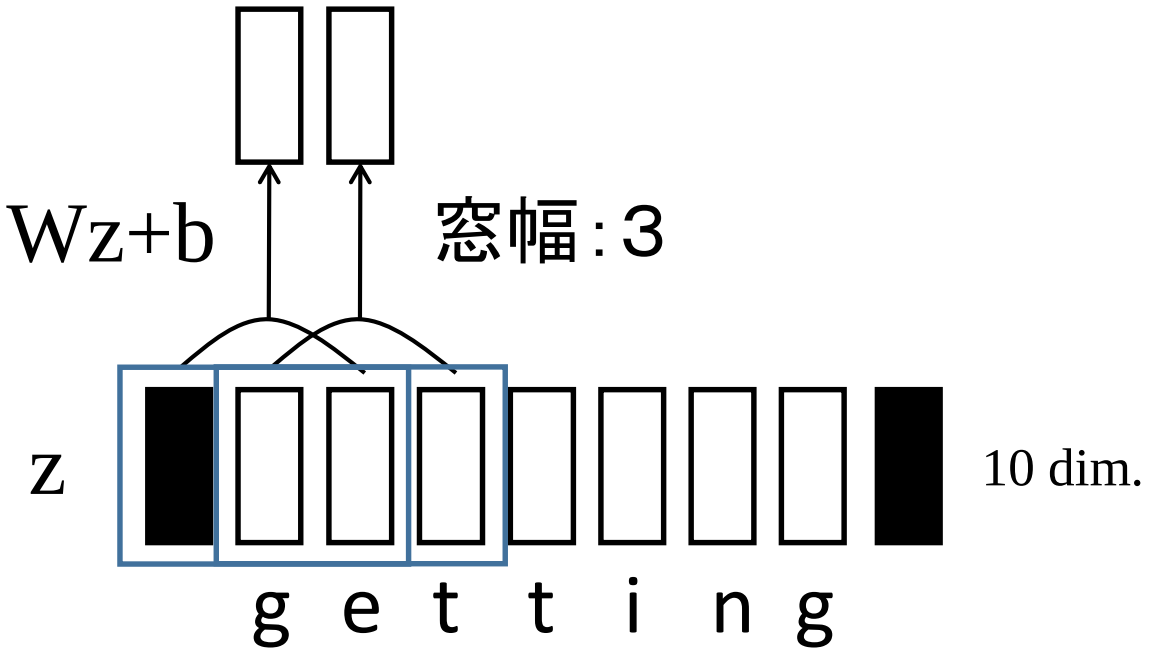
特徴ベクトルの計算


複単語
getting ... back
単語
in getting their money back
畳み込み
ニューラル
ネットワーク
畳み込み
ニューラル
ネットワーク
文字
i n g e t t i n g
b a c k

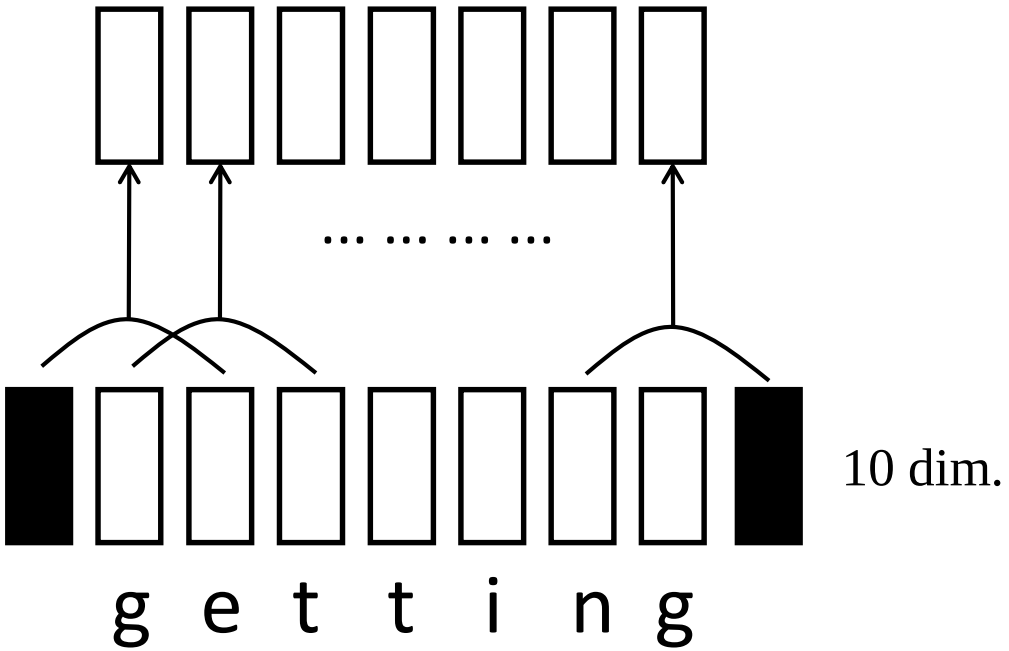
文字レベルの特徴ベクトル



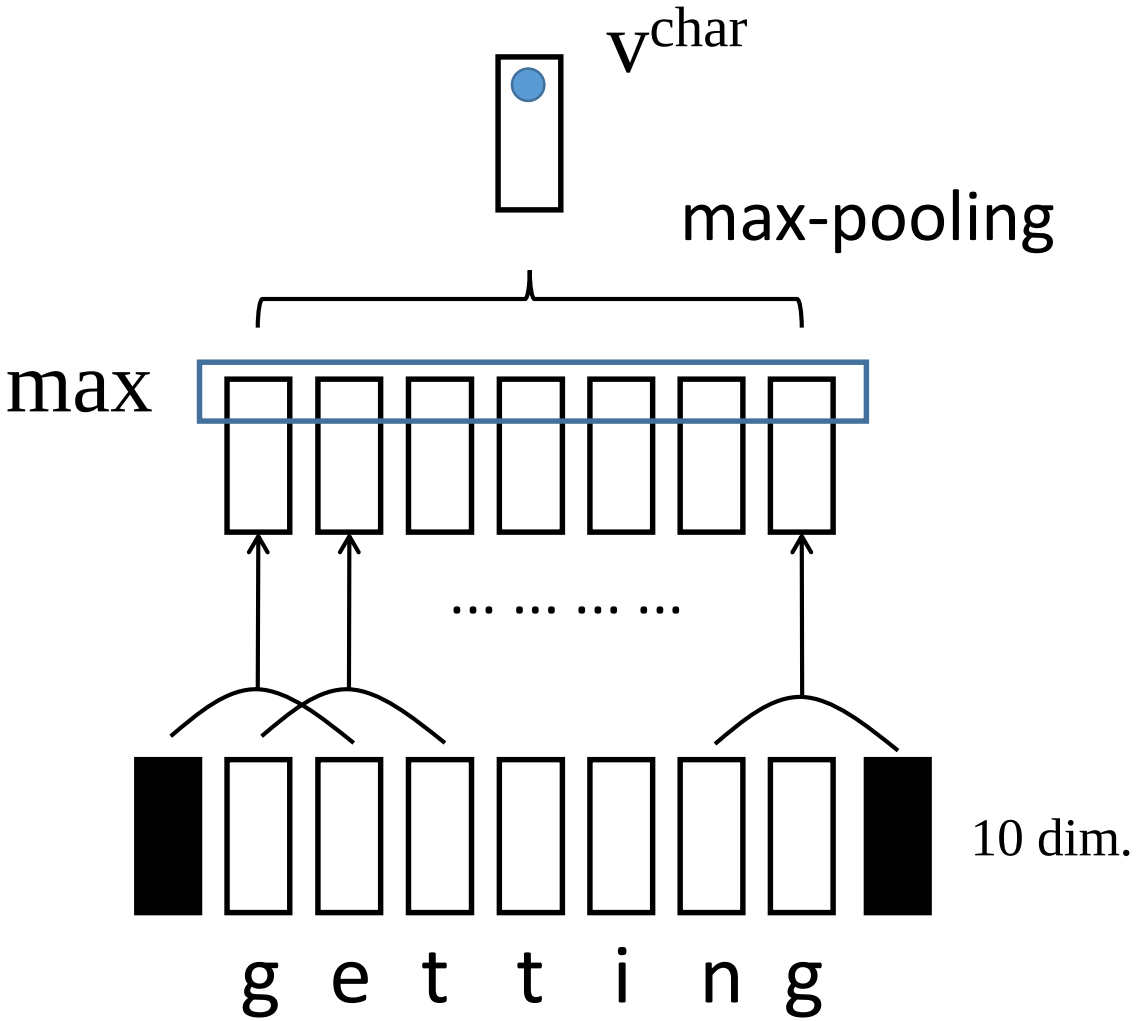
文字レベルの特徴ベクトル



文字レベルの特徴ベクトル

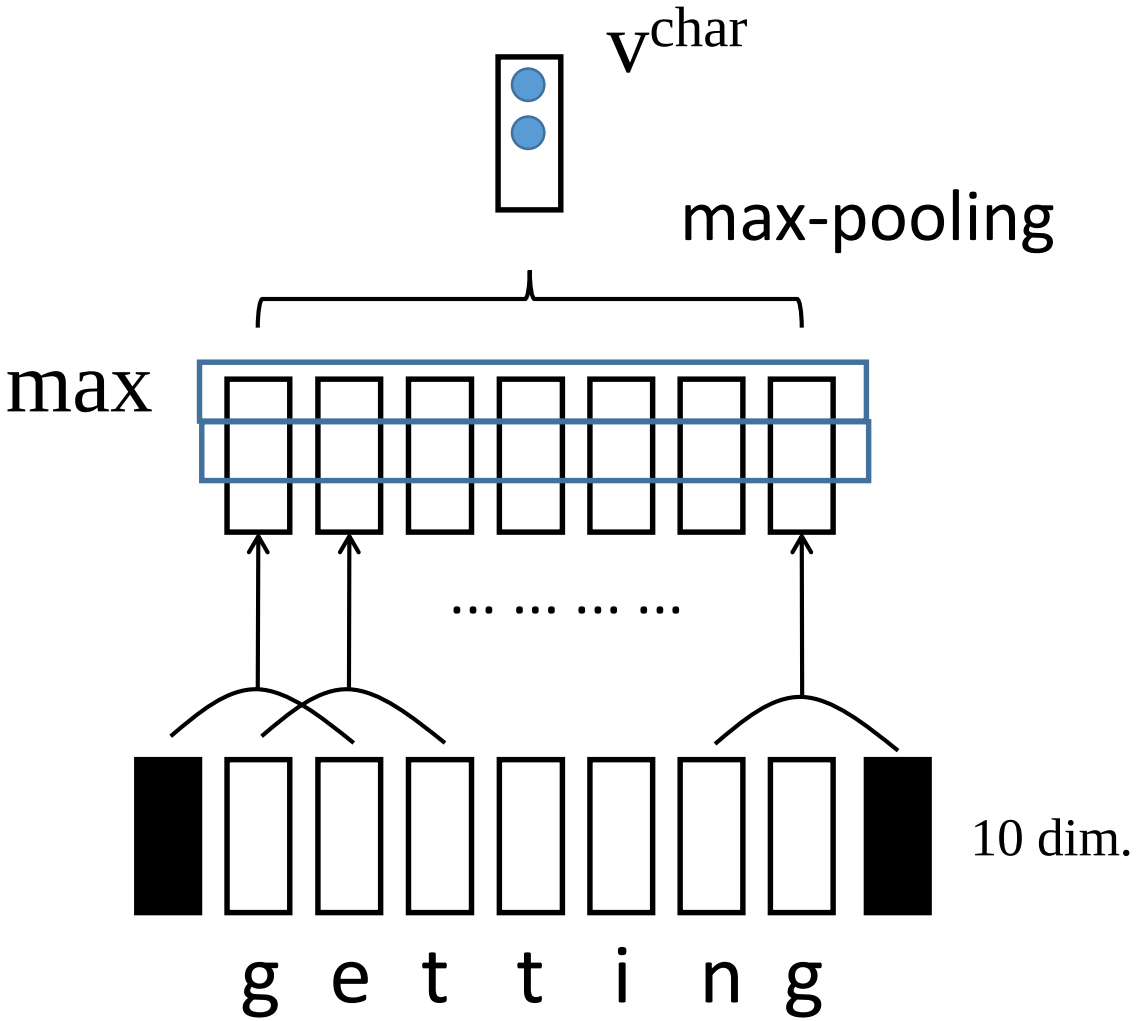


文字レベルの特徴ベクトル



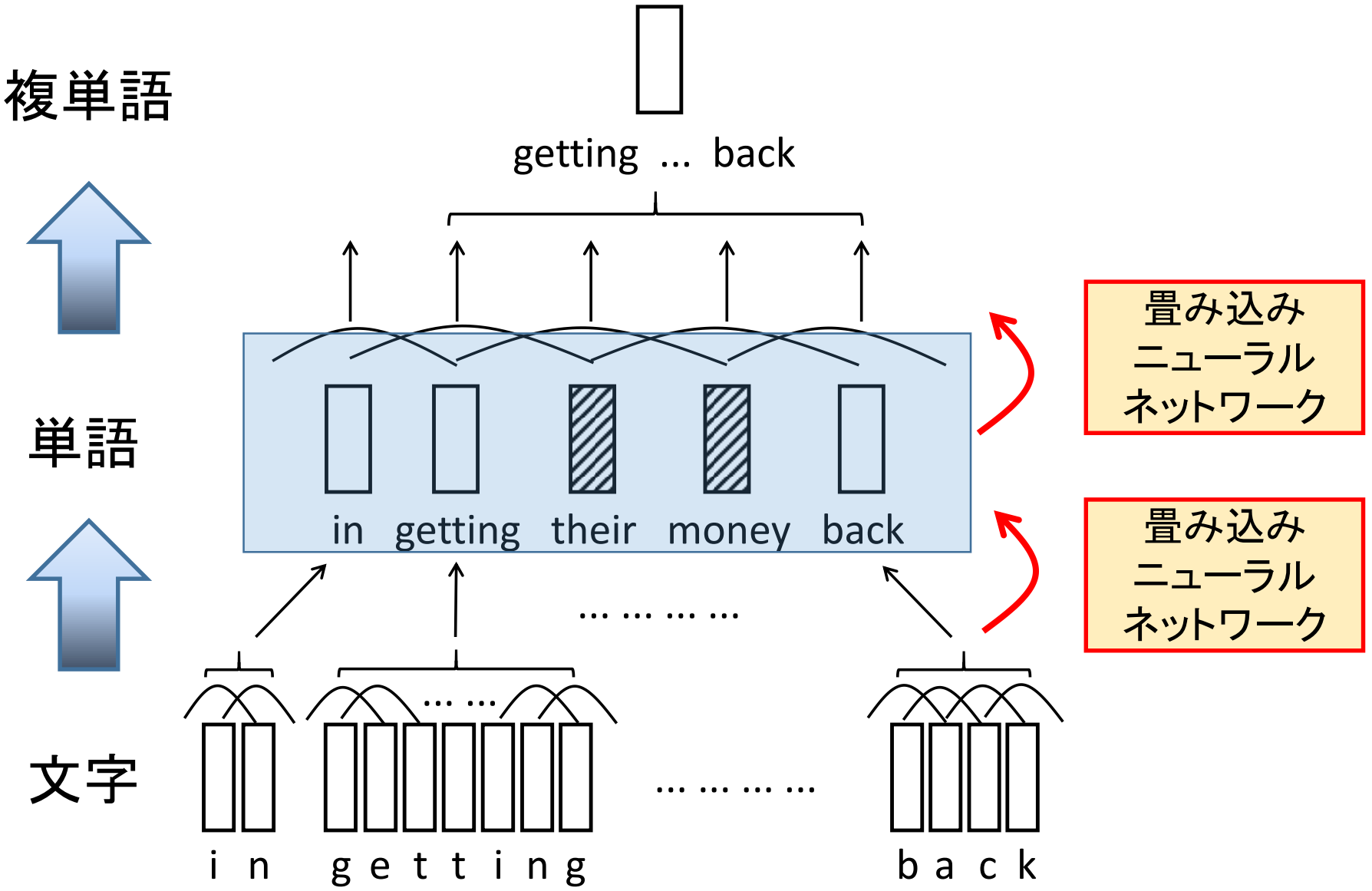
重要な特徴量
を抽出

文字レベルの特徴ベクトル



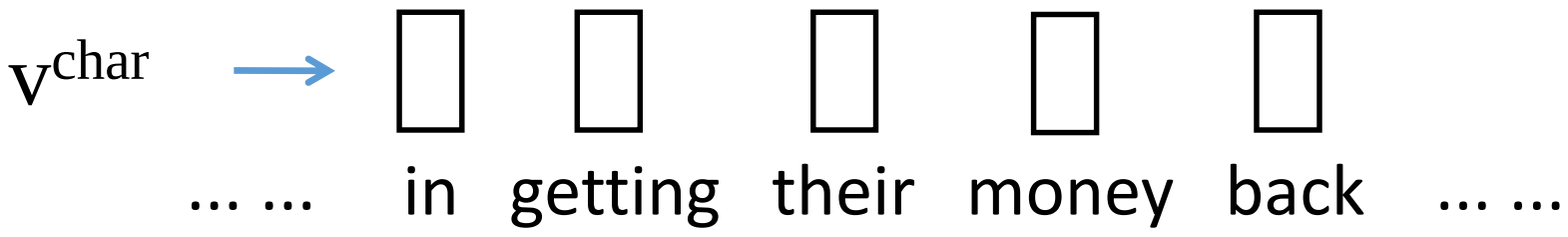
重要な特徴量
を抽出

特徴ベクトルの計算


複単語
getting ... back
単語
in getting their money back
畳み込み
ニューラル
ネットワーク
畳み込み
ニューラル
ネットワーク
文字
i n
g e t t i n g
b a c k

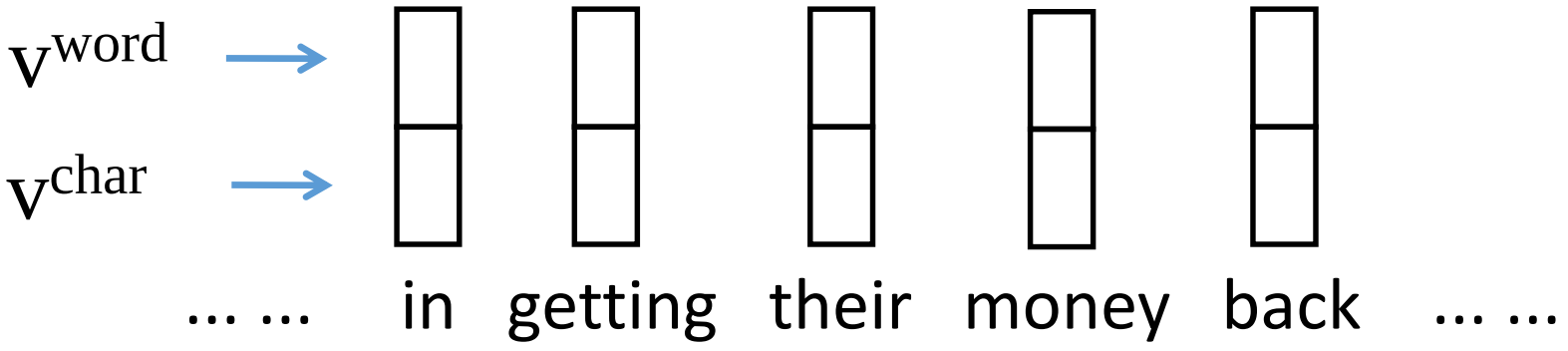
単語レベルの特徴ベクトル

文字ベクトルと，単語ベクトルを連結する

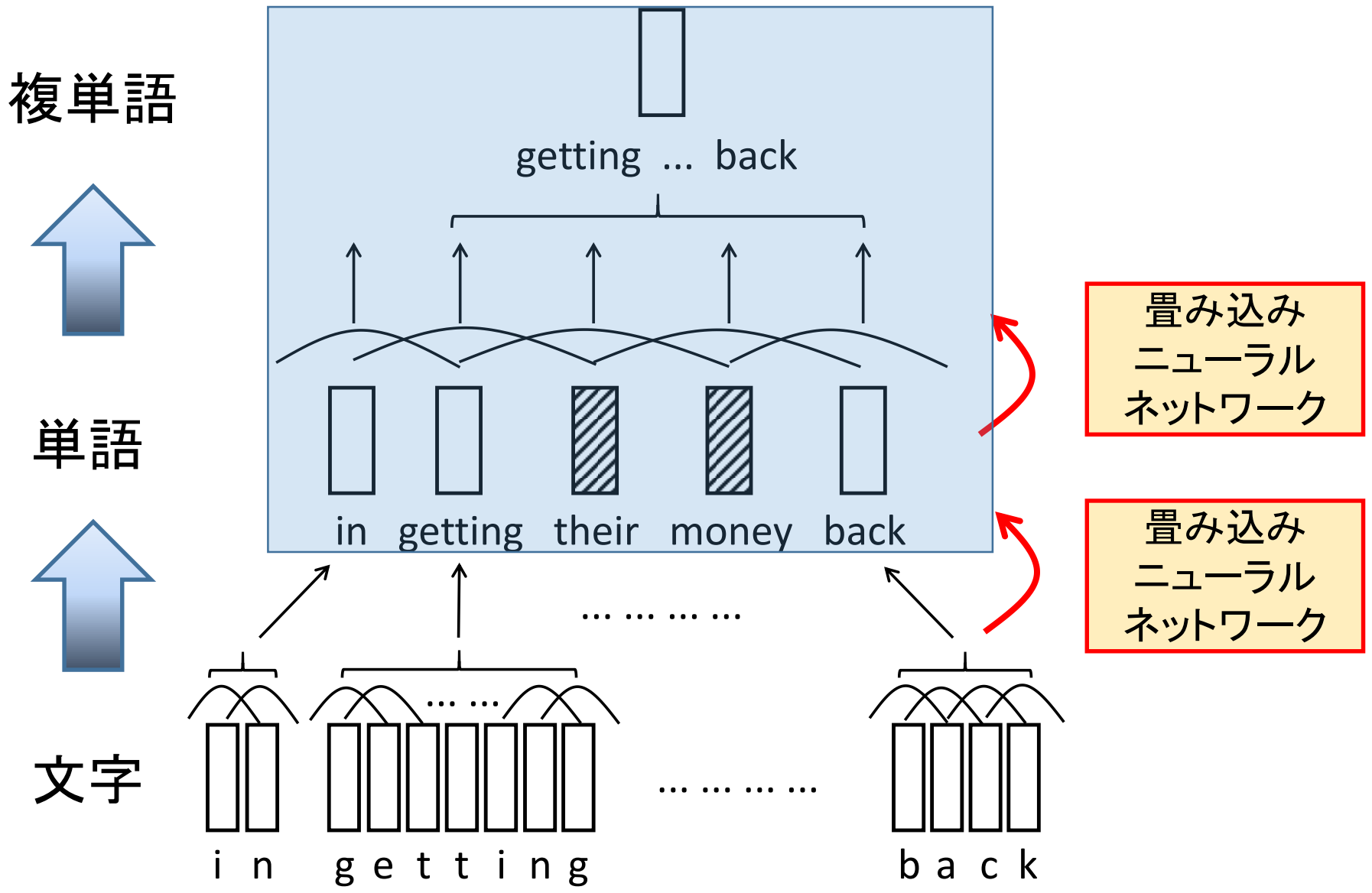


単語レベルの特徴ベクトル

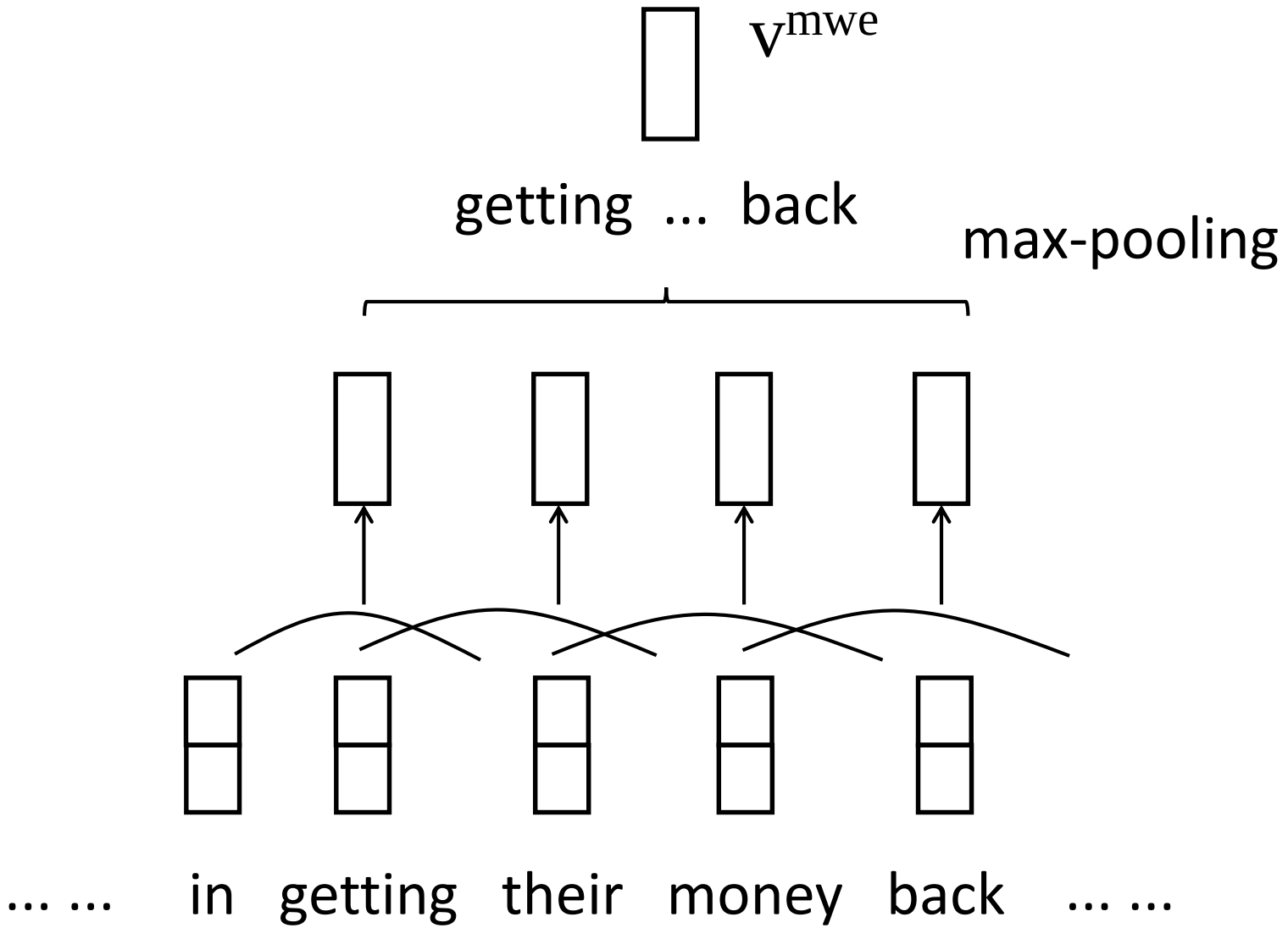
文字ベクトルと，単語ベクトルを連結する



特徴ベクトルの計算

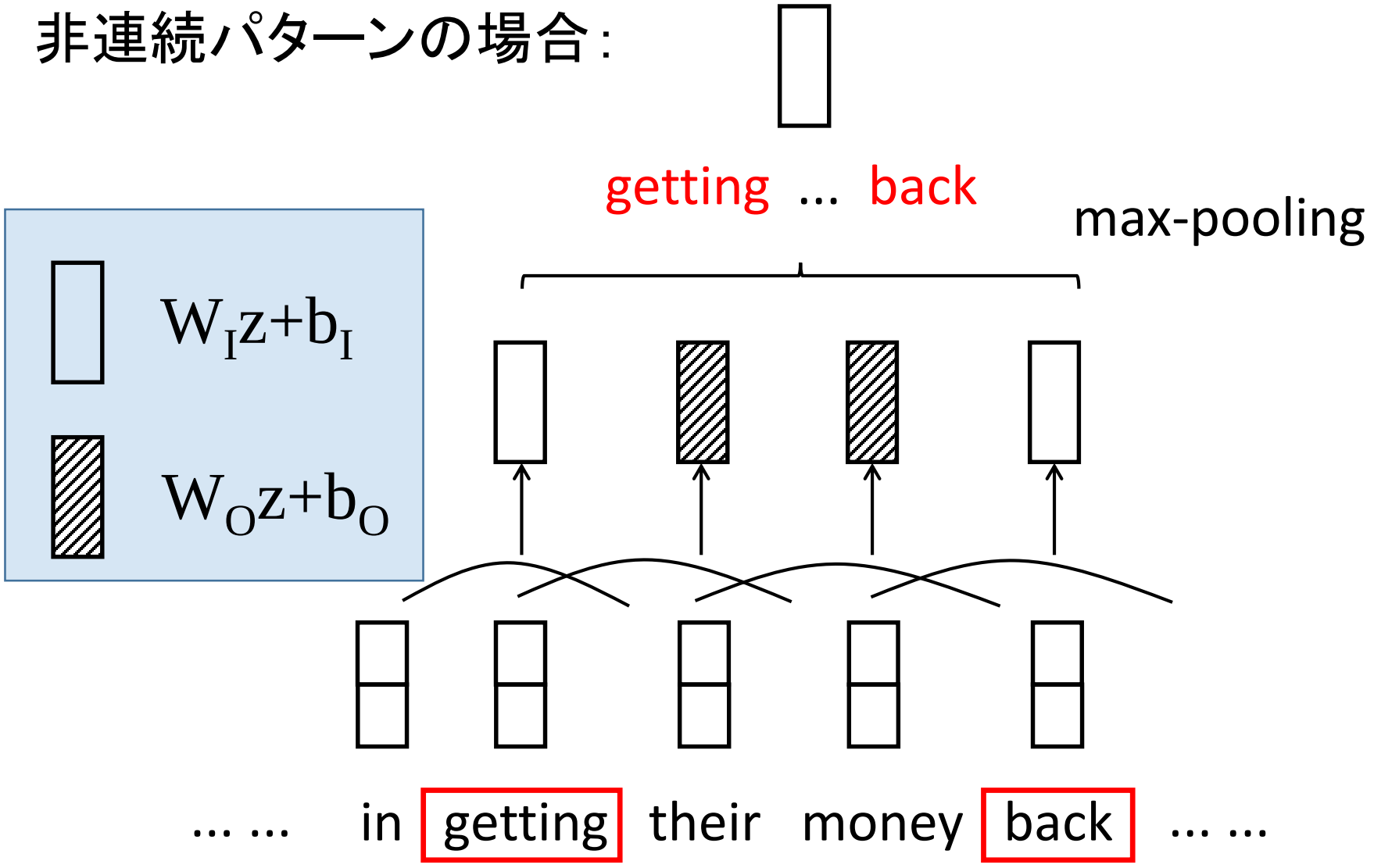

複単語
単語
文字
getting ... back
in getting their money back
畳み込み
ニューラル
ネットワーク
畳み込み
ニューラル
ネットワーク
i n
g e t t i n g
b a c k

複単語レベルの特徴ベクトル

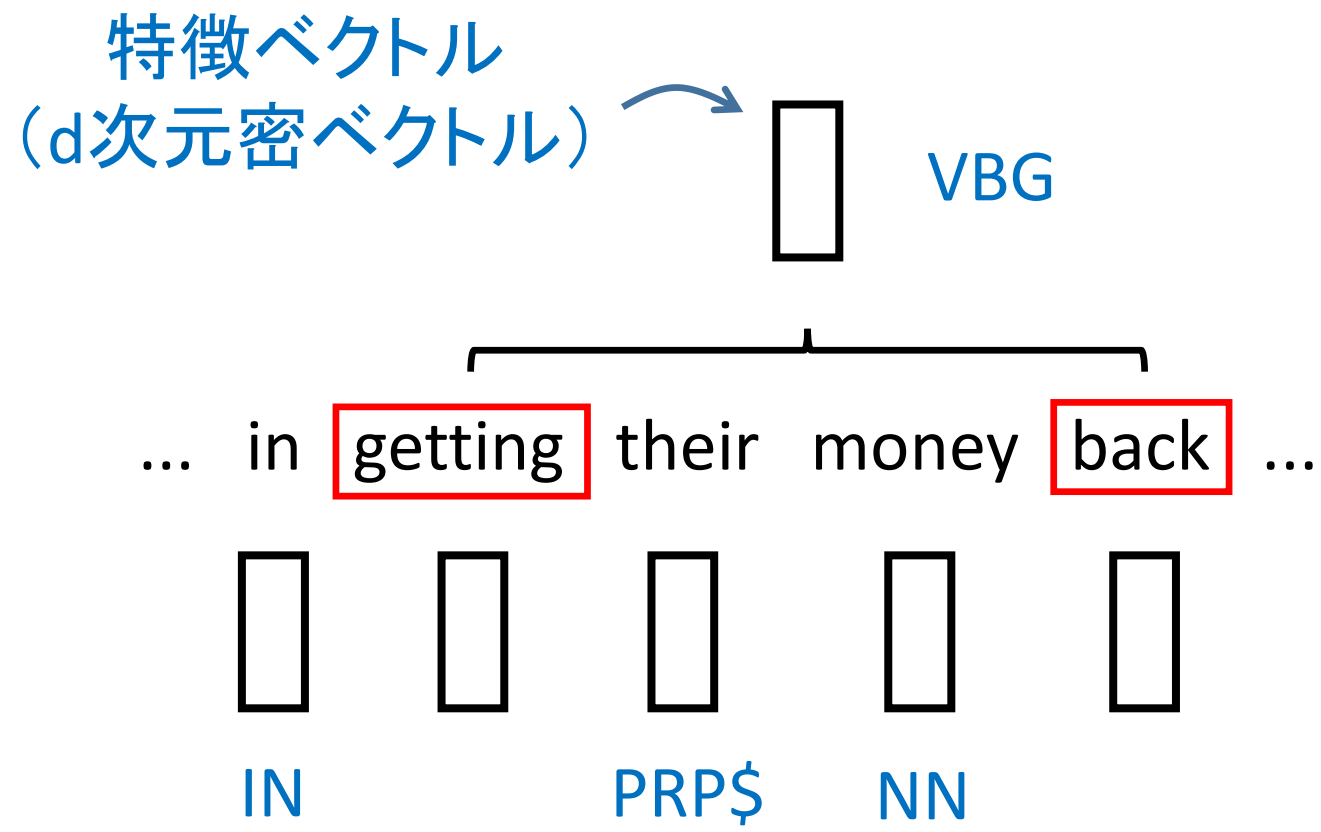


複単語レベルの特徴ベクトル

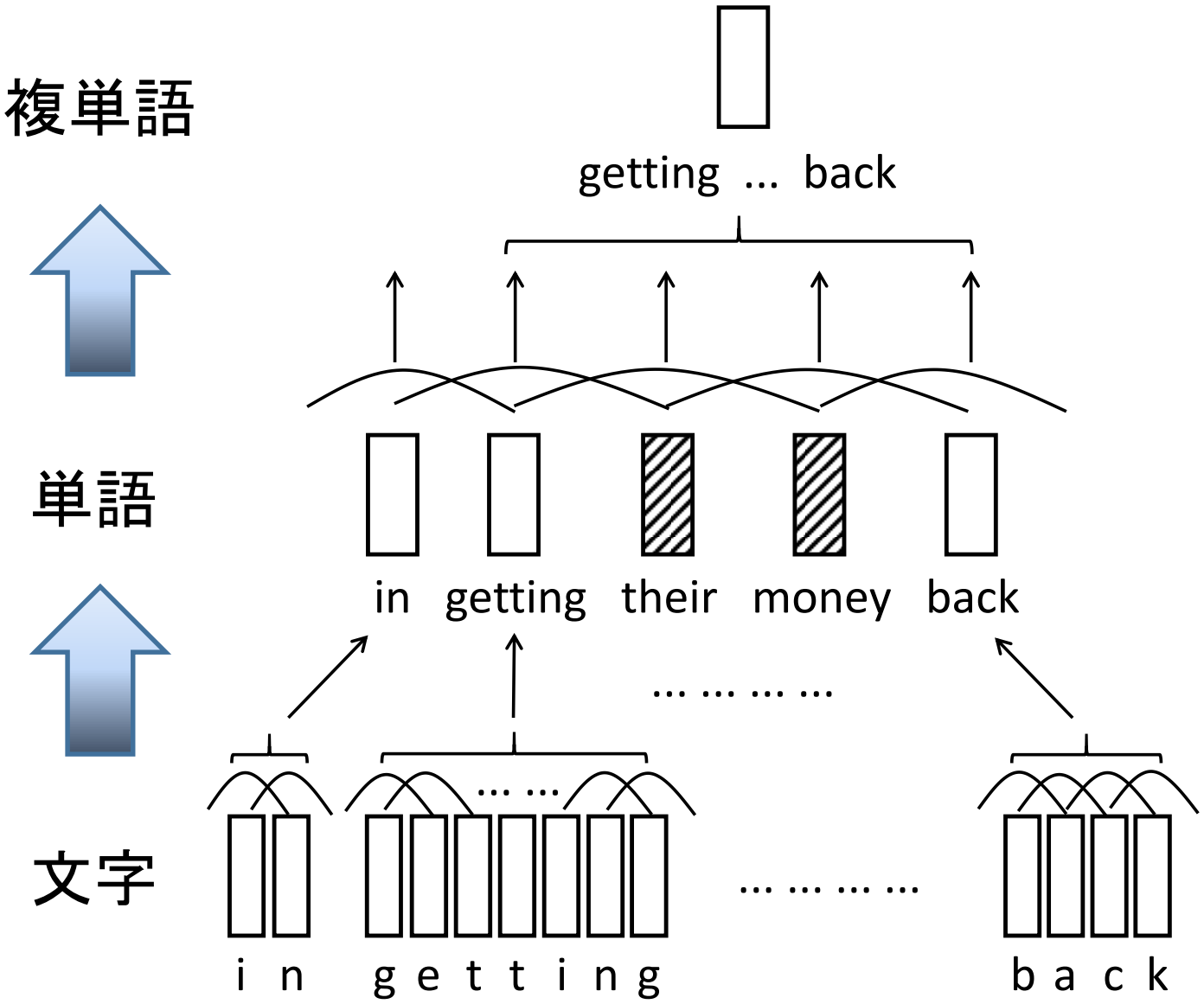
非連続パターンの場合：



解析の流れ



特徴ベクトルの計算


複単語
getting ... back
単語
in getting their money back
文字
... ...
... ...
... ...
i n
g e t t i n g
b a c k

実験

実験設定

データセット

OntoNotesコーパス（約37000文）

複単語表現の注釈： 約1400種類，12000事例

学習データ： セクション02-21，テストデータ：セクション23

評価

トークンと品詞タグの完全一致基準で適合率，再現率，F値

提案モデルの学習

- クロスエントロピー損失関数
- 確率的勾配法（SGD）

実験設定

提案手法の主なハイパーパラメータ

- 文字ベクトル次元: 10
- 単語ベクトル次元: 150
- 複単語ベクトル次元: 300
- 文字CNNの窓幅: 5
- 単語CNNの窓幅: 5

実験設定

比較手法

1. 規則ベース

   if　複単語表現の候補が文中で連続　then　正例
   else　負例

2. 拡張BIO系列ラベリング [Schneider et al. '14]

   固有表現抽出の手法を，非連続パターンを扱えるように拡張

   ... in getting their money back ...

   O B o o I

1と2は，品詞タグを事前に与える必要がある
→ Stanford Tagger + 10-fold jack-knife法で品詞を付与

実験結果

全トークンに対する評価結果

	適合率	再現率	F 値
規則ベース	95.9	96.7	96.3
拡張 BIO 系列ラベリング [10]	96.8	96.7	96.7
提案手法	**97.3**	**97.3**	**97.3**

複単語表現（2単語以上）のみの評価結果

	適合率	再現率	F 値
規則ベース	76.1	92.3	83.4
拡張 BIO 系列ラベリング [10]	93.3	90.0	91.6
提案手法	**92.2**	**93.5**	**92.8**

まとめ

タスク:

複単語表現の同定と品詞タグ付け

手法:

階層的畳み込みニューラルネットワーク

文字，単語，複単語の特徴ベクトルを自動的に学習

結果:

既存手法を上回る解析精度