知能機械と自然言語処理
第14回：
自然言語処理の応用5：
現代の自然言語処理

David Ramamonjisoa

目次

- 分散表現(単語埋め込み：Word embeddings)
- テキスト分類
- Word2Vecを用いた類語の取得、足し算、引き算
- テキストから情報抽出

分散表現(単語埋め込み:Word embeddings)

- 分散表現(あるいは単語埋め込み)とは、単語を高次元の実数ベクトルで表現する技術です
- 単語の分散表現は一般に 200次元などの高い次元のベクトルで表される
- 近い意味の単語を近いベクトルになる
- ベクトルの足し算が意味の足し算に対応する
- 大きなコーパスからの学習と加法構成性を特徴としている
- ディープラーニングと自然言語処理の応用

単語埋め込みレヤー (Embedding Layer)

- One-hot表現
  - 一番シンプルなEmbedding手法。まず表現したい語彙リストを作成し、各単語を表現する次元を準備する。表現したい文に含まれている単語に対応する次元を1に、それ以外を0にする方法。文のベクトルを作るBag-of-words（BoW）表現に利用される。
  - 欠点：未知語を扱うことができない、次元数が膨大となる
- ニューラルネットワークの実装(Keras Embedding)

Word2Vec

- word2vecは、大量のテキストデータを解析し、各単語の意味をベクトル表現化する手法。単語をベクトル化することで、単語同士の意味の近さを計算したり、単語同士の意味を足したり引いたりすることが可能。
- Word2vecには下記2つの方法が存在:CBOWとSkip-Gram

CBOWとSkip-Gram

- cbowモデルは、コンテキストに基づいて現在の単語を予測することにより、埋め込みを学習します。
- Skip-Gramは、ある単語の前後に出現する単語の出現確率を計算することでベクトル化をする方法。これは、意味が近い（≒単語ベクトルの距離が近い）単語は、周辺の単語も似ているはずという仮説に基づきます。

Glove

- Word2vec翌年の2014年に発表されたアルゴリズム。Glove はグローバルな情報を用いる count-basedな手法とローカルな文脈の情報を用いる predictive な手法を組み合わせたもの。いわゆる良いとこどりをした手法らしい。学習が速い、精度が高い、そして小さいコーパスでも動作可能とのこと。

fastText

- 2016年にFacebookが発表した手法。これまでの手法は単語をベースとするため、未知語に対応するのが難しい。これを克服するためのアイデアとして、単語より小さな単位でEmbedding を行う。
- 文字レベルのN-gram (Character N-gram) であるsub-wordを用いる。 Word2Vecでは、活用形が考慮されない（goとgoes、going、これらは全て「go」だが、字面的には異なるので別々の単語として扱う）。これに対してfastTextでは、単語を構成要素に分解（goesならgoとes)し、字面の近い単語同士により意味のまとまりをもたせる。

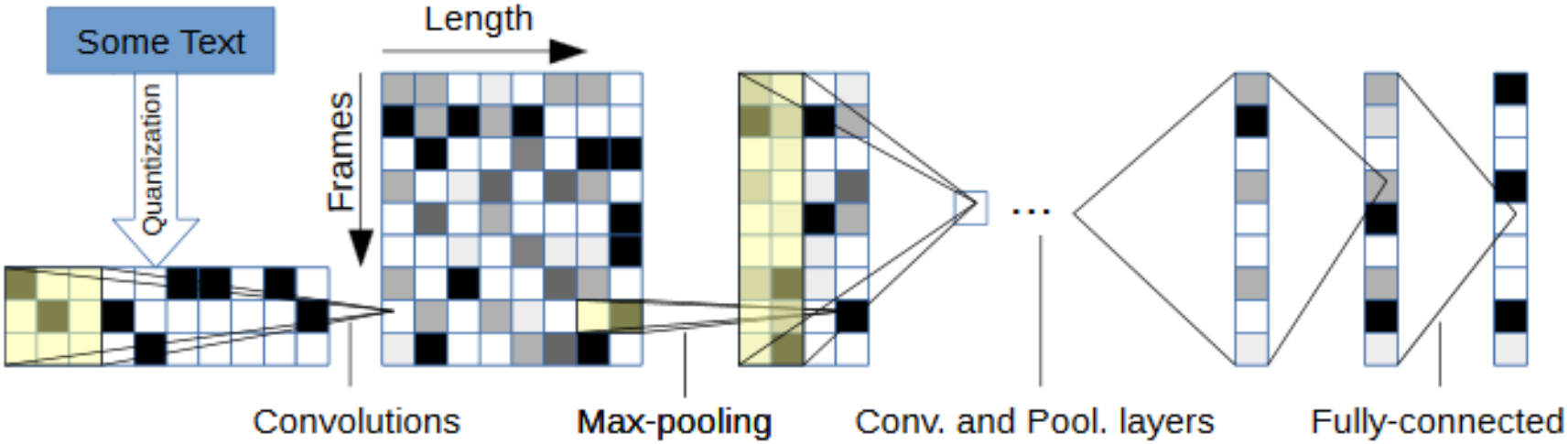
文字ベースで分散表現

- sub-wordよりも小さい単位である文字ベースでのEmbedding (Character-based Embedding) も存在している。近年は特に RNNを用いた文章生成などで用いられている。
- 日本語や中国語のような漢字文化圏では、文字種数のオーダーが2 桁以上異なるために個々の文字がより複雑な単語的意味をもっており，文字レベルでの Embedding が研究されている。漢字の場合は、その字形自体も意味をエンコードしていると。まぁ、言われてみれば本当にその通りです。字形を画像として捉え、CNNを用いて視覚的特徴のEmbeddingを作成している例もあるとのこと。

分散表現+CNN = テキスト分類

- 論文：Yoav Goldberg “A Primer on Neural Network Models for NLP” 2015

Zhang et al.:”Character-level CNN for Text Classification.” 2015



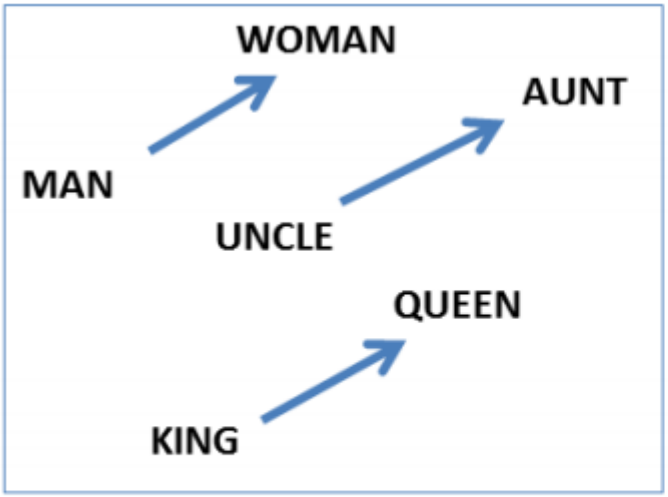
Word2Vecを用いた類語の取得、足し算、引き算

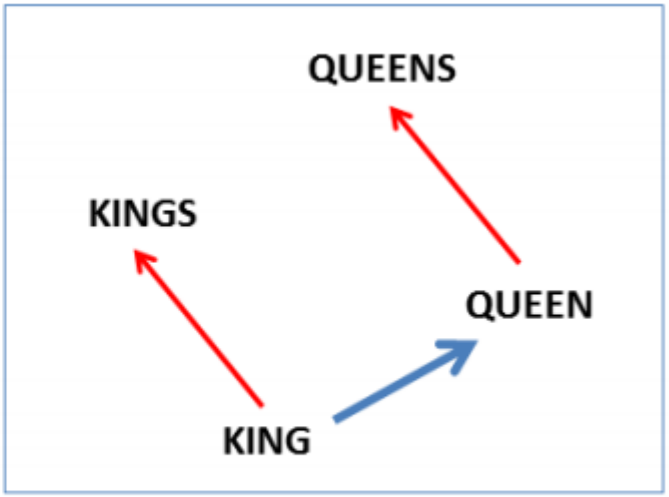
- 知識データを活用することでできること
  - “アメリカ”→“米国”, “フランス”→“仏国”,
  - キーワードの類似度測定に使う
  - ベクトル同士のコサイン類似度を計算する
  - 例えば、コーパス中に
    - 「かわいい犬を飼い始めた」と「かわいい猫を飼い始めた」
    - 文脈が同じであるため、「犬」と「猫」は似ている
    - これは「分布仮説」と呼ぶ

アナロジーの計算

- KING-MAN+WOMAN = QUEEN
- MAN-UNCLE+WOMAN = AUNT
- $y - x + x' = y' \equiv y + x' = y' + x \equiv y - x = y' - x', KING + WOMAN = QUEEN + MAN, KING - MAN = QUEEN - WOMAN$

単語ベクトル





(Mikolov et al., NAACL HLT, 2013)

テキストからの情報抽出

- 関係のアノテーション
- 正規表現を用いた関係抽出
- 係り受け構造をお用いた関係抽出

データの中から特定の情報を抜き出し、抜き出した情報を整理すること

大学名へのアノテーション付与

リンクを作成する：原因→結果

系列ラベリング

- データ列として単語列が与えられ、それぞれの単語に対して固有名詞をラベル付けを加工する
- LSTMニューラルネットワークなどを用いて学習を行う
- CRF++条件付き確率場を用いた学習などがある