文献紹介

長岡技術科学大学
松本宏

文献

- Ganitkevitch, J., Van Durme, B., & Callison-Burch, C.
- 2013, June
- PPDB: The Paraphrase Database.
- In HLT-NAACL (pp. 758-764).

概要

- 大規模パラフレーズ・データベースの公開
- ピボット法をつかってパラフレーズの取得
- より類似性を考慮したスコア付け

パラフレーズの構築

- 今までにも多くの構築手法は提案されてきた
  - DIRT: Lin and Pantel, 2001
  - MSR paraphrase corpus: Dolan et al., 2004
- ピボット手法での構築も
  - Bannard and Callison-Burch, 2005
  - Zhou et al. 2006
  - Riezler et al. 2007
- 構築された資源は一般公開されていない

ピボット手法

- Callion-Burchのピボット手法に従う
  - 英語文字列e1, e2、どちらも外国語fに翻訳可能
  - このときe1とe2は同じ意味を成すと考える
  - このように e1 → f → e2 とfを介して <e1, e2>の換言対を得るのでピボット法と呼ばれる
- この手法で多様な換言対の取得が可能
  - ノイズも多く含まれる

SCFG + 統語ラベル

- Synchronous Context Free-Grammarの文法に以下の様な文法がある

$$\mathbf{r} \stackrel{\text{def}}{=} C \rightarrow \langle f, e, \sim, \vec{\varphi} \rangle$$

  - C: 非終端
  - f, e: 終端、非終端記号
  - ~: f-e関係
    - 共通非終端記号をもつ
  - Φ: 素性関数ベクトル
  - r: ルール

重み付け

- (Zhao et al. 2008) に習い
  - 素性関数ベクトルφを線形モデル化し、コスト計算

$$cost(\mathbf{r}) = -\sum_{i=1}^{N} \lambda_i \log \varphi_i.$$

パラフレーズ文法の構築

- パラレルコーパスより翻訳訳文法取得

$$\mathbf{r}_1 \stackrel{\text{def}}{=} C \rightarrow \langle f, e_1, \sim_1, \vec{\varphi}_1 \rangle$$

$$\mathbf{r}_2 \stackrel{\text{def}}{=} C \rightarrow \langle f, e_2, \sim_2, \vec{\varphi}_2 \rangle,$$

- 2つを組み合わせ

$$\mathbf{r}_p \stackrel{\text{def}}{=} C \rightarrow \langle e_1, e_2, \sim_p, \vec{\varphi}_p \rangle,$$

- e1, e2で共通非終端記号をもつ

例

- 提案手法により得られるパラフレーズ例

$$NP \rightarrow \text{ the } NP_1 \text{ of } NNS_2 \mid \text{the } NNS_2 \text{ 's } NP_1$$

Scoring Paraphrases Using Monolingual Distributional Similarity

- 分布類似度を用いたスコア付け
  - 単言語資源完結ではノイズを拾いやすかった
    - Ex) DIRT Lin and Pantel, 2001
  - ピボット手法には有効
    - Chane et al. 2011
  - 素性として線形モデルに組み込む

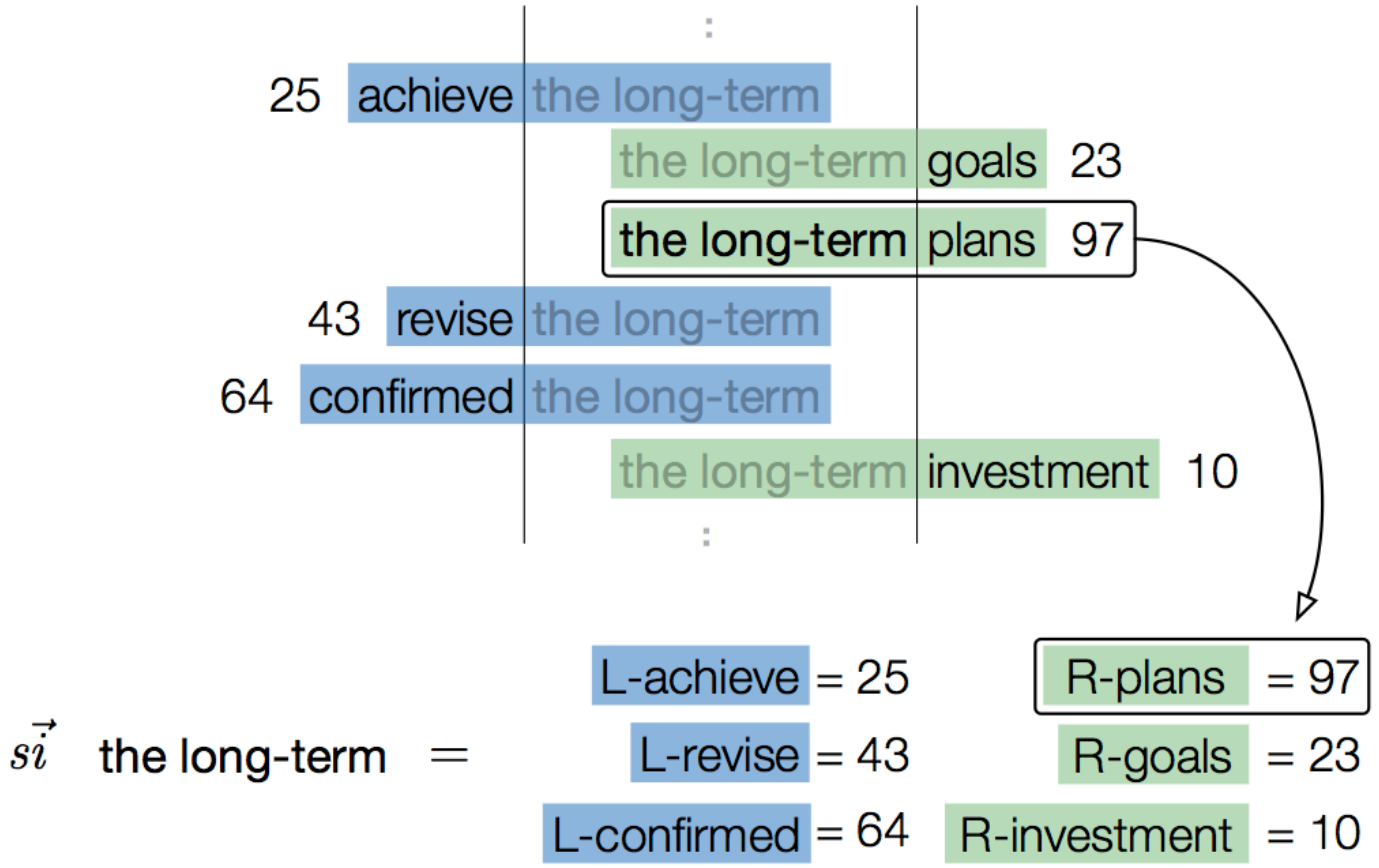
フレーズ・ベクトル化 &
パラフレーズ類似度

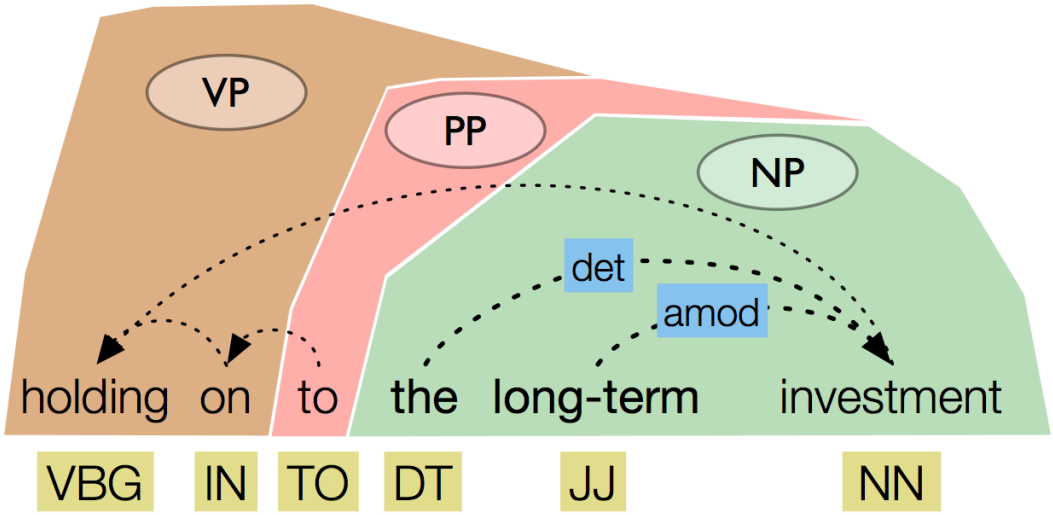
- フレーズのベクトル化
  - フレーズを数値化するため文脈ベクトル化
- 素性:
  - 依存関係、品詞、見出し語などがよく使われる
- 素性@PPDBでは
  - 周辺単語情報を利用

PPDBの素性

- 周辺三単語: ngram corpusより
  - 位置情報をもった語彙
  - 見出し語ベース
  - 品詞
  - 固有表現クラスのユニグラム、バイグラム
- フレーズに対してIN/OUTの依存関係
  - 先/元の語彙, 見出し語, 品詞
- 統語素性
  - フレーズの主要素
  - 組み合わせ範疇文法に似せたラベル

例


:
25 achieve the long-term
the long-term goals 23
the long-term plans 97
43 revise the long-term
64 confirmed the long-term
the long-term investment 10
:
$\vec{si}$ the long-term =
L-achieve = 25
L-revise = 43
L-confirmed = 64
R-plans = 97
R-goals = 23
R-investment = 10



$\vec{si}$ the long-term $=$

lex-R-investment lex-L-on-to

pos-L-IN-TO pos-L-TO lex-L-to

dep-det-R-investment pos-R-NN

dep-amod-R-investment

dep-det-R-NN dep-amod-R-NN

syn-gov-NP syn-miss-L-NN

English Paraphrases – PPDB: Eng

- 利用コーパス:
  - aa
- 内容:
  - *lexical*
    - 単単語
  - *phrasal*
    - 単語列
  - *syntactic* paraphrases
    単語、非終端を含めた表現

調査

- 1,900 述語パラフレーズのサンプル
- 著者らが 1 to 5 の五段階評価