【ICDE2015 勉強会】

Session 11: Strings and Texts (論文2,3)

担当：欅惇志（東工大）

Session 11: Strings and Texts

- 論文紹介
  - Short Text Understanding Through Lexical-Semantic Analysis
    - Wen Hua, Zhongyuan Wang (Renmin University of China), Haixun Wang (Google Research), Kai Zheng, Xiaofang Zhou (The University of Queensland)
    - **Best Paper Award**
    - 既存の NLP ツールでは short text に対しての解析精度が低いため，語と語の “関連度” を測って上手くやる
  - Generating Reading Orders over Document Collections
    - Georgia Koutrika, Lei Liu, Steve Simske (HP Labs)
    - General な文書から specific な文書へ読み進める

※ 図表の一部は論文からコピー

Short Text Understanding Through Lexical-Semantic Analysis

- 目的
  - short text のテキスト分割，品詞付与，コンセプトラベリング
    - NLP のタスクのうち下段の処理
- Short text
  - 例
    - クエリ (5 words 未満)
    - ツイート (140 characters 以下)
  - 文法規則に従っていない
    - 既存の NLP ツールでは分析できない
  - 短い
    - 統計的アプローチを使えない
  - 上記より，さまざまな曖昧性持つ

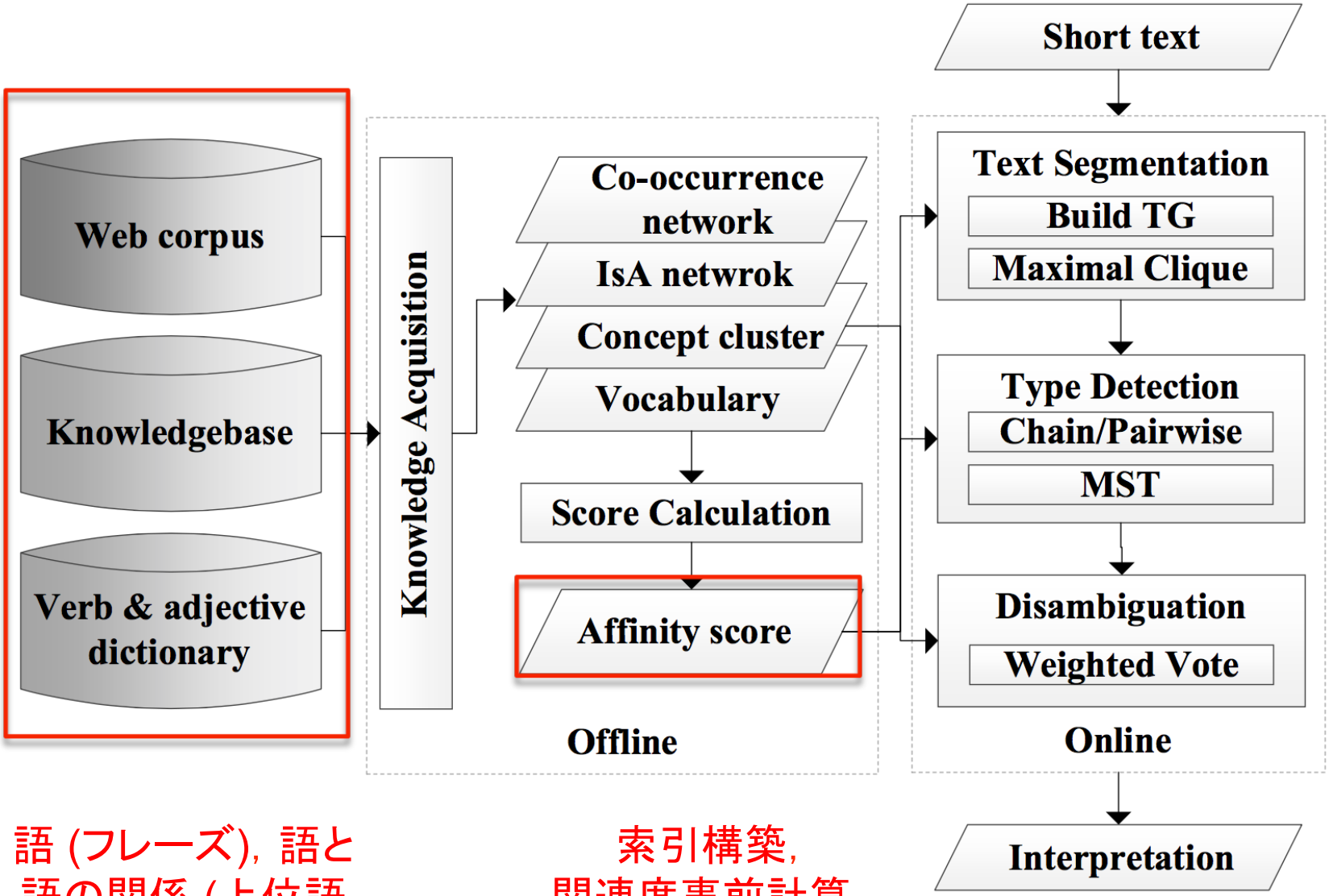
曖昧性の具体例

- テキストの分割点が曖昧
  - “vacation april in paris” VS. “april in paris lyrics”
    - 時期　地名　曲名
  - 辞書語の最長単語を選択するだけでは解決できない
- 品詞が曖昧
  - “watch$_{[v]}$ free movie” VS. “watch$_{[c]}$ omega”
  - 非文 (文法規則に従わない) ため，ルールベース使えない
  - Short text だと統計的機械学習手法も上手くいかない
    - (実質，素性として “ふわっとした” 文法規則なようなものも使うため)
- コンセプトが曖昧
  - “hotel calfornia eagles$_{[e](band)}$” VS. “jaguar$_{[e](brand)}$”
  - 知識がないと判断できない
    - “hotel carfornia” は，animal の eagles より band の eagles と関連が強い

知識 -> short text の理解には必須

提案手法の流れ

- short text
  - “book disneyland hotel calfornia”
- 分割
  - book disneyland hotel calfornia
- 品詞付与
  - $\text{book}_{[v]}$ $\text{disneyland}_{[e]}$ $\text{hotel}_{[c]}$ $\text{california}_{[c]}$
- コンセプトラベル付与
  - $\text{book}_{[v]}$ $\text{disneyland}_{[e](\text{park})}$ $\text{hotel}_{[c]}$ $\text{california}_{[c](\text{state})}$

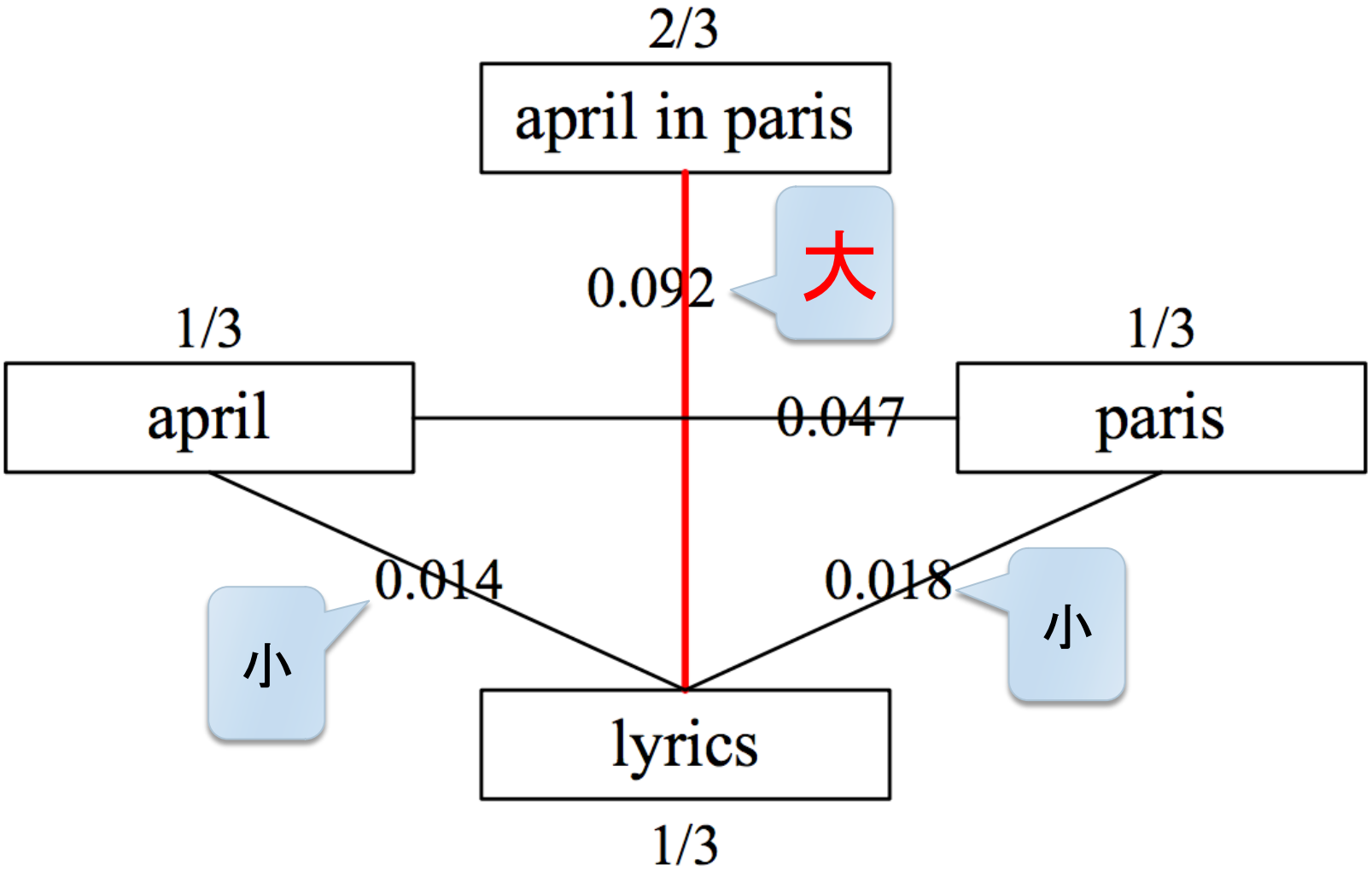


語 (フレーズ), 語と語の関係 (上位語, 下位語 etc…,)

索引構築, 関連度事前計算

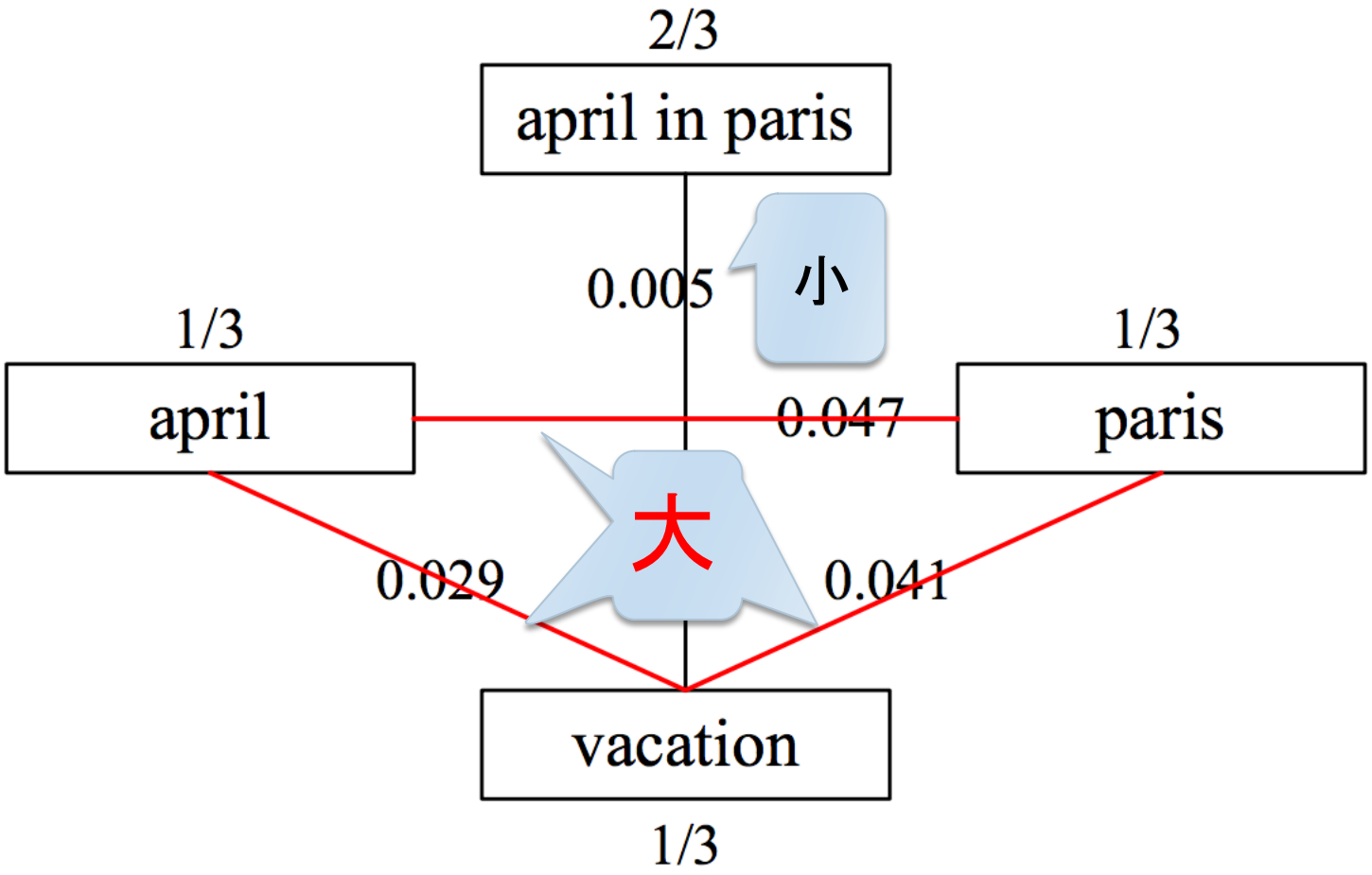
分割方法

- 関連の強い語の組ができるように語を選択する



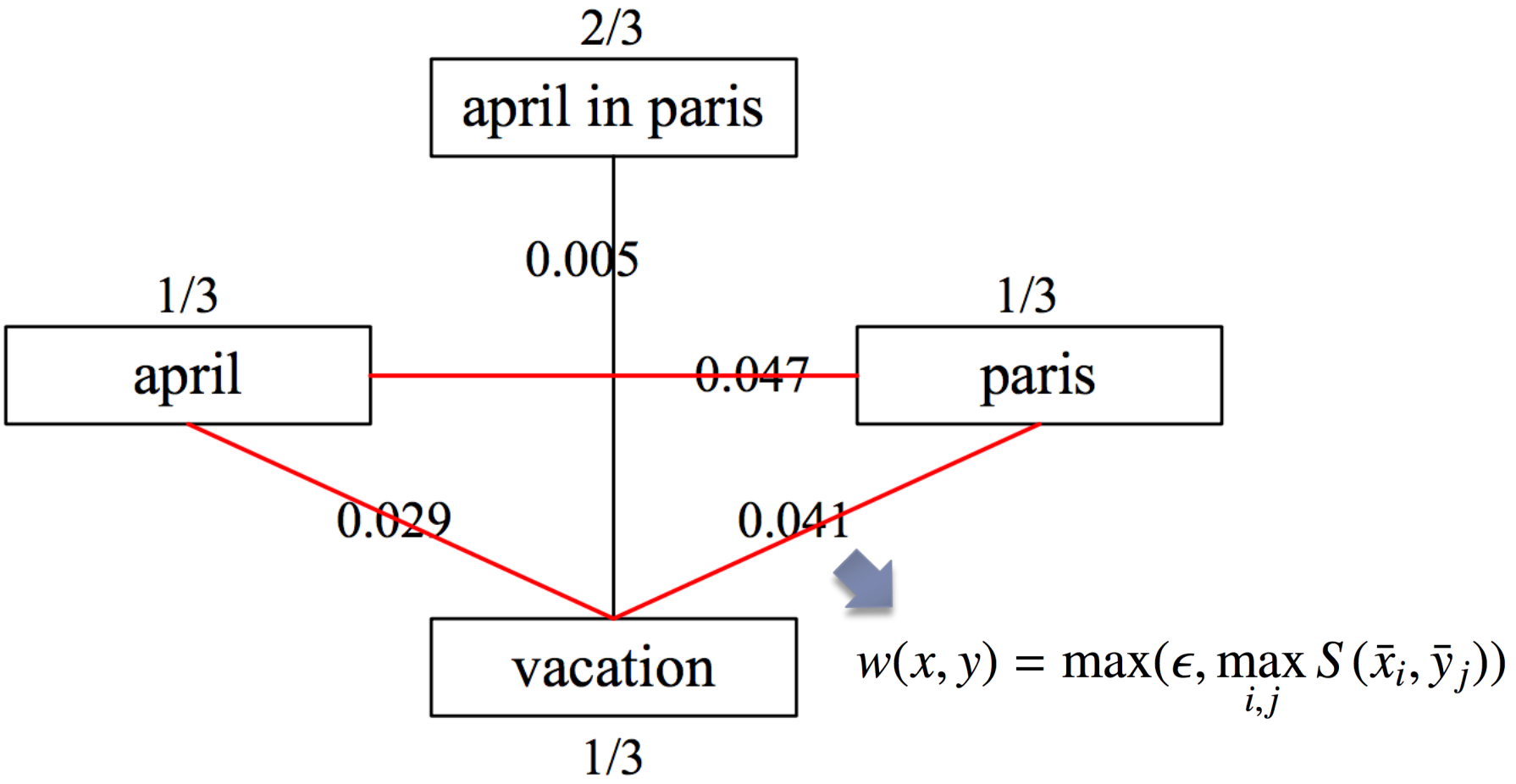
分割方法

- 関連の強い語の組ができるように語を選択する



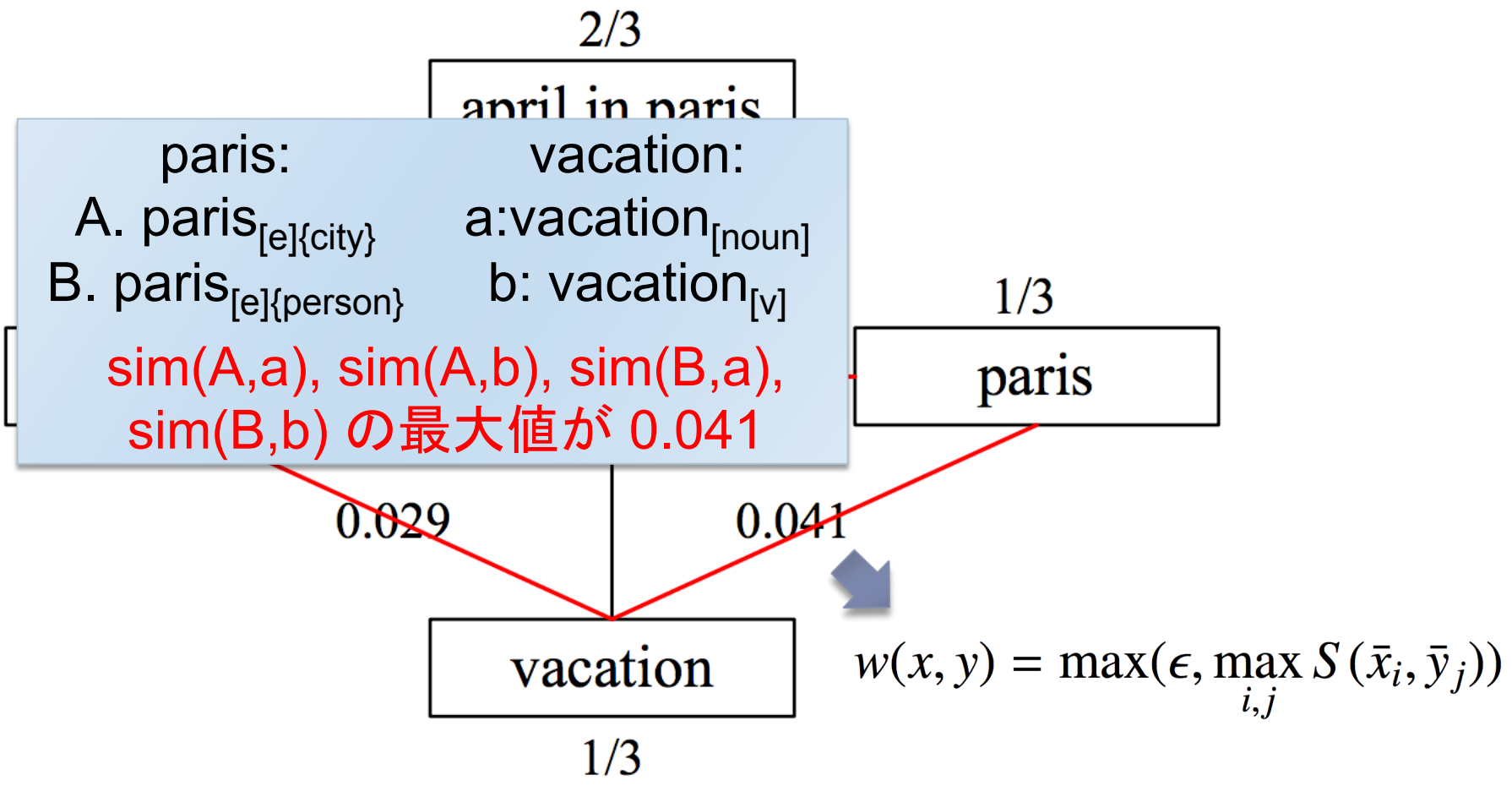
分割方法

- 関連の強い語の組ができるように語を選択する



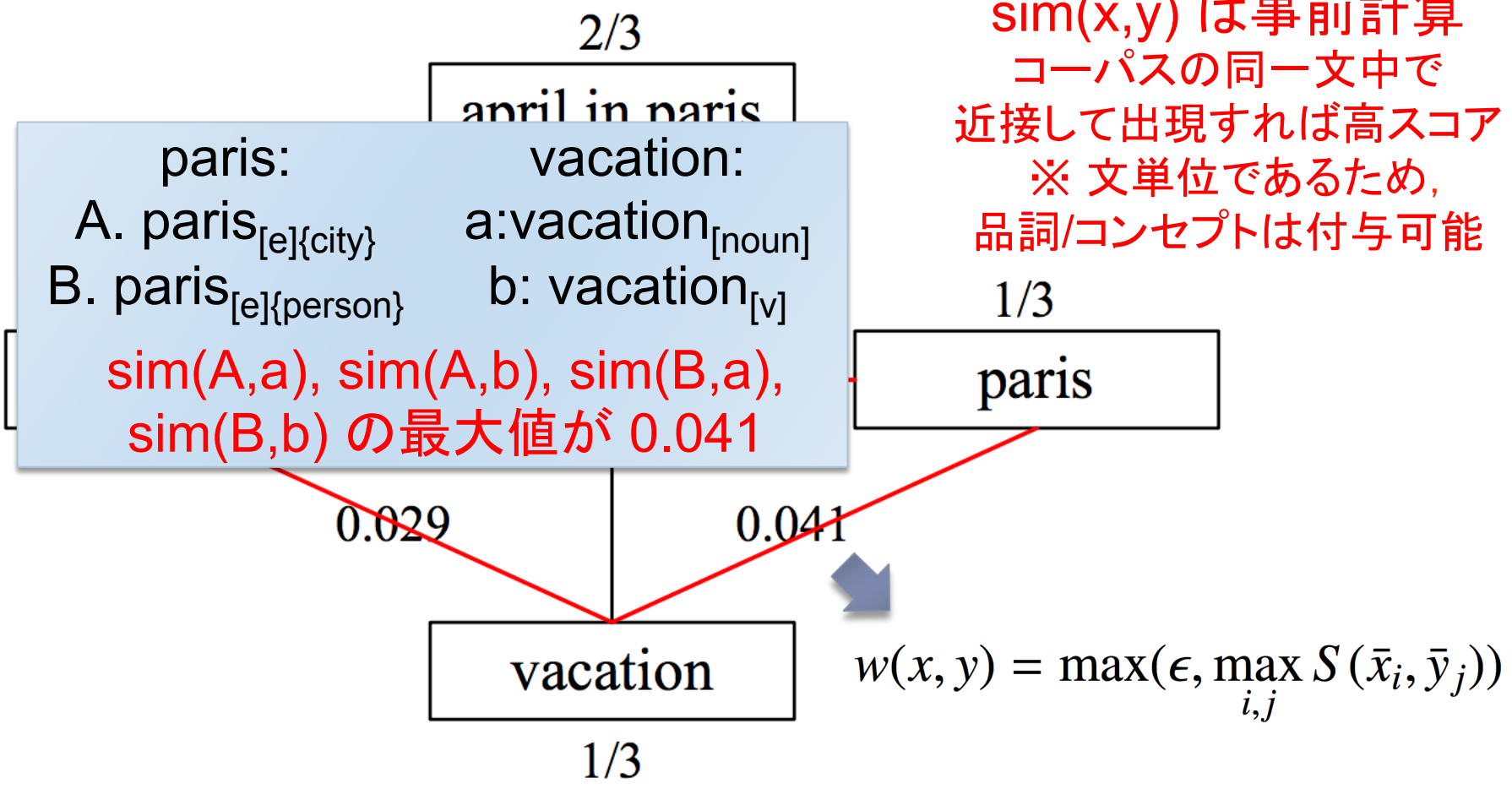
分割方法

- 関連の強い語の組ができるように語を選択する



分割方法

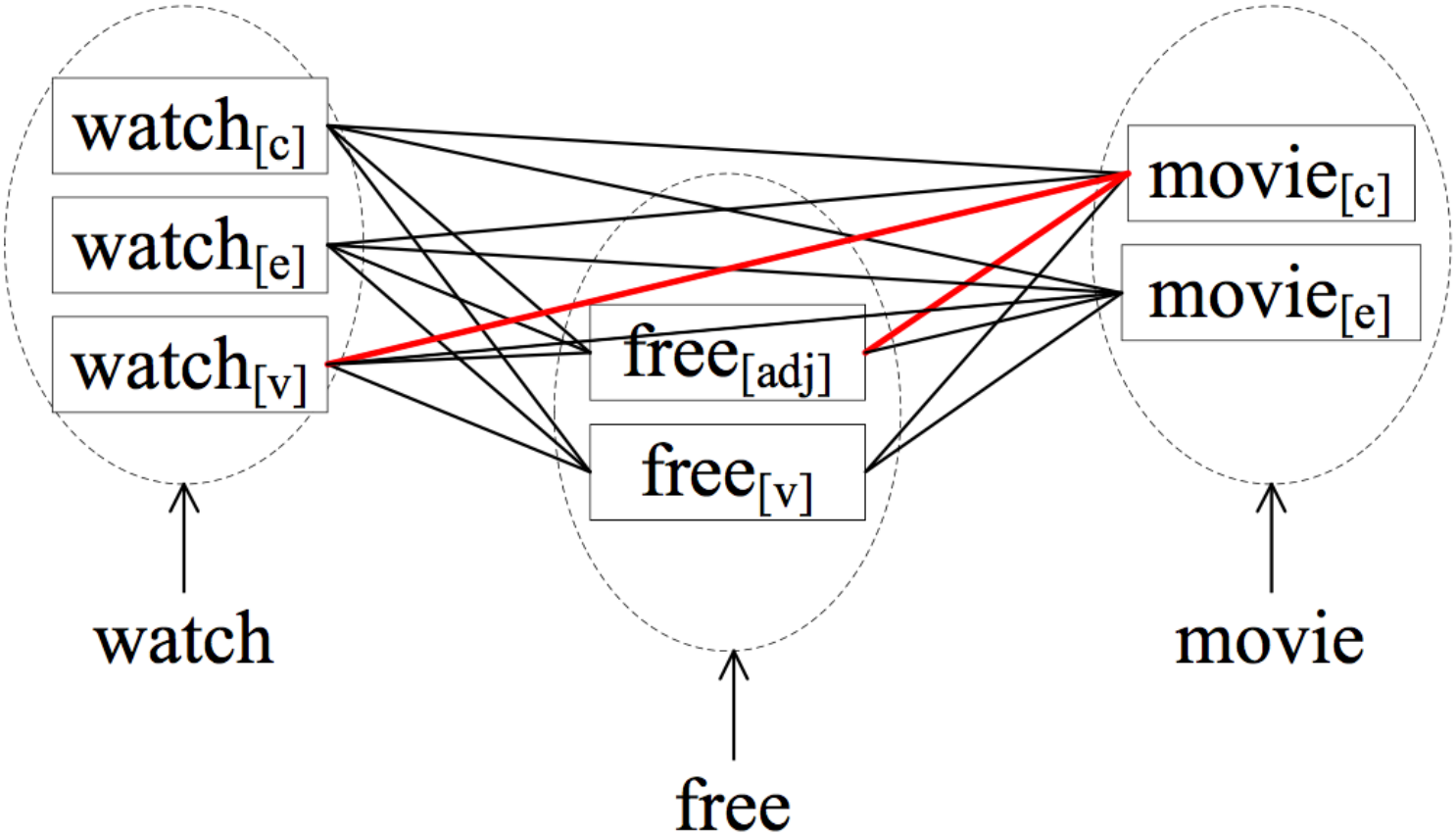
- 関連の強い語の組ができるように語を選択する



sim(x,y) は事前計算
コーパスの同一文中で
近接して出現すれば高スコア
※ 文単位であるため，
品詞/コンセプトは付与可能

品詞付与，コンセプトラベリング

- エッジの重みの総和が最大になる組合せを選択
  - 品詞が複数のコンセプトラベルを持つ場合は，もっとも重みが大きくなるコンセプトを選ぶ



評価実験

- 分割精度

	Longest-Cover	MaxCBF	MaxCMC
accuracy	0.954	**0.984**	0.979

- 品詞付与精度

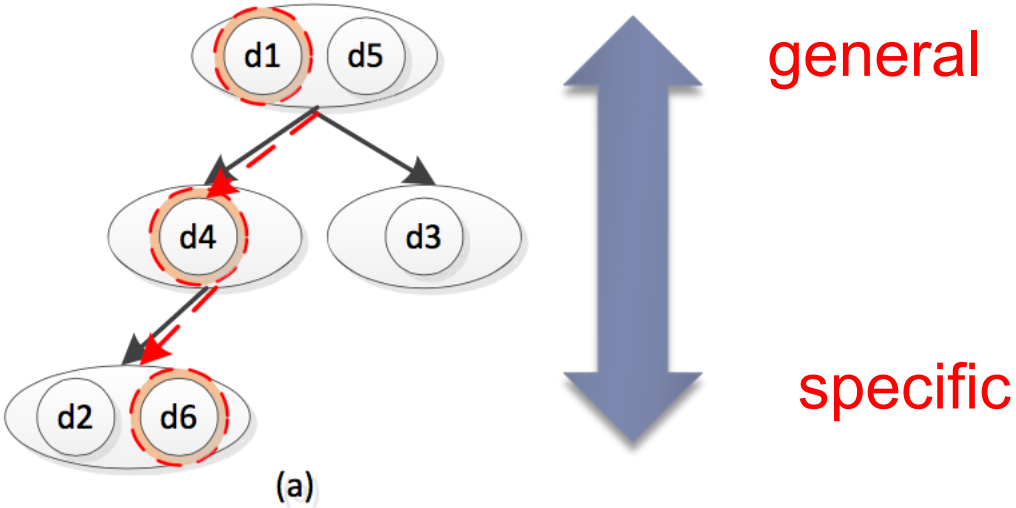
	ST	CM	PM
lexical-level	0.865	0.967	**0.978**
semantic-level	0.944	0.969	**0.973**
term-level	0.932	0.968	**0.974**
query-level	0.876	0.955	**0.967**

- コンセプトラベリング精度

	Song	Kim	Our Approach
term-level	0.694	0.701	**0.943**
query-level	0.525	0.526	**0.890**

Generating Reading Orders over Document Collections

- 一般的な検索結果
  - クエリの適合度の降順に並んだリスト
  - あるページ○○を探したい場合には問題なし
  - 複数のページを何らかの基準で閲覧するのは無理
- 本研究の検索結果
  - General なページが root，specific なページが leaf になるに構築された tree
  - 文書のトピックの粒度が等しければ同じ階層



ルール

- Generality score of document d: $g(d)$
  - さまざまなトピックについて書かれているほど高スコア
    - General 度高
  - トピックモデルで，各文書の各トピックに属する確率を算出
- Document overlap of $d1$ and $d2$: $o(d1, d2)$
  - Jaccard 係数

- $|g(d1) - g(d2)| < t$
  - Equivalent (階層化しない)
- $g(d1) - g(d2) > k$
  - $d2$ を $d1$ の子階層へ
- $g(d2) - g(d1) > k$
  - $d1$ を $d2$ の子階層へ

評価実験

- 人出で ground truth の tree を構築

Overlap の閾値のパラメータ

General 度の閾値のパラメータ

	$\tau = 0.5$	$\tau = 0.7$	$\tau = 0.9$
$\kappa = 0.001$	0.1739	0.1279	0.1558
$\kappa = 0.005$	0.1366	0.1214	0.1625
$\kappa = 0.01$	0.1632	0.1738	0.1456