Multilingual Synchronization focusing on Wikipedia

2011-02-27

Introduction

- Wikipedia: Multilingual encyclopedia
  - Supports over 270 languages
    - English, German, Spanish, French, Chinese, Arabic, …
    - Allows cross-lingual navigation with *inter-language link*
      - Inter-language links: hyperlinks from any page in one Wikipedia language edition to one or more nearly equivalent or exactly equivalent pages in another Wikipedia language editions
  - Different quantity of data on each languages
    - Wikipedia other language editions often suffer from lack of information compared to the English version
      - Multilingual stat on Feb. 2011
        - » English: 3.5 million articles (Most dominant)
        - » French: 1 million articles (3rd)
        - » Korean: 156,290 articles (22nd)

Goal of M-Sync

- Multilingual Synchronization
  - Synchronizing contents of Wikipedia from multiple different languages
    - *Linking* among multiple language contents
    - *Combining* them to synthesis
  - The various Wikipedia editions from different languages
    - can offer more precise and detailed information based on different intentions/backgrounds/cultures
    - can fill the gap between different languages and to acquire the integrated knowledge

Sub-Task (NOW)

- Goal:
  - Finding correlated terms from hypertexts using multilingual topical synthesis

    - Comparison: sum(page length)

Languages	Domain = Disease	Domain = Settlement
English	7,726,724	43,555,917
French	3,761,923	21,270,331
Spanish	3,739,162	15,265,574
Chinese	1,472,109	10,496,202
Korean	842,463	5,813,650
Union	17,542,381	96,401,674

Sub-Task

- Hypothesis
  - X is correlated with Y in L_1 → X' should be correlated with Y' in L_2
    - Where Y' is a corresponding term to Y in different language
      - Assumption
        - » Inter-language links are accurate links to connect two pages about the same entity or concept in different languages
    - Where X' is a translating term to X in different language
  - X is correlated with Y according to its strength using topical synthesis

Outline of Method

- Input
  - Pages in multiple languages
- Output
  - Ordered(ranked) correlated-term sets from page

    or
  - Weighted graph with titles and links as vertices and the co-relatedness between these vertices as edges

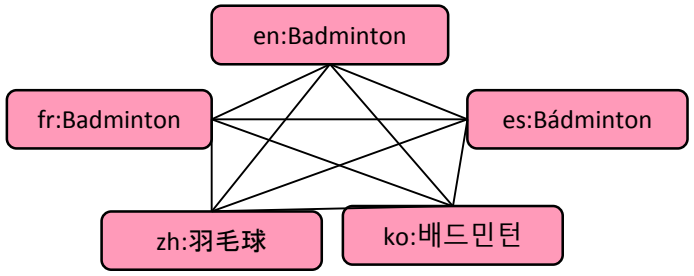
Outline of Method

- Process
  - Selecting target
    - Languages
      - Selecting target languages to used as synchronization sources
        - » English, Spanish, French, Chinese, Korean
      - Selecting target languages to used as synchronization targets
        - » Korean & English
    - Pages
      - 5-clique pages from the above languages
      - Domain: settlement(the largest), disease(neutral)
  - Extracting correlated-terms from hypertexts
    - Extracting links from pages
    - Extracting links from history versions in a temporal manner
  - Translating correlated-terms into target languages
  - Computing weights of co-relatedness using multilingual topic synthesis

Preprocessing: Selecting Target

- Source languages(5)
  - English, Spanish, French, Chinese, Korean

    A subset of UN official languages
- Extracting target pages with a 5-clique by inter-language links
  - Assumption:
    - Pages founded in all 5 languages are key pages and the target to sync
    - Enforcing consistency of a link path
      - If a path from X(L_1) to X'(L_2) founded once,
        its inverse path (X', X) is automatically added to the output



Extracting correlated-terms from hypertexts

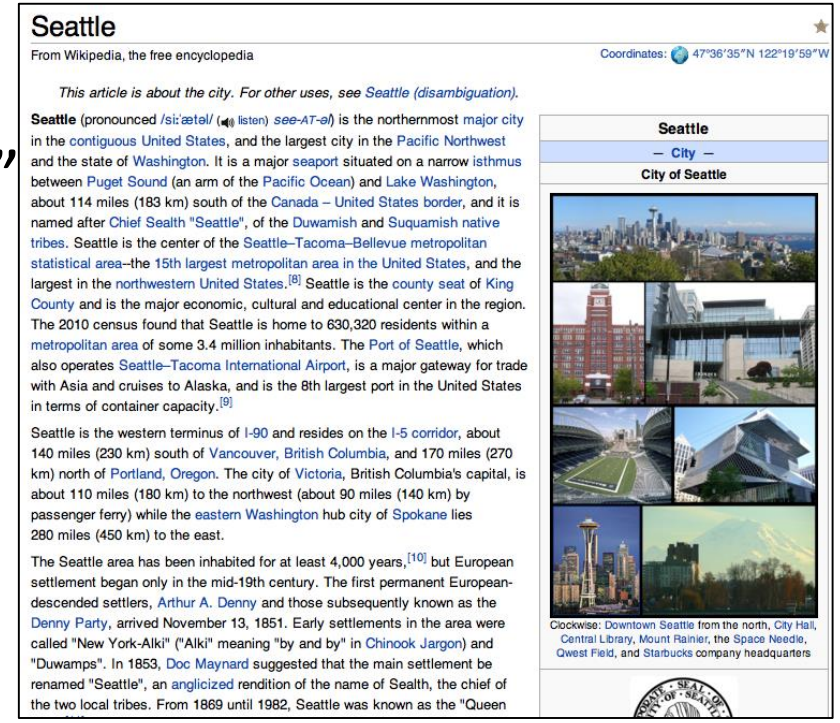
- Hypertext pages
  - Containing links to other pages
    - links
      - navigate to a web page with more detailed information
      - point to previously published web pages with similar or related content
  - Connectivity between pages often proven to play an important role in determining the relevance

Link types of Wikipedia

- **internal links** to other pages in the wiki
  - Syntax usage: [[Main Page]]
- **external links** to other websites
- **interwiki links** to other websites registered to the wiki in advance
  - Unlike internal links, interwiki links do not use page existence detection
  - Syntax usage: [[wikipedia:Sunflower]]
- **Interlanguage links** to other websites registered as other language versions of the wiki

Example of hyperlinks

- Example links(out-going) of ***Seattle***:
  - “northwestern United States”
  - “Washington”
  - “Lake Washington”
  - “Michael McGinn (mayor)”
  - …



Seattle

From Wikipedia, the free encyclopedia

Coordinates: 47°36′35″N 122°19′59″W

This article is about the city. For other uses, see Seattle (disambiguation).

Seattle (pronounced /siːˈætəl/ (listen) *see-AT-əl*) is the northernmost major city in the contiguous United States, and the largest city in the Pacific Northwest and the state of Washington. It is a major seaport situated on a narrow isthmus between Puget Sound (an arm of the Pacific Ocean) and Lake Washington, about 114 miles (183 km) south of the Canada – United States border, and it is named after Chief Sealth "Seattle", of the Duwamish and Suquamish native tribes. Seattle is the center of the Seattle–Tacoma–Bellevue metropolitan statistical area–the 15th largest metropolitan area in the United States, and the largest in the northwestern United States.[8] Seattle is the county seat of King County and is the major economic, cultural and educational center in the region. The 2010 census found that Seattle is home to 630,320 residents within a metropolitan area of some 3.4 million inhabitants. The Port of Seattle, which also operates Seattle–Tacoma International Airport, is a major gateway for trade with Asia and cruises to Alaska, and is the 8th largest port in the United States in terms of container capacity.[9]

Seattle is the western terminus of I-90 and resides on the I-5 corridor, about 140 miles (230 km) south of Vancouver, British Columbia, and 170 miles (270 km) north of Portland, Oregon. The city of Victoria, British Columbia's capital, is about 110 miles (180 km) to the northwest (about 90 miles (140 km) by passenger ferry) while the eastern Washington hub city of Spokane lies 280 miles (450 km) to the east.

The Seattle area has been inhabited for at least 4,000 years,[10] but European settlement began only in the mid-19th century. The first permanent European-descended settlers, Arthur A. Denny and those subsequently known as the Denny Party, arrived November 13, 1851. Early settlements in the area were called "New York-Alki" ("Alki" meaning "by and by" in Chinook Jargon) and "Duwamps". In 1853, Doc Maynard suggested that the main settlement be renamed "Seattle", an anglicized rendition of the name of Sealth, the chief of the two local tribes. From 1869 until 1982, Seattle was known as the "Queen

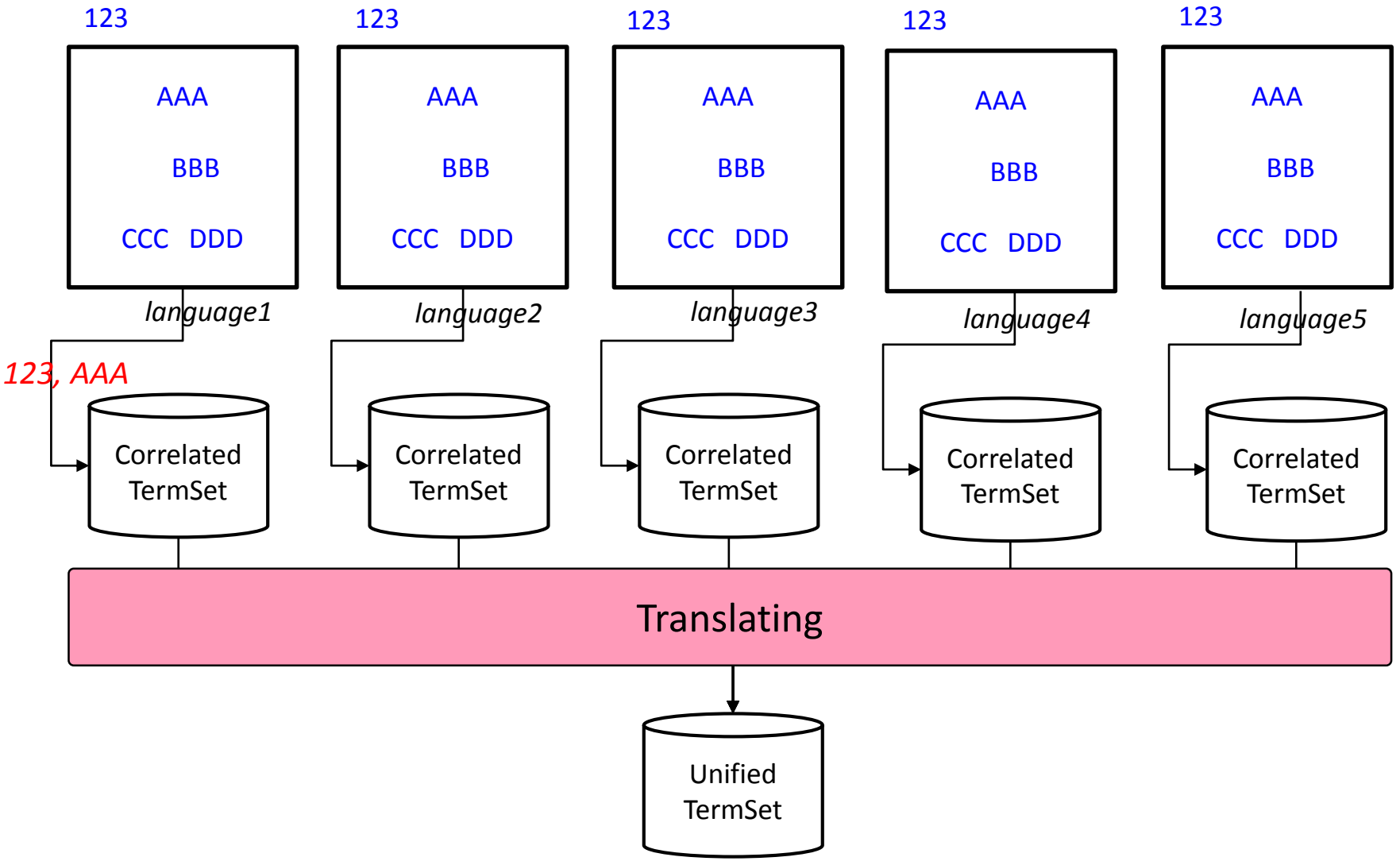
Seattle

— City —

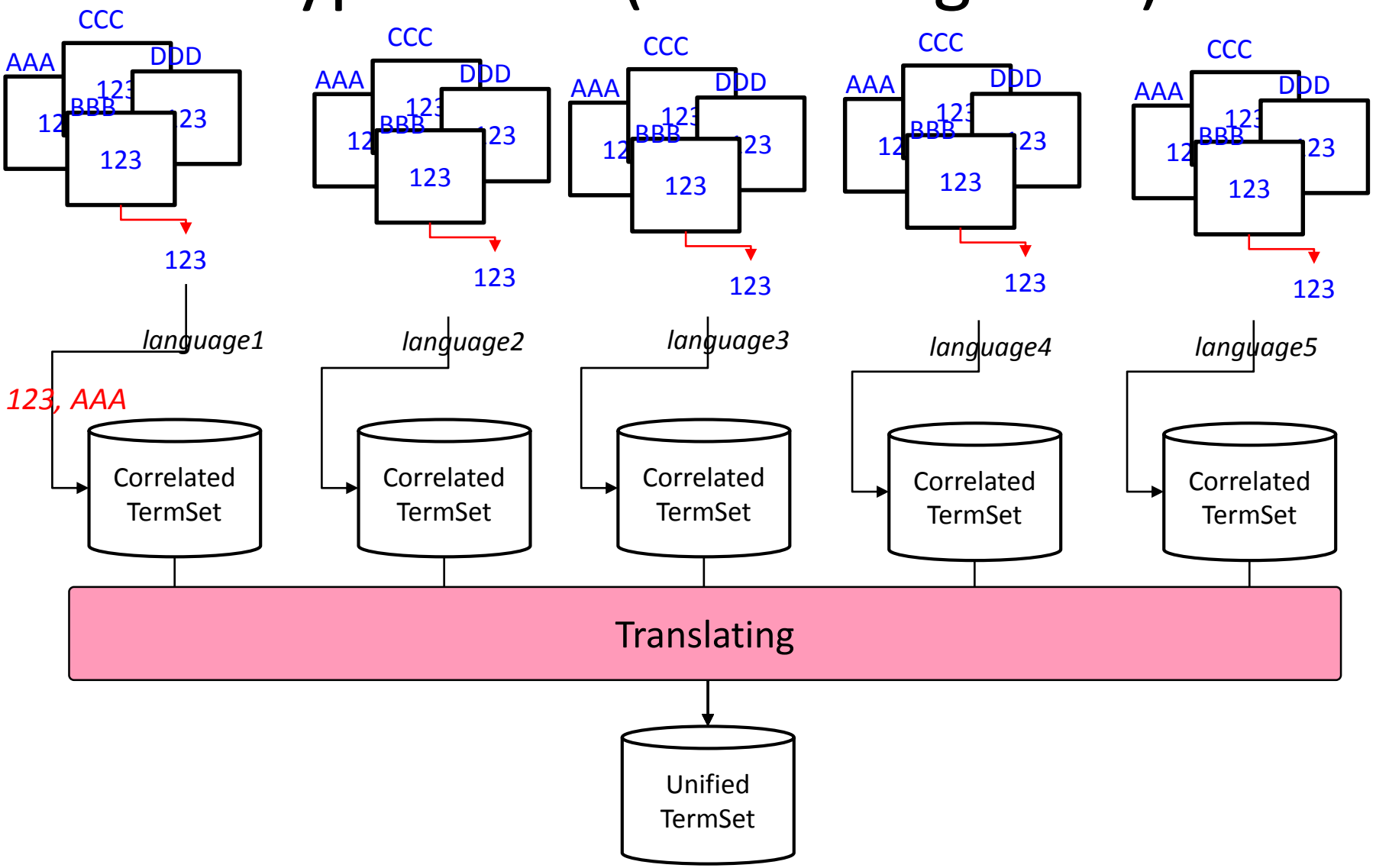
City of Seattle

Clockwise: Downtown Seattle from the north, City Hall, Central Library, Mount Rainier, the Space Needle, Qwest Field, and Starbucks company headquarters

Extracting correlated-terms from hypertexts (out-going links)



Extracting correlated-terms from hypertexts (in-coming links)



Translating terms

- Method
  - Wikipedia dictionary-based
    - We have collected the cross-lingual term pars to build bilingual word pairs
      - 4 dictionaries are available: EN-KO, ES-KO, FR-KO, ZH-KO
    - Weakness
      - Lack of vocabularies
  - Google translation API-base
    - Weakness
      - Terms(keywords) are too short to solve the word sense disambiguation using MT

Computing weights of co-relatedness using multilingual topic synthesis

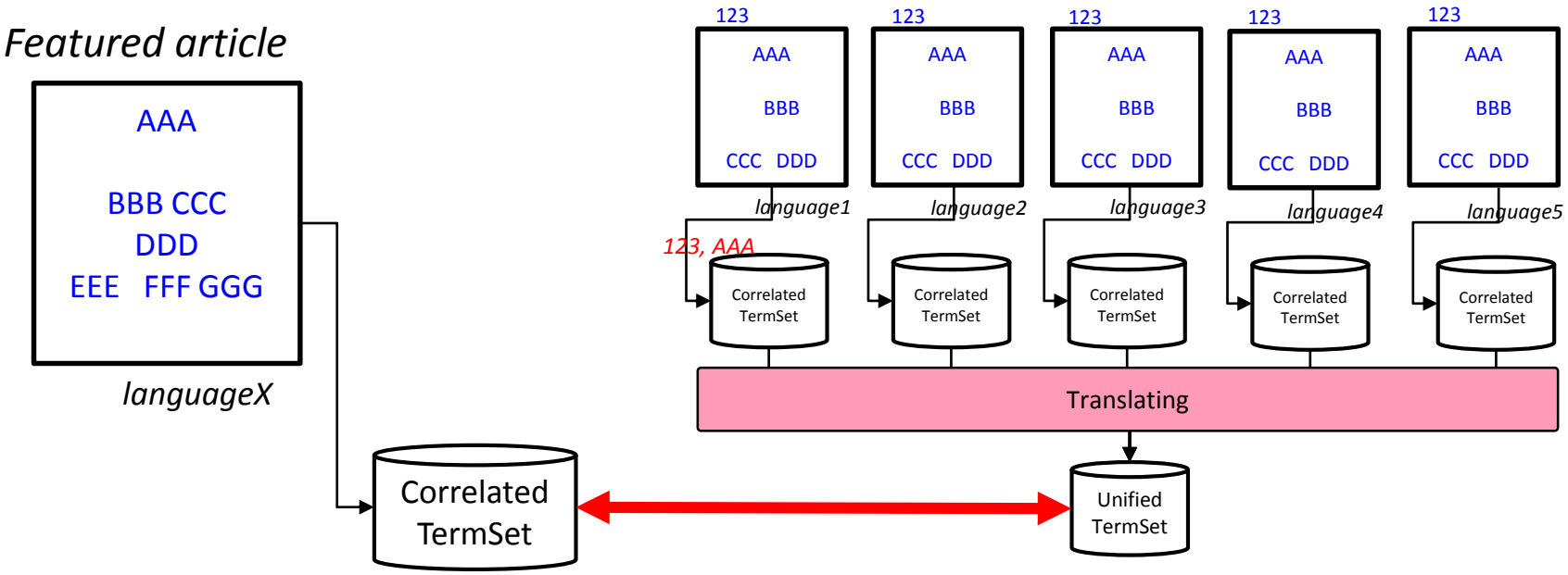
- Weight of correlations
  - Baseline: frequency-based method
  - However, different Wikipedia has different viewpoints and concerns
  - We should give different weight of synthesized correlated term sets according to different lingual usages and frequencies
    - Our proposed solution:
      - analyzing the topical distributions on each languages and
      - Computing weights of correlated terms by each topics' interest
      - Approach
        - » LDA-based topic distribution using links
        - » Align cross-lingual topic clusters
          - Intersection: common topic
          - Difference: unique topic

Evaluation

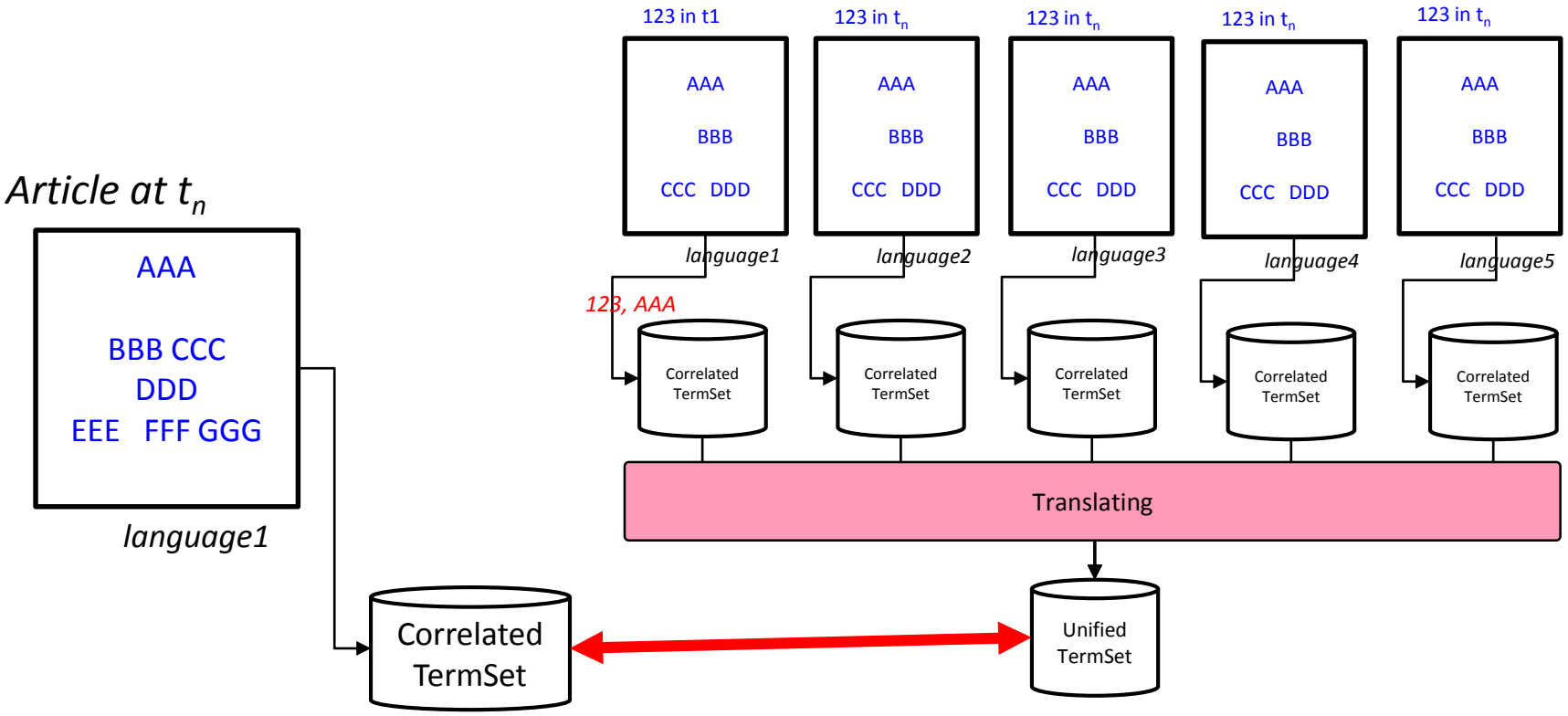
- What
  - Comparison with
    - Discovered new correlated terms without topical synthesis
      - Simple union approach
    - Discovered new correlated terms with topical synthesis
      - Ranked union using calculating the strength of relatedness
- How
  - Public measures
    - Co-occurrence
    - Mutual information
    - Normalized Google distance
  - Wikipedia oriented measures
    - Comparison with the featured articles
    - Comparison with the temporal manner

Comparison with featured articles

- Featured articles:
  - are considered to be the best articles in Wikipedia, as determined by Wikipedia's editors



Comparison with temporal manner



Contributions

- To support the seed data (seed keywords) to complete articles in a multilingual manner, or to guide users in generating new articles in Wikipedia
- To find unknown correlated words using various sources