Efficient Web-Based Linkage of Short to Long Forms

Yee Fan Tan¹, Ergin Elmacioglu², Min-Yen Kan¹, and Dongwon Lee²

¹National University of Singapore, Singapore
²The Pennsylvania State University, USA

{tanyeefa,kanmy}@comp.nus.edu.sg
{ergin,dongwon}@psu.edu
Linkage of Short to Long Forms

For each short form, what long forms does it correspond to?

We use a search engine to perform linkage

Short forms (SF)

ACSAC
KDD
KDID
UbiComp
WebDB

Long forms (LF)

Annual Computer Security Applications Conference
Asia-Pacific Computer Systems Architecture Conference
Knowledge Discovery and Data Mining
Knowledge Discovery in Inductive Databases
Ubiquitous Computing
International Workshop on Web and Databases
Framework: Web-Based (Record) Linkage

Design of queries
1. “sf ∧ lf”
   - e.g., Oh and Isahara (2008)
   - not scalable

2. “sf” or “lf”, or both
   - e.g., Cimiano et al. (2005)
   - linear number of queries

Search engine evidence
- Each call returns 10 results
- Contains page title, snippet, URL
- Can also download web pages
- How to postprocess?

Our focus!

We only consider “sf” or “lf” queries

- For each short form sf
  - For each long form lf
    - Obtain information for sf and lf using search engine
    - Compute $score_{sf}(lf)$ using obtained information
    - Rank the long forms according to $score_{sf}(lf)$

×

√
Related Work: Short/Long Form Extraction

- Extraction of short and long form pairs from full text articles e.g., Schwartz and Hearst (2003)

Scenario template creation (STC) is the problem of generating a common semantic representation from a set of input articles. [...] Here, we use the term salient aspect (SA) to refer to any of such slot fillers that people would regard as important to describe a particular scenario. Figure 1 shows such a manually-built scenario template in which details about important actions, actors, time and locations are coded as slots.

STC is an important task that has tangible benefits for many downstream applications. In the Message Understanding Conference (MUC), manually-generated STs were provided to guide Information Extraction (IE). An ST can also be viewed as regularizing a set of similar articles as a set of attribute/value tuples, enabling multi-document summarization from filled templates.
Contributions

• Unify related threads of research in framework

• Method
  • First to exploit the Web for linking short form to long forms
  • Proposed an effective count-based method for this task

• Efficiency
  • Adaptively combine a query probing approach with our count-based method

• Evaluation
  • Three datasets on three different domains

Done!
Our Count-based Methods

1. \( \text{count}(\text{sf} \rightarrow \text{lf}) \)

Example: \( \text{sf} = \) “HGP”, \( \text{lf} = \) “Human Genome Project”

Human Genome Project - Wikipedia, the free encyclopedia

More on the sequencing of the human genome
The international Human Genome Project (HGP) and Celera Genomics ... Uses of the HGP genome assemblies in Celera genome assemblies and impact on assembly. ...

More on the sequencing of the human genome -- Waterston et al. 100 ...
Approximately 60% of the underlying sequence data and 100% of the mapping data used in Celera’s analysis came from the HGP, and the HGP genome assembly ...

2. \( \text{count}(\text{sf} \leftarrow \text{lf}) \)

• Interchange the roles of \( \text{sf} \) and \( \text{lf} \) in \( \text{count}(\text{sf} \rightarrow \text{lf}) \)

3. \( \text{count}(\text{sf} \leftrightarrow \text{lf}) \)

• Sum of \( \text{count}(\text{sf} \rightarrow \text{lf}) \) and \( \text{count}(\text{sf} \leftarrow \text{lf}) \)

\( \text{count}(\text{sf} \rightarrow \text{lf}) = 2 \)
Comparison with Other Methods

**DBLP Computer Science Conferences and Workshops**
- KDD: Knowledge Discovery and Data Mining
- KDID: Knowledge Discovery in Inductive Databases
- UbiComp: Ubiquitous Computing
- WebDB: International Workshop on Web and Databases

**NASDAQ Composite**
- AAPL: Apple Inc.
- CSCO: Cisco Systems, Inc.
- DELL: Dell Inc.
- MSFT: Microsoft Corporation

**Human Genome Acronym List**
- MALDI: matrix-assisted laser desorption ionization
- Mb: megabase
- MHC: major histocompatibility complex
- MIT: Massachusetts Institute of Technology

**Schwartz and Hearst (2003)**
- Non-web based method that scans full text articles for “lf (sf)” patterns

**Inverse Host Frequency (Tan et al., 2006)**
- Similarity between sf and lf vectors of URL hostnames

**Sahami and Heilman (2006)**
- Variation of tf-idf cosine similarity on snippets or web pages

Search engine: Google
Y. F. Tan, E. Elmacioglu, M.-Y. Kan, and D. Lee - Efficient Web-Based Linkage of Short to Long Forms

### Results (10 snippets)

<table>
<thead>
<tr>
<th>Method</th>
<th>DBLP Recall</th>
<th>Ranked Precision</th>
<th>NASDAQ Recall</th>
<th>Ranked Precision</th>
<th>GENOMES Recall</th>
<th>Ranked Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>count(sf ← lf)</td>
<td>0.942</td>
<td>0.965</td>
<td>0.999</td>
<td>0.992</td>
<td>0.954</td>
<td>0.848</td>
</tr>
<tr>
<td>count(sf → lf)</td>
<td>0.627</td>
<td>0.961</td>
<td>0.587</td>
<td>0.603</td>
<td>0.553</td>
<td>0.420</td>
</tr>
<tr>
<td>count(sf ↔ lf)</td>
<td>0.918</td>
<td>0.934</td>
<td>1.000</td>
<td>0.999</td>
<td>0.905</td>
<td>0.762</td>
</tr>
<tr>
<td>IHF</td>
<td>0.686</td>
<td>0.532</td>
<td>0.240</td>
<td>0.226</td>
<td>0.449</td>
<td>0.426</td>
</tr>
<tr>
<td>Schwartz &amp; Hearst</td>
<td>0.603</td>
<td>0.244</td>
<td>0.240</td>
<td>0.226</td>
<td>0.449</td>
<td>0.426</td>
</tr>
<tr>
<td>Sahami &amp; Heilman</td>
<td>0.868</td>
<td>0.532</td>
<td>0.240</td>
<td>0.226</td>
<td>0.449</td>
<td>0.426</td>
</tr>
</tbody>
</table>

- `count(sf ↔ lf)` consistently produced best results!
- `count(sf ← lf)` much better than `count(sf → lf)`
- `count(sf ↔ lf)` is slightly better than Sahami and Heilman
  - `count(sf ↔ lf)` is simpler and faster for same number of queries
- Schwartz and Hearst, and IHF are not as competitive
Adaptive Combination with Query Probing

- How to reduce the number of time-consuming search engine queries?

Adaptive combination
- For each short form
  - Let faster but weaker method ($M_w$) rank the long forms
  - If heuristic determines that $M_w$ gives a poor ranking
    - Have stronger but slower method ($M_s$) rank the long forms

- $M_w$: Query probing
- $M_s$: Count-based method

Query probing
- Joint Conference on Digital Libraries
- European Conference on Digital Libraries
- Digital Libraries

Query probing
- Query frequently occurring $n$-grams from long form list
  - $score_p(sf, lf) =$ number of results containing both $sf$ and $lf$

Heuristic
- Any $lf$ with $score_p(sf, lf) > 0$?
Results

DBLP
Recall: -0.033 to +0.016
Ranked precision: -0.066 to +0.021

NASDAQ
Recall: -0.023 to -0.004
Ranked precision: -0.044 to -0.014

GENOMES
Recall: -0.036 to +0.009
Ranked precision: -0.059 to +0.004

Reduced number of search engine queries while maintaining linkage performance!
Conclusion

- **Efficient Web-Based Linkage of Short to Long Forms**

- **Effective**
  - Count-based method

- **Efficient**
  - Adaptive combination with query probing
  - Reduce the number of search engine queries needed

- **Evaluation**
  - Shown effective on three datasets of different domains
Thank You