Domain Cartridge: Unsupervised Framework for Shallow Domain Ontology Construction from Corpus

Subhabrata Mukherjee
Jitendra Ajmera, Sachindra Joshi

Max Planck Institute for Informatics
IBM India Research Lab

CIKM 2014

November 17, 2014
Motivation: Domain Term Discovery

Usefulness for Parsing. Consider the examples:

- “use sprint zone”
  - Parse w/o domain knowledge — use/noun sprint/verb zone/noun
  - Parse with domain knowledge — use/verb {sprint zone}/noun

- “transfer files via usb cable”

Parser generates noisy or incomplete parse without the domain knowledge

- ‘sprint’ and files’ are not verbs
- “sprint zone, usb cable” are multi-word concepts
Motivation: Domain Term Discovery

Usefulness for Parsing. Consider the examples:

- “use sprint zone”
  - Parse w/o domain knowledge — use/noun sprint/verb zone/noun
  - Parse with domain knowledge — use/verb {sprint zone}/noun

- “transfer files via usb cable”

Parser generates noisy or incomplete parse without the domain knowledge

- ‘sprint’ and files’ are not verbs
- “sprint zone, usb cable” are multi-word concepts
Motivation: Domain Term Discovery

Usefulness for Parsing. Consider the examples:

- “use sprint zone”
  - Parse w/o domain knowledge — use/noun sprint/verb zone/noun
  - Parse with domain knowledge — use/verb {sprint zone}/noun

- “transfer files via usb cable”

Parser generates noisy or incomplete parse without the domain knowledge

- ‘sprint’ and files’ are not verbs
- “sprint zone, usb cable” are multi-word concepts
Motivation: Domain Relation Discovery

- Interactive dialogue systems
  - For user query “battery of my device depletes fast”, the knowledge ‘battery’ is a Feature-Of ‘device’ enables system to clarify about Type-Of device

- Query expansion
  - E.g. Consider Synonyms along with original query, ‘battery’ is a Feature-Of ‘phone’ as well as ‘tablet’ ‘device’

- Query re-formulation
  - For user query “screen freezes E5150”, the knowledge ‘E5150’ is a Type-Of ‘Error’ results in query re-formulation “screen freezes error E5150"
Motivation: Domain Relation Discovery

- Interactive dialogue systems
  - For user query “battery of my device depletes fast”, the knowledge ‘battery’ is a Feature-Of ‘device’ enables system to clarify about Type-Of device

- Query expansion
  - E.g. Consider Synonyms along with original query, ‘battery’ is a Feature-Of ‘phone’ as well as ‘tablet’ ‘device’

- Query re-formulation
  - For user query “screen freezes E5150”, the knowledge ‘E5150’ is a Type-Of ‘Error’ results in query re-formulation “screen freezes error E5150"
Motivation: Domain Relation Discovery

- Interactive dialogue systems
  - For user query “battery of my device depletes fast”, the knowledge ‘battery’ is a Feature-Of ‘device’ enables system to clarify about Type-Of device

- Query expansion
  - E.g. Consider Synonyms along with original query, ‘battery’ is a Feature-Of ‘phone’ as well as ‘tablet’ ‘device’

- Query re-formulation
  - For user query “screen freezes E5150”, the knowledge ‘E5150’ is a Type-Of ‘Error’ results in query re-formulation “screen freezes error E5150"
Unsupervised Framework

- Typically for a domain, a lot of knowledge articles, manuals, tutorials etc. are available in a variety of formats.

- Most of these documents have less hyperlink and table (info-box as in Wikipedia) information, or extraction is difficult (E.g. pdf).

- Challenge is to learn a shallow ontology from raw unannotated plain text.
Unsupervised Framework

- Typically for a domain, a lot of knowledge articles, manuals, tutorials etc. are available in a variety of formats.

- Most of these documents have less hyperlink and table (info-box as in Wikipedia) information, or extraction is difficult (E.g. pdf).

- Challenge is to learn a shallow ontology from raw unannotated plain text.
Unsupervised Framework

- Typically for a domain, a lot of knowledge articles, manuals, tutorials etc. are available in a variety of formats

- Most of these documents have less hyperlink and table (info-box as in Wikipedia) information, or extraction is difficult (E.g. pdf)

- Challenge is to learn a shallow ontology from raw unannotated plain text
Domain Cartridge as a Graph
Domain Cartridge as a Graph

device

handset

android

Operating system

sim

insert

card

Samsung

Samsung Galaxy victory

Samsung array
Domain Cartridge as a Graph

- handset
- android
- Operating system
- sim
- insert
- card
- Samsung Galaxy victory
- Samsung array
- device
- blackberry
- install
- samsung
- Domain term
- Domain process
- Action-On

Diagram showing relationships and actions involving domain terms.
Roadmap

- Unsupervised framework for shallow domain ontology construction:
  - Domain Term Discovery (DTD)
  - Improvement of Parser performance by DTD
  - Domain Relation Discovery (DRD)

- Use-Case: Improvement of an in-house Question-Answering system


- Conclusions
Roadmap

- Unsupervised framework for shallow domain ontology construction:
  - Domain Term Discovery (DTD)
  - Improvement of Parser performance by DTD
  - Domain Relation Discovery (DRD)

- Use-Case: Improvement of an in-house Question-Answering system


- Conclusions
Roadmap

- Unsupervised framework for shallow domain ontology construction:
  - Domain Term Discovery (DTD)
  - Improvement of Parser performance by DTD
  - Domain Relation Discovery (DRD)

- Use-Case: Improvement of an in-house Question-Answering system


- Conclusions
Roadmap

- Unsupervised framework for shallow domain ontology construction:
  - Domain Term Discovery (DTD)
  - Improvement of Parser performance by DTD
  - Domain Relation Discovery (DRD)

- Use-Case: Improvement of an in-house Question-Answering system


- Conclusions
Corpus: Knowledge articles, manuals, tutorials etc.

Domain Cartridge: Framework
Parsing

Domain Cartridge: Framework
Parsing

“Turn the wi-fi radio on or off”

English Slot Grammar (ESG) parser used. 50 - 100 times faster than Charniak parser
Prismatic Relations

Shallow semantic relationship (SSR) annotation over ESG parser output generates normalized parser relation

E.g., “Samsung has a battery” and “Samsung’s battery died” both generate the same relation ‘nnMod:samsung_battery’
Prismatic Relations

Shallow semantic relationship (SSR) annotation over ESG parser output generates normalized parser relation

E.g., “Samsung has a battery” and “Samsung’s battery died” both generate the same relation ‘nnMod:samsung_battery’
Lucene Index – For efficient retrieval of relations, documents, positional information, proximity based queries etc.
Domain Cartridge: Framework

Diagram:
- Corpus
- ESG Parser
- Prismatic Relations
- Random Index
- Secondary Index
- Primary Index
- Domain Terms
- HITS
- Synonym Extractor
- Type-Of Extractor
- Feature-Of Extractor
- Action-On Extractor
- Domain Ontology
Domain Term Discovery

ESG parser maintains a domain term lexicon of multi-word concepts. E.g. “touch screen, sprint navigation”

Noun Phrase Chunking on *document titles* to extract frequently occurring concepts as domain words

![Diagram of noun phrase chunking](https://via.placeholder.com/150)
Domain Term Discovery

ESG parser maintains a domain term lexicon of multi-word concepts. E.g. “touch screen, sprint navigation”

Noun Phrase Chunking on document titles to extract frequently occurring concepts as domain words

Turn

obj

comp

or

radio

nadj

nadj

the

wi-fi

lconj

rconj

on

off

the wi-fi radio
Domain Term Discovery

- Enrich lexicon and bootstrap parser
- Parser generates refined output

High precision but low recall — as titles are precise, clean but short

To extract more fine-grained domain terms HITS is used on parser output
Domain Term Discovery

- Enrich lexicon and bootstrap parser
- Parser generates refined output

High precision but low recall — as titles are precise, clean but short

To extract more fine-grained domain terms HITS is used on parser output
HITS

- Any Shallow Semantic Relation (SSR) from ESG parser is a *hub* generating domain terms

- Any domain term is an *authority* influenced by incoming features from hubs

- Good authorities incorporated in Parser Domain Term Lexicon

- Parser is re-run, refined relations generated, and previous steps iterated until convergence
HITS

▶ Any Shallow Semantic Relation (SSR) from ESG parser is a *hub* generating domain terms

▶ Any domain term is an *authority* influenced by incoming features from hubs

▶ Good authorities incorporated in Parser Domain Term Lexicon

▶ Parser is re-run, refined relations generated, and previous steps iterated until convergence
HITS

- Any Shallow Semantic Relation (SSR) from ESG parser is a hub generating domain terms

- Any domain term is an authority influenced by incoming features from hubs

- Good authorities incorporated in Parser Domain Term Lexicon

- Parser is re-run, refined relations generated, and previous steps iterated until convergence
HITS

- Any Shallow Semantic Relation (SSR) from ESG parser is a *hub* generating domain terms
- Any domain term is an *authority* influenced by incoming features from hubs
- Good authorities incorporated in Parser Domain Term Lexicon
- Parser is re-run, refined relations generated, and previous steps iterated until convergence
Domain Adaptation for IE and IR

Domain Term Discovery

Domain Relation Discovery

Experiments

Hubs

rel:svo:phone_offer_subscription

rel:nnMod:phone_screen

rel:npo:phone_on_battery

rel:dm_obj:use_phone

rel:dm_obj:charge_phone

rel:dm_obj:charge_battery

Authorities

cpt:verb:offer

cpt:noun:phone

cpt:noun:iPhone

cpt:noun:battery

cpt:verb:charge
Feedback

Domain Cartridge: Framework
Parser Performance Improvement

Number of incomplete parses went down by 73% after incorporating domain terms in the parser lexicon
Domain Terms

software-version htc-evo wi-fi memory-card microsoft-exchange lg-optimus samsung-m400 samsung-galaxy-victory software-updates samsung-array text-messaging touch-screen blackberry-bold

Table: Snapshot of multi-word domain terms extracted by NP Chunking.

optimus-g set-up novatel-wireless e-mail sierra-wireless apple-id google-maps play-music mobile-network 10-digit internet-explorer slacker-radio caller-id google-search address-book my-computer software-update blackberry-id as-well-as windows-update terms-of-service drop-down pro-700 add-on scp-2700 mac-os device-manager voice-mail non-camera

Table: Snapshot of multi-word domain terms extracted by HITS (not found by NP Chunking).
Domain Cartridge: Framework

Diagram showing the flow of data through the framework, starting with a corpus, then passing through an ESG parser and a prismatic relations process, leading to secondary and primary indices, and finally to domain terms extracted through random index, synonym extractor, type-of extractor, feature-of extractor, and action-on extractor.
Random Indexing (RI)

For computing word similarity and dimensionality reduction

RI considers “term X term” co-occurrence, as opposed to “term X document” matrix — allowing for incremental learning of context information, scaling up with the corpus size

Relational Distributional Similarity — Two terms are similar if they appear in a similar context with similar Shallow Semantic Relations

Random Index Vector Update — Neighborhood constitutes of syntactic relations between target term and neighboring terms
Random Indexing (RI)

For computing word similarity and dimensionality reduction

RI considers “term X term” co-occurrence, as opposed to “term X document” matrix — allowing for incremental learning of context information, scaling up with the corpus size

*Relational Distributional Similarity* — Two terms are similar if they appear in a similar context with similar Shallow Semantic Relations

Random Index Vector Update — Neighborhood constitutes of syntactic relations between target term and neighboring terms
Random Indexing (RI)

For computing word similarity and dimensionality reduction

RI considers “term X term” co-occurrence, as opposed to “term X document” matrix — allowing for incremental learning of context information, scaling up with the corpus size

*Relational Distributional Similarity* — Two terms are similar if they appear in a similar context with similar Shallow Semantic Relations

Random Index Vector Update — Neighborhood constitutes of syntactic relations between target term and neighboring terms
Domain Cartridge: Framework
Synonym Discovery

Random Index gives top $N$ similar terms for a given term

HITS gives dominant domain terms and domain (SSR) relations

$$Sim(w_i, w_j) = \frac{\sum_p \mathbb{1}_{l_i = l_j, k_i = k_j} (f_{w_{k_i}} p, f_{w_{k_j}} p')}{\sum_p \sum_r \mathbb{1}_{l_i = l_r, k_i = k_r} (f_{w_{k_i}} p, f_{w_{k_r}} p')}$$

Numerator — #Freq. of common (dominant) words in both neighborhood with similar dominant SSR relations

Denominator — #Freq. of the common word in any other neighborhood with similar SSR relation
Synonym Discovery

Random Index gives top $N$ similar terms for a given term

HITS gives dominant domain terms and domain (SSR) relations

$$Sim(w_i, w_j) = \frac{\sum_p \sum_{l_i = l_j, k_i = k_j} (f_{w_{k_j}, p}, f_{w_{k_j}, p'})}{\sum_p \sum_r \sum_{l_i = l_r, k_i = k_r} (f_{w_{k_r}, p}, f_{w_{k_r}, p'})}$$

Numerator — #Freq. of common (dominant) words in both neighborhood with similar dominant SSR relations

Denominator — #Freq. of the common word in any other neighborhood with similar SSR relation
Synonym Discovery (RI)
Domain Cartridge: Framework
Relation Discovery

ESG SSR relations exploited to discover domain relation between two words

Feature-Of typically marked by noun-noun modifications and subject-object relations

Relation Discovery

Action-On marked by “dm” and verb-object relations


Type-Of marked by Hearst patterns like “or, especially” and SSR relations like “svo:include, npo:like, npo:such-as, npo:as”

E.g. “rel:svo:devices_include_HTC, rel:npo:applications_such-as_WhatsApp, rel:npo:features_like_call, rel:npo:contact_such-as_address”.
Relation Discovery

Action-On marked by “dm” and verb-object relations


Type-Of marked by Hearst patterns like “or, especially” and SSR relations like “svo:include, npo:like, npo:such-as, npo:as”

E.g. “rel:svo:devices_include_HTC, rel:npo:applications_such-as_WhatsApp, rel:npo:features_like_call, rel:npo:contact_such-as_address”.
Domain Term Evaluation

5000 articles, tutorials and manuals from the smartphone domain

We used the Back-of-the-Book Index (BOI) of manuals, to create ground truth for domain term discovery

Baselines:


- **BabelNet** (R. Navigli and S. P. Ponzetto. BabelNet: Building a very large multilingual semantic network. ACL '10.)

- **Yago** (F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. WWW '07.)
## Domain Term Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>WordNet</td>
<td>22.62%</td>
</tr>
<tr>
<td>NP Chunking on Titles</td>
<td>32.45%</td>
</tr>
<tr>
<td>HITS</td>
<td>40.87%</td>
</tr>
<tr>
<td>Yago</td>
<td>43.77%</td>
</tr>
<tr>
<td>BabelNet</td>
<td>53.74%</td>
</tr>
</tbody>
</table>

**Table:** Domain term evaluation.
Recall of a Question-Answering System

<table>
<thead>
<tr>
<th>Recall@N</th>
<th>With Domain Term Lexicon</th>
<th>Without domain term lexicon</th>
</tr>
</thead>
<tbody>
<tr>
<td>recall@1</td>
<td>0.40</td>
<td>0.33</td>
</tr>
<tr>
<td>recall@2</td>
<td>0.49</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Table: Performance of a QA system with and without domain term lexicon.

Incorporation of domain terms in parser lexicon improves QA system performance

---

Domain Relation Evaluation

2000 word pairs (500 for each of four categories) are manually annotated by two annotators

<table>
<thead>
<tr>
<th>System</th>
<th>Type-Of</th>
<th>Feature-Of</th>
<th>Action-On</th>
</tr>
</thead>
<tbody>
<tr>
<td>BabelNet, WordNet</td>
<td>19.27%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Yago</td>
<td>25.12%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Domain Cartridge</td>
<td>77%</td>
<td>85.7%</td>
<td>68%</td>
</tr>
</tbody>
</table>

Table: Recall comparison of systems for 3 relations.
Synonym Discovery: Distributional Similarity Comparison

<table>
<thead>
<tr>
<th>System</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yago</td>
<td>38%</td>
<td>32%</td>
<td>34.37%</td>
</tr>
<tr>
<td>BabelNet, WordNet</td>
<td>83%</td>
<td>31%</td>
<td>45.14%</td>
</tr>
<tr>
<td>Domain Cartridge (DC)</td>
<td>58%</td>
<td>41%</td>
<td>47.60%</td>
</tr>
<tr>
<td>DC + WordNet</td>
<td>62%</td>
<td>40%</td>
<td>49.00%</td>
</tr>
<tr>
<td>DC + ESG Parser Features</td>
<td>65%</td>
<td>39%</td>
<td>49.14%</td>
</tr>
</tbody>
</table>

Table: Precision-Recall comparison of Domain Cartridge (random-indexing, HITS and sim. eqn.) with other systems.
Synonym Discovery: Comparison with Distributional Similarity Measures in WordNet

<table>
<thead>
<tr>
<th>WordNet</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCH</td>
<td>0.22</td>
</tr>
<tr>
<td>RES</td>
<td>0.31</td>
</tr>
<tr>
<td>JCN</td>
<td>0.42</td>
</tr>
<tr>
<td>PATH</td>
<td>0.42</td>
</tr>
<tr>
<td>LIN</td>
<td>0.43</td>
</tr>
<tr>
<td>WUP</td>
<td>0.43</td>
</tr>
<tr>
<td>LESK</td>
<td>0.45</td>
</tr>
<tr>
<td>Domain Cartridge</td>
<td>0.49</td>
</tr>
</tbody>
</table>

**Table**: F-Score comparison of WordNet similarity measures with Domain Cartridge.
Conclusions

- Unsupervised framework for shallow domain ontology construction, without using manually annotated resources
- Multi-words form an important component of Domain Term Discovery
- Incorporation of domain terms in parser lexicon results in 73% reduction in incomplete parses, improving performance of an in-house QA system by upto 7%
- Synonym discovery approach, using Relational Distributional Similarity, RI, HITS etc., performs better than other existing approaches