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**

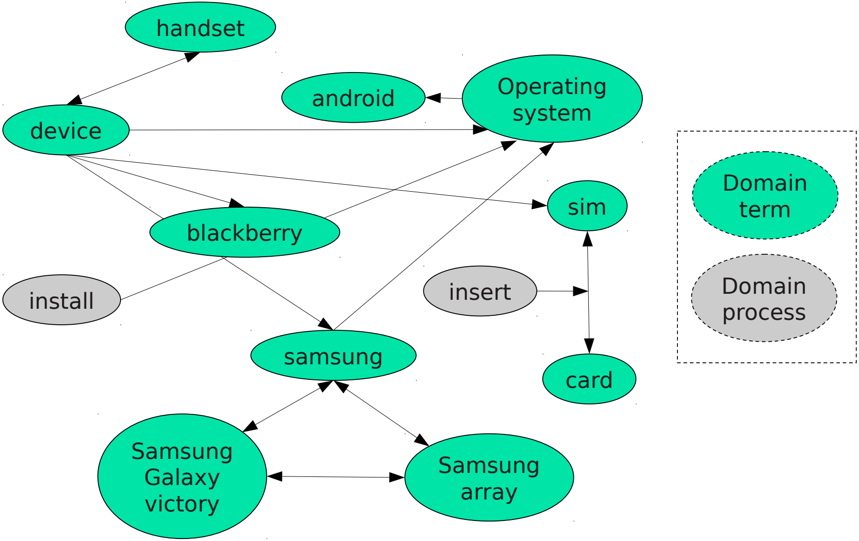
Unsupervised Framework

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

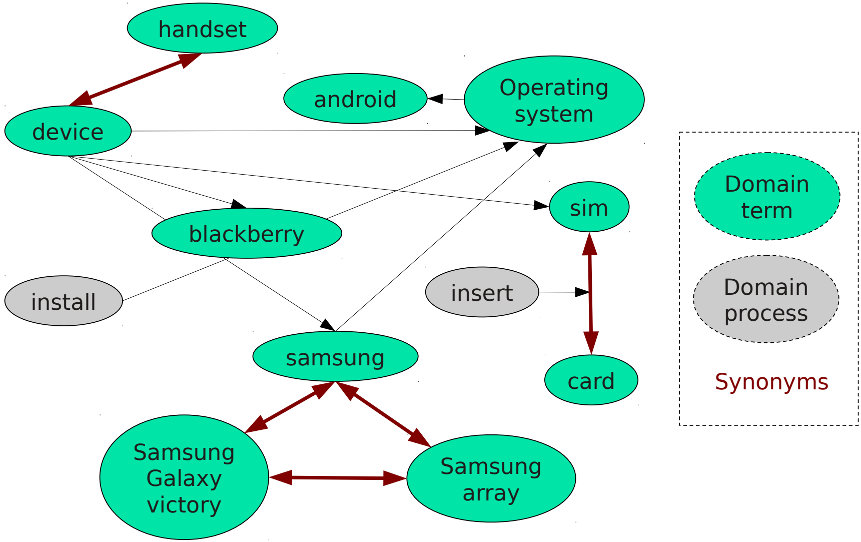
- Most of these documents have less hyperlink and table (info-box as in Wikipedia) information, or extraction is difficult (E.g. pdf)

- Challenge is to learn a *shallow* ontology from raw unannotated **plain text**

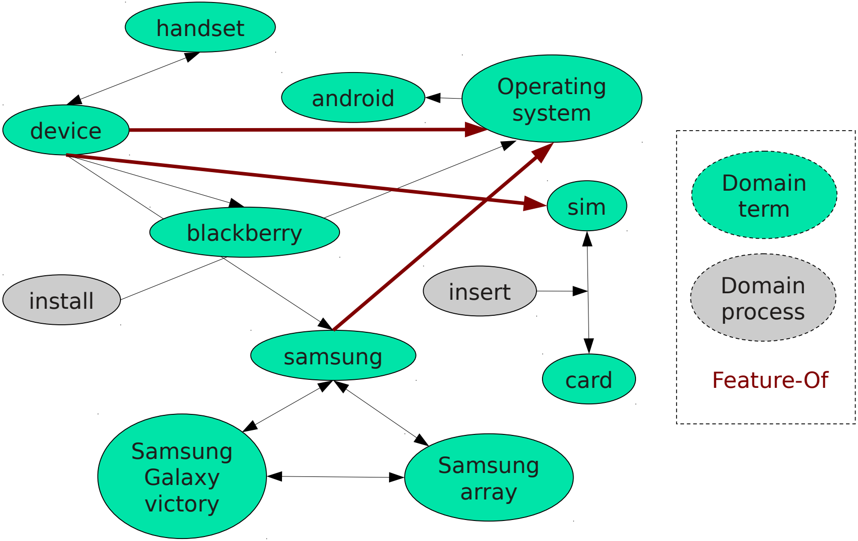
Domain Cartridge as a Graph



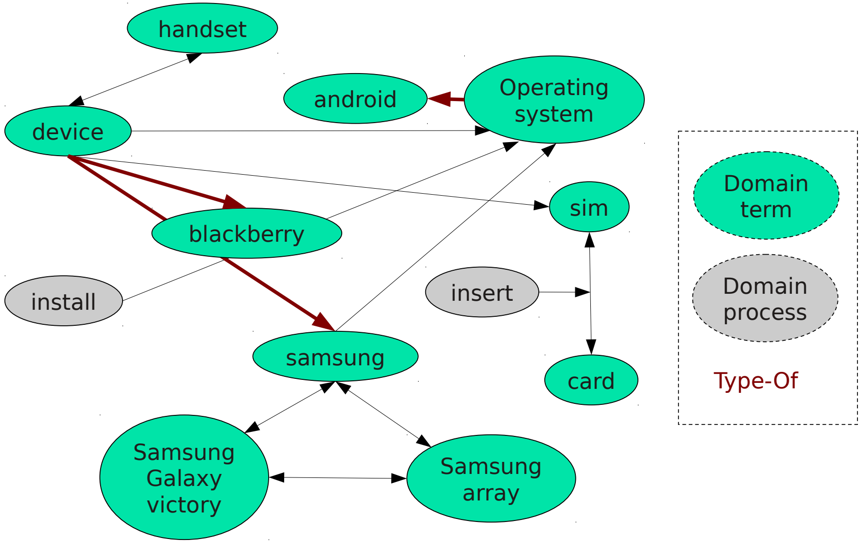
Domain Cartridge as a Graph



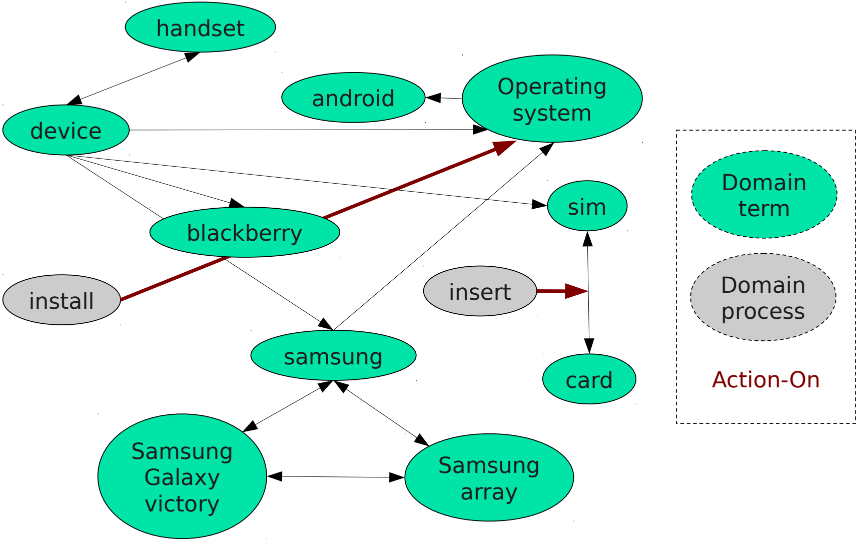
Domain Cartridge as a Graph



Domain Cartridge as a Graph



Domain Cartridge as a Graph



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
- Experiments: Manual Evaluation, Comparison with BabelNet, WordNet, Yago
- 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
- Experiments: Manual Evaluation, Comparison with BabelNet, WordNet, Yago
- Conclusions

Roadmap

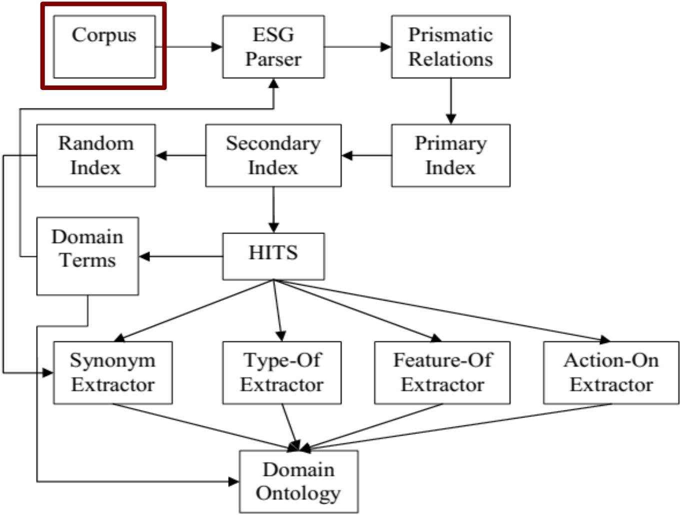
- Unsupervised framework for shallow domain ontology construction:
  - Domain Term Discovery (DTD)
  - Improvement of Parser performance by DTD
  - Domain Relation Discovery (DRD)
- Use-Case: Improvement of an in-house Question-Answering system
- Experiments: Manual Evaluation, Comparison with BabelNet, WordNet, Yago
- 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
- Experiments: Manual Evaluation, Comparison with BabelNet, WordNet, Yago
- Conclusions

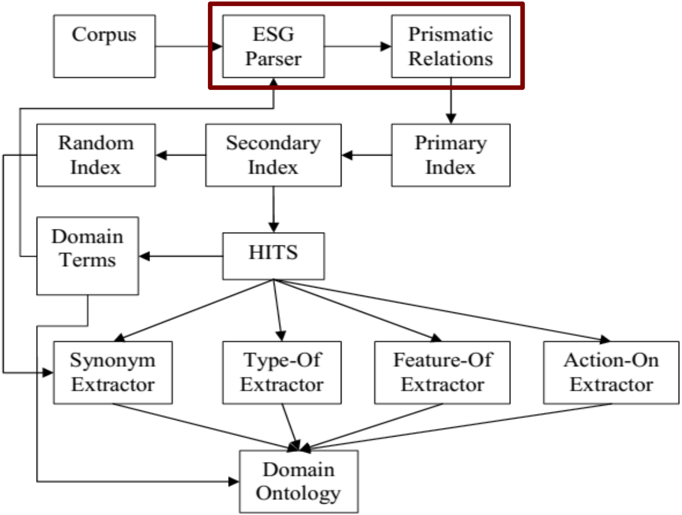
Corpus: Knowledge articles, manuals, tutorials etc.

Domain Cartridge: Framework



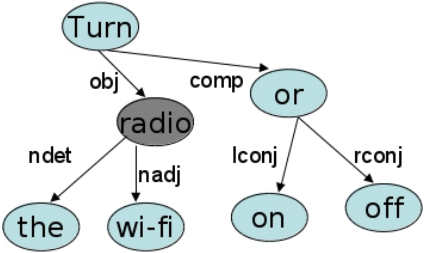
Parsing

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

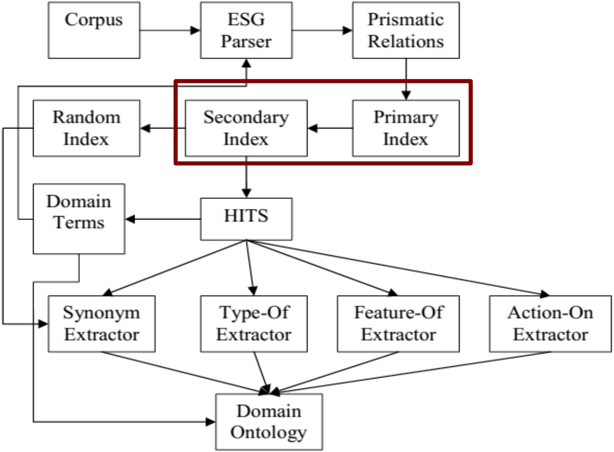
E.g., “**Samsung has a battery**" and “**Samsung’s battery died**" both generate the same relation ‘nnMod:samsung_battery’

Prismatic Relations

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

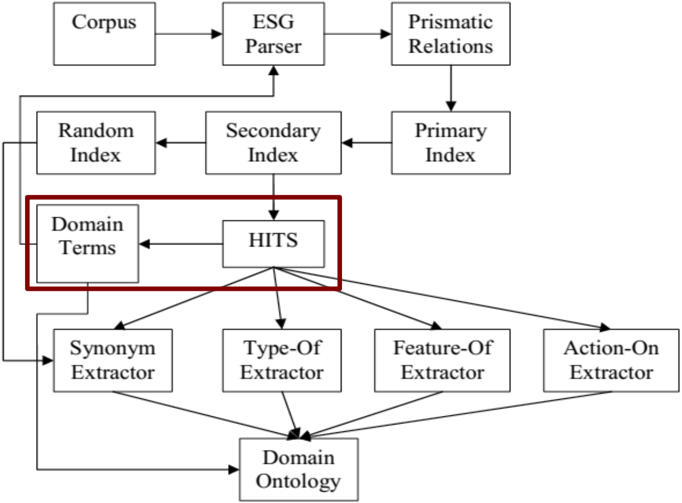
E.g., “**Samsung has a battery**" and “**Samsung’s battery died**" both generate the same relation ‘nnMod:samsung_battery’

Domain Cartridge: Framework



Lucene Index – For efficient retrieval of relations, documents, positional information, proximity based queries etc.

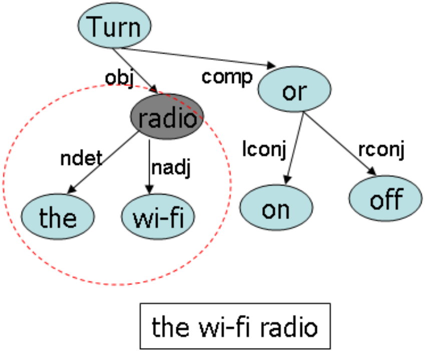
Domain Cartridge: Framework



Domain Term Discovery

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

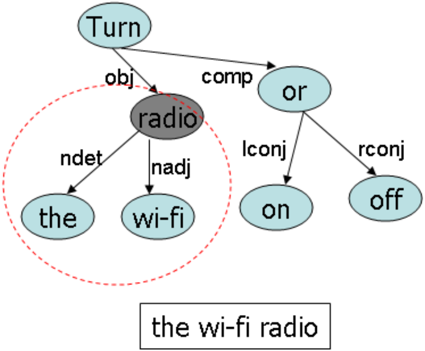
Noun Phrase Chunking on *document titles* to extract frequently occuring concepts as domain words



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 occuring concepts as domain words



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

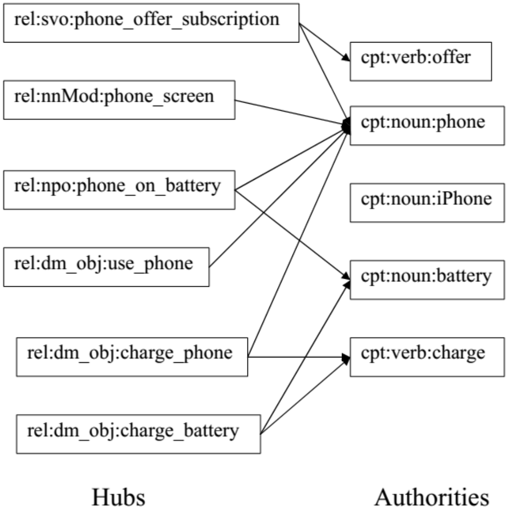
- Any Shallow Semantic Relation (SSR) from ESG parser is a *hub* generating domain terms
- Any domain term is an *authority* influenced by incoming features from hubs
- Good authorities incorporated in Parser Domain Term Lexicon
- Parser is re-run, refined relations generated, and previous steps iterated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

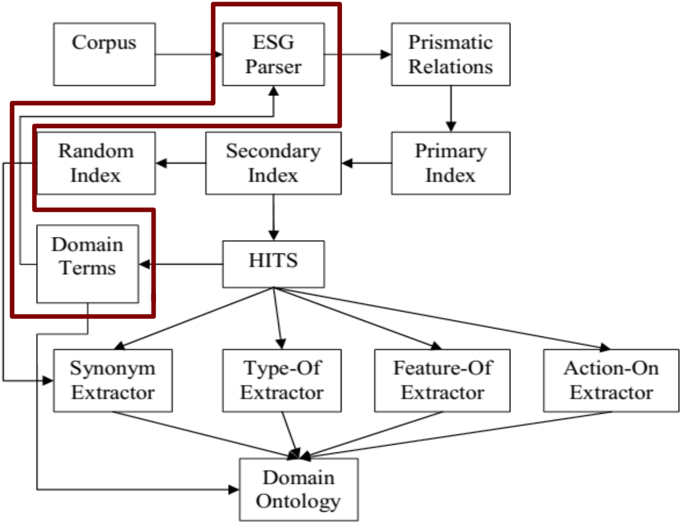
HITS

- 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


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
cpt:verb:offer
cpt:noun:phone
cpt:noun:iPhone
cpt:noun:battery
cpt:verb:charge
Hubs
Authorities

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

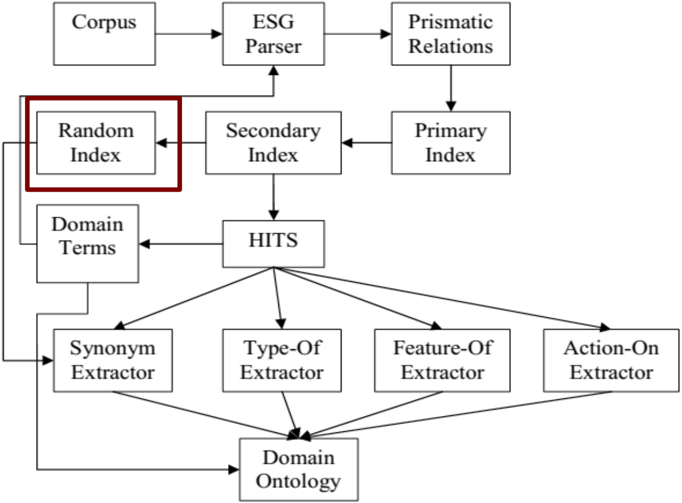
software-version htc-evo wi-fi memory-card microsoft-exchange lg-optimus samsung-m400 samsung-galaxy-victory software-updates samsung-array text-messaging touch-screen blackberry-bold

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



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)

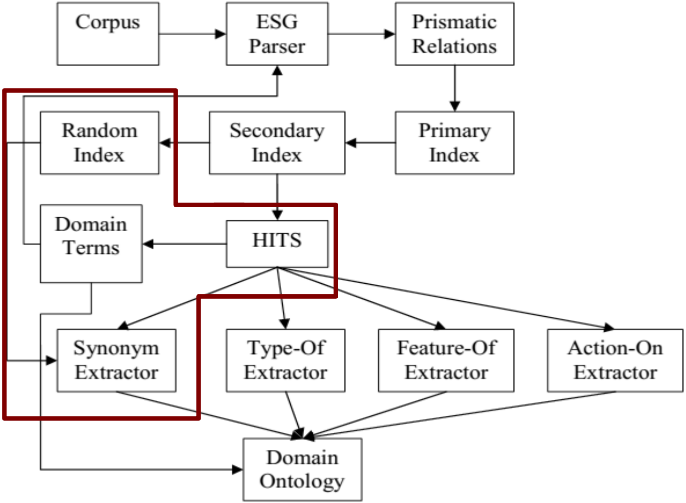
For computing word similarity and dimensionality reduction

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

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

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

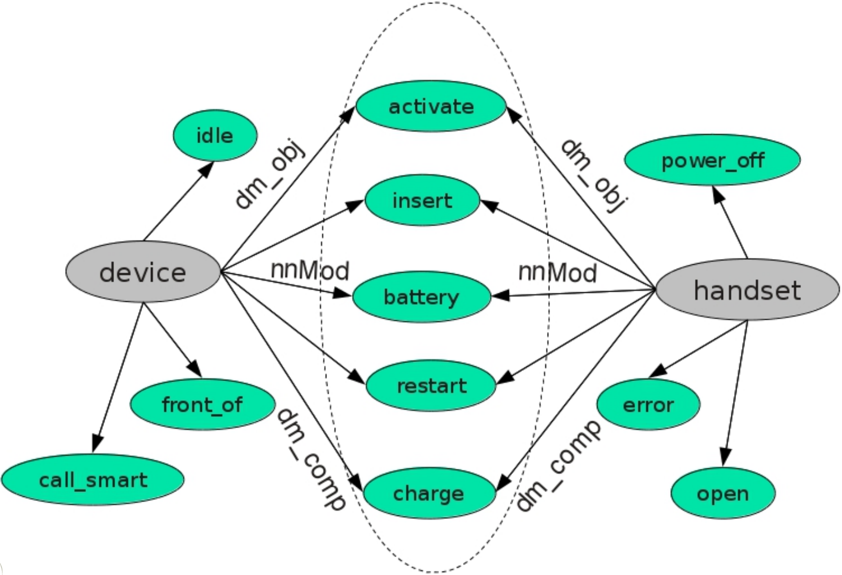
HITS gives dominant domain terms and domain (SSR) relations

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

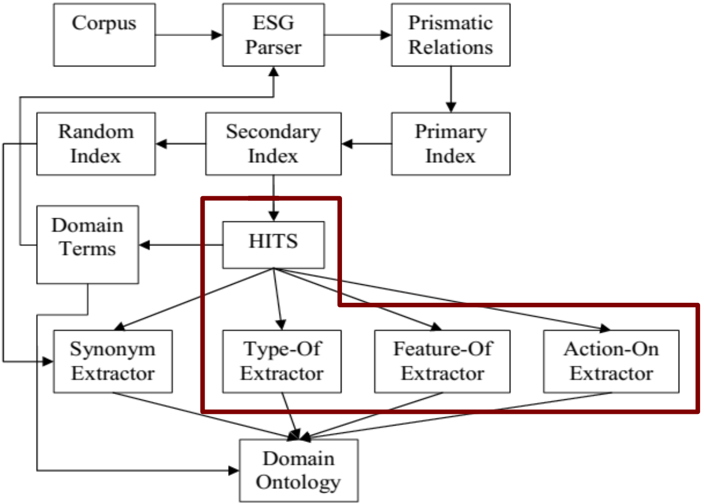
Numerator — #Freq. of common (dominant) words in both neighborhood with similar *dominant* SSR relations

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

Synonym Discovery (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

“*rel:nnMod:**network_life**, rel:nnMod:**account_settings**, rel:svo:**phone**_access_**internet*** etc."

Relation Discovery

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

E.g. “*rel:svo:tap_**add_account***, *rel:dm_obj:**activate_device***, *rel:svo:mobile_**sync_phone***, *rel:svo:account_**use_phone** etc.*"

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

E.g. “*rel:svo:tap_**add_account***, *rel:dm_obj:**activate_device***, *rel:svo:mobile_**sync_phone***, *rel:svo:account_**use_phone** etc.*"

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:

- WordNet (G. A. Miller. Wordnet: A lexical database for english. COMMUNICATIONS OF THE ACM, 38, 1995.)
- 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

Method	**Recall**
WordNet	22.62%
NP Chunking on Titles	32.45%
HITS	40.87%
Yago	43.77%
BabelNet	53.74%

Table: Domain term evaluation.

Recall of a Question-Answering System

Recall@N	**With Domain Term Lexicon**	**Without domain term lexicon**
recall@1	0.40	0.33
recall@2	0.49	0.45

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

Incorporation of domain terms in parser lexicon improves QA system performance

[1] D. Gondek et al. A framework for merging and ranking of answers in DeepQA. IBM Journal of Research and Development, 56(3), 2012.

Domain Relation Evaluation

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

System	**Type-Of**	**Feature-Of**	**Action-On**
BabelNet, WordNet	19.27%	-	-
Yago	25.12%	-	-
Domain Cartridge	77%	85.7%	68%

Table: Recall comparison of systems for 3 relations.

Synonym Discovery: Distributional Similarity Comparison

System	Precision	Recall	F-Score
Yago	38%	32%	34.37%
BabelNet, WordNet	83%	31%	45.14%
Domain Cartridge (DC)	58%	41%	47.60%
DC + WordNet	62%	40%	49.00%
DC + ESG Parser Features	65%	39%	49.14%

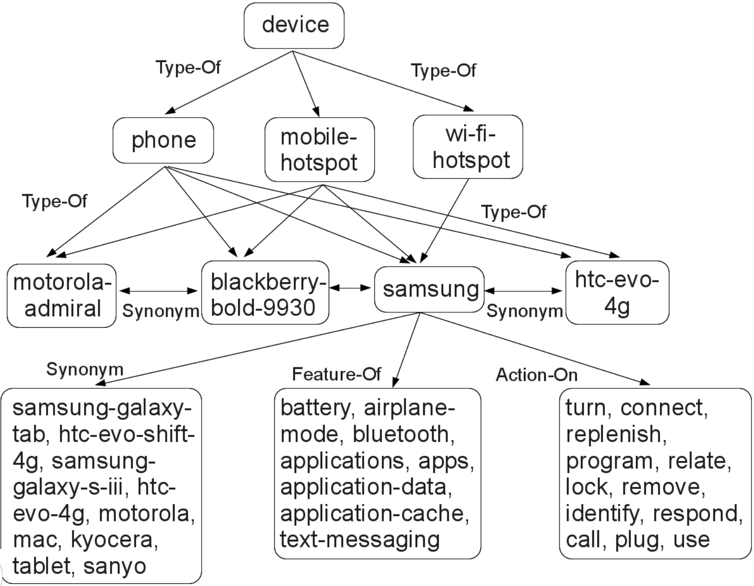
Table: Precision-Recall comparison of Domain Cartridge (random-indexing, HITS and sim. eqn.) with other systems.

Synonym Discovery: Comparison with Distributional Similarity Measures in WordNet

WordNet	**F-Score**
LCH	0.22
RES	0.31
JCN	0.42
PATH	0.42
LIN	0.43
WUP	0.43
LESK	0.45
Domain Cartridge	0.49

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

Domain Cartridge Ontology Snapshot



device

Type-Of

Type-Of

phone

mobile-hotspot

wi-fi-hotspot

Type-Of

Type-Of

motorola-admiral

Synonym

blackberry-bold-9930

samsung

Synonym

htc-evo-4g

Synonym

Feature-Of

Action-On

samsung-galaxy-tab, htc-evo-shift-4g, samsung-galaxy-s-iii, htc-evo-4g, motorola, mac, kyocera, tablet, sanyo

battery, airplane-mode, bluetooth, applications, apps, application-data, application-cache, text-messaging

turn, connect, replenish, program, relate, lock, remove, identify, respond, call, plug, use

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