max planck institut
informatik

Unsupervised Approach for Shallow Domain Ontology Construction from Corpus

Subhabrata Mukherjee[‡], Jitendra Ajmera[†], Sachindra Joshi[†]

[‡]Max-Planck-Institut für Informatik (Germany), [†]IBM Research Lab (India)

Introduction

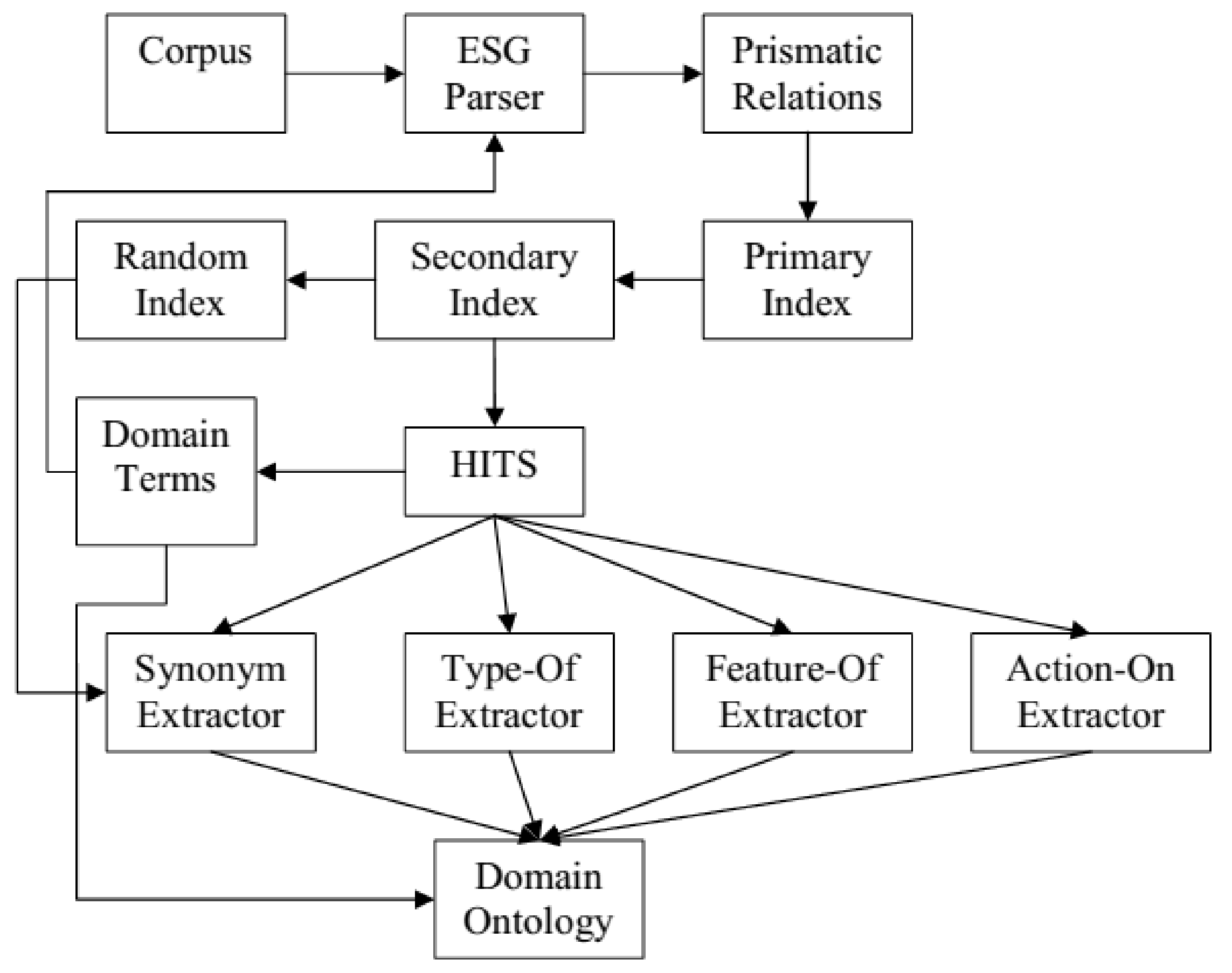
- Ontology is a knowledge base of structured list of concepts and relations
- A domain ontology consists of *domain-specific* concepts and relations
- In this work, we focus on 4 primary relations : *Feature-Of, Type-Of, Action-On* and *Synonyms*
- A domain ontology incorporates domain awareness in an IR system to account for the domain semantics of terms and their relationships
- We propose an approach to create such ontology from corpus *automatically* without using any manually annotated resource or supervision
- Input to the system is a corpus consisting of a set of html or knowledge articles and pdf manuals

Domain Term Discovery

- First step in ontology construction is to discover important domain concepts, *especially* multi-word terms such as *Samsung-Galaxy-Tab, call-log, 4g-connection etc.*
- *We use the parse tree structure of a slot grammar parser output for this purpose*
- *Noun phrase chunking is done on the parser output to discover domain terms by finding frequent subtrees of noun-nodes*
- A *bipartite* graph of parser relations and associated (multi-word) domain terms, extracted by the above step, is constructed
- HITS algorithm is run over the graph to identify important domain relations (hubs) generating significant domain terms (authorities)
- The discovered domain terms are fed into the parser lexicon (making it aware of multi-word domain terms) and the above steps iterated
- The parser performance is improved generating better semantic relations
- The above process of domain term discovery achieves a recall improvement of **18%** over WordNet
- It improves the performance of an existing Question-Answering System by **7%**, as it becomes aware of the domain
- Extracted Domain Terms Snapshot :

samsung blackberry device software novatel software-version application htc-evo wi- memory-card bluetooth motorola kyocera browser voicemail microsoft-exchange lg-optimus

System Block Diagram

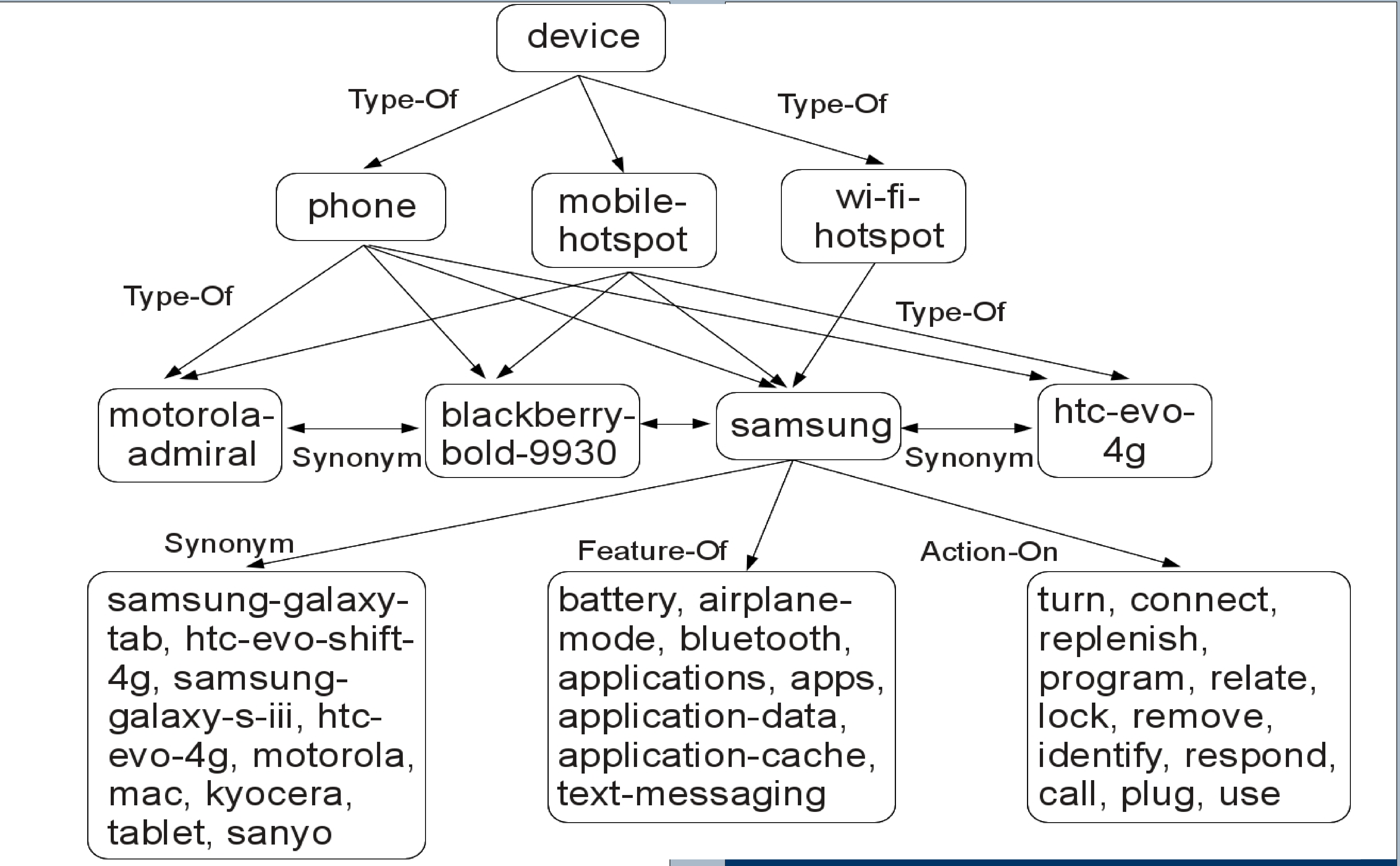


Domain Relation Discovery

- *Shallow semantic relationship* (SSR) annotation is done over the parser output, consisting of rules to generate projections for all the frames in the corpus and generate normalized parser relations like *svo, npo, nnMod, dm etc.*

1. *svo* depicts a subject-verb-object tuple. E.g. *rel:svo:phone-offer-feature, rel:svo:phone-show-message etc.*
2. *nnMod* depicts noun-noun modications. E.g. *rel:nnMod:iPhone-battery, rel:nnMod:screen-icon etc.*
3. *dm* depicts actions on entities. *E.g. rel:dm-obj:use-phone, rel:dm-comp:plug-iPhone etc.*
4. *npo* depicts terms connected by prepositions. *E.g. subscription-to-service, battery-on-phone etc.*

- Action-On ontology relation represents any activity (method) on a given domain term. The SSR *dm* and *svo* help in Action-On identication.
- Type-Of relations depict Is-A hierarchy. To discover Type-Of clues, the *svo* and *npo* SSR's are used in conjunction with *Hearst* patterns (E.g. verbs like *include*, prepositions *like*, *such-as* and *as*, *etc.*). E.g. *apps-include-WhatsApp, rel:npo:features-like-call etc*
- Feature-Of relations depict components or functionalities of a domain term. To discover Feature-Of relations we use SSR's *nnMod* and *svo*.
- We follow the notion of *relational distributional similarity*, and define two words to be Synonyms if they appear in a similar context with similar SSR relations in the neighborhood
- We use *Random Indexing* (RI) for dimensionality reduction as well as similarity computation.


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

System Framework

- The corpus is parsed using English Slot Grammar Parser and SSR relations are generated
- The Primary Index stores all parser output
- The Secondary Index stores only SSR relations
- HITS is run over SSR relations and associated domain terms, in conjunction with NP chunking
- Extracted multi-word domain terms and relations are fed to the parser lexicon and steps iterated
- RI is a word co-occurence based approach to statistical semantics allowing for incremental learning of context information
- RI retrieves a set of similar candidates for a word based on similar *SSR* distribution in the corpus
- Domain Terms and significant relations from HITS are used to extract *Action-On*, *Type-Of* and *Feature-Of* ontology relations
- Random Index helps in Synonym identification using *relational distributional similarity hypothesis*

Evaluation Results

- Evaluated on 5000 articles, tutorials and manuals from the smartphone domain. 2000 word-pairs are manually annotated (500 for each relation)
- WordNet could only discover **1** word-pair for Feature-Of (subset of *Meronymy* and *Holonymy*) and **74** word-pairs for Type-Of (corr. to *Hyponymy* and *Hypernymy*)
- WordNet does *not* contain any Action-On reln. type

Relation	Our Approach	
	Precision	Recall
Feature-Of	74.9%	85.7%
Action-On	63.88%	68%
Type-Of	57%	77%

WordNet Similarity Measures	F-Score Synonyms
LCH	0.22
RES	0.31
JCN	0.42
PATH	0.42
LIN	0.43
WUP	0.43
LESK	0.45
Our Approach	0.49