Unsupervised Approach for Shallow Domain Ontology Construction from Corpus

Subhabrata Mukherjee
Max-Planck-Institut für Informatik
smukherjee@mpi-inf.mpg.de

Jitendra Ajmera
IBM India Research Lab
jajmera1@in.ibm.com

Sachindra Joshi
IBM India Research Lab
jsachind@in.ibm.com

ABSTRACT

In this work we propose an unsupervised approach to construct a domain-specific ontology from corpus. It is essential for *Information Retrieval* systems to identify important domain concepts and relationships between them. We identify important *domain terms* of which *multi-words* form an important component. Our approach identifies 40% of the domain terms, compared to 22% identified by WordNet on manually annotated smartphone data. We propose an approach to construct a shallow ontology from discovered domain terms by identifying *four* domain relations namely, *Synonyms ('similar-to'), Type-Of ('is-a'), Action-On ('methods') and Feature-Of ('attributes')*, where we achieve an F-Score of 49.14%, 65.5%, 65% and 80% respectively.

Categories and Subject Descriptors

H.0 [**Information Systems**]: General

1. INTRODUCTION

Ontology is a knowledge base of structured lists of concepts, and their relations. Such knowledge representation is useful for the purpose of a variety of text analysis problems such as document similarity computation, search result re-ranking and interactive dialogue systems. In this paper, we present an approach to automatically construct a shallow ontology from a domain corpus. Such corpus typically consists of a set of html or knowledge articles and pdf manuals. We view this domain ontology as a graph, where the nodes represent domain concepts and edges represent the relations among these concepts. We extract 4 types of relations namely, *Feature-Of*, *Action-On*, *Type-of* and *Synonyms*. Figure 1 shows a snapshot of the constructed smartphone domain ontology using our approach.

Such domain ontology can be used to induce domain awareness in an information retrieval system, so that it takes into account the domain semantics of terms and their relationships, compared to the simple lexical matching of terms. Our work differs from related works as we focus to create such ontology from corpus *automatically* without using *any* manually annotated resource like *WordNet* or supervision.

Our approach starts with finding important domain concepts where we exploit the parse tree structure of a slot grammar parser output. This is explained in Section 2.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577350.

Next, we consider the shallow semantic relationships (SSR) present among these domain concepts for finding the 4 ontology relations as explained in Section 3.

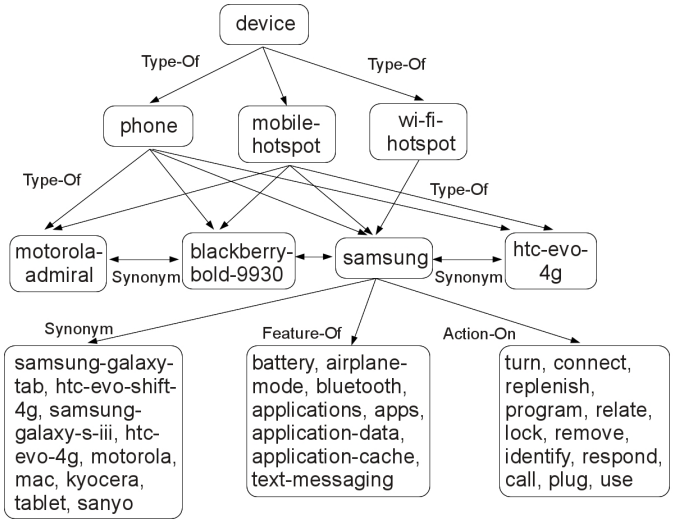

Figure 1: Snapshot of Constructed Smartphone Domain Ontology

Such domain specific discovery is required since manually-constructed resources like WordNet typically miss domain specific concepts and their relations. In our analysis we found that only 22.62% percentage of domain concepts in the smartphone domain figured in the WordNet. Furthermore, only 10.53% of domain relations were present.

2. DOMAIN TERM DISCOVERY

The first step towards gathering insights about a new domain is to discover a list of important domain concepts, especially the multi-word terms such as 'Samsung-Galaxy-Tab', 'Call-log', '4g-connection' etc.

We make use of the parse tree structure of a slot grammar parser [2] output for this purpose. All the documents in the corpus are parsed using the slot grammar parser. Noun phrase chunking is done on the parser output to discover domain terms. This is achieved by finding frequent subtrees of noun-nodes. A frequency thresholding step is performed to remove all the unnecessary and noisy entries in this list. Table 1 shows a snapshot of the domain terms discovered using the noun phrase chunking approach. The next step involves finding the four types of ontology relations among these discovered domain terms. To facilitate this step, the

parser is made aware of the domain specific concepts by providing it a domain lexicon as input.

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

Table 1: Snapshot of Multi-Word Domain Terms Discovered using Noun Phrase Chunking

3. DOMAIN RELATION DISCOVERY

Our approach to find the four types of ontology relations is based directly on the SSR relations extracted from the parser output. The SSR relations are of the following form:
1. *svo* depicts a subject-verb-object tuple. For example: *rel:svo:phone_offer_feature, rel:svo:phone_show_message etc.*
2. *nnMod* depicts noun-noun modifications. For example: *rel:nnMod:iPhone_battery, rel:nnMod:screen_icon etc.*
3. *dm* depicts actions on entities. For example: *rel:dm_obj:use_phone, rel:dm_comp:plug_iPhone etc.*
4. *npo* depicts terms connected by prepositions. For example: *subscription_to_service, battery_on_phone etc.*

Action-On ontology relation represents any activity (method) on a given domain term. For example, 'charge' and 'display' are activities on 'battery' and 'menu', respectively. By definition, an Action-On relation pair consists of a 'verb' that acts on a 'noun'. The SSR *dm and svo* help in Action-On identification. E.g. "*rel:svo:tap_**add_account**, rel:svo:phone_**access_internet**, rel:svo:mobile_**sync_phone***"

Type-Of relations depict *Is-A* hierarchy *i.e.* a parent-child relation. For example: "Samsung is a Type-Of mobile", "Internet Explorer is a Type-Of browser, Angry Birds is a Type-Of application *etc.*". In order to discover the Type-Of clues, the *svo* and *npo* SSR's are used in conjunction with the Hearst [1] patterns (e.g. verbs like *include*, prepositions like *like, such-as* and *as*, *etc.*). E.g. "*rel:svo:**devices**_include_**HTC**, rel:npo:**applications**_such-as_**WhatsApp**, rel:npo:**features**_like_**call***".

Feature-Of relations depict components or functionalities of a domain term. For example: "screen is a Feature-Of mobile", "wi-fi is a Feature-Of network", "life is a Feature-Of battery" *etc.* In order to discover Feature-Of relations we use the SSR's *nnMod* and *svo*. E.g."*rel:nnMod:**network_life**, rel:nnMod:**iPhone_battery**, rel:svo:**motorola**_run_**device-software**, rel:svo:**router**_decrease_**signal-strength** etc.*"

We define two words to be *Synonyms* if they appear in a similar context. Here, we follow the notion of *relational distributional similarity* [3]. Since it is computationally very expensive to go over all the word pairs to compute the distributional similarity, we use *Random Indexing* for dimensionality reduction as well as similarity computation.

Random Indexing (RI) [6] is a word co-occurence based approach to statistical semantics. RI uses statistical approximations of the full word co-occurence data to achieve dimensionality reduction, resulting in much quicker running time and fewer required dimensions. This facilitates fast computation of similarity between candidate domain terms as well as different relation discovery. It is scalable and allows for the incremental learning of context information.

However, as opposed to most of the previous works that consider raw neighborhood of a term for random indexing, we use only those neighboring terms to define the context for a target term that share a syntactic dependency (given by the slot grammar parser) with the target term. Random index is used to get a set of similar candidates for a word based on similar SSR distribution in the corpus.

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

Table 2: Precision-Recall for 3 Relations

WordNet	F-Score
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

Table 3: F-Score Comparison of WordNet Similarity Measures with Our Approach for Synonyms

4. EXPERIMENTAL EVALUATION

We collected 5000 articles, tutorials and manuals from the smartphone domain. 500 word pairs for each of the four relations, resulting in 2000 word pairs, were manually annotated. Table 2 shows the precision-recall figures for Feature-Of, Action-On and Type-Of.

In our work, we use WordNet as the baseline for relation discovery. WordNet [4] could only discover 1 word-pair for Feature-Of (subset of the relations *Meronymy* and *Holonymy*) and 74 word-pairs for Type-Of (corresponding to the relations *Hyponymy* and *Hypernymy*). WordNet does not contain any relation corresponding to Action-On.

A number of similarity measures are defined over the WordNet taxonomy that exploit distributional similarity to find the relatedness of 2 concepts. We considered 7 similarity measures from [5] as our baseline for Synonym discovery approach. Table 3 shows the F-score comparison of different wordnet similarity measures with our approach.

5. CONCLUSIONS

In this work, we propose an unsupervised approach to construct a shallow domain ontology from corpus. Unlike other existing approaches, we do not make use of manually annotated resources like WordNet or any mode of supervision, and still obtain better performance over WordNet.

6. REFERENCES

[1] M. A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proc. of 14th Conference on Computational Linguistics*, 1992.

[2] M. C. McCord and A. Bernth. Using slot grammar. *IBM Deutschland Informationssysteme GmbH Scientific Center Institute for Logic and Linguistics*, 1992.

[3] S. Mcdonald and M. Ramscar. Testing the distributional hypothesis. In *Proc. of 23rd Annual Conference of Cognitive Science Society*, 2001.

[4] G. A. Miller. Wordnet: A lexical database for english. *COMMUNICATIONS OF THE ACM*, 38, 1995.

[5] T. Pedersen and S. Patwardhan. Wordnet::similarity - measuring the relatedness of concepts. 2004.

[6] M. Sahlgren. Introduction to random indexing. *Methods and Applicatons of Semantic Indexing Workshop*, 2005.