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

General Terms
Knowledge Extraction, Ontologies

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. 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
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:


4. *npo* depicts terms connected by prepositions. For example: *rel:npo:subscribe_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:top_add_account, rel:svo:phone_access_internet, rel:svo:mobile_sync_phone"

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 the Feature-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"

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"

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


