Volume 7, Issue 5, May 2017 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Extension of Knowledge Structure with the Help of Data Gathering

Vaishali Bhargava
M.Tech Scholar, Department of Computer Science Engg.,
AKTU, Uttar Pradesh, India

Sapna Singh
Faculty, Department of Computer Science& Engg.,
AKTU, Uttar Pradesh, India

Abstract-This Work focuses on essential phases of knowledge base population by using: extracting facts from raw data ,validating the extracted information, enhancing system and utilizing knowledge bases to improve real world application. Knowledge bases encrypt real world data in machine readable data formats. The goal of knowledge base population is to decrypt this knowledge and extending knowledge bases. Information Extraction is the task of discovering different types of facts and patterns in unstructured data. Firstly slot filling task is investigated. Slot filling gathers information from a collection of web ,news or other sources to determine a set of slot values for a given person and organization entities. Second consider how to estimate the correctness of the extracted slot values. we propose an estimation model based on maximum entropy framework and demonstrate the effectiveness of this model to improve the slot filling aggregation through weighted voting strategy. Third we take rich annotation learning to fill the gap between an expert annotator and a feature engineer. We introduce an algorithm to enrich features with the guidance of all levels of rich annotation from human annotators. It is studied that with little additional annotation time , we can significantly improve the performance. Finally , we utilize knowledge bases in a real world application. We propose a systematic study to show the effectiveness of semantic knowledge encrypted in the entities on modeling user interests, by utilizing the entity information from knowledgebases.

Keywords-personal knowledge graph population, personal assertion detection, relation detection, slot filling, user profiling for content recommendation, rich annotation guided learning.

I. INTRODUCTION

A knowledge base is used to gather and manage knowledge in the form of logical statement. Knowledge base stores factual information in the form of relationships between entities. Knowledge bases follow the Resource Description Framework to represent facts in the form of binary relationship in particular (entity, predicate ,value) triples, where entity represent person/organization, predicate indicates the relationship and a value can be another entity a type, an attribute, and other factual information. Knowledge bases provide structured information that is interpretable by computer.

The Knowledge Base Population is process of discovering facts about entities from large collection of data and used it to enhance knowledgebase. KBP is an active research task, aims to promote research in discovering information about entities and augmenting a knowledge base with this information. Knowledge base population consists of these knowledge base population tasks:

A. Entity linking

Overall goal of KBP is to identifies entities and link them to corresponding knowledge base entries in a provided document, then discover attributes about the entities and finally expand the KB with new attribute

B. Slot filling

The goal of slot filling is to collect from the corpus information regarding certain attributes of an entity. This task is to complete all known information about a given query entity.

II. PERSONAL KNOWLEDGE GRAPH POPULATION

In this statistical language understanding is proposed with key language understanding components.

A. Personal assertion detection

This term classifies the spoken words into binary classes. We formulate this problem as a binary classification task and apply support vector machines framework to perform classification.

We use linear kernals as provided in SVM package, since they are efficient. The features include the ngrams, stems, part of speech tags and their combination. The output of this stage provide us with coarse grained information.

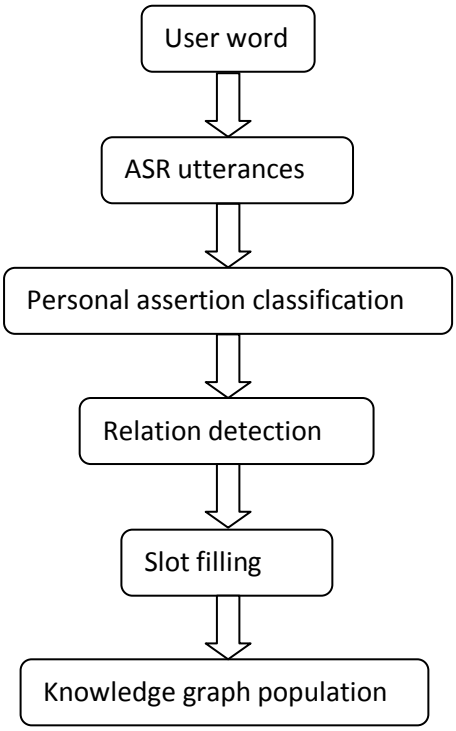

Figure1: framework of personal graph population

B. Relation detection

The goal of relation detection is to determine that which relation in the part of knowledge graph related to the utterance have been invoked in the user utterances. One utterance can invoke more than one relation. We apply the SVM package to classify each utterance into one or more classes. We construct r SVM models where r is the number of relation classes. The ith SVM is trained with all the examples with negative labels. Then we apply all k SVM models to each utterance to determine which relation are invoked in it.

C. Slot filling

The goal of Slot Filling is to collect from the mass information regarding precise attributes of an entity, which may be a person or some type of organization. Each query in the Slot Filling task contains the name of the entity, its type (person or organization), a background document containing the name its node ID and the attributes which need not be filled. Attributes are prohibited if they are already filled in the reference data base and can only take on a single value. Along with each slot fill, the system must provide the ID of a document which supports the correctness of this fill. If the mass does not provide any information for a given attribute, the system should generate a NIL response (and no document ID). KBP2010 defined 26 types of attributes for persons and 16 types of attributes for organizations. Some of these attributes are specified as only taking a single value, while some can take multiple values (e.g., top employees).The reference KB includes hundreds of thousands of entities based on articles from an October 2008 dump of English Wikipedia which contains 818,741 nodes. The source collection includes 1,286,609 newswire documents, 490,596 web documents. To score Entity Linking, one take each query and check whether the KB node ID (or NIL) returned by a system is correct or not. Then compute the Micro-averaged Accuracy, computed across all queries. Slot Filling task is to complete all known information about a given query entity. birthplace, birthdate, occupation, spouse, etc.

A key feature of this is relation release is the classification of a sentence and two entities in the sentence to a relation of interest. The Slot Filling task required to automatically filter information from the document collection which fills missing KB attributes for distinct entities. The slot-filling task is a mixture of traditional IE (Information Extraction) and QA (Question Answering).

Slot filling architecture:

The basic Slot filling architecture contains three phases:

1. Document/passage retrieval:- Fetching passages involving the queried entity.
2. Answer extraction:- Getting specific answers from the retrieved passages.
3. Answer combination:- Joining and selecting among the answers extracted.

III. USER PROFILING FOR CONTENT RECOMMENDATION

Web plays crucial role in the distribution of data from different sources to the users. The problem of Data Overload necessitates the use of the content recommendation techniques to help choose the best items matching users' interests. Thus users get better response to meet their needs without wasting much time distilling returned data. Most of these content recommendation systems face various difficulties while identifying and providing high-quality items to users. This is our motivation for conducting this analysis on content recommendation, and modeling user interests is a

key and challenging module for personalized content recommendation. Meeting user requirements implicates a thorough understanding of their interests expressed explicitly through search queries.

IV. RICH ANNOTATION GUIDED LEARNING

In a regular annotation interface, a human annotator is only asked to provide the final labels. We call this basic annotation 'Level 0'. We can see that among these elements, little study has been conducted on encompassing rich annotations from human annotators. In most cases it was not the obligation of the human annotators to write down their evidence or comments during annotation. In contrast, the human learning scenario involves more interactions. However, we can assume that any annotator is able to verify and comment on his/her judgment.

Some human annotators are on various intense levels:

Level 1: Ask an annotator to verify a label by providing surface evidence.

Level 2: Ask an annotator to verify a label by providing deep evidence.

Level 3: Ask an annotator to provide comments about linguistic features or resources that might be helpful for system development. The algorithm aims to extensively incorporate all comments from an old development data set into an automatic correction component. This assessor can be applied to improve the results for a new test data set (i.e., "new homework" in human learning).

The algorithm can be summarized as follows.

1. The pipeline starts by running the baseline system to generate results. In this step we can also add the outputs from other systems or even human annotators. We will present one case study on slot filling which incorporates these two additional elements.
2. We obtain comments from human annotators on a small development set Di. Each time we ask a human annotator to pick N ¹ random results and provide a new comment on each result. We found that most of the expert comments are rather implicit and even require global knowledge. Nonetheless these comments represent general solutions to reduce the common errors from the baseline system.
3. We encode these comments into features then train a Maximum Entropy(MaxEnt) based automatic assessor Ai using these features. For each response generated from the baseline system, Ai can classify it as correct or incorrect. We choose a statistical model instead of rules because heuristic rules may over fit a small sample set and highly dependent on the order. In contrast, a MaxEnt model has the power of incorporating all comments into a uniform model by assigning weights automatically. In this way we can integrate assessment results tightly with comments during MaxEnt model training.
4. Finally, Ai is applied as a post-processing step to any new data set Di+1, and filters out those results judged as incorrect.

V. CONCLUSION

Knowledge bases are resources that encode world knowledge in machine readable formats. Knowledge base population aims at understanding this knowledge and augmenting knowledge bases with more semantic information.

a.) Statistical language understanding approach is introduced that make knowledge graph from conversational dialog. Personal assertion classification identifies the user utterances that are relevant with facts. Relation detection classification identifies personal assertion utterance into one of the predefined relation classes. Slot filling labels the attribute of relation.

b.) Rich annotation guided learning framework is presented to fill the gap between an expert annotator and a feature engineer.

REFERENCES

[1] Adomavicius, Gediminas and Tuzhilin, Alexander (2005). "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Exten-sions". In: IEEE Trans. Knowl. Data Eng. 176, pp. 734–749. doi: 10.1109/TKDE. 2005.99.

[2] Agarwal, Deepak, Chen, Bee-Chung, and Elango, Pradheep (2009). "Spatio-temporal models for estimating click-through rate". In: Proceedings of the 18th International Con-ference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009, pp. 21–30. doi: 10.1145/1526709.1526713.

[3] Agarwal, Deepak, Chen, Bee-Chung, Elango, Pradheep, and Wang, Xuanhui (2012). "Personalized click shaping through Lagrangian duality for online recommendation". In: The 35th International ACM SIGIR conference on research and development in Informa-tion Retrieval, SIGIR '12, Portland, OR, USA, August 12-16, 2012, pp. 485–494. doi: 10.1145/2348283.234.

[4] Agichtein, Eugene (2006). "Confidence estimation methods for partially supervised rela-tion extraction". In: In SDM 2006

[4] Auer, Sören, Bizer, Christian, Kobilarov, Georgi, Lehmann, Jens, Cyganiak, Richard, and Ives, Zachary G. (2007). "DBpedia: A Nucleus for a Web of Open Data". In:

[5] The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007. Pp. 722–

[6] Bach, Nguyen, Huang, Fei, and Al-Onaizan, Yaser (2011). "Goodness: A Method for Measuring Machine Translation Confidence". In: The 49th Annual Meeting of the As-

[7] Belleau, François, Nolin, Marc-Alexandre, Tourigny, Nicole, Rigault, Philippe, and Moris-sette, Jean (2008). "Bio2RDF: Towards a mashup to build bioinformatics knowledge systems". In: Journal of Biomedical Informatics 4 1.5, pp. 706–716.doi: 10.1016/j.jbi. 2008.03.004.

[8] Bentivogli, Luisa, Clark, Peter, Dagan, Ido, and Giampiccolo, Danilo (2011). "The Sev-enth PASCAL Recognizing Textual Entailment Challenge". In: Proceedings of the

[9] Fourth Text Analysis Conference, TAC 2011, Gaithersburg, Maryland, USA, November 14-15, 2011.

[10] Berger, Adam L., Pietra, Stephen Della, and Pietra, Vincent J. Della (1996). "A Max-imum Entropy Approach to Natural Language Processing". In: Computational Lin-guistics 2 2.1, pp. 39–7 1.

[11] https://www.ijarcsse.com/docs/papers/Volume_5/2_February2015/V5I2-0461.pdf

