Towards a thematic role based target identification model for question answering

Rivindu Perera[1] *Udayangi Perera*[1]
(1) Informatics Institute of Technology, Colombo 06, Sri Lanka
rivindu.perera@hotmail.com, udayangi@iit.ac.lk

ABSTRACT

Target identification plays a crucial role in web based question answering. But still current approaches are not matured enough to extract the exact target of any given question and therefore leads the system to low precision. To address this gap in the current researches we propose thematic role based methodology to extract the target type of the question. Proposed solution is fully wrapped in the shallow semantic processing of the question rather directing it to the deep parsing. Research employs dative alternation of the question thus providing strict rule based approaches to be implemented to elicit the target with high confidence. Furthermore, the proposed solution can be extended with semantically rich target types by mapping concepts identified in question to semantic categories. This extensibility exhibits that our new approach is scalable and can be tweaked to achieve high precision level that current methods are incapable to achieve.

KEYWORDS: Question answering, target identification, shallow semantic processing, thematic roles

Proceedings of the Workshop on Question Answering for Complex Domains, pages 39–46,
COLING 2012, Mumbai, December 2012.

1 Introduction

Question answering is the process of extracting the exact answer for a natural language inspired query which usually lies in the Natural Language Processing (NLP) and the Information Retrieval (IR) domains. To extract the answer with high precision, target of the question must be identified in pre-processing stages. Current approaches used in target identification are based on pattern matching approaches and rule based approaches identified through the usage (Shtok et al., 2012). But drawback noticed in this approach is that such techniques cannot be extended with semantically analyzed structures for target identification.

Due to absence of semantic structures in target identification, question answering process may be subjected to several unseen issues during answer extraction. Among these issues, inability to extract the answer though there is enough information in knowledge base is considered as one of the critical issue to be fixed in future question answering. This issue is placed in even more complex stage when question taxonomies are developed with the use of learning process which extracts question target types while processing questions formed by users (Hartrumpf, 2006). Furthermore, inaccurate target identification can also lead the question answering systems to formulate incorrect answer patterns when presenting the final answer for the user thus leading them to have low confidence rates.

Therefore, we propose a solution where target identification in question answering is powered by identified thematic roles in questions. We design our heuristic in a way that future researches can also incorporate the method by extending the structure with any thematic role that need to be incorporated.

To evaluate this new paradigm we have used Scholar - question answering system (Perera, 2012) which is designed with the proposed target identification method by this research. This paper will unwrap all steps taken to develop this novel method with an empirical viewpoint of each and every approach we have employed during implementation.

2 Background of the study

2.1 Target identification in question answering

Bilotti and Nyberg (Bilotti and Nyberg, 2008) argue that question answering can be taken in to a level that can challenge human abilities only through a better extraction technique which can get the exact answer for the given query. However, in their research which warps around the OpenEphyra question answering system, shows that passage ranking is not the most important task in question answering. Ramakrishnan et al. (Ramakrishnan et al., 2003) also support this concept showing that high quality answer can only be extracted through the proper understanding of the target required by the end user. But Whittaker et al. (Whittaker et al., 2006) bring out that factoid question answering cannot be implemented with a pre-processed set of target types which can be selected by the end user rather this research shows the importance of dynamic target type identification in answer extraction can lead question answering systems to be more flexible and useful when such systems are used in open domain question answering.

Kato et al. (Kato et al., 2006) show a practical target identification method using 4 different target types which are responsible to generate answers using categorization of answer type. Table

1 below, shows the syntactic classification of user utterances and its distribution found by Kato and his team.

Syntactic form	
Wh-type Question	87.7% (544)
Yes-no Question	9.5% (59)
Imperative (Information request)	2.6% (16)
Declarative (Answer to clarification)	0.2% (1)

Table 1: Syntactic classification of user utterances from (Kato et al., 2006)

According to these findings it is noted that Wh-type questions are the main type of questions that any particular question answering systems should be able to answer. But this type of a distribution cannot be considered as accurate in all the scenarios that must be handled through a open domain question answering system. Sacaleanu & Neumann (Sacaleanu and Neumann, 2006) show that in cross-language question answering, target of the question cannot be determined by simple rule based approach rather need to be analysed thoroughly through semantically rich aspects.

2.2 Thematic roles

Pighin et al. (Pighin et al., 2007) introduce a two-steps supervised strategy for the identification and classification of thematic roles. In this approach presented by Pighin and his team, wide variety of themes are considered providing better overview of the recognition of thematic roles and classification in a complex and wide area of natural text. However this research does not employ the verb sense information in classification stage. Therefore, in a question answering system this approach cannot be used with original structure as question answering needs verbs to be defined with high precision considering the sense they provide.

Liu and Soo (Liu and Soo, 1993) carried out a research in the area of knowledge acquisition considering thematic role based approach. In this novel method proposed and evaluated by this research, syntactic clues are incorporated to get the exact role to the acquisition phase. But the drawback noticed in this research is that need of extensive syntactic resources to determine the knowledge to be acquired. Therefore when applied to a question answering system this method should be trained with large amount data to make this heuristic available for all sorts of questions.

3 Method

In our approach target identification is entirely based on the thematic role identified which shows the type of the answer to be extracted. This novel paradigm is also inspired from the research carried out by Yang et al. (Yang et al., 2006) which introduces contextual question answering using relevancy recognition. But to transform this question answering process to a flexible state we also introduce the method that users are given the chance to select the thematic role that they need. However, if such thematic role is absence terms used in the question, its structure and the semantic representation are considered to extract the thematic role.

3.1 Thematic role identification

In the target identification process the first task is to identify the thematic role to be identified which later transformed in to a target type. In our approach, seven different thematic roles are incorporated and these are listed in Table 2 with their applicability in the question context. These thematic roles are inspired from the seminal work carried out by Jurafsky & Martin (Jurafsky and Martin, 2000).

Thematic role	Applicability
Agent	To get the agent role of a question. This may incorporate any object type if specified object is involved in the act playing the role of agent. Ex: *Who* found the Google?
Instrument	If the question is related with an event, instruments used in the event are classified under this role What is the *chemical substance* he used to make NaOH?
Goal	Goal thematic role can show any type of a objective such as a location, event or some other result which is carried out to invoke a different type of an event To *where* he travelled?
Patient	Object type of a event is categorized under this thematic role Ex: What *company* did Sergy Brin start?
Beneficiary	Beneficiary of a question is the person or thing that gets some benefit from the event. For *whom* he made the aircraft?
Source	When questions are associated with transfer events, then origin of the subjected object is considered as source. *Where* did he come from?
Result	When questions are associated with result of an event. Ex: *What* did he build?

Table 2: Thematic roles

To identify the thematic role of a question, we employ rule based approach determined by the considered set of thematic roles. As the first task question is represented in a tagged form using Hidden Markov Model (HMM) tagger. Reason behind to use this stochastic tagging procedure is that HMM tagger selects the best sequence of tags for the entire question processed(Jurafsky and Martin, 2000). Basically, bigram-HMM tagger we employed will therefore assign the tag considering the sequence as a whole as expressed in fundamental theorem in (1),

$$t_i = \arg\max_j P(t_j|t_{i-1})P(w_i|t_j) \tag{1}$$

where t_i represents current tag to be determined and w_i as the current word considered. But we have also considered several other approaches like Brill tagging as well. But earlier mentioned reason inspired us to utilize this HMM tagging. Tagged question is used to invoke the basic analysis of the structure of the question. But our main task of thematic role assignment is done via predefined model which consume the tagged question to map the appropriate model. Simply, once question is tagged with appropriate tag sets, it is easy identify what lexical context it represents as a formal description is available for the question. This formal description is used to select the thematic role from predefined collection of thematic role to abstract formal description

matching. Once the thematic role is identified it is associated with the specified question to support the answer extraction process.

3.2 Thematic role assignment and metadata processing

Identified thematic role will be assigned to the specified question showcasing the answer type required to be extracted. But with the thematic role several other metadata can also be attached to the question to make the answer extraction process more accurate and fast. If thematic role required represent any type of supported named entity them the named entity type will also be attached to the question. For an example for a question like "who is the founder of Google" will be assigned with the "agent" thematic role. But in question processing it can be identified that this agent type is actually mapped to a "person" named entity type. Therefore, rather assigning the generic theme of agent as a metadata representation "person" named entity will also be attached to the question to support answer extraction by reducing the search space. We currently consider six such types of named entities in our approach, person, location, currency, city, date and organization.

3.3 Answer extraction

When the thematic role is assigned to a question, answer extraction process can be stated focusing answers which represent the type required by the thematic role and which are compatible with the named entity type specified. After the extraction process, confidence level can be assigned to the extracted answer by analyzing the compatibility that answer carries with thematic role and metadata associated with the question being processed.

4 Results and discussion

To evaluate the proposed novel approach, we employ 280 questions from past TREC (Voorhees, 2001) series (TREC-8 and TREC-9). We manually categorized these 280 questions into 7 main classes representing all major thematic roles we are defining in this research. Important factor we have noticed is that for some thematic roles, population of questions is not sufficient. But as TREC is defining its own standard of question formulation and as future researches in the same track need to compare result with our approach, we have used the original collection without adding our own questions to populate classes with fewer questions.

Question class based on thematic role	With correct target	With incorrect target	Correctly answered
Agent	68	3	62
Instrument	58	6	51
Goal	10	9	6
Patient	51	5	43
Beneficiary	12	3	8
Source	23	1	22
Result	24	7	17
Total	246	34	219

Table 3: Evaluation result using TREC question set

Table 3 expresses the result of evaluation expressing three factors, the number of questions with correctly identified targets, the number of questions with incorrectly identified targets and number of questions where correct answers are acquired using identified thematic role.

According to the evaluation results it is noted that systems have achieved 78.20% average accuracy level considering correctly answered questions. When comparing with other systems which are tested with same TREC question sets it can be determined that this accuracy level is better than such system have achieved (Zheng, 2002) (Voorhees, 2003). But importantly it can be noticed that error rate of target identification is lying in the 12.14% which is quite acceptable and therefore shows high accuracy level in target identification.

Though our approach has shown excellent accuracy as an average rate, it can be clearly identified that for some individual thematic roles, low accuracy levels are also displayed. According to our preliminary analysis of this behaviour several reasons are uncovered. Firstly, target identification greatly depends on the steady structure of the questions. This encompasses that if question structure is leading to the answer, for an example through agent type or patient type, then it is easy to assign thematic role rather mining it deeper. Furthermore, it is found that short questions which can be directly formed into a grammatical representation can ended up with high accuracy levels in thematic role assignment.

5 Conclusion and future work

In this paper we illustrated an approach to determine the target type of a question by analyzing the thematic role of the question to be processed. As thematic roles are based on the semantic representation of the natural text this approach can be extended to support several semantic processing tasks. Furthermore, in several stages we have employed rule based approaches to process the question as probabilistic approaches cannot be applied with semantic representation with high accuracy.

To evaluate this novel heuristic we have used the question answering system- Scholar which uses the same strategy to identify the target. During evaluation we achieved excellent accuracy which inspires us to develop this model as an independent library to incorporate with other question answering systems. In future our focus is entirely placed on the implementation of this heuristic as a library and to apply several other semantic processing methodologies to increase the accuracy level of this novel paradigm.

References

Bilotti, M.W., Nyberg, E., 2008. Improving text retrieval precision and answer accuracy in question answering systems, Coling 2008: Proceedings of the 2nd workshop on Information Retrieval for Question Answering. Association for Computational Linguistics, Manchester, UK, pp. 1-8.

Hartrumpf, S., 2006. Adapting a semantic question answering system to the web, Proceedings of the Workshop on Multilingual Question Answering. Association for Computational Linguistics, pp. 61-68.

Jurafsky, D., Martin, J.H., 2000. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall PTR.

Kato, T., Masui, F., Fukumoto, J.i., Kando, N., 2006. WoZ simulation of interactive question answering, Proceedings of the Interactive Question Answering Workshop at HLT-NAACL 2006. Association for Computational Linguistics, New York City, NY, pp. 9-16.

Liu, R.-L., Soo, V.-W., 1993. An empirical study on thematic knowledge acquisition based on syntactic clues and heuristics, Proceedings of the 31st annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, Columbus, Ohio, pp. 243-250.

Perera, R., 2012. Scholar: Cognitive Computing Approach for Question Answering, Department of Computer Science, Informatics Institute of Technology. University of Westminster.

Pighin, D., Moschitti, A., Basili, R., 2007. RTV: tree kernels for thematic role classification, Proceedings of the 4th International Workshop on Semantic Evaluations. Association for Computational Linguistics, Prague, Czech Republic, pp. 288-291.

Ramakrishnan, G., Jadhav, A., Joshi, A., Chakrabarti, S., Bhattacharyya, P., 2003. Question Answering via Bayesian inference on lexical relations, Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering - Volume 12. Association for Computational Linguistics, Sapporo, Japan, pp. 1-10.

Sacaleanu, B., Neumann, G., 2006. Cross-cutting aspects of cross-language question answering systems, Proceedings of the Workshop on Multilingual Question Answering. Association for Computational Linguistics, pp. 15-22.

Shtok, A., Dror, G., Maarek, Y., Szpektor, I., 2012. Learning from the past: answering new questions with past answers, Proceedings of the 21st international conference on World Wide Web. ACM, Lyon, France, pp. 759-768.

Voorhees, E.M., 2001. Question answering in TREC, Proceedings of the tenth international conference on Information and knowledge management. ACM, Atlanta, Georgia, USA, pp. 535-537.

Voorhees, E.M., 2003. Evaluating the evaluation: a case study using the TREC 2002 question answering track, Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1. Association for Computational Linguistics, Edmonton, Canada, pp. 181-188.

Whittaker, E.W.D., Hamonic, J., Yang, D., Klingberg, T., Furui, S., 2006. Monolingual web-based factoid question answering in Chinese, Swedish, English and Japanese, Proceedings of the Workshop on Multilingual Question Answering. Association for Computational Linguistics, pp. 45-52.

Yang, F., Feng, J., Fabbrizio, G.D., 2006. A data driven approach to relevancy recognition for contextual question answering, Proceedings of the Interactive Question Answering Workshop at HLT-NAACL 2006. Association for Computational Linguistics, New York City, NY, pp. 33-40.

Zheng, Z., 2002. AnswerBus question answering system, Proceedings of the second international conference on Human Language Technology Research. Morgan Kaufmann Publishers Inc., San Diego, California, pp. 399-404.