Construction of Vietnamese Argument Annotated Dataset for Why-Question Answering Method

Chinh Trong Nguyen and Dang Tuan Nguyen(✉)

Faculty of Computer Science, University of Information Technology, VNU-HCM,
Ho Chi Minh City, Vietnam
{chinhnt,dangnt}@uit.edu.vn

Abstract. In this paper, the method of building a Vietnamese Argument Annotated Dataset (VAAD) is presented. This dataset contains argumentative data which can be used to answer the why-questions. Therefore, it is important to discover the characteristics of the answers of why-questions to develop why-question answering method by using causal relations between texts. In addition, this dataset can be used to generate the testing dataset for evaluation of answering method. In order to build the dataset, a process of four steps is proposed after studying relevant problems. To briefly evaluate the method, an experiment is conducted to show the applicability of the method in practice.

Keywords: Discourse analysis · Why-question answering · Vietnamese Argument Annotated Dataset

1 Introduction

At present, the development of question answering systems for Vietnamese language can be founded on researched solutions of answering the factoid questions [13–16]. These solutions are mostly based on knowledge mining techniques therefore they need a large annotated corpus to train, to evaluate and to develop.

Although why-questions are rarely asked, 5 % of all questions asked according to the observation of Hovy [1], they seem to be the important type of question because their answers, found by causal relations in discourse structures instead of the bag of words in texts, provide the reasons about problems. Therefore, building a why-question answering (why-QA) system for Vietnamese language has been conducted. However, the Vietnamese corpus for researching why-question answering methods is lacked. Although TREC has developed testing datasets for question answering systems for many years, the datasets mostly contain factoid questions and they are written in English. At present, it is important to build a large Vietnamese annotated dataset for researching and testing why-QA.

For the above reasons, a Vietnamese Argument Annotated Dataset (VAAD) for why-questions should be built to develop why-QA answering methods. The dataset should be suitable for developing many answering methods and evaluation. In this paper, the process of building VAAD for why-questions is presented in five sections. Section 1 introduces the exigence of developing VAAD. Section 2 explores some problems related

P.C. Vinh et al. (Eds.): ICTCC 2016, LNICST 168, pp. 124–132, 2016.
DOI: 10.1007/978-3-319-46909-6_12

to building the dataset. According to these problems, the annotation format of Vietnamese VAAD and the building process is presented Sect. 3. Then, the experiment of building the dataset is presented in Sect. 4. At the end, some conclusions are drawn in Sect. 5.

2 Related Works

The methods of question answering can be divided into two approaches that are knowledge mining, as in [13–16], and knowledge annotation, as in [17]. The methods based on knowledge mining techniques have the advantage of information redundancy from the internet. The redundancy of information can be utilized to propose question answering methods which do not need to use complex natural language processing techniques. Therefore, many researches in question answering have focused on this approach.

According to the knowledge mining approach, developing question answering methods need large datasets to discover the patterns which are used to find the candidate answers. These datasets are also used to test the question answering methods. These datasets should be not only collected but also annotated into a specific format. The format of a dataset depends on the feature analyzed by the researching methods. For example, Saint-Dizier's dataset in [12] is annotated by using Rhetorical Structure Theory (RST) [7] because the question answering method is based on the argumentation which is identified in discourse structure of the document.

In why-QA, the question answering method can be divided into two types: cue-based method and discourse-based method. The cue-based methods are developed with clues as in [11] or with cue words and paragraph retrieval techniques as in [2]. They have the simplicity in analysis but the results are quite low because the semantic features have not been analyzed yet. In contrast, the discourse-based methods are developed with discourse structure of the document as in [4–6, 12]. In this type, the methods have to use the context of the sentences in a document to build the relations between them. These relations express the intention of the writer. Among these relations, the causal relations between sentences form the writer's argument structures. The discourse-based methods need more complicated analysis but their results are more relevant to the questions than the cue-based ones. Despite of the differences, these types of answering method need why-QA datasets for training and testing. These datasets have to be built for each research project because there are no appropriate dataset for all purposes.

In discourse structure of document, there are two approaches of representation. In the RST representation [7], a document is a "tree of spans". Each span, which can be a clause, a sentence or a paragraph, links to another span following rhetorical relations to form a larger span. These spans are still presented in text therefore they are easy to search. In the Discourse Representation Theory (DRT) [18], a document is a set of Discourse Representation Structure (DRS) which is a group of first-order logic expressions. These representations can be used to reason in order to find new information, however it is complex to build a set of DRS from a document in natural language.

In other aspect of discourse structure, the visual structure of a document also affects its discourse structure as Power shown in [8].

3 Building VAAD for Developing Why-QA Method

The purpose of building the VAAD is to develop why-QA methods. These methods can be cue-based or discourse-based approaches therefore the dataset should be annotated in a simple format so that it can be used easily. In addition, the dataset can be used to generate testing sets by transforming the result parts of causal relations in to why-questions. For example, the causal relation "Tom is not allowed to ride a bicycle because Tom is young" has the result part "Tom is not allowed to ride a bicycle". Thus, a why-question "why Tom is not allowed to ride a bicycle?" can be built by transforming the result part. In order to make more complex why-questions, synonyms or similar semantic phrases can be used to expand the original result parts.

The process of building VAAD dataset has four steps that are documents collecting, argument annotating, patterns extracting and argument annotated fragments collecting.

3.1 Documents Collecting

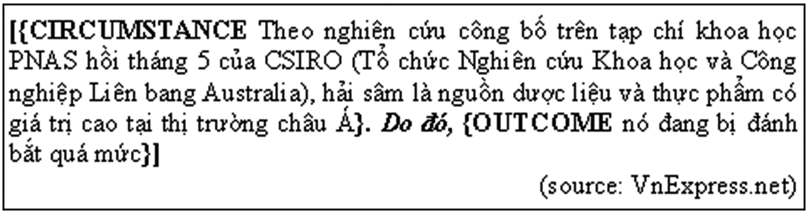
During the process of collecting documents containing arguments, the observations show that there are many news posts or comments without any arguments in them. These news posts or comments are often about new products, instructions, sports news. In order to collect documents containing arguments, Google[1] is used to search for document containing phrases which are more likely to appear in an argument, such as "tại sao" ("why"), "công dụng của" ("the use of"), "hạn chế của" ("the disadvantages of"). Then, the links in google search results are extracted and used to download the origin web pages. After that, the scripts, banners, etc. of the web pages are eliminated and the texts of main content of the web pages are extracted. These texts form a dataset for annotating in the next step.

3.2 Argument Annotating

According to the simplicity of the RST representation, the dataset is annotated follow these rules:

- All spans which are not in any argument are unchanged.
- Spans, which are in a certain argument, are place in a pair of symbols "[" and "]"
- A span which is an argument is annotated as follow: causal part and result part are place in a pair of symbols "{" and "}" in which they follow a notation of their role in the argument; the cue phrase which informs the type of causal relation is unchanged. Figure 1 illustrates an annotated argument fragment.

[1] https://www.google.com.

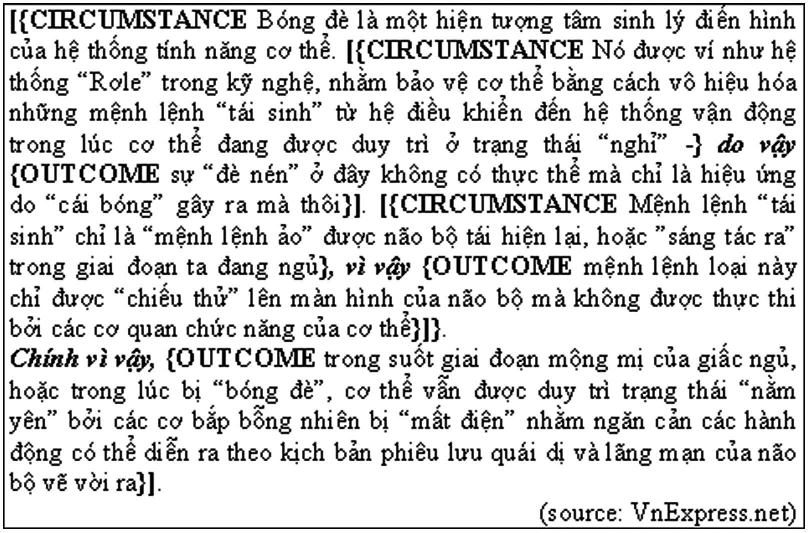


[{**CIRCUMSTANCE** Theo nghiên cứu công bố trên tạp chí khoa học PNAS hồi tháng 5 của CSIRO (Tổ chức Nghiên cứu Khoa học và Công nghiệp Liên bang Australia), hải sâm là nguồn dược liệu và thực phẩm có giá trị cao tại thị trường châu Á}. ***Do đó,*** {**OUTCOME** nó đang bị đánh bắt quá mức}]

(source: VnExpress.net)

Fig. 1. A structure of an argument annotated fragment. The bold words are the roles of two parts in a causal relation (CIRCUMSTANCE - OUTCOME). The bold, italic words, "Do đó" (therefore) is a cue phrase indicates the circumstance - result relation.

– An argument can be a part of another argument as shown in Fig. 2.



[{**CIRCUMSTANCE** Bóng đè là một hiện tượng tâm sinh lý điển hình của hệ thống tính năng cơ thể. [{**CIRCUMSTANCE** Nó được ví như hệ thống "Role" trong kỹ nghệ, nhằm bảo vệ cơ thể bằng cách vô hiệu hóa những mệnh lệnh "tái sinh" từ hệ điều khiển đến hệ thống vận động trong lúc cơ thể đang được duy trì ở trạng thái "nghỉ" -} ***do vậy*** {**OUTCOME** sự "đè nén" ở đây không có thực thể mà chỉ là hiệu ứng do "cái bóng" gây ra mà thôi}]. [{**CIRCUMSTANCE** Mệnh lệnh "tái sinh" chỉ là "mệnh lệnh ảo" được não bộ tái hiện lại, hoặc "sáng tác ra" trong giai đoạn ta đang ngủ}, ***vì vậy*** {**OUTCOME** mệnh lệnh loại này chỉ được "chiếu thử" lên màn hình của não bộ mà không được thực thi bởi các cơ quan chức năng của cơ thể}]}.

Chính vì vậy, {**OUTCOME** trong suốt giai đoạn mộng mị của giấc ngủ, hoặc trong lúc bị "bóng đè", cơ thể vẫn được duy trì trạng thái "nằm yên" bởi các cơ bắp bỗng nhiên bị "mất điện" nhằm ngăn cản các hành động có thể diễn ra theo kịch bản phiêu lưu quái dị và lãng mạn của não bộ vẽ vời ra}].

(source: VnExpress.net)

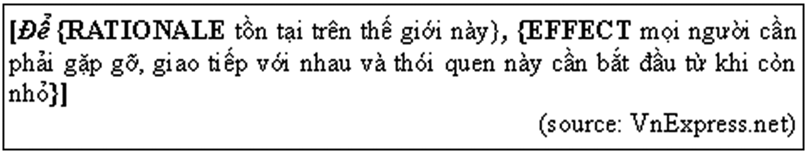
Fig. 2. An argument can be a part of another argument. In this figure, the first paragraph is the causal part and the second paragraph is the result part of an argument. There are two arguments in the first paragraph.

By using these rules, the arguments in document are easy to extract. In addition, if there is any further language analysis needed, it can be applied easily to discover more precise patterns. In this format, the causal relations in RST is divided into four types according to [4]: rationale - effect, purpose - outcome, circumstance - outcome and means - outcome.

3.3 Patterns Extracting

After identifying arguments by annotating the causal relations. The patterns containing cue phrases and some specific marks such as periods, commas, new-lines are also identified. A causal relation can be an inner-sentence, an inter-sentence or an inter-paragraph relation.

In an inner-sentence relation, as in Fig. 3 all parts of the relation are bounded in two periods and they do not contain any period. In an inter-sentence relation, as in Fig. 1 above, there is only one period; and in an inter-paragraph relation, as in Fig. 2 above, there are one more new-line symbols.



[***Để*** {**RATIONALE** tồn tại trên thế giới này}, {**EFFECT** mọi người cần phải gặp gỡ, giao tiếp với nhau và thói quen này cần bắt đầu từ khi còn nhỏ}]

(source: VnExpress.net)

Fig. 3. The inner-sentence relation in which all parts of the relation are bounded in two periods and there is no period in all parts of the relation.

In this step, the cue phrases are used as core feature to identify the argument because the cue phrase have stably meaning of discourse function as shown in [7, 9]. Therefore, the patterns are manually identified and used to extract arguments having the same patterns in websites to enrich the dataset.

3.4 Argument Annotated Fragments Collecting

By using the patterns discovered in step 3, a crawler is used to fetch the news posts on websites to extract the argument annotated fragment. By using the crawler, the process of building VAAD is reduced greatly in cost of manually collecting and annotating. However, this method has a disadvantage of not collecting arguments of new patterns. The extracted arguments of collected news posts are automatically annotated with the proposed format according to the patterns which are used to extract them.

4 Experiment

In order to evaluate the method of building VAAD, 34 articles are collected according to step 1 and annotated as describing in step 2. Then, the 49 argument fragment patterns, as shown in Table 1 are manually identified. Then, these patterns are represented in regular expressions to collect argument fragments.

After identifying argument fragment patterns, a set of 608 articles downloaded from internet using crawler are process with the patterns to generate 2609 fragments. The cue phrases associated with these fragments are presented in Table 2 to show which cue phrases are frequently used. In order to evaluate the precision of the argument identification method, 250 fragments are randomly selected in 2609 fragments. These 250 fragments are then manually check if they are argument fragments. After checking, there are 195 fragments are argument fragments which yield the precision of 0.78.

Table 1. The list of manually identified cue phrases.

Phrase	Relation type
... . Vì vậy,	inter-sentence
... . Bởi vậy, ...	inter-sentence
... . Vì thế ...	inter-sentence
... . Điều này làm cho ...	inter-sentence
... Do đó, ...	inter-paragraph
... do ...	inner-sentence
Nhờ ..., ...	inter-sentence
... .Thế nên	inter-sentence
... . Kết quả ...	inter-sentence
... .Vì vậy ...	inter-sentence
... . Do vậy, ...	inter-sentence
Để ..., ...	inner-sentence
..., chính vì vậy ...	inner-sentence
... . Do vậy ...	inter-sentence
... do vậy ...	inner-sentence
... . Vì lẽ đó, ...	inter-sentence
... là nguyên nhân chính dẫn tới ...	inner-sentence
Do ... mà ...	inner-sentence
... . Điều này khiến ...	inter-paragraph
... là do ...	inner-sentence
... cho nên ...	inner-sentence
..., do vậy ...	inner-sentence
... Chính vì vậy, ...	inter-paragraph
... Vậy, ...	inter-paragraph
... dẫn đến ...	inner-sentence
... vì ...	inner-sentence
... , vì vậy ...	inner-sentence
... . Điều này dẫn đến ...	inter-sentence
... . Đây là lý do ...	inter-sentence
... , đây là lý do tại sao ...	inner-sentence
Bởi vì ... nên ...	inner-sentence
... là nhờ ...	inner-sentence
Nguyên nhân ... do ...	inner-sentence
Với ... , ...	inner-sentence
Nhờ ... mà ...	inner-sentence
... để ...	inner-sentence
... với mục đích ...	inner-sentence
... nên ...	inner-sentence
... gây ...	inner-sentence
... Như vậy, ...	inter-paragraph
... ảnh hưởng tới ...	inner-sentence
Vì ... nên ...	inner-sentence
Bởi ... , ...	inner-sentence
... . Và đó là lý do ...	inter-sentence
để ... thì ...	inner-sentence
... cho thấy ...	inner-sentence
... khiến ...	inner-sentence
... bằng cách ...	inner-sentence
... bởi ...	inner-sentence

Table 2. The list of cue phrases used to extract 2609 fragments and their number of use.

Phrase	Number of use
... để ...	923
... do ...	328
... nên ...	277
... vì ...	240
... khiến ...	158
... gây ...	163
... bởi ...	114
... cho thấy ...	91
... . Vì thế, ...	69
... nhằm ...	55
... biến thành ...	47
... bằng cách ...	30
... . Kết quả ...	21
... dẫn đến ...	21
... ảnh hưởng đến ...	21
... . Do đó ...	14
... . Vì vậy, ...	13
Để ... , ...	6
... , vì thế ...	3
... . Nhờ đó, ...	3
... là nhờ ...	3
... làm cho ...	3
... với mục đích ...	3
... . Do vậy, ...	1
... cho nên ...	1
... nguyên nhân chính ...	1

The reasons of the wrong identifying argument fragments are the ambiguity of the cue phrase and the misidentifying inter-paragraph relation. The ambiguity of cue phrase such as, “để” (in order to) and “để” (to put), can be overcome by POS tag process before identifying patterns and extracting argument fragments. The misidentifying inter-paragraph relation is more difficult to overcome. It requires a completely RST structure of the document to identify which paragraphs form a span in RST. However, the number of inter-paragraph argument fragments collected are not very large. Therefore this method can be used to build VAAD for developing a why-QA method.

The experiment result shows that the proposed method can be applied in practice with the higher precision by applying POS tagging task.

5 Conclusions and Future Works

In this paper, the research on building VAAD for developing why-QA method is presented. This dataset is important to find out the characteristics of argument of text fragments to answer the why-questions in Vietnamese. In addition, the testing dataset for why-QA method can be generated from this dataset. The testing dataset is also important to evaluate the answering method. Because the arguments are some kinds of

RST relations, this paper proposes a method of automatically identifying argument fragments from news posts in the internet using cue phrases. The cue phrases are used in this method because their linguistic functions of discourse are stable. Therefore, the process of four steps which are collecting documents, argument annotating, patterns extracting and argument annotated fragments collecting is proposed to build the dataset.

According to the proposed process, an experiment has been conducted and it shows that the process can be apply to automatically build the practical VAAD for developing why-QA method after POS tagging the documents for extracting patterns and collecting argument fragments.

In future, Vietnamese RST parser should be developed to overcome the misidentifying inter-paragraph causal relation to enrich VAAD.

References

1. Hovy, E.H., Hermjakob, U., Ravichandran, D.: A question/answer typology with surface text patterns. In: 2nd International Conference on Human Language Technology Research, California, pp. 247–251 (2002)
2. Verberne, S., Boves, L., Oostdijk, N., Coppen, P.: Using syntactic information for improving why-question answering. In: 22nd International Conference on Computational Linguistics, Manchester, United Kingdom, pp. 953–960 (2008)
3. Verberne, S.: Developing an approach for why-question answering. In: 11th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, Trento, Italy, pp. 39–46 (2006)
4. Delmonte, R., Pianta., E.: Answering why-questions in closed domains from a discourse model. In: Conference on Semantics in Text Processing, pp. 103–114. ACL, Stroudsburg (2008)
5. Oh, J., Torisawa, K., Hashimoto, C., Sano, M., Saeger, S. D.: Why-question answering using intra- and inter-sentential causal relations. In: 51st Annual Meeting of the Association for Computational Linguistics, pp. 1733–1743. ACL Anthology, Sofia (2013)
6. Higashinaka, R., Isozaki, H.: Corpus-based question answering for why-questions. In: 3rd International Joint Conference of Natural Language Processing, Hyderabad, India, pp. 418–425 (2008)
7. Mann, W.C., Thompson, S.A.: Rhetorical structure theory: towards a functional theory of text organization. Text **3**(8), 243–281 (1988)
8. Power, R., Scott, D., Bouayad-Agha, N.: Document Structure. Comput. Linguist. **29**(2), 211–260 (2003)
9. Marcu, D.: The rhetorical parsing of natural language texts. In: 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics, pp. 96–103. ACL, Stroudsburg (1997)
10. Hwee, T.N., Leong, H.T., Lai, J.P.K.: A machine learning approach to answering questions for reading comprehension tests. In: The 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, pp. 124–132. ACL, Stroudsburg (2000)
11. Riloff, E., Thelen, M.: A rule-based question answering system for reading comprehension tests. In: The 2000 ANLP/NAACL Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems, pp. 13–19. ACL, Stroudsburg (2000)

12. Saint-Dizier, P.: Processing Natural Language Arguments with the <TextCoop> Platform. Argument Comput. **3**(1), 49–82 (2012). Taylor and Francis
13. Zheng, Z.: AnswerBus question answering system. In: The 2nd International Conference on Human Language Technology Research, pp. 399–404. Morgan Kaufmann Publishers Inc., San Francisco (2002)
14. Clarke, C., Cormack, G., Kemkes, G., Laszlo, M., Lynam, T., Terra, E., Tilker, P.: Statistical selection of exact answers (multitext experiments for TREC 2002). In: TREC, pp. 823–831. NIST (2002)
15. Brill, E., Dumais, S., Banko, M.: An analysis of the AskMSR question-answering system. In: The ACL 2002 Conference on Empirical Methods in Natural Language Processing, pp. 257–264. ACL, Stroudsburg (2002)
16. Buchholz, S., Daelemans, W.: Shapaqa: shallow parsing for question answering on the world wide web. In: Euroconference Recent Advances in Natural Language Processing, Tzigov Chark, Bulgaria, pp. 47–51 (2001)
17. Katz, B., Felshin, S., Yuret, D., Ibrahim, A., Lin, J.J., Marton, G., McFarland, A.J., Temelkuran, B.: Omnibase: uniform access to heterogeneous data for question answering. In: Andersson, B., Bergholtz, M., Johannesson, P. (eds.) NLDB 2002. LNCS, vol. 2553, pp. 230–234. Springer, Heidelberg (2002)
18. Kamp, H.: Discourse representation theory. In: Gabbay, D., Guenthner, F. (eds.) Handbook of Philosophical Logic, vol. 15, pp. 125–394. Springer, Netherlands (2011)