Domain Specific Automatic Question Generation from Text

Katira Soleymanzadeh
International Computer Institute
Ege University
Izmir, Turkey
Katirasole@gmail.com

Abstract

The goal of my doctoral thesis is to automatically generate interrogative sentences from descriptive sentences of Turkish biology text. We employ syntactic and semantic approaches to parse descriptive sentences. Syntactic and semantic approaches utilize syntactic (constituent or dependency) parsing and semantic role labeling systems respectively. After parsing step, question statements whose answers are embedded in the descriptive sentences are going to be formulated by using some predefined rules and templates. Syntactic parsing is done using an open source dependency parser called MaltParser (Nivre et al. 2007). Whereas to accomplish semantic parsing, we will construct a biological proposition bank (BioPropBank) and a corpus annotated with semantic roles. Then we will employ supervised methods to automatic label the semantic roles of a sentence.

1 Introduction

"Cognition is the mental action or process of acquiring knowledge and understanding through thought, experience, and the senses." (Stevenson, 2010) . Thought is triggered by asking questions and attempt to find answer of questions cause knowledge acquisition. Researches indicate that questioning is a powerful teaching technique. Lecturers benefit from questions for students' knowledge evaluation, student's stimulation to thinking on their own and encourage students to self-learning. Also, students can review and memorize information previously learned by questioning themselves.

Generating questions manually need much time and effort for lecturers. Moreover, student face considerable problems exercising and memorizing lessons. To address these challenges, Automatic question generation (AQG) systems can provide sample questions to alleviate lecturer's effort and help students in self-learning.

Our motivation in generating questions automatically is to facilitate lecturer effort and help students to practice on course materials more efficiently. Our goal in my thesis is building a system for question generation from Turkish biological text. We take biology text as input of our system and generate questions which will rank based on questions quality.

AQG is one of the challenging problems in natural language processing especially when semantic analysis is needed to generate comprehensive questions like how and why. To the best of our knowledge, AQG approaches in Turkish have been proposed by Cabuk et al. (2003) and Orhan et al. (2006). Both of these studies just have used syntactic approach without any semantic analysis for generating questions. However, generating questions from biological text, which contain complex process, cannot rely on syntactic approach merely. Relation between entities in a biological process make it difficult to analyze in syntactic level. Understanding these process needs some level of semantic analysis. In my proposal, we plan to generate comprehensive questions like how and why in addition to when, where, who and whom. Therefore, we need syntactic and semantic analysis of descriptive sentences.

Syntactic analysis of a sentence determines the structure of phrases of a text and converts it into a more structured representation, the parse tree. Characterizing "who" did "what" to "whom," "where," "when," "how" and "why" is semantic analysis of a sentence. Semantic role labeling (SRL) is a task of automatically identifying

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics- Student Research Workshop, pages 82–88
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-3014

semantic relations between predicate and its related arguments in the sentence. Assigning pre-defined set of semantic roles such as Agent, Patient and Manner to arguments is defined as predicate-argument structure (PAS) identification problem.

Lexical resources like PropBank (Palmer et al. 2005) and FrameNet (Baker et al. 1998) are needed to label semantic role of arguments. The Turkish lexical semantic resource (TLSR) were built by Isguder Şahin and Adalı (2014). TLSR is in general domain and does not cover biological field. Moreover, size of TLSR is small compared to PropBank in other languages. At present the number of annotated verb frame and sense are 759 and 1262 respectively. Domain sensitivity of SRL systems have been emphasized by many researchers (Albright et al. 2013; Carreras & Màrquez 2005; Johansson & Nugues 2008; Pradhan et al. 2008). Pradham et al. (2008) showed that the performance of SRL systems dropped dramatically by almost 10% when domain of testing data is different from training data. Albraight et al (2013) indicated the accuracy enhancement of SRL systems with the existence of in-domain annotations of data. Therefore, to automatically generating questions from biological text using semantic parsing, we first need to build an SRL system in the biological domain. To this end we will construct a lexical resource for the biology domain along with a corpus annotated with semantic roles in semi-automatic manner. Furthermore, there is not automatic SRL system in Turkish yet. So, we plan to design a supervised SRL system too.

In AQG step, we parse descriptive sentence using syntactic and semantic parser. Automatic SRL system which will construct in the first phase of my thesis, will employ to parse descriptive sentence semantically. Syntactic parsing of descriptive sentence will do by an open source dependency parser called MaltParser (Nivre et al. 2007). Semantic role labels and syntactic tags will use to identify content to generate relevant question (i.e. if semantic role label is “Arg0” then the question type will be “who”). In the question formation step, some predefined rules and template will utilize. The quality of the generated questions will measure based on its syntactic and semantic correctness and its relevancy to the given sentence.

2 Background

In order to generating interrogative sentences from descriptive sentences, syntactic and semantic approaches are taken. Constituency or dependency parser are used to parse a descriptive sentence in syntactic approach. Afterward, with respect to the label of phrase, appropriate type of question is selected. There are several AQG system that have utilized syntactic approach. Mitkov et al. (2006) proposed multiple choice question generation system to assess students’ grammar knowledge by utilizing syntactic approach. Heilman and Smith (2009) described a syntactic and rule based approach to automatically generate factual questions to evaluate students’ reading comprehension. Liu et al. (2012) developed template based AQG system by using syntactic approach, called G-Asks, to improve students’ writing skill. Cabuk et.al. (2003) employed a syntactic parser to get stem, derivational and inflectional affixes of words of sentence. Predefined rules were used to identify phrases of sentence. In the last step questions were generated by transforming rules based on identified phrases of previous step. Orhan et al. (2006) generate template based math questions for students of elementary school.

In order to generate questions using semantic approach, semantic role of arguments is labeled firstly. Then proper question type is selected according to the semantic labels. Mannem et al. (2010) utilized SRL and Named Entity Recognition (NER) system to generate rule based questions. Lindberg et al. (2013) generated template based questions for educational purpose by using a semantic approach. By the use of a semantic approach, Mazidi and Nielsen (2014) generated questions in specific domains such as chemistry, biology and earth science. After analyzing text by the SRL and constituency parsing system, relevant questions are generated based on predefined templates.

Lecturer assess students’ reading comprehension by utilizing questions. Generating pedagogical questions are time consuming and a lot of lecturer effort is needed. The main goal of my thesis is to automatically generate question using both of syntactic and semantic approach to alleviate these efforts. To the best of our knowledge, generating questions by employing semantic approach will do for the first time in Turkish. My thesis is similar to Mazidi and Nielsen’s work in terms of utilizing

semantic approach but is different in question formation step.

Due to the need for an SRL system in semantic question generation systems, we plan to design a supervised SRL system. Supervised, unsupervised and semi-supervised machine learning methods are applied in building SRL systems. In supervised method, after extracting features from training data, a 1-N (N is number of roles), a classifier (such as support vector machine (SVM), Maximum entropy (MaxEnt) and Naïve Bayes (NB)) is used to label semantic roles. Garg and Henderson (2012) used Bayes method to SRL where dependency parses are used to extract features. Albright et al. (2013) constructed a corpus annotated with semantic roles of clinical narratives that is called MiPAQ. Monachesi et al. (2007) extracted features from dependency parser to use in supervised K-nearest neighbor algorithms to SRL.

In semi-supervised methods, a small amount of data is annotated with their semantic roles that is called seed data. The classifier is trained using the seed data. Unlabeled data is classified using this system and the most confident predictions are added to expand the initial training data. This expansion is carried out iteratively a few times. Semi-supervised self-training and co-training methods were used in many SRL research (Do Thi et al. 2016; Kaljahi & Samad 2010; Lee et al. 2007) recently and they showed their performance in in-domain data. In those study standard supervised algorithms was used as classifier and the features were extracted by constituency parser.

The features extracted from constituency parses defined by Gildea and Jurafsky (2002) are used as basic features in most SRL system. Predicate, phrase type, headword, constituency parse path, phrase position and voice of predicate are some basic features. They mentioned that using syntactic parses is necessary for extracting features.

A role-annotated corpus together with lexical resources in PropBank and FrameNet, are used as training data in many supervised SRL systems in English. Semantic roles of all verbs and their several senses in the Penn Treebank corpus was annotated in the PropBank corpus. Basic roles such as Agent and Patient are listed by Arg0, Arg1, …, Arg5 and adjunct roles like Time and Location are labeled as ArgM (ArgM-TMP, ArgM-LOC, …). Table 1 show the basic and adjunct semantic roles defined in PropBank with their related question type. Since sentences in PropBank is taken from Wall Street Journal [WSJ], then the performance of supervised classifier outside the domain of the WSJ is decreased. Several methods are utilized to construct a semantically annotated corpus: direct annotation, using parallel corpus and using semi-supervised methods. Bootstrapping approach is applied by Swier and Stevenson (2004) to annotate verbs in general domain. Pado and Lapata (2009) exploited translation of English FrameNet to construct relevant corpus in another language. Monachesi et al. (2007) used semi-supervised method and translation of English PropBank to construct corpus in Dutch. Afterwards annotated sentences was corrected by annotators to use as training corpus in supervised methods.

	Argument	Question type
Basic	Arg0	Who?
	Arg1	Whom?
	Arg2	What?
Adjunct	Arg-TMP	When?
	Arg-LOC	Where?
	Arg-MNR	How?
	Arg-PRP/CAU	Why?

Table 1: PropBank's some basic and adjunct semantic roles.

Since accuracy of SRL system drop dramatically in outside the domain of the annotated corpus domain in English, building comprehensive lexical resources in biology domain will improve SRL system in Turkish for biological text. Due to the lots of effort to construct such lexical resource, we will build it in semi-automatic manner by employing self-training semi supervised method where dependency parses will use to extract features. In my proposal, we will use standard supervised method (SVM, MaxEnt and NB) to build SRL system to evaluate their performance in Turkish.

3 Methodology

Before diving in to automatic generating questions in biology domain, we will construct a semantically annotated corpus and SRL system. The following sections will describe our proposed methods in detail to do these issues.

3.1 Corpus Construction

We first consider the annotation of semantic roles in biology domain. To address this issue, first we

collect biology texts from different sources like article, textbook and etc. Articles and textbooks will take from "Journal of Biyolojik Çeşitlilik ve Koruma"[1] and "Biyoloji ders kitabı 9, 10, 11, 12"[2], respectively. Afterwards we tag the part of speech (POS) of sentences to identify the predicate and then create lexical recourses with their predicate-argument structure (PAS). Kisla's tool (2009) is employed to POS tagging and morphologic analyzing of sentence. The predicates are selected by their frequency and their importance in domain. English PropBank structure and guidelines are used as reference structure to annotate PAS in Turkish. As a pilot study, we chose 500 sentences from biology high school textbook and tagged their POS. After identifying predicate, we ranked them based on their frequency. Some of selected predicates and their PAS are shown in tables 2 and 3 respectively.

Verb	Frequency
Sentezlemek (Synthesize)	23
İnceltmek (Thinning)	22
Adlandırmak (Naming)	20

Table 2: Some of the selected verbs

Since the annotation process is expensive and time consuming, we address this problem with using self-training method to create corpus in semi-automatic manner. The aim of semi-supervised method is to learn from small amount of annotated data and use large amount of unannotated data to develop the training data. SRL is a done in three steps: predicate identification, argument identification and argument classification. In first step we use POS tagging to identify predicate and its sense will be decided with some filtering rules. In Turkish "-imek, etmek, eylemek, olmak ve kılmak" (to do, make, render, to be) are auxiliaries that give predicate role to some noun words and are called auxiliary verbs. When encountering these verbs, we consider this verb with its preceding word as a predicate. For example, "sentezlenmiş olmak" (is synthesized) is the predicate of "Substrat düzeyinde fosforilasyonla 2 ATP de sentezlenmiş olur.". (2 ATP is synthesized by phosphorylation at the substrate level.)

To accomplish argument identification following rules are applied to select candidate arguments:

- Phrases are considered as argument if there is a dependency relation between them and the predicate.
- Existence of collocation is examined to consider as a candidate argument.

Note that these assumptions will not cover all candidate argument, but will be improved during this thesis.

Roleset id: Sentez.01 , (Synthesize) (kimya)
Element veya başka maddeleri bir arayı getirerek yapay olarak bileşik cisimler oluşturma, bireşim "(create)"
Roles:
Arg0-PAG: oluşturan (creator)
Arg1-PRD: oluşan şey (thing created)
Arg2-VSP: kaynak (source)

Table 3: Annotation of semantic role of predicate "sentez" (Synthesize)

Argument classification is done by self-training. Yarowsky and Florian (2002) utilized self-training for word sense disambiguation problem in 1995. Yarowsky's experimental results showed that the performance of self-training method is almost as high as supervised methods. Our intuition is that by utilizing self-training method, the effort to label semantic roles will reduce substantially. Self-training method is implemented in the following steps. First of all, seed data that is annotated manually by expert is used to train the classifier. After initial training step, all unlabeled data are classified and more appropriate data are selected to add to seed data to improve classifier performance by using more training data. Standard machine learning classifiers, SVM, MaxEnt and NB are used in the self-training method. In our proposal, we do following steps to select more accurate labeled data to expand training data: All unlabeled data are classified using three different classifiers. When two of them are agree about argument label and assigned probability of label is above predetermined threshold value, then this label is considered as true label and added to initial training data. If previous condition is not satisfied, then true label is the one which its assigned probability is maximum among the others and above predefined

[1] http://www.biodicon.com/
[2] http://www.eba.gov.tr/

threshold. Semi-automatic labeled data is corrected by annotators afterwards.

Determining effective and convenient features play an essential role in building SRL systems. These features drive from syntactic or semantic parsing systems. In our proposal we will use dependency parser to extract features. In our study we define features shown in Table 4 along with base features defined by Gildea and Jurafsky (2002). The effect of more features such as NE and biology terms will examine to improve performance of SRL system.

3.2 Automatic Question Generation

AQG is performed in three steps: content selection (which part of sentence must be asked), determine question type based on selected content and construct question. In my thesis, first the declarative sentence is labeled by our proposed SRL system. Based on labeled roles, content and question type are selected. In QG step, predetermined templates and rules are applied. We plan to generate templates manually as well as automatically. “Niye <X> <yüklem>?” (why <X> <predicate>?) and “Ne zaman <X> <yüklem>?” (when <X> <predicate>) are examples of templates. If there are no proper template for generating a question, then a rule based method is applied. In rule based method, Turkish question structure is considered to form question. In the first step, the selected content will be removed from the sentence. Then question type is chosen depending on the identified semantic role. For example, “kim” (who) is used if the semantic role label is Arg0. In the third step, selected content is replaced by question word. Finally, the grammar of generated question will be checked. In QG phase, to avoid generating vague question like “canlı dağılımı için ne önemlidir?” (what is important for live distribution?) from sentence “Bu canlı dağılımı için önemlidir.” (This is important for live distribution.) some filtering rules will apply. As an example, the sentences which begin with “Bu, Şu, O” (this, that, it) will not considered as descriptive sentence to generate question. Moreover, to add complexity to question we will use paraphrase of phrases.

4 Evaluations

To evaluate the SRL system, precision, recall, F1 and accuracy will be calculated. The following components are evaluated for the quality of the whole system:

- Argument identification performance
- Argument classification performance when arguments are known
- Performance of system when training data is in news domain and test data is in biology domain and vice versa.
- Performance of self-training method in news and biology domain

Rus at al. (2010) evaluated generated questions with the parameters, relevance, question type, syntactic correctness and fluency, ambiguity and variety. All parameters are among 1 and 4 which 1 is the best and 4 is the worst score. In my thesis we will evaluate generated questions by these parameters and the parameters that will define. ‘questions importance in education’ can be one of these parameters. We will ask three experts to evaluate generated questions manually.

5 Conclusion

Questions are used to assess the level of students’ understanding of the given lecture by the lecturer from pedagogical view. Therefore, automatically generating question alleviate lecturer’s effort to generate interrogative sentences. Moreover, tutoring system and question answering are some applications which benefit from questions too.

In my thesis, we propose syntactic and sematic approach to generate questions from descriptive sentences. To do this, a three-phase approach will take. Since generating question in semantic approach needs semantic analysis of sentences, we will construct a lexical semantic resource along with a semantically annotated corpus in biology domain, firstly. In the second phase, we built an SRL system to parse a sentence semantically. Finally, syntactically and semantically parsed descriptive sentences will be used to generate interrogative sentences. It is the first time that semantic approach is utilized for AQG in Turkish. Semantically annotated corpus in biology domain can use in several applications such as information extraction, question answering and summarization. Investigating the performance of biology corpus encourage researcher to transfer our proposed methodology to construct such semantic corpus in other domains like chemistry, geography and etc.

Feature	Description
YükKök	Stem of predicate
YükPOS	Predicate's POS tag
Voice	Predicate's voice
YükYapı	Compound, simple, derivative structure of predicate
Konum	Position of head-word
Baş-ad	Argument's head-word
Baş-adPOS	POS tag of argument's head-word
Baş-ad Kök	Stem of argument's head-word
Bağlılık İlişkisi	Dependency relation of argument's head-word
Bağlılık Yolu	Dependency path between arguments and predicate
Bağlılık Yolunun uzunluğu	Length of dependency path between arguments and predicate
SolPOS	POS tag of the argument's leftmost word
SağPOS	POS tag of the argument's rightmost word
Kelimeye eklenen ek	Morphologic analysis of head-word
Pos Yolu	POS tags between arguments and predicate

Table 4: The features will be used in argument classification

References

Albright, Daniel, Arrick Lanfranchi, Anwen Fredriksen, William F Styler, Colin Warner, Jena D Hwang, Jinho D Choi, Dmitriy Dligach, Rodney D Nielsen & James Martin. 2013. Towards comprehensive syntactic and semantic annotations of the clinical narrative. Journal of the American Medical Informatics Association 20.922-30.

Baker, Collin F, Charles J Fillmore & John B Lowe. 1998. The berkeley framenet project. Paper presented to the Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1, 1998. http://aclweb.org/anthology/P98-1013.

Çabuk, Hüseyin, Çiğdem Yüksel, Zeki Mocan, Banu Diri & M Fatih Amasyalı. 2003. Metin Analizi Ve Sorgulama (MAvS). Koç Üniversitesi İstanbul 11.

Carreras, Xavier & Lluís Màrquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. Paper presented to the Proceedings of the Ninth Conference on Computational Natural Language Learning, 2005. http://aclweb.org/anthology/W05-0620.

Do Thi, Ngoc Quynh, Steven Bethard & Marie-Francine Moens. 2016. Facing the most difficult case of Semantic Role Labeling: A collaboration of word embeddings and co-training. Paper presented to the Proceedings of the 26th International Conference on Computational Linguistics, 2016. http://aclweb.org/anthology/C16-1121.

Garg, Nikhil & James Henderson. 2012. Unsupervised semantic role induction with global role ordering. Paper presented to the Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, 2012. http://aclweb.org/anthology/P12-2029.

Gildea, Daniel & Daniel Jurafsky. 2002. Automatic labeling of semantic roles. Computational linguistics 28.245-88. http://aclweb.org/anthology/J02-3001.

Heilman, Michael & Noah A Smith. 2009. Question generation via overgenerating transformations and ranking. DTIC Document.

Isguder Sahin, Gozde Gul & Esref Adalı. 2014. Using Morphosemantic Information in Construction of a Pilot Lexical Semantic Resource for Turkish. In Proceedings of the 21st International Conference on Computational Linguistics.929-36. http://aclweb.org/anthology/W14-5807.

Johansson, Richard & Pierre Nugues. 2008. The effect of syntactic representation on semantic role labeling. Paper presented to the Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, 2008. http://aclweb.org/anthology/C08-1050.

Kaljahi, Zadeh & Rasoul Samad. 2010. Adapting self-training for semantic role labeling. Paper presented to the Proceedings of the ACL 2010 Student Research Workshop, 2010. http://aclweb.org/anthology/P10-3016.

Kisla, T. 2009. An integrated method for morphological analyse and part of speech tagging in Turkish. Doctoral dissertation, Ege University, Izmir, Turkey

Lee, Joo-Young, Young-In Song & Hae-Chang Rim. 2007. Investigation of weakly supervised learning for semantic role labeling. Paper presented to the Advanced Language Processing and Web Information Technology, 2007. ALPIT 2007. Sixth International Conference on, 2007.

Lindberg, David, Fred Popowich, John Nesbit & Phil Winne. 2013. Generating natural language questions to support learning on-line.

Liu, Ming, Rafael A Calvo & Vasile Rus. 2012. G-Asks: An intelligent automatic question generation system for academic writing support. Dialogue & Discourse 3.101-24.

Mannem, Prashanth, Rashmi Prasad & Aravind Joshi. 2010. Question generation from paragraphs at UPenn: QGSTEC system description. Paper presented to the Proceedings of QG2010: The Third Workshop on Question Generation, 2010.

Mazidi, Karen & Rodney D Nielsen. 2014. Linguistic Considerations in Automatic Question Generation. Paper presented to the ACL (2), 2014. http://aclweb.org/anthology/P14-2053.

Mitkov, Ruslan, Li An Ha & Nikiforos Karamanis. 2006. A computer-aided environment for generating multiple-choice test items. Natural Language Engineering 12.177-94.

Monachesi, Paola, Gerwert Stevens & Jantine Trapman. 2007. Adding semantic role annotation to a corpus of written Dutch. Paper presented to the Proceedings of the Linguistic Annotation Workshop, 2007. http://aclweb.org/anthology/W07-1513.

Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov & Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. Natural Language Engineering 13.95-135.

Orhan, Zeynep, Ceydanur Öztürk & Nihal Altuntaş. 2006. SınavYazar: Đlköğretim için Otomatik Sınav ve Çözüm Üretme Aracı SınavYazar: A Tool for Generating Automatic Exams and Solutions for Primary Education.

Padó, Sebastian & Mirella Lapata. 2009. Cross-lingual annotation projection for semantic roles. Journal of Artificial Intelligence Research 36.307-40.

Palmer, Martha, Daniel Gildea & Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. Computational linguistics 31.71-106. http://aclweb.org/anthology/J05-1004.

Pradhan, Sameer S, Wayne Ward & James H Martin. 2008. Towards robust semantic role labeling. Computational linguistics 34.289-310. http://aclweb.org/anthology/J08-2006.

Rus, Vasile, Brendan Wyse, Paul Piwek, Mihai Lintean, Svetlana Stoyanchev & Cristian Moldovan. 2010. Overview of the first question generation shared task evaluation challenge. Paper presented to the Proceedings of the Third Workshop on Question Generation, 2010. http://aclweb.org/anthology/W10-4234.

Stevenson, A. (2010). *Oxford dictionary of English*. Oxford University Press, USA.

Swier, Robert S & Suzanne Stevenson. 2004. Unsupervised semantic role labelling. Paper presented to the Proceedings of EMNLP, 2004. http://aclweb.org/anthology/W04-3213.

Yarowsky, David & Radu Florian. 2002. Evaluating sense disambiguation across diverse parameter spaces. Natural Language Engineering 8.293.