Automatic Question Generation Using Semantic Role Labeling for Morphologically Rich Languages

Daniel VASIĆ*, Branko ŽITKO, Hrvoje LJUBIĆ

Abstract: In this paper, a novel approach to automatic question generation (AQG) using semantic role labeling (SRL) for morphologically rich languages is presented. A model for AQG is developed for our native speaking language, Croatian. Croatian language is a highly inflected language that belongs to Balto-Slavic family of languages. Globally this article can be divided into two stages. In the first stage we present a novel approach to SRL of texts written in Croatian language that uses Conditional Random Fields (CRF). SRL traditionally consists of predicate disambiguation, argument identification and argument classification. After these steps most approaches use beam search to find optimal sequence of arguments based on given predicate. We propose the architecture for predicate identification and argument classification in which finding the best sequence of arguments is handled by Viterbi decoding. We enrich SRL features with custom attributes that are custom made for this language. Our SRL system achieves F_1 score of 78% in argument classification step on Croatian hr 500k corpus. In the second stage the proposed SRL model is used to develop AQG system for question generation from texts written in Croatian language. We proposed custom templates for AQG that were used to generate a total of 628 questions which were evaluated by experts scoring every question on a Likert scale. Expert evaluation of the system showed that our AQG achieved good results. The evaluation showed that 68% of the generated questions could be used for educational purposes. With these results the proposed AQG system could be used for possible implementation inside educational systems such as Intelligent Tutoring Systems.

Keywords: automatic question generation; morphologically rich languages; natural language processing; semantic role labeling

1 INTRODUCTION

The process of converting unstructured data into structured "machine-readable" data is a very complex task, which involves the retrieval of essential concepts and their systematic editing from a text written in natural language. Since the manual data annotation is a labor-intensive and time-consuming task, there is motivation to automate it. Concept identification in a particular domain is a very important first step. Identifying concepts and defining their semantic relationships seems challenging, but different natural language understanding techniques have already been developed. This paper focuses on identifying semantics in a text, a task known as Semantic Role Labeling (SRL), the widely studied challenge of recovering predicate-argument structure for natural language words, typically verbs.

This task covers advanced methods of understanding the meaning of a sentence used in many natural language processing tasks such as automatic question generation (AQG), machine translation, knowledge extraction and text summarization. The automatic recognition of semantic roles for the inflective Croatian language is a demanding task due to its free word order. The application of SRL tools can greatly contribute to the development of Intelligent Tutoring Systems (ITS) [1] based on natural language processing. This is a special type of ITS in which the communication with the student is carried in natural language.

There are several systems that try to solve the communication problem in ITSs, and by far the most famous one is AutoTutor [2]. Research shows that AutoTutor has produced learning gains that are about 0.8 standard deviation units above controls that read static instructional materials for an equivalent amount of time [3]. AutoTutor consists of five different components and each of them is responsible for a different stage of learning. This research focuses on Curriculum script repository that defines the content associated with a question or problem.

It consists of ideal answers, expected answers, a set of misconceptions, a set of keywords and synonyms and markup language rules for speech and gesture generator that can be easily generated with an authoring tool called the AutoTutor Script Authoring Tool [4]. This means that the process of creating new course material is not fully automated. A system similar to AutoTutor but created in the Croatian language is called CoLaBTutor (Controlled Language-Based Tutor) [5]. The system uses various natural language processing techniques to enable the learner to communicate in natural language. CoLabTutor uses a variety of lexical resources and custom-made rules to enable communication in the process of teaching and testing. The time and cost of content development for such system are much higher than the development of traditional teaching material used in teaching today. This is the biggest drawback of implementing ITS systems, and methodologies to accelerate this process are currently an active topic of research. An SRL component would allow implementation of an automatic tool for development of instructional content. This allows such a system automatic generation of sentences and questions, the verification of text similarity and the knowledge extraction from text written in natural language. The system would receive unstructured text as input, which would be transformed into content that the system could easily manipulate.

The rest of this paper is structured as follows: Section 2 shows related work on SRL. Section 3 describes the architecture of the proposed SRL system for our native speaking language. Section 4 shows the SRL evaluation results of our proposed approach. Section 5 suggests an approach for the AQG model. In section 6 we show the results of our qualitative analysis of the generated questions and comparison to similar AQG systems in other languages. In section 6 we also present some suggestions for the correction of errors that were noticed by our experts during evaluation. In section 7 conclusion is provided and some suggestions for future work.

2 RELATED WORK
2.1 SRL Task Description

The term semantic role denotes the relationships that verbs in a sentence have with other words. Verbs express the semantics of an event, which is described as relational information among participants in that event and projects a syntactic structure that encodes that information. The verbs are also highly variable, displaying a rich array of semantic and syntactic behavior. Verb classifications help natural-language word processing systems organize verbs into groups that share basic semantic and syntactic features. Semantic roles basically describe the conceptual relationships between participants in a given sentence. They illustrate the basic "Who, What, Where, How, and When" information within a sentence. The concept of semantic roles that first appeared in modern linguistics in the late-1960s is called case grammar [6]. Case grammar is a system of linguistic analysis and focuses on finding a connection between the valence of the verb and the context in which the verb is located. According to this grammar, the morphological and syntactic structures of all languages are derived from "hidden" semantic categories, not from syntactic categories as stated in the theory of generative grammar.

Semantic roles can be used to assist in various advanced text processing methods such as question generation, plagiarism detection, and summarization of multiple documents [7-10]. Therefore, it is very important to develop precise methods for machine recognition of semantic roles. Semantic roles provide a layer of abstraction over the syntactic dependencies of words in a sentence. These tags hide information that is insensitive to syntactic changes and provide some level of semantics, which is why this task is often called shallow semantic parsing. Machine recognition of semantic roles within sentences can be viewed through two machine learning approaches.

The first approach to SRL includes a classification task, where a semantic role for each word in a sentence, depending on its predicate, is determined. Typically, such approach uses syntactic information to "learn" semantic roles through the text tagged with features. Other approaches are based on the rich set of features extracted from texts using the body of expert knowledge [11, 12], or features extracted automatically using deep learning algorithms [13-15]. Such approaches require a voluminous text labeled with semantic roles and are suitable for resource-rich languages. Also, their main disadvantage is that they are limited to the domain in which they are trained. Other approaches present semantic role labeling as a task of grouping words and sentences [16-19]. Identifying semantic roles in a text using unsupervised machine learning does not produce as good results as methods based on rich lexical resources and different classification tasks. Another disadvantage of unsupervised machine learning methods is that they make rigorous assumptions about data such as an assumption that the semantic arguments of the predicate remain consistent despite variation in syntactic function. Furthermore, unlike supervised methods, they rely on simple sentential features, which are not adequate for the development of the SRL system for a very free word order language.

2.2 SRL Dataset for Croatian Language

Although there are many approaches that use parallel corpora to train multilingual tools for SRL [16, 20], they do not include the Croatian language. A corpus annotated with semantic roles in Croatian is currently under construction. The only annotated corpus to be used as training data for supervised machine learning systems [21] was developed within the project Semantic Role Labeling in Slovene and Croatian. The tag set for the Croatian SRL is developed following an approach from Prague Dependency Treebank (PDT) [22]. It contains a total of 87,387 tagged tokens. 3,003 sentences are used for training and 754 sentences for testing. All sentences are annotated with syntactic and semantic tags as shown in Fig. 1.

In the sentence "*Kosovo ozbiljno analizira process privatizacije u svjetlu učestalih pritužbi*" all words are labeled with their respective part-of-speech according to the MULTEXT-East specification, dependency relations using Universal Dependencies[1] and the semantic roles are labeled with respect to the predicate "analizira" (Eng. analyze). Semantic arguments are the actor of the action (ACT), the patient of the action (PAT) and the manner in which the action is held (REG).
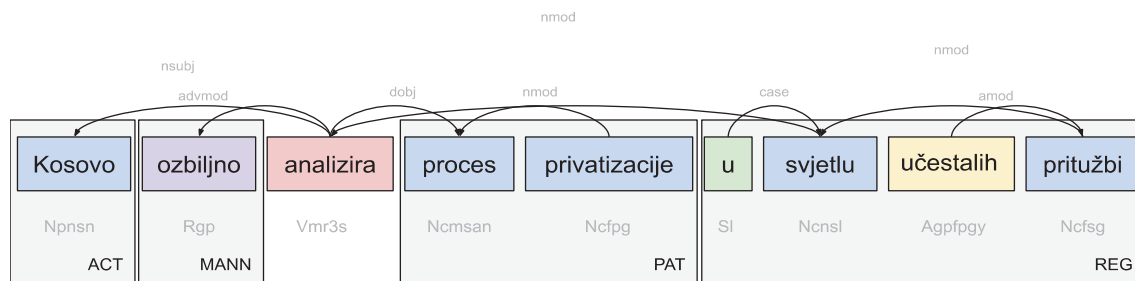

Figure1 Schematic representation of predicates, semantic frames and syntactic information such as dependency tree and part-of-speech tags in the Croatian language

3 APPROACH TO SEMANTIC ROLE LABELING OF MORPHOLOGICALLY RICH LANGUAGES

For languages that have a rich set of hand-tagged data, supervised machine learning methods that can "learn" from training data to distinguish individual classes are a logical choice. Most natural language processing tasks can be reduced to labeling words in a sentence, whether it is syntactic or semantic processing. Let us define $W = \{w_1, w_2, \ldots, w_n\}$ as a set of all words in a language, and

---

[1] https://universaldependencies.org/

$$S = \left\{ s_1^T, s_2^T, ..., s_m^T \right\}, w_i \in s_j^T, w_i \in W, j = 1,...,m, i = 1,...,n$$

as a set of all sentences.

Let $T = \{t_1, t_2, …, t_n\}$ be the set of all tags, and let

$$O = \left\{ o_1^T, o_2^T, ..., o_k^T \right\}, t_f \in o_g^T, t_h \in T, g = 1,...,i, h = 1,...,k$$

be the set of all tag vectors for each sentence. The function that assigns a vector to each sentence from the set O needs to be found using an optimization algorithm. Good results can be achieved by defining a function that extracts a number of features from each word as shown in the following Eq. (1).

$$\phi : W \to F, F = \left\{ f_1^T, f_2^T, ..., f_n^T \right\}, \vee f = \phi(w_i),$$
$$w_i \in W, i = 1,...,n \tag{1}$$

Table 1 Features for predicate identification and argument classification for morphologically rich languages

Predicate identification		Argument classification	
Feature name	Feature description	Feature name	Feature description
BOS	If a word is at the beginning	PredWord	The predicate of the current word
EOS	If a word is at the end	PredPos	POS tag of the current predicate
PredWord	Predicate word	PredDeprel	Predicate dependency relation
PredPos	Predicate part of speech	PredFets	Predicate features
PredDeprel	Predicate dependency relation	ArgWord	Current word
PredType	Predicate type feature	ArgPos	POS tag of the current word
PredGender	Gender feature of predicate	ArgDeprel	Dependency relation of the current word
PredNumber	Number feature of predicate	ArgFets	Features of the current word
PredCase	Case of the predicate	DeprelPath	Dependency path from the current word to the predicate
PredDefiniteness	Definiteness of verb if exists	PosPath	POS of the path from predicate to current word
PredVForm	Verb form if exists	Position	The position with respect to the predicate
PredPerson	Person feature of predicate	PredParentWord	Predicate parent word from the dependency tree
PredParentWord	Predicate parent word from the dependency tree	PredParentPos	Predicate parent part of speech from the dependency tree
PredParentPos	Predicate parent part of speech from the dependency tree	PredParentFeats	Predicate parent features
PredParentFeats	Predicate parent features	DepSubCat	Subcategorization frame of the current word
PredParentType	Predicate parent type	ChildDepSet	Children of the word with their respective dependency relation
PredParentDegree	Predicate parent degree feature	ChildWordSet	Children of the word
PredParentGender	Predicate parent gender feature	ChildPOSSet	Children of the word with their respective POS tags
PredParentNumber	Predicate parent number feature	LeftSiblingWord	Word before the current word
PredParentCase	Predicate parent case	LeftSiblingPos	POS of the word before the current word
PredParentDef	Predicate parent definiteness feature	LeftSiblingFeats	Feature of the word before the current word
PredParentVForm	Predicate parent verb form if exists	RightSiblingWord	Word after current word
PredParentPerson	Predicate parent person feature	RightSiblingPos	POS of the word after the current word
PredParentDeprel	Predicate parent dependency relation	RightSiblingFeats	Feature of the word after the current word
DepSubCat	Subcategorization frame of the current word	LeftWord	Leftmost dependent of the current word
ChildDepSet	Children of the word with their respective dependency relation	LeftPos	Leftmost dependent POS of the current word
ChildWordSet	Children of the word	LeftFeats	Leftmost dependent features of the current word
ChildPOSSet	Children of the word with their respective POS tags	RightWord	Rightmost dependent of the current word
ChildCaseSet	Children of the word with their respective case feature	RightPos	Rightmost dependent POS of the current word
ChildTypeSet	Children of the word with their respective type feature	RightFeats	Rightmost dependent features of the current word

The feature function consists of a series of word transformations that extract informative features specific to the machining task to be performed. Set of all features is encoded into numerical representations that are suitable for classification algorithms. A general approach for development of the SRL system set by [12, 23] embodies a couple of steps. The first step is predicate identification where the system identifies predicate of a sentence. The second step is argument identification in which the system identifies if a given word is semantically related to the predicate. The last step of the SRL system is the argument classification in which given predicate-argument pairs system identifies the label that connects those pairs.

Strategies for SRL can be diverse and most systems use three step approach. In this approach argument identification is implemented as a binary classification and argument classification is multi-label classification. The argument identification step is used to filter the words to reduce the number of non-argument labels that can overfit the argument classification model. This is the main problem for under resourced languages that do not have large corpus of hand annotated data.

SRL in the Croatian language was performed on the corpus shown in section 2 using features shown in Tab. 1. SRL strategy used in this paper is a model based completely on a sequence tagging approach. In predicate identification step CRF is used to find sequence of binary values that indicate predicate and non-predicate words. Custom handcrafted sets of features are used that are defined by [11] but they are modified to take into account morphological information from the words. We do this by removing **Feats** feature defined in [11] and adding type, case, gender and number feature to every predicate candidate. The same is implemented in argument classification where we generate multiple sequences based

on the predicate used morphological information from the given argument candidate.

There are two types of classification models: independent models and joint models. Joint models find the best overall sequence of arguments for given predicate while independent models predict individual argument label for given argument predicate pairs. Independent models are prone to inconsistencies so many SRL approaches are based on joint model approaches. This is usually implemented by using beam search over the predicted labels. In this article we used Viterbi decoding algorithm on conditional probabilities inferred by CRF to find optimal sequence of SRL arguments.

CRF is a type of statistical modeling method used to label data sequences. CRF is said to be the precursor to today's recurrent neural networks. It is a commonly used tool in natural language processing for part-of-speech tagging [24, 25] tagging named entities [26], and other sequential tagging tasks. CRF is a non-directional graphical model whose vertices can be divided into two disjunctive sets. In sequence modeling, a chain is a graph we are usually interested in. The input set of variables F represents a series of observations while T represents the hidden states, i.e. the states to be obtained through the input parameters. The conditional dependence of input and output variables is defined through a series of functions of significance $f(i, T_{i-1}, T_i, F)$.

This function actually defines which variables at the input define the probability of each possible occurrence of the output variable T_i. For each feature, the model assigns a numerical weight value and combines them to determine the "tag" probabilities for the input values. Learning of these parameters Ω is done using maximum probability for $p(T_i|F_i, \Omega)$; for exponential distributions this problem can be solved by using gradient descent or the quasi-Newton methods.

4 EVALUATION OF THE PROPOSED SRL METHOD

We compared the results obtained with CRF method and custom-defined features with the results obtained with the mate-tools [11] using default German features as the German language of all of the supported is the closest to the Croatian language. The system was trained and evaluated using train-test split obtained from the hr 500k corpora.

Our system[2] was trained using the L-BFGS method for 100 iterations, also L_1 and L_2 regularization is used with L_1 coefficient 0.2009, and L_2 coefficient 0.0284. The CRF transition generator is used for data augmentation and transition that did not occur in training data but, based on available transitions are more likely to occur. The classifier evaluated using 3-fold cross validation and randomized search was used to find optimal hyper parameters.

Tab. 2 shows the results obtained with our CRF method. We also tested linear classifier, but CRF method gave slightly better results.

We compared results obtained by our CRF method to benchmark results obtained by mate-tools[3]. The official benchmark results used gold tags for the predicate identification stage. We rerun the same tool on the same

dataset but included predictions from predicate identification stage. The results of our evaluation show that mate-tool gave results with F_1 score of 98.53 for predicate identification stage and F_1 score of 71.87% for argument classification stage. The CRF method compared to mate-tools gave better results with increase in F_1 score of 0.65% for predicate identification and increase in F_1 score of 6% for argument classification. This is an improvement when compared to the results of the state-of-the-art SRL models for the Croatian language that have been published recently [27].

Table 2 CRF method evaluation results

Predicate identification				
CRF				
	P	**R**	**F_1**	**Total**
Y	0.96	0.97	0.97	2021
	0.99	0.99	0.99	14614
Total	**0.99**	**0.99**	**0.99**	**16635**
Argument classification				
ACMP	0.75	0.52	0.62	23
ACT	0.9	0.94	0.92	962
AIM	0.74	0.51	0.61	45
CAUS	0.62	0.25	0.35	53
COND	0.54	0.54	0.54	13
CONT	0.56	0.36	0.43	14
DUR	0.73	0.49	0.59	88
EVEN	0.62	0.58	0.6	26
FREQ	0.91	0.67	0.77	15
GOAL	0.54	0.3	0.39	43
LOC	0.56	0.6	0.58	97
MANN	0.56	0.48	0.52	100
MEAN	0.38	0.57	0.45	21
MODA	0.81	0.99	0.89	115
MWP	0.93	0.37	0.53	35
ORIG	0.58	0.59	0.58	64
PAT	0.82	0.81	0.82	1052
PHRA	0.17	0.08	0.11	12
QUAN	0.77	0.53	0.63	43
REC	0.8	0.7	0.75	117
REG	0.47	0.36	0.41	47
RESL	0.81	0.83	0.82	562
REST	1.0	0.43	0.6	7
SOUR	0.5	0.18	0.27	11
TIME	0.74	0.81	0.77	238
Total	**0.79**	**0.78**	**0.78**	**3803**

The most discriminative features for predicate identification were **PredWord** feature, **ChildCaseSet** feature and **PredType** among others. The analysis shows that if **ChildCaseSet** feature value is nominative and locative there is a high probability that the next element of the sequence is the predicate, or if the **PredType** is main there is a very high probability that the next word in the sequence is again predicate. In argument classification step the **DepPath** and **PosPath** features are the most informative thought, **ArgFeatsType** is also highly informative especially for identification of TIME, LOC and DUR arguments. In both predicate identification and argument classification, morphological features such as the case, number, gender and type are shown to be most informative in inferring correct classification.

---

[2]https://github.com/danielvasic/CroatianSRL.git

[3]https://github.com/clarinsi/bilateral-srl

5 AUTOMATIC QUESTION GENERATION USING SEMANTIC ROLE LABELING

With the development of a method for semantic role labeling in this article, we also propose a system for automatic question generation (AQG) for the Croatian language. Croatian language opposite to the English language is a highly inflected language. In this section, we present a simple model for question generation using semantic roles. Due to the complexity of Croatian language, building a robust question generation system is a very complex task. We propose rules that can be implemented over the SRL graph, bearing in mind that there is room for improving the quality of the questions using other syntactic structures such as part-of-speech tags, dependency tree, named entities and co reference resolution. Process of building AQG system for Croatian language starts with the mapping of the constituent question types to respective semantic roles. Templates for question generation that are used to generate constituent questions are shown in Tab. 3.

Table 3 Question generation template for constituent questions

SRL argument	Question template	Question description
ACT	Tko PRED PAT?	Who
RESTR, RESLT	Kako ACT PRED PAT?	How
TIME	Kada ACT PRED PAT?	When
LOC	Gdje ACT PRED PAT?	Where
QUANT	Koliko PRED ACT?	How much
AIM	U koju svhu PRED ACT?	Which cause
DUR	Do kada PRED ACT?	Until when
GOAL	Gdje PRED ACT?	Where
EVENT	U kakvom PRED ACT?	In what
FREQ	Koliko PRED ACT?	How many

The predicate of the sentence is known to define the action expressed in the sentence. Our proposed system generates questions based on the predicates in the sentence. We use predicate arguments to determine which type of constituent question can be generated. If the predicate contains arguments required by the template, the system replaces the argument inside the template with its respective value. We adapt the approach for question generation from [28] but map the arguments from PropBank style annotations [29] to Prague Dependency Treebank style annotations. We also suggest using rich morphological information from part-of-speech tags and dependency parse to further improve the quality of questions. Implemented rules such as getting auxiliary and compound verbs for each predicate, greatly improve the quality of questions though there is a room for improvement.

6 EVALUATION OF THE PROPOSED AUTOMATIC QUESTION GENERATION METHOD

The total of 758 sentences from hr 500k testing set sentences for SQL evaluation were used to generate in total 628 question and answer pairs. The questions were evaluated by three experts, university level teachers, because of the implementation of this system for educational purposes. Each question is evaluated on three different levels, grammar, semantics a relevance as proposed in [9] using Likert scale from one to five where

the scale is represented as follows 1) dissatisfied, 2) not-satisfied, 3) satisfied, 4) very satisfied and 5) extremely satisfied.

Upon completion we measured inner-evaluator agreement with Fleiss' Kappa. This analysis showed fair consensus between evaluators' ratings on all three levels of evaluation. We also evaluated results of grading on 3 different scales. First scale is standard Likert five grade scale, second scale is four grade scale that was obtained by concatenating last and second to last grades, and third scale was obtained by concatenation of second and third grade. We concatenated the grades randomly although other approaches for concatenation could also be used and evaluated. The results of Fleiss' Kappa evaluation are shown in Tab. 4.

Table 4 Fleiss' Kappa results for three different scales

Grammar	Semantics	Relevance
0.34	0.24	0.19
0.38	0.28	0.22
0.38	0.45	0.33

In the first column Fleiss' Kappa for grammar is shown for five, four and three grades. In second column same scores are shown for semantics and last column shows the relevance results. These results show that evaluators have fair and moderate agreement about grammar, semantics and relevance of the generated questions. Fleiss' Kappa score increased substantially when three and four grades were introduced especially in semantics and relevance categories which show that maybe for these categories three level grades are more appropriate.

Table 5 Average ratings for different type of constituent questions

	Grammar	Semantics	Relevance	Total
ACT	3.27	3.69	3.74	10.70
AIM	3.36	4.08	4.31	11.75
DUR	3.27	3.66	3.39	10.32
FREQ	2.30	2.70	2.40	7.40
GOAL	3.63	4.15	4.02	11.79
LOC	4.11	4.61	4.42	13.13
QUANT	4.33	4.42	3.67	12.42
RESLT	3.25	3.35	3.65	10.25
RESTR	1.50	1.50	1.93	4.93
TIME	3.87	4.61	4.40	12.88
Total	**3.35**	**3.81**	**3.80**	**10.97**

In Tab. 5 we show average ratings for different types of constituent questions. Most of the questions are actor type (72%) and there are very few frequency type questions (0.007%). From the table we can see that the quality of location, quantity and time type questions are rated with the best average scores, but are scored less relevant than actor type questions, which is scored as satisfactory. Also, frequency type questions are scored as grammatically incorrect but satisfactory understandable and not very relevant. To present overall system performance we used the average of all ratings of the AQG system. The evaluation results show that our system, by the evaluation of the experts, generates questions with satisfactory results. The worst scores on all levels are achieved by RESTR type of questions. This could mean that the template for this type of question is badly defined.

Our proposed system achieved average rating of 3.36 for grammar category, for semantic category the system achieved average of 3.81 and for relevance category the

system achieved average score of 3.80. This means that the system for AQG could be used for automatically generating questions that could be used for educational purposes. In that regard we evaluated to which extent the generated questions could be used for educational purposes. Questions that have total scoring higher than 10 are considered as useful. This analysis shows that 71.3% of the generated questions could be used for educational purposes. These achieved results show us that we can streamline question generation inside authoring tools of ITS systems and achieve good results. Such a component can reduce the time for the creation of course content by generating questions from unstructured texts.
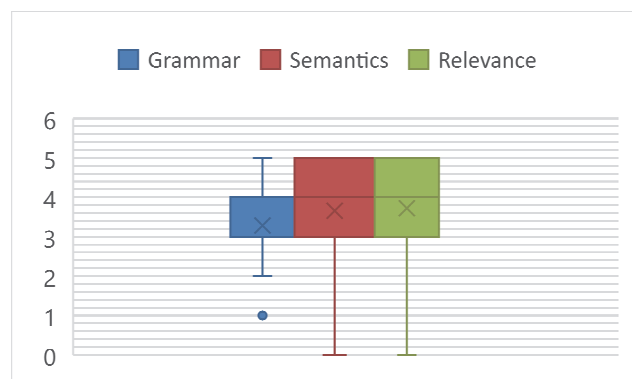

Figure 2 Average ratings and standard deviations for AQG system values range from 1-5 for grammar, semantics and relevance but 0-15 for total

7 CONCLUSION

In this article we presented a new approach for SRL of the texts written in Croatian language. We present results of the new SRL model that is based on CRF and custom defined features. This approach showed very competitive results with overall increase of 6% in F_1 score compared to the benchmark results. Using SRL component we also presented AQG model where we defined custom templates for the questions. Using this model, we generated and qualitatively evaluated 628 questions. The questions were evaluated for use in educational systems by expert university level teachers. The result of evaluation show that system generates questions that are very satisfactory for use in educational systems.

In the generated questions we have noticed that pronouns greatly affect the quality of the questions. One of the solutions to this problem is the use of the coreference resolution tool. We also noticed that the system is having problems generating Who or What type of questions. The solution to this problem could be by using word sense disambiguation algorithm (WSD) such is Lesk algorithm [30] and finding WordNet [31] or in case of Croatian language CROWN [32] entries to determine the animacy of the argument. Using auxiliary verbs from original sentence with predicate improves grammatical quality of questions. The system generates very dissatisfactory questions when copular predicate is present. This could be fixed by applying some constraints on dependency tree of the sentence.

Using morphosyntactic information, background knowledge resources, named entities and coreference resolution the quality of the generated questions could be further improved. Compared with the system for AQG in English language presented in [9] our system performed very well, by generating 71.3% usable questions for educational purposes. This is very good result taking into account that Croatian is resource poor language that is highly inflected when compared to English language. Due to this fact any system that automatically structures textual information is great contribution to this language and could be used in any system that makes decisions based on knowledge.

Such system is ITS that could receive unstructured text at the input, which is transformed into content that the system can easily manipulate. With such content, the system can generate questions within the knowledge testing subsystem and generate instructional material within the teaching subsystem. Based on this we can see how SRL and AQG in ITS can be an active topic for research. For future research we will try to improve our AQG system by using other natural language processing tasks and knowledge base. We also plan to integrate our AQG system into ITS system and evaluate the system in real educational environment.

Acknowledgements

8 REFERENCES

[1] Psotka, J., Massey, L. D., & Mutter, S. A. (1998). Intelligent tutoring systems: lessons learned. *Lawrence Erlbaum Associates Inc.*

[2] Graesser, A. C., Chipman, P., Haynes, B. C., & Olney, A. (2005). Auto tutor: An intelligent tutoring system with mixed-initiative dialogue, *IEEE Trans. Educ., 48*(4), 612-618. https://doi.org/10.1109/TE.2005.856149

[3] Graesser A. C., Rus V., Cai Z., & Hu X. (2011). Question answering and generation in Applied Natural Language Processing: Identification, Investigation and Resolution. *IGI Global,* 1-16. https://doi.org/10.4018/978-1-60960-741-8.ch001

[4] Susarla, S. C., Adcock, A. B., Van Eck, R. N., Moreno, K. N., & Graesser, A. (2003). Development and evaluation of a lesson authoring tool for AutoTutor.

[5] Žitko, B. (2010). *Model inteligentnog tutorskog sustava zasnovan na obradi kontroliranog jezika nad ontologijom.* Doctoral dissertation, University of Zagreb.

[6] Fillmore, C. J. (1967). The Case for Case. *Texas Symp. Lang. Universals*, 109.

[7] Osman, A. H., Salim, N., Binwahlan, M. S., Alteeb, R., & Abuobieda, A. (2012). An improved plagiarism detection scheme based on semantic role labeling. *Appl. Soft Comput., 12*(5), 1493-1502. https://doi.org/10.1016/j.asoc.2011.12.021

[8] Osman, A. H., Salim, N., Binwahlan, M. S., Twaha, S., Kumar Y. J. & Abuobieda A. (2012). Plagiarism detection scheme based on semantic role labeling. *Proceedings - 2012 International Conference on Information Retrieval and Knowledge Management, CAMP'12*, 30-33. https://doi.org/10.1109/InfRKM.2012.6204978

[9] Flor, M. & Riordan, B. (2018). A Semantic Role-based Approach to Open-Domain Automatic Question Generation. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 254-263. https://doi.org/10.18653/v1/W18-0530

[10] Keklik, O., Tuglular, T., & Tekir S. (2019). Rule-Based Automatic Question Generation Using Semantic Role Labeling. *IEICE Transactions on Information Systems,*

102(7), 1362-1373.
https://doi.org/10.1587/transinf.2018EDP7199

[11] Björkelund, A., Hafdell, L., & Nugues, P. (2009). Multilingual Semantic Role Labeling. *Proceedings of Thirteen Conference on Computational Natural Language Learning (CoNLL 2009) Shared Task*, 43-48. https://doi.org/10.3115/1596409.1596416

[12] Gildea, D. & Jurafsky, D. (2002). Automatic Labeling of Semantic Roles. *Computational Linguistics, 28*(3), 245-288. https://doi.org/10.1162/089120102760275983

[13] Roth, M. & Woodsend, K. (2014). Composition of Word Representations Improves Semantic Role Labelling. *Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing*, 407-413. https://doi.org/10.3115/v1/D14-1045

[14] Fitzgerald, N., Täckström, O., Ganchev, K., & Das, D. (2015). Semantic Role Labeling with Neural Network Factors. *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 960-970. https://doi.org/10.18653/v1/D15-1112

[15] Xia, Q. Li, Z., Zhang, M., Zhang, M., Fu, G., Wang, R., & Si, L. (2019) Syntax-aware Neural Semantic Role Labeling. *Proceedings of AAAI Conference on Artificial Intelligence, 33*, 7305-7313. https://doi.org/10.1609/aaai.v33i01.33017305

[16] Padó, S. & Lapata, M. (2009). Cross-lingual Annotation Projection of Semantic Roles. *Journal of Artificial Intelligence Research, 36*, 307-340. https://doi.org/10.1613/jair.2863

[17] Lang, J. & Lapata, M. (2014). Similarity-driven semantic role induction via graph partitioning. *Computational Linguistics,* 633-669. https://doi.org/10.1162/COLI_a_00195

[18] Titov, I. & Klementiev, A. (2012). A Bayesian approach to unsupervised semantic role. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics,* 12-22.

[19] Lang, J. & Lapata, M. (2011) Unsupervised semantic role induction with graph partitioning. *Proceedings of 2011 Conference on Empirical Methods in Natural Language Processing*.

[20] Akbik, A. & Li, Y. (2016). POLYGLOT: Multilingual Semantic Role Labeling with Unified Labels. *Proceedings of ACL-2016 System Demonstrations*, 1-6. https://doi.org/10.18653/v1/P16-4001

[21] Ljubešić, N., Agić, Ž., Klubička, F., Batanović, V., & Erjavec, T. (April 2018). Training corpus hr 500k 1.0. Retrieved from https://github.com/nljubesi/hr500k

[22] Hajič, J. (2018). Prague Dependency Treebank 3.5. Retrieved from http://ufal.mff.cuni.cz/pdt3.5

[23] Quirk, C., Choudhury, P., Gao, J., Suzuki, H., Toutanova, K., Gamon, M., Wen-tau, S., Lucy, Y., & Cherry, V. C. (2012). MSR SPLAT, a language analysis toolkit. *Proceedings Conference of the North American Chapter of the Association for Computational Linguistics*, 21-24.

[24] Silfverberg, M., Ruokolainen, T., Lindén, K., & Kurimo, M. (2014). Part-of-Speech Tagging using Conditional Random Fields: Exploiting Sub-Label Dependencies for Improved Accuracy. https://doi.org/10.3115/v1/P14-2043

[25] Krishnapriya, V., Sreesha, P., Harithalakshmi, T. R., Archana, T. C., & Vettath, J. N. (2014). Design of a POS tagger using conditional random fields for Malayalam. *Proceedings First International Conference on Computational Systems and Communications (ICCSC)*, 370-373. https://doi.org/10.1109/COMPSC.2014.7032680

[26] Senevirathne, K. U., Attanayake, N. S., Dhananjanie, M. H., Weragoda, S. U., Nugaliyadde, A., & Thelijjagoda, S. (2015). Conditional Random Fields based Named Entity Recognition for Sinhala. *2015 IEEE 10th International Conference on Industrial and Information Systems (ICIIS)*, 302-307. https://doi.org/10.1109/ICIINFS.2015.7399028

[27] Gantar, P., Despot, K. Š., Krek, S., & Ljubešić, N. (2018). Towards Semantic Role Labeling in Slovene and Croatian. *Proceedings Conference on Language Technologies & Digital Humanities in Ljubljana,* 93-98.

[28] Mazidi, K. & Tarau, P. (2016). Automatic question generation: From NLU to NLG. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9684*, 23-33. https://doi.org/10.1007/978-3-319-39583-8_3

[29] Palmer, M., Kingsbury, P., & Gildea, D. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics, 31*(1), 71-105. https://doi.org/10.1162/0891201053630264

[30] Toutanova, K., Haghighi, A., & Manning, C. D. (2008). A Global Joint Model for Semantic Role Labeling. *Computational Linguistics, 34*(2), 161-191. https://doi.org/10.1162/coli.2008.34.2.161

[31] G. A. Miller (1995). WordNet: A Lexical Database for English. *Comunications of ACM*, 38(11), 39-41. https://doi.org/10.1145/219717.219748

[32] Raffaelli, I., Bekavac, B., Agić, Ž., & Tadić, M. (2008), Building Croatian WordNet. *Proceedings of the Fourth Global WordNet Conference*, 349-359.

Contact information:

Daniel VASIĆ
(Corresponding author)
University of Mostar, Faculty of Science and Education,
88000 Mostar, Bosnia and Herzegovina
E-mail: daniel.vasic@fpmoz.sum.ba

Branko ŽITKO, PhD
University of Split, Faculty of Science,
10000 Split, Croatia
E-mail: bzitko@pmfst.hr

Hrvoje LJUBIĆ
University of Mostar, Faculty of Science and Education,
88000 Mostar, Bosnia and Herzegovina
E-mail: hrvoje.ljubic@fpmoz.sum.ba