Natural Language Engineering (2023), **29**, pp. 693–719
doi:10.1017/S1351324922000390

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Recognition of visual scene elements from a story text in Persian natural language

Mojdeh Hashemi-Namin, Mohammad Reza Jahed-Motlagh* and Adel Torkaman Rahmani

Iran University of Science and Technology, Tehran, Iran
*Corresponding author. E-mail: jahedmr@iust.ac.ir

(Received 29 March 2020; revised 28 May 2022; accepted 30 May 2022; first published online 24 August 2022)

Abstract

Text-to-scene conversion systems map natural language text to formal representations required for visual scenes. The difficulty involved in this mapping is one of the most critical challenges for developing these systems. The current study mapped Persian natural language text as the headmost system to a conceptual scene model. This conceptual scene model is an intermediate semantic representation between natural language and the visual scene and contains descriptions of visual elements of the scene. It will be used to produce meaningful animation based on an input story in this ongoing study. The mapping task was modeled as a sequential labeling problem, and a conditional random field (CRF) model was trained and tested for sequential labeling of scene model elements. To the best of the authors' knowledge, no dataset for this task exists; thus, the required dataset was collected for this task. The lack of required off-the-shelf natural language processing modules and a significant error rate in the available corpora were important challenges to dataset collection. Some features of the dataset were manually annotated. The results were evaluated using standard text classification metrics, and an average accuracy of 85.7% was obtained, which is satisfactory.

Keywords: Text-To-Scene Conversion system; Visual scene generation; Conceptual scene model; Persian natural language text; Conditional random fields

1. Introduction

The research on Text-To-Scene Conversion (TTSC) dates back to 1980 (Adorni, Di Manzo, and Giunchiglia 1984). Most TTSC systems take a natural language text in English as their input, but TTSC systems in other languages such as Swedish (Johansson, Nugues, and Williams 2004), French (Kayser and Nouioua 2009), Chines (Lu and Zhang 2002), Korean (Hong *et al*. 2018), Hindi (Jain *et al*. 2018), Japanese (Takahashi, Ramamonjisoa, and Ogata 2007), Russian (Ustalov and Kudryavtsev 2012), and Indonesian (Helfiandri, Zakhralativa Ruskanda, and Khodra 2020) have been developed. To the best of the authors' knowledge, no TTSC system from Persian natural language has previously been reported. The current study develops one of the initial steps in a Persian text-to-scene conversion system called PERSIS MEANS (PERSIan Story to MEaningfully ANimated Scene). The final PERSIS MEANS which will be developed in the future steps of the current study will take an input story in Persian natural language text and produce meaningful animation based on that. This study recognizes those tokens of the input story text that have a visualization in the final animation (i.e., the output of PERSIS MEANS) and then fills a conceptual scene model by these tokens as output. These conceptual scene models correspond to the scenes in the final animation and contain descriptions of visual elements of those scenes. The proposed

conceptual scene model and how to fill it using the input text is the main contribution of the current study.

The difficulty of this mapping task from input natural language text to formal representation required for visual scenes (like conceptual scene model in the current study) is one of the main challenges in developing TTSC. This mapping task was done deterministically in the elementary stages of well-known TTSC systems, such as WordsEye[a] (Coyne and Sproat 2001), SceneSeer[b] (Chang, Savva, and Manning 2014b; Zeng, Tan, and Ren 2016; Pardhi *et al*. 2021, and Yadav, Sathe, and Chandak 2020) which focus on obtaining and correctly visualizing the spatial relations between the objects of a scene. The relatively simple and constrained structure of the input sentences in all of these systems enable their developers to map input tokens to formal representation deterministically. The domain of the current study was literary realistic stories about the prophets. Story sentences were not focused on the objects or their properties, but on the events occurring in the context of encounters by each prophet with his tribe. In the first attempt of mapping, these Persian stories to the conceptual scene model, deterministic rules based on the part-of-speech (POS) tags, word-sense-disambiguation (WSD), and semantic role labeling (SRL) of input text were designed. However, the structural challenges of Persian natural language, especially in literary stories of the above mentioned domain, resulted in low accuracy in recognizing the elements of the conceptual scene model. Persian literary text usually forms long sentences. The average token number per sentence in this study was 17.63. Most of the sentences had more than one verb. This means that almost all of the sentences had nested inner sentence/sentences. The average number of verbs per sentence was 2.34. In many sentences, key semantic roles for the SRL were omitted because of syntactic or semantic symmetry, especially roles, which pointed to the actor of the verb or the thing that was acted upon (Shamsfard 2011). These challenges of Persian literary sentences prevented the deterministic development of the system in the first attempt.

WordsEye and CONFUCIUS have used advanced semantic processing levels (Ma 2006; Coyne *et al*. 2010) in which mapping takes place during the process (Jackendoff 1990). The NLP in Persian language has made acceptable progress in morphological and syntax processing (Shamsfard, Jafari, and Ilbeygi 2010b), but semantic analysis of Persian text is in the initial stages and no off-the-shelf module is available (Shamsfard 2011). In recent years, some SRL corpora have been built in academic research labs. These corpora need to be checked by experts, and their accuracy needs to be improved. One of these corpora was used in the current study and will be described below.

Other TTSC systems have developed mapping through machine learning techniques. Glass and Bangay (2008) used hierarchical rule matching and generalization to identify semantic categories of concepts in fiction books. Rouhizadeh (2013) enriched location information of VigNet, which is in the core of WordsEye, using crowd-sourcing data. Chang, Savva, and Manning (2014a) expanded the deterministically extracted set of explicit spatial relations in SceneSeer with implicit spatial relations not specified in the text using learned spatial priors. Chang *et al*. (2015) also used a machine learning approach for the lexical grounding problem. To map Persian natural language text to a conceptual scene model, the elements of the conceptual scene model were learned from story tokens through machine learning techniques in the current study. The proper dataset preparations and choosing the best model to fit the problem at hand are the two success factors for solving a problem using machine learning models. Because there is no TTSC system in Persian to the authors' knowledge, no such a dataset exists.

To prepare the required dataset, the input text must be processed using NLP pipeline modules. The POS tagger and syntax analyzer modules are available in Persian, but no off-the-shelf module for semantic analysis and WSD is available (Shamsfard 2011). An SRL corpus containing only

[a]http://www.wordseye.com.
[b]https://dovahkiin.stanford.edu/fuzzybox/text2scene.html.

https://doi.org/10.1017/S1351324922000390 Published online by Cambridge University Press

7656 tokens is available (Mesgar *et al*. 2014). The limited number of sentences with SRL tags available limited the total number of tokens in the prepared dataset to 7946 tokens in 451 sentences. None of these tokens was disambiguated through lexicons; thus, the WSD tags were prepared manually. Sense granularity, the derivational and generative nature of the Persian language were some challenges faced during the disambiguation of a dataset's tokens. To recognize conceptual scene model elements from story tokens in a supervised manner, each token should be tagged with a class label of the type of scene model element. These tags were manually prepared. The dataset prepared in the current study, especially its WSD tags, can be used by other researchers to learn models for different tasks and produce larger datasets.[c]

Using sentences to produce a conceptual scene model makes the nature of the problem sequential. Independent modeling of each scene element (as in traditional machine learning modeling) results in discarding important information that exists in the sequence of tokens in a sentence. Therefore, in the current study, the conditional random field (CRF) model was selected for the sequence labeling task (Lafferty, McCallum, and Pereira 2001). The imbalanced nature of the collected data motivated the use of CRF. The proportion of different elements of the conceptual scene model in the prepared dataset was diverse. CRF can handle this imbalanced data and learn the elements with a small number of samples, as in named-entity recognition (NER) problem modeling (Sutton and McCallum 2012). As a second attempt for the mapping task at hand, learning with traditional non-sequential machine learning models was tested, resulting in 76.58% average accuracy. Using CRF as the third and the last attempt, despite limitations on the number of samples in the dataset, resulted in acceptable accuracy. The average accuracy was 85.7%. Standard evaluation metrics of labeling are provided in detail.

The main contributions of this study consist of the development of the first TTSC system with Persian natural language text as input, the modeling of the problem of mapping Persian natural language text to the formal representation required for visual scene as a sequence-tagging problem and learning a CRF model for this task, preparation of a dataset for the mapping task, and tackling the problem of a lack of the required off-the-shelf NLP modules to prepare the dataset.

Section 2 presents works related to the current study in detail. Section 3 introduces the elements of the conceptual scene model, as proposed in the current study. Section 4 provides detailed information about the process of dataset collection. The machine learning model used to label the visual elements of the conceptual scene model is addressed in Section 5. Section 6 provides and discusses the evaluation results of labeling visual scene elements in the story text, and Section 7 presents conclusions about the mapping problem at hand.

2. Related work

The design methodology and language understanding approaches used for mapping the input natural language text to a formal representation in TTSC systems are reviewed in this section. These design methodologies can be classified as deterministic or automatic. The automatic approaches can further be classified as rule-based or data-driven (Hassani and Lee 2016). The data-driven approach was applied in the current study by learning a CRF model based on the collected dataset. Both syntactic and semantic analyses of the input text were used to produce the dataset required for the mapping task.

WordsEye is an early and frequently cited TTSC system that converts text to a static 3D scene (Coyne and Sproat 2001). Its input sentences consist of simple position words, color, size, and distance of objects like a dog is on the table. The relatively limited structure of their input text enabled them to use deterministic rules for mapping. It consists of two components: linguistic analysis and a scene depictor. The linguistic analysis component parses the input text and constructs a

[c]https://github.com/hasheminamin/PERSIS-MEANS-DATASET.

https://doi.org/10.1017/S1351324922000390 Published online by Cambridge University Press

dependency structure. This structure is then utilized to construct a semantic representation deterministically. This semantic representation is a formal representation in which objects, actions, and relations are represented in terms of semantic frames (Fillmore 1982). This component uses a POS tagger and associates the words with noun POS tags with 3D objects. It applies a set of predefined spatial patterns based on the dependency structure and captures the spatial relations. The words with verb POS tags are associated with a set of parametrized functions by the component.

Their focus is on correctly visualizing the senses of the words; thus, they have advanced the semantic processing of the input text in recent developments. They have incorporated lexical, semantic, and contextual knowledge to propose Scenario-Based Lexical Knowledge Resource (SBLR) (Coyne *et al*. 2010). SBLR is a lexical knowledge base customized to represent the lexical and common-sense knowledge for TTS conversion purposes and is derived from WordNet (Miller 1995) and FrameNet (Ruppenhofer *et al*. 2010). The SBLR developers have augmented the derived lexical semantic information to include finer-grained relations and properties of entities required to depict scenes and capture the different senses of properties related to those properties and relations (Coyne *et al*. 2010). The researchers used both syntactic and semantic analysis of natural language text to visualize a scene based on the input text correctly. They integrated deterministic, semantic processing and data-driven approaches to fulfill the mapping task.

CONFUCIUS is a text-to-animation conversion system that receives a natural language sentence in English and visualizes it as 3D animation (Ma 2006). Lexical visual semantic representation (LVSR) was proposed for this system to represent a relationship between the lingual and visual meanings. LVSR is based on the lexical conceptual structure (LCS) (Jackendoff 1990). The need for a method of mapping syntax onto semantics and vice versa is the core of LCS. Jackendoff proposed a set of entities (conceptual primitives) for conceptual structure and a sound foundation onto which communication rules between syntax and semantics (these conceptual entities) could be built. LVSR adapted LCS for the purpose of language visualization and provided finer ontological categories of concepts for generating humanoid character animation. It also related arguments in the conceptual structure to arguments in syntax through a set of rules. They integrated advanced levels of semantic processing and rule-based techniques to solve the mapping problem.

Glass and Bangay (2008) converted natural language fiction books to a 3D animated virtual environment. Initially, the natural language text in English was converted to an intermediate representation, and then, this intermediate representation was converted to a populated 3D environment. This intermediate representation consists of the original text that is annotated in different categories of concepts (Glass and Bangay 2009). A hierarchical rule-based learning system was created to learn patterns that were used to create annotations. The patterns are tree structures that abstract the input text according to the structural (token, phrase, sentence) and syntactic (parts-of-speech, syntactic function) categories. They stated that a supervisor must slightly manipulate such annotated text before being converted to animation (Glass and Bangay 2009). In this research, a semi-automatic rule-based learning technique was applied to the mapping problem. No semantic analysis was used to understand language.

SceneSeer is an interactive text-to-3D scene generation system that allows users to design 3D scenes using natural language (Chang *et al*. 2014b). The developers parsed the textual description of a scene and deterministically converted it to a scene template to generate the 3D scene described by the input text. This scene template includes a set of constraints on the objects present and the explicit spatial relations between them. For each object o_i, properties like category label, color, material, and the number of occurrences in the scene are identified based on the phrase in which the object is mentioned. For this rule-based approach, the text's sentences are first syntactically analyzed using the Stanford CoreNLP pipeline.[d] Headwords of noun phrases were identified as candidate objects. They filtered each noun using WordNet so that they only include physical

[d]http://nlp.stanford.edu/software/corenlp.shtml.

https://doi.org/10.1017/S1351324922000390 Published online by Cambridge University Press

Table 1. The design methodology and the language understanding approaches that have been used for mapping the input natural language text to a formal representation in some TTSC systems

System	Focus	Design methodology	Language understanding approaches
WordsEye (basic mapping)	Visualizing spatial relations	Deterministic	Syntactic & semantic analysis
WordsEye (improved mapping)	correctly visualizing the senses of words	Advanced semantic processing	Syntactic & semantic analysis
CONFUCIUS	Total TTSC	Advanced semantic processing/rule-based	Syntactic & semantic analysis
Glass	Animating fiction books	Rule-based learning	Syntactic analysis
SceneSeer (basic mapping)	Extracting explicit spatial relations	Deterministic	Syntactic analysis
SceneSeer (improved mapping)	Inferring implicit spatial relations	Data-driven	Syntactic analysis
PERSIS MEANS	Mapping input natural language text to a formal representation	Data-driven	Syntactic & semantic analysis

objects (excluding locations). Coreference resolution was done using the Stanford Coreference system. To extract the properties of each object, they assessed adjectives and other nouns in noun phrases containing the name of that object. The explicit spatial relations between objects were extracted using dependency patterns.

The developers of SceneSeer then improved on the TTSC system and expanded this deterministically extracted set of explicit spatial relations to include implicit spatial relations not specified in the text using learned spatial priors (Chang *et al*. 2014a). For this purpose, they collected a set of texts describing spatial relations between two objects in 3D scenes by running an experiment on Amazon's Mechanical Turk (AMT) (Fort, Adda, and Cohen 2011). AMT is an online crowdsourcing framework for data collection using human intelligence tasks (HITs). As an interactive text-to-3D scene generation system, the users of SceneSeer use textual commands to interactively refine the automatically created scene by adding, removing, replacing, and manipulating objects (Chang *et al*. 2017). The SceneSeer developers generally used deterministic rules to map input natural language text to the scene template and then used data-driven techniques to enrich the scene template. Only syntactic analysis of natural language text was used in their study for the mapping task, and their focus was on spatial information of the scene.

The information comparing these TTSC systems is shown in Table 1. Table 2 also shows the different types of information used in each of these systems. A syntactic analysis of the input text and WSD were used in all of these TTSC systems, but the semantic analysis was used in only some of them, such as WordsEye and CONFUCIUS.

Sequential modeling of sentences through CRF has been used by NLP studies in the Persian language, similar to studies in the English language. Arian and Sabbagh (2017) applied CRF to learn semantic role labeling of Persian sentences and had acceptable success.

3. Elements of conceptual scene model

Miaoulis and Plemenos (2009) stated that conceptual scene models are scene models that have been modeled in a declarative expression in natural language documents or based on a semantic network structure. Such models are generic and result in the creation of a group of less

https://doi.org/10.1017/S1351324922000390 Published online by Cambridge University Press

Table 2. Different types of information used in TTSC systems for the mapping task (Hassani and Lee 2016)

System	Syntactic analysis	Word-sense-disambiguation (WSD)	Semantic analysis
WordsEye	Statistical parsing	VigNet (extracted from FrameNet) & WordNet	Dependency
CONFUCIUS	Dependency parsing	WordNet	Lexical
Glass	Regular expression	WordNet	–
SceneSeer	Statistical parsing	WordNet	–
PERSIS MEANS	Statistical parsing	FarsNet (the Persian WordNet)	Semantic Role Labeling

Table 3. The elements of the conceptual scene model, which are mapped from the input text

Scene model part	Scene element name		Scene element description
	ROLE		human characters
Specifications of the		ROLE-ACTION	actions performed by the ROLE
elements present in the		ROLE-STATE	states of the ROLE
scene		ROLE-INTENT	intents of the ROLE
	ANIMATED-OBJECT		objects that can perform an action
		ANIMATED-OBJECT-STATE	states of the ANIMATED-OBJECT
		OBJECT-ACTION	actions performed by the ANIMATED-OBJECT
	STATIC-OBJECT		objects that cannot perform an action
		STATIC-OBJECT-STATE	states of the STATIC-OBJECT
Specifications of scene generalities	LOCATION TIME		location of the scene time of the scene

abstract models. They can be converted to 3D visualization of the final scene by adding spatial arrangements and determining 3D models of the scene elements. The elements of the conceptual scene model, which the input text is mapped to them in the current study, are introduced in Table 3.

This conceptual scene model consists of two main parts. The first defines the specifications of the elements present in the scene, and the second includes the information about general specifications of the scene, such as the location or time which the scene occurs in. In the first part, main scene elements can take three forms: ROLE, ANIMATED-OBJECT, and STATIC-OBJECT. ROLE represents the human characters, which are among the most important elements of the scene. Each human character in the input text is mapped to a ROLE in the scene model. The non-human elements of the story are objects. If the objects are animated, in the sense that they can perform an action, they are categorized as ANIMATED-OBJECTs. Animals are the most typical ANIMATED-OBJECTs and can perform acts such as running, sitting, and eating. Non-animated objects fall into the category of STATIC-OBJECTs.

Each of these three main scene elements can have more specific properties, which should be modeled in the conceptual scene model. For example, the different states of each object (ANIMATED/STATIC) are modeled as the ANIMATED-OBJECT-STATEs/STATIC-OBJECT-STATEs. The actions performed by each ANIMATED-OBJECT/ROLE based on the input story

https://doi.org/10.1017/S1351324922000390 Published online by Cambridge University Press

are represented by OBJECT-ACTION/ROLE-ACTION. The state and intent of a character in a scene are represented by ROLE-STATE and ROLE-INTENT, respectively.

Although ROLE-STATE and ROLE-INTENT properties of a ROLE have similarity, but they have different applications in the scene model. ROLE-STATEs of a ROLE usually reflect in the visualization directly, despite ROLE-INTENTs, which usually do not directly reflect in the scene and have an implicit effect on other scene elements instead. For example, the word "تهمت" ("slander") in the phrase "تهمت زدن" ("to slander"), the word "معارفه" ("introduction") in the phrase "مراسم معارفه" ("introduction ceremony"), and the word "صحبت" ("talk") in the phrase "صحبت کردن" ("to talk") have no direct visualization, but they reflect in the state of the ROLE described by these words (mapped to ROLE-INTENTs). In all these 3 cases, the ROLE performs the "talking" ROLE-ACTION but in the case of "تهمت" ("slander") probably talks with an angry face, in the case of "معارفه" ("introduction") probably talks with courtesy or enthusiasm, and in the case of "صحبت" ("talk") probably talks with no special state. When a token maps to ROLE-INTENT, it can be used to infer ROLE-STATEs in this study's future steps to produce meaningful animation. Another distinction between ROLE-STATE and ROLE-INTENT is that the existence of a ROLE-INTENT in a scene of a story affects the ROLE-STATEs and other scene elements of the current and also the following scenes, but the presence of a ROLE-STATE in a scene does not have such a long-lasting effect.

The second part of the conceptual scene model includes general specifications such as LOCATION and TIME of the current scene. Each scene (plan) of a story occurs in one specific location according to the division of the story into scenes.

The authors tried to design the scene model elements from the screenwriter perspective. From this perspective, the scene elements can be divided into two separate parts: actors and non-actors. The actors are those elements that can perform an action on the scene (as well as a change in the state). The actors of a scene (plan) maps to ROLE and ANIMATED-OBJECT scene elements. Other non-actor elements of a scene are modeled by STATIC-OBJECT. Both actor and non-actor scene elements can have change in their states in a scene. This modeling is rational from screenwriter perspective and led to 13 proposed scene elements.

Although the actors in the input story could be animals (modeled by ANIMATED-OBJECT), the domain of the selected corpus was literary realistic stories about the prophets. This means that the actors of all stories were human. Hence, the adequacy and coverage of the scene model for the stories with non-human actors were not tested. Selecting the literary stories to collect the dataset means that the adequacy of the proposed scene model for the stories with non-literary writings was not tested.

3.1 Comparison with scene models in other TTSC Systems

Since WordsEye (Coyne and Sproat 2001) and SceneSeer (Chang *et al*. 2014b) focused on obtaining and correctly visualizing the spatial relations between the objects of a scene, the objects and their properties were the key elements of their scene model. However, the scene model in CONFUCIUS (Ma 2006) and (Glass and Bangay 2008) tried to cover different elements that might have been present in a visual scene. So, they are comparable to the scene model proposed in the current study. Table 4 indicates that there are many elements in common with the scene models in these systems.

The ROLE element in the proposed scene model is modeled with the HUMAN in CONFUCIUS and the Avatar in the Glass's model. CONFUCIUS models the non-animated objects (STATIC-OBJECT) as OBJ, and humans, animals, plants, and all other animated objects as HUMAN. Glass makes no distinction between an ANIMATED-OBJECT and a STATIC-OBJECT. Both the ROLE-ACTIONs and the OBJECT-ACTIONs are modeled with the EVENT

https://doi.org/10.1017/S1351324922000390 Published online by Cambridge University Press

Table 4. Comparison of the conceptual scene model elements between Ma (2006), Glass and Bangay (2008) and the current study

Our Script model	CONFUCIUS	Glass
ROLE	HUMAN	Avatar
ROLE-ACTION	EVENT	Transition
ROLE-STATE	PROPERTY, STATE	–
ROLE-INTENT	PROPERTY	–
ANIMATED-OBJECT	HUMAN	Object
ANIMATED-OBJECT-STATE	PROPERTY, STATE	–
OBJECT-ACTION	EVENT	Transition
STATIC-OBJECT	OBJ	Object
STATIC-OBJECT-STATE	PROPERTY, STATE	–
LOCATION	PATH, PLACE	Setting, Relation
TIME	TIME	Setting
–	AMOUNT	–

in CONFUCIUS, but Glass only models ROLE-ACTION with Transition. The STATE element in CONFUCIUS is a static situation, which does not involve changes, and usually refers to a fact. The static properties of both HUMAN and OBJ are modeled with the STATE, and their other variable properties are modeled with the PROPERTY. Glass does not model neither the features of the objects nor humans. The LOCATION element is modeled with the PATH and PLACE in CONFUCIUS and with the Setting and Relation (explicit description of a spatial relation) in Glass's model. The Setting element in Glass's model also covers the time of the scene, which is modeled with the TIME element in CONFUCIUS and the proposed scene model.

Neither CONFUCIUS nor Glass distinguish between ROLE-STATE and ROLE-INTENT elements. CONFUCIUS has the AMOUNT element, which specifies the quantity of OBJs, which has no equivalent in the proposed model, nor in the Glass's model.

3.2 A sample scene model mapped from a Persian sentence

Figure 1 shows two snapshots of a sequence from the film "Saint Mary",[e] that is the visualization of the sentence "هر وقت زکریا به دیدار او می‌رفت غذاهای مخصوصی در کنار محراب او مشاهده می‌کرد که باعث شگفتی‌ش می‌شد." ("Each time Zakariya visited her, he noticed that she was provided with special food in her sanctuary, which caused him to wonder"). In the mapping of this sentence to the proposed scene model "زکریا" ("Zakariya") and "او" ("she"), which refers to "Saint Mary," mapped to ROLE; the actions "می‌رفت" ("was going") and "می‌کرد مشاهده" ("noticed") mapped to ROLE-ACTION; and the states of "دیدار" ("visit"), "مشاهده" ("notice"), and "شگفتی" ("wonder") mapped to ROLE-STATE. The token "غذاهای" ("food") is modeled as a STATIC-OBJECT with "مخصوصی" ("special") as its STATIC-OBJECT-STATE. This STATIC-OBJECT-STATE is reflected by non-seasonal fruits, which are bright and shiny in the visual scene. The tokens "کنار" (near) and "محراب" (sanctuary) mapped to the LOCATION of the event.

[e] http://shahriarbahrani.ir/ فیلم ها

https://doi.org/10.1017/S1351324922000390 Published online by Cambridge University Press

Figure 1. The visualization of the Persian sentence: "هر وقت زکریا به دیدار او می‌رفت غذاهای مخصوصی در کنار محراب او مشاهده می‌کرد که باعث شگفتی‌ش می‌شد." ("Each time Zakariya visited her, he noticed that she was provided with special food in her sanctuary, which caused him to wonder"). The tokens mapped to the scene model elements are marked in the visualization.

4. Collecting the dataset

In the current study, learning the elements of the conceptual scene model from story tokens was done to map the Persian natural language text to the conceptual scene model. To learn this mapping, a dataset was required which contains the proposed scene model elements mapped to its tokens. The assignment of WSD tags to dataset tokens was also required, as in many TTSC systems (Table 2). To the authors' knowledge, this required dataset does not exist in the Persian language, so it was collected. To collect the dataset, apart from the tokens of the sentences and their mapped conceptual scene model elements (labels), other information about each token is required.

To prepare these types of information, off-the-shelf Persian NLP modules were required. Syntax analyzer module was available (Shamsfard *et al*. 2010b), but there was no semantic analyzer module in Persian (Shamsfard 2011); only SRL corpora produced in academic research labs were available. Although these corpora were in the development phase and needed to be checked by experts to improve their accuracy, they comprised all of the available data containing semantic analysis information in the Persian language. The sentences in all of these corpora (except one) had general subjects and were collected from news and general texts. Literary real stories about the prophets were the subject of only one of these corpora (Mesgar *et al*. 2014), which was developed as part of the Qur'anic Question and Answer Project (Iran Telecommunication Research Center 2014) by the Iran Telecommunication Research Center[f] (ITRC). The proposed conceptual scene model elements were designed to model the visualization information of a scene and will be used to produce meaningful animation in this ongoing study. The interdependent sentences of the stories, which will be converted into interdependent visual scenes, were preferred for future steps of the current study over independent sentences with general subjects; thus, the ITRC SRL corpus was selected for dataset production purposes.

The ITRC corpus annotated each story token with syntactic and semantic information. The syntactic and semantic analysis methods used in the development of this corpus have been published as the Syntactic Manual of Style (Qur'anic Question and Answer Project 2014b) and Semantic Manual of Style (Qur'anic Question and Answer Project 2014a). The Syntactic Manual of Style was developed using dependency grammar in the Persian language (Tabibzadeh 2006), and the Semantic Manual of Style was adapted from Propbank 3.0 (Palmer, Gildea, and Kingsbury 2005) and adjusted to the Persian language.

To prepare the intended dataset, WSD tags and mapped scene elements (label tags that the model must learn) should be added to the ITRC corpus. To the authors' knowledge, no WSD tagged corpus was available in Persian, but these tags must be included in the required dataset. Table 2 shows that all TTSC systems have used a WSD tag in the mapping task. The significant

[f]https://www.itrc.ac.ir.

https://doi.org/10.1017/S1351324922000390 Published online by Cambridge University Press

Table 5. The tags prepared for the annotators to annotate the WSD and the conceptual scene model element tags. These tags were obtained from the ITRC corpus, which was in CoNLL 2008 format (Surdeanu *et al*. 2008)

No.	Name	Description
1	ID	Token counter, starting at 1 for each new sentence
2	FORM	Unsplit word form or punctuation symbol
3	GPOS	Gold part-of-speech tag from the Treebank
4	DEPREL	Syntactic dependency relation to the HEAD
5	HEAD	Syntactic head of the current token, which is either a value of ID or zero (0)
6	PRED	Role sets of the semantic predicates in this sentence
7	ARG	Columns with argument labels for each semantic predicate following textual order

role of the WSD information in both the first and second attempts at solving the mapping problem (designing deterministic mapping rules and learning traditional non-sequential machine learning models, especially rule-based models) confirmed the necessity of the WSD tag. To produce the dataset, the WSD tag and mapped scene model element (label tag) of each token were annotated manually.

4.1 Manual annotation task

In the annotation process, the ITRC corpus containing 7656 tokens in 451 sentences was given to the annotator. She added the WSD tag and the scene model element that are mapped to each token of a sentence. Then, two other annotators checked and corrected the output of the first annotator to minimize the probability of errors. In the corpus provided by ITRC, each token of stories was annotated in CoNLL 2008 format (Surdeanu *et al*. 2008). Not all CoNLL 2008 tags were intended for inclusion in the dataset. The unwanted tags of each token were eliminated from the corpus, and the remaining tags were used by the annotators. Table 5 shows the tags prepared for the annotators.

Figure 2 shows the annotation Interface used by the annotators. The columns 1–5 of Figure 2 are the same as rows 1–5 of Table 5. The 6th and 7th rows of Table 5 are multiplied by the number of verbs in each sentence and form the final columns of Figure 2 for each sentence. These columns were in a text file that was given to the annotators (as annotation interface) to manually fill in the wanted columns.

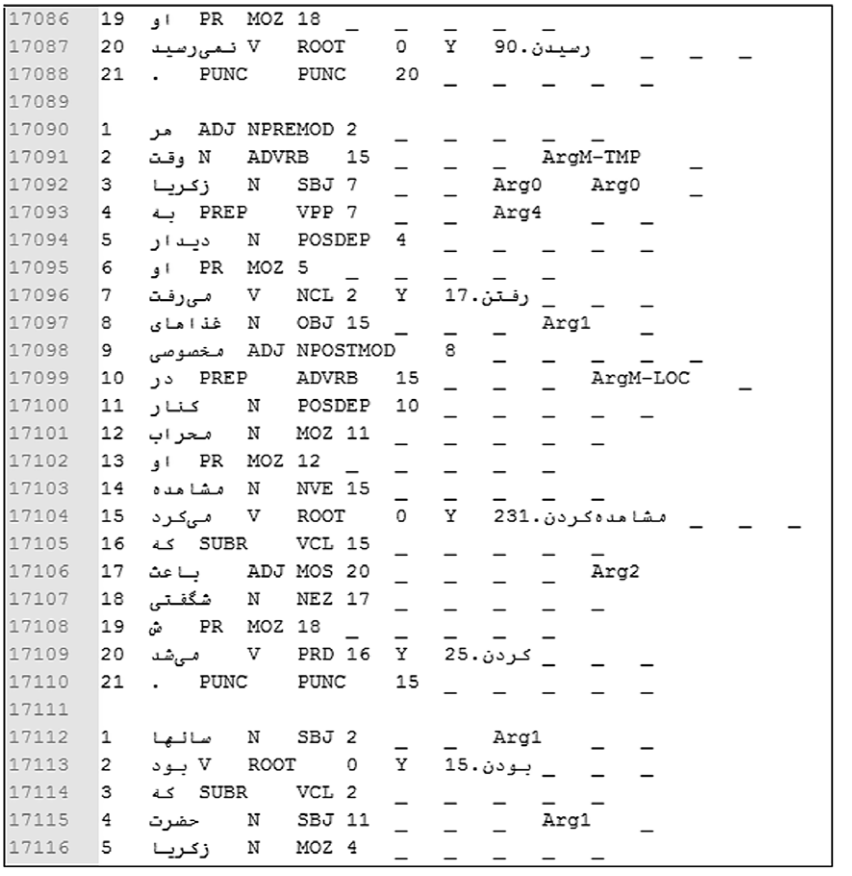
4.1.1 WSD annotation and its challenges

The lexicon used for WSD tagging was FarsNet 1.0[g] (Shamsfard *et al*. 2010a). FarsNet is designed to include a Persian WordNet containing about 17,000 synsets in the first phase. The inner language relations established between the senses and synsets of FarsNet are the same as those in WordNet 2.1. This lexicon was developed by the Information Technology Faculty of the Cyberspace Research Institute[h] in cooperation with Shahid Beheshti University.[i]

The manual annotation of the WSD tag in Persian stories presents several challenges (Shamsfard 2011). To select the appropriate sense, considering sense granularity in FarsNet, great attention should be paid, so it is a time-consuming task. The frequency of words per synset and

[g] http://farsnet.nlp.sbu.ac.ir/Site3/Modules/Public/Default.jsp.
[h] http://www.csri.ac.ir.
[i] http://www.sbu.ac.ir.

https://doi.org/10.1017/S1351324922000390 Published online by Cambridge University Press



17086	19	او	PR	MOZ	18	_	_	_	_	_		
17087	20	نمی‌رسید	V	ROOT	0	Y	90.رسیدن		_	_	_	
17088	21	.	PUNC	PUNC	20	_	_	_	_	_		
17089												
17090	1	هر	ADJ	NPREMOD	2	_	_	_	_	_		
17091	2	وقت	N	ADVRB	15	_	_	_	ArgM-TMP		_	
17092	3	زکریا	N	SBJ	7	_	_	Arg0		Arg0	_	
17093	4	به	PREP	VPP	7	_	_	Arg4		_	_	
17094	5	دیدار	N	POSDEP	4	_	_	_	_	_		
17095	6	او	PR	MOZ	5	_	_	_	_	_		
17096	7	می‌رفت	V	NCL	2	Y	17.رفتن		_	_	_	
17097	8	غذاهای	N	OBJ	15	_	_	_	Arg1		_	
17098	9	مخصوصی	ADJ	NPOSTMOD		8	_	_	_	_	_	
17099	10	در	PREP	ADVRB	15	_	_	_	ArgM-LOC		_	
17100	11	کنار	N	POSDEP	10	_	_	_	_	_		
17101	12	محراب	N	MOZ	11	_	_	_	_	_		
17102	13	او	PR	MOZ	12	_	_	_	_	_		
17103	14	مشاهده	N	NVE	15	_	_	_	_	_		
17104	15	می‌کرد	V	ROOT	0	Y	231.مشاهده‌کردن		_	_	_	
17105	16	که	SUBR	VCL	15	_	_	_	_	_		
17106	17	باعث	ADJ	MOS	20	_	_	_	_	Arg2		
17107	18	شگفتی	N	NEZ	17	_	_	_	_	_		
17108	19	ش	PR	MOZ	18	_	_	_	_	_		
17109	20	می‌شد	V	PRD	16	Y	25.کردن		_	_	_	
17110	21	.	PUNC	PUNC	15	_	_	_	_	_		
17111												
17112	1	سالها	N	SBJ	2	_	_	Arg1		_	_	
17113	2	بود	V	ROOT	0	Y	15.بودن		_	_	_	
17114	3	که	SUBR	VCL	2	_	_	_	_	_		
17115	4	حضرت	N	SBJ	11	_	_	_	Arg1		_	
17116	5	زکریا	N	MOZ	4	_	_	_	_	_		

Figure 2. The annotation interface used by the annotators.

senses per word is 1.78 and 1.37, respectively.[j] Emerging words built by concatenating words and affixes are frequent in Persian due to its derivational and generative nature. These words do not exist in FarsNet and they increased the missing values in WSD annotation. Nearly 60% of the 7656 tokens of the corpus had no equivalent sense, and 50% of these are stop words, with another 10% (non-stop words) missing their WSD tag in FarsNet. It is described in Section 5 that unlike common computational linguistic systems, the stop words have not been removed from the corpus. To limit the set of WSD tag values in the prepared dataset, principal concepts were selected from the top of the concept hierarchy in FarsNet. These principal concepts are shown in Table 6. During manual annotation, the annotators disambiguated each token by determining its equivalent sense in FarsNet (WSD tag) and its hypernym word from the principal concepts list (super-WSD tag).

4.1.2 Scene elements annotation and its challenges

The last tag that should be manually added to the corpus to provide the information required for the desired dataset is the conceptual scene model element tag (label tag to be learned). It should be determined as the mapped scene element of each token from the list: ROLE, ROLE-ACTION, ROLE-STATE, ROLE-INTENT, ANIMATED-OBJECT, ANIMATED-OBJECT-STATE, OBJECT-ACTION, STATIC-OBJECT, STATIC-OBJECT-STATE, LOCATION, TIME, JUNK, and NO. The

[j] http://farsnet.nlp.sbu.ac.ir/Site3/Modules/Public/Statistics.jsp.

https://doi.org/10.1017/S1351324922000390 Published online by Cambridge University Press

Table 6. Top principal concepts of FarsNet and their equivalent entry in WordNet

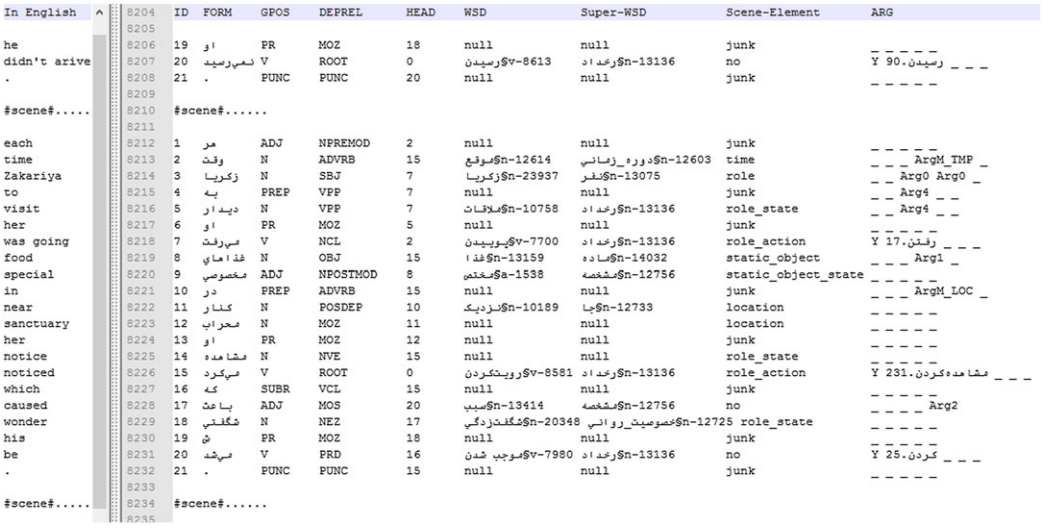
No.	Top principal concepts of FarsNet	Its equivalent WordNet entry
1	هست§n-12706	entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
2	نفر§n-13075	human (any living or extinct member of the family Hominidae characterized by superior intelligence, articulate speech, and erect carriage)
3	مسبب§n-13036	causing, causation (the act of causing something to happen)
4	جانور§n-12239	animal, animate being, beast, brute, creature, fauna (a living organism characterized by voluntary movement)
5	جان-دار§n-10393	alive, live (possessing life)
6	ساختار§n-12875	(object of) construct, build, make (make by combining materials and parts)
7	جا§n-12733	location (a point or extent in space)
8	شیء§n-12703	object, physical object (a tangible and visible entity; an entity that can cast a shadow)
9	ماده§n-14032	material, stuff (the tangible substance that goes into the makeup of a physical object)
10	مشخصه§n-12756	feature, characteristic (a prominent attribute or aspect of something)
11	خصوصیت-روانی§n-12725	psychological feature (a feature of the mental life of a living organism)
12	دوره-زمانی§n-12603	time period, period of time, period (an amount of time)
13	اندازه§n-12768	measure, quantity, amount (how much there is or how many there are of something that you can quantify)
14	رخداد§n-13136	event (something that happens at a given place and time)
15	رابطه§n-13628	relation (an abstraction belonging to or characteristic of two entities or parts together)
16	گروه§n-12753	group, grouping (any number of entities (members) considered as a unit)
17	خبر§n-12766	declaration (a statement that is emphatic and explicit (spoken or written)

definition of these scene elements is provided in Table 3. To map each token of a sentence to the proper scene element, the definition of each scene element was given to the annotators, and they were told to imagine the visual scene that is related to each sentence of every scene of each story. The annotators then mapped every token of each sentence to a scene element label according to their imagined visualization.

The JUNK label was designed to model tokens in sentences, which usually belong to the stop word list. The NO label was designed to model tokens in sentences, which are not visualized in the converted visual scene. For example, tokens "می‌دانست", "تعبیر", and "رویا" (translated as "knew," "interpretation," and "dream") in the sentence "او تعبیر رویا را می‌دانست." ("He knew the interpretation of the dream.") have no corresponding visualization elements; thus, they have been mapped to the NO label. Although tokens with a JUNK label usually have no visualization, their separation from the tokens with a NO label was rational and improved the results' accuracy. The tokens with NO labels are at times referential nouns, which head noun phrases, while JUNK tokens are not.

The ITRC corpus consists of 12 stories about the prophets. Each story consists of multiple sentences. In order to annotate scene elements, the annotators separated the different scenes of each

https://doi.org/10.1017/S1351324922000390 Published online by Cambridge University Press



In English		ID	FORM	GPOS	DEPREL	HEAD	WSD	Super-WSD	Scene-Element	ARG
	8205									
he	8206	19	او	PR	MOZ	18	null	null	junk	_ _ _ _ _
didn't arive	8207	20	نمی‌رسید	V	ROOT	0	رسیدن$v-8613	رخداد$n-13136	no	Y 90.رسیدن _ _ _
.	8208	21	.	PUNC	PUNC	20	null	null	junk	_ _ _ _ _
	8209									
#scene#.....	8210	#scene#.....								
	8211									
each	8212	1	هر	ADJ	NPREMOD	2	null	null	junk	_ _ _ _ _
time	8213	2	وقت	N	ADVRB	15	موقع$n-12614	دوره_زمانی$n-12603	time	_ _ _ ArgM_TMP _
Zakariya	8214	3	زکریا	N	SBJ	7	زکریا$n-23937	نفر$n-13075	role	_ _ Arg0 Arg0 _
to	8215	4	به	PREP	VPP	7	null	null	junk	_ _ Arg4 _ _
visit	8216	5	دیدار	N	VPP	7	ملاقات$n-10758	رخداد$n-13136	role_state	_ _ Arg4 _ _
her	8217	6	او	PR	MOZ	5	null	null	junk	_ _ _ _ _
was going	8218	7	می‌رفت	V	NCL	2	پوییدن$v-7700	رخداد$n-13136	role_action	Y 17.رفتن _ _ _
food	8219	8	غذاهای	N	OBJ	15	غذا$n-13159	ماده$n-14032	static_object	_ _ _ Arg1 _
special	8220	9	مخصوص	ADJ	NPOSTMOD	8	مختص$a-1538	مشخصه$n-12756	static_object_state	_ _ _ _ _
in	8221	10	در	PREP	ADVRB	15	null	null	junk	_ _ _ ArgM_LOC _
near	8222	11	کنار	N	POSDEP	10	نزدیک$n-10189	جا$n-12733	location	_ _ _ _ _
sanctuary	8223	12	محراب	N	MOZ	11	null	null	location	_ _ _ _ _
her	8224	13	او	PR	MOZ	12	null	null	junk	_ _ _ _ _
notice	8225	14	مشاهده	N	NVE	15	null	null	role_state	_ _ _ _ _
noticed	8226	15	می‌کرد	V	ROOT	0	رویت‌کردن$v-8581	رخداد$n-13136	role_action	Y 231.مشاهده‌کردن _ _ _
which	8227	16	که	SUBR	VCL	15	null	null	junk	_ _ _ _ _
caused	8228	17	باعث	ADJ	MOS	20	سبب$n-13414	مشخصه$n-12756	no	_ _ _ _ Arg2
wonder	8229	18	شگفتی	N	NEZ	17	شگفت‌زدگی$n-20348	خصوصیت_روانی$n-12725	role_state	_ _ _ _ _
his	8230	19	ش	PR	MOZ	18	null	null	junk	_ _ _ _ _
be	8231	20	می‌شد	V	PRD	16	موجب شدن$v-7980	رخداد$n-13136	no	Y 25.کردن _ _ _
.	8232	21	.	PUNC	PUNC	15	null	null	junk	_ _ _ _ _
	8233									
#scene#.....	8234	#scene#.....								
	8235									

Figure 3. The completed corpus; WSD, Super-WSD, and scene element tags added for the sentence in Figure 1. The column added to the left of the figure shows the English translation of each token.

story. A scene is a plan or shot as specified by the standard terminology of filmmaking. According to this terminology, each script consists of several sequences and each sequence is composed of different plans (Nazari 2006). A plan is one take, that is, the film recorded in the interval from when the camera is turned on and when it is turned off. A plan is the smallest part of a film and structurally is equivalent to a single word in the text. A sequence is a thematic grouping of events in a movie, such as a firing or wedding sequence. Each plan or scene occurs in one specific location, so the factor to separate the scenes of a story is a change in location in which the scene occurs. Figure 3 shows the completed corpus in which WSD, super-WSD, and mapped scene model elements are added to each token of sentences, in columns 6–8, respectively, and different scenes of all stories are marked. A column is added to the left of the figure to show the English translation of each token.

The manual annotation of the scene element tag faced some challenges. An inaccurate mental visualization of a sentence or even misunderstanding of the sentences could cause the tagging to be incorrect. Additionally, ambiguous sentences are likely to exist in each story. These are sentences that can have multiple interpretations or visualizations. Confusion between scene elements such as ROLE-STATE and ROLE-INTENT are more probable because of the similarity of their meanings. A high degree of accuracy should be applied by the annotators to correctly recognize the mapped scene element in order to prevent inappropriate use of the NO label.

The LOCATION and TIME scene elements were two of the most difficult ones to assign. The ArgM-LOC and ArgM-TMP SRL tags of a sentence (if available) show the locative and temporal tokens in a sentence, respectively. These tags can mislead the mapping task because they sometimes denote an abstract (not physical) location or time of the event occurring in that sentence. The "در" token (translated as "in") in the sentence "او مردی را در خواب دید." ("He saw a man in his dream.") has the ArgM-LOC SRL tag, but this tag does not actually refer to a physical location for the event "dream." The token "وقتی" ("when") in the phrase "وقتی او به شهر رسید" ("when he arrived in town") has an ArgM-TMP SRL tag, but it does not refer to the actual time at which the "arrive" event occurred. The token denoting the time of a scene may have been present in one or more previous scenes, which could cause the imagining and mapping task of the TIME tag even harder. The correct division of a story to its scenes by the annotators is another challenge, which affects

https://doi.org/10.1017/S1351324922000390 Published online by Cambridge University Press

the accuracy of the manual annotation process, especially the LOCATION and TIME tags of a scene.

An additional problem exists when mapping the tokens of a story to scene element tags. The SRL corpus provided by the ITRC was produced as part of a research project and its accuracy must be improved. This corpus originally contained nearly 15,000 tokens. Nearly half of the sentences were wrongly labeled with SRL tags and had to be eliminated from further processing. This decreased the total number of usable tokens of the corpus to 7656 tokens in 451 sentences. These wrongly labeled sentences in the ITRC corpus were scattered throughout the corpus and their elimination meant that many scenes in every story missed some sentences. Although the elimination of these sentences did not harm other sentences, the integrity of the scene that missed some of its describing sentences decreased. This problem, in some cases, made the imagining and mapping task difficult but tolerable for the human annotators.

4.2 The completed corpus and dataset

The manual annotation task of 7656 tokens took nearly ten days for each annotator because finding the WSD tag of each token from FarsNet was a time-consuming task, and the mapping of tokens to scene elements was even more challenging. Table 7 shows an example of a sentence that was manually annotated and its translation into English (column 3). Columns 1–2, 4–6, and 10 are from the ITRC corpus. Columns 7–9 were added by annotators. Column 7 is the WSD tag using the FarsNet lexicon (Shamsfard *et al*. 2010a). Column 8 is the hypernym of the token from the principal concepts list (Table 6), and column 9 is the scene element tag mapped to each token. The average pairwise inter-annotator agreement (IAA) score was 85.5%, 87.9%, and 75.6%, according to kappa value, for column 7–9, respectively. The IAA score is interpreted as substantial agreement if it was between 61.0% and 80.0%, and as almost perfect if it was between 0.81% and 0.99% based on what was reported in Landis and Koch (1977). The annotators had an almost perfect agreement on WSD and Super-WSD tags and substantial agreement on the scene element tags, which means it was more difficult for them to decide between different scene elements of a scene. The "completed" corpus in the current study, especially its WSD tags, can be used by other researchers to learn models for different NLP tasks.

After completion of the manual annotation task of the corpus, the dataset was produced based on the completed corpus. The FORM, GPOS, DEPREL, WSD, Super-WSD, Scene-Element, and ARG tags of each token of the corpus (columns 2, 4, 5, 7, 8, 9, and 10) were selected for the dataset. The tokens having more than one ARG in the column 10 were duplicated, and only one ARG has been written for every instance of the token. The seven features of each token have been written in one line separated by tab character, and the sentences were separated by a blank line. The different scenes of each story and different stories, which have been marked in the completed corpus, were not marked in the dataset. The dataset has 7946 tokens in 451 sentences taken from 12 stories. The number of tokens in the produced dataset is more than the number of tokens in the completed corpus because of the duplication of the tokens having more than one SRL argument. Figure 4 shows the part of the dataset, which was based on the part of the completed corpus shown in Table 7. The third and fourth tokens in the main sentence of Figure 4 are both related to the third token of Table 7 because it has two SRL arguments.

5. Conditional Random Fields

A CRF is a popular probabilistic method for structured prediction (Sutton and McCallum 2012). Structured prediction methods are a combination of graphical modeling and classification. They are able to compactly model multivariate data as graphical models and perform predictions using large sets of input features as classification methods. Traditional classification models predict only

https://doi.org/10.1017/S1351324922000390 Published online by Cambridge University Press

Table 7. The sentence in Figure 1 manually annotated and its translation into English (column 3). Columns 1–2, 4–6, and 10 are from the ITRC corpus. Columns 7–9 were added by annotators; the WSD tag, the hypernym of the token from the principal concepts list, and the scene element tag mapped to each token, respectively

ID	FORM	In English	GPOS	DEPREL	HEAD	WSD	Super-WSD	Scene-Element	ARGs
1	هر	each	ADJ	NPREMOD	2	–	–	JUNK	- - -
2	وقت	time	N	ADVRB	15	§n-12614موقع	§n-12603دوره–زمانی	TIME	- ArgM-TMP -
3	زکریا	Zakariya	N	SBJ	7	§n-23937زکریا	§n-13075نفر	ROLE	Arg0 Arg0 -
4	به	to	PREP	VPP	7	–	–	JUNK	- Arg4 -
5	دیدار	visit	N	POSDEP	4	§n-10758ملاقات	§n-13136رخداد	ROLE-STATE	- - -
5’	دیدار	visit	N	VPP	7	§n-10758ملاقات	§n-13136رخداد	ROLE-STATE	Arg4 - -
6	او	her	PR	MOZ	5	–	–	JUNK	- - -
7	می‌رفت	was going	V	NCL	2	§v-7700پوییدن	§n-13136رخداد	ROLE-ACTION	- - -
8	غذاهای	food	N	OBJ	15	§n-13159غذا	§n-14032ماده	STATIC-OBJECT	- Arg1 -
9	مخصوصی	special	ADJ	NPOSTMOD	8	§a-1538مختص	§n-12756مشخصه	STATIC-OBJECT-STATE	- - -
10	در	in	PREP	ADVRB	15	–	–	JUNK	- ArgM-LOC -
11	کنار	near	N	POSDEP	10	–	–	LOCATION	- - -
12	محراب	sanctuary	N	MOZ	11	–	–	LOCATION	- - -
13	او	her	PR	MOZ	12	–	–	JUNK	- - -
14	مشاهده	seeing	N	NVE	15	–	–	ROLE-STATE	- - -
15	می‌کرد	saw	V	ROOT	0	§v-8581رویت کردن	§n-13136رخداد	ROLE-ACTION	- - -
16	که	which	SUBR	VCL	15	–	–	JUNK	- - -
17	باعث	caused	ADJ	MOS	20	§n-13414سبب	§n-12756مشخصه	NO	- - Arg2
18	شگفتی	wonder	N	NEZ	17	§n-20348شگفت‌زدگی	§n-12725خصوصیت–روانی	ROLE-STATE	- - -
19	ش	his	PR	MOZ	18	§n-23937زکریا	§n-13075نفر	JUNK	- - -
20	می‌شد	be	V	PRD	16	–	–	NO	- - -
21	.	.	PUNC	PUNC	15	–	–	JUNK	- - -

a single class variable, but graphical models are powerful in the sense that they can model many interdependent variables. The sequential relation between output variables is perhaps the simplest form of dependency between output variables in graphical models. CRF predicts an output vector $\mathbf{y} = \{y_0, y_1, ..., y_T\}$ of random variables given an observed feature vector x when these random variables depend on each other. In the mapping problem at hand, the input is a sentence of a story, which is a sequence of tokens $\mathbf{x} = \{x_0, x_1, ..., x_T\}$. Each x_s is a feature vector of a token positioned at index s. The output is a sequence of scene model elements $\mathbf{y} = \{y_0, y_1, ..., y_T\}$ where each y_s is the scene element mapped to the token positioned at index s of the input sequence (sentence).

Scene model elements mapped from sentence tokens are interdependent. If one token in a sentence maps to ROLE in its corresponding conceptual scene model, it is more likely that other tokens of that sentence will map to the properties of that ROLE, properties such as ROLE-ACTION, ROLE-STATE, or ROLE-INTENT. In Persian language sentences, if a token is mapped

https://doi.org/10.1017/S1351324922000390 Published online by Cambridge University Press

Table 8. The Persian sentence "او گل زیبایی را دید." (translated as "He/She saw a beautiful flower.") to show the interdependencies of scene elements in a sentence. The tokens are ordered according to the sentence in the Persian language

No.	Token	In English	Scene model element
1	او	he/she	ROLE
2	گل	flower	STATIC-OBJECT
3	زیبایی	beautiful	STATIC-OBJECT-STATE
4	را	–	JUNK
5	دید	saw	ROLE-ACTION
6	.	.	JUNK

In English		FORM	GPOS	DEPREL	ARG	WSD	Super-WSD	Scene-Element
	8029							
each	8030	هر	ADJ	NPREMOD	null	null	null	junk
time	8031	وقت	N	ADVRB	ArgM_TMP	موقع$n-12614	دوره_زماني$n-12603	time
Zakariya	8032	زكريا	N	SBJ	Arg0	زكريا$n-23937	نفر$n-13075	role
Zakariya	8033	زكريا	N	SBJ	Arg0	زكريا$n-23937	نفر$n-13075	role
to	8034	به	PREP	VPP	Arg4	null	null	junk
visit	8035	ديدار	N	VPP	Arg4	ملاقات$n-10758	رخداد$n-13136	role_state
her	8036	او	PR	MOZ	null	null	null	junk
was going	8037	مي‌رفت	V	NCL	null	پيوستن$v-7700	رخداد$n-13136	role_action
food	8038	غذاهاي	N	OBJ	Arg1	غذا$n-13159	ماده$n-14032	static_object
special	8039	مخصوصي	ADJ	NPOSTMOD	null	مختص$a-1538	مشخصه$n-12756	static_object_state
in	8040	در	PREP	ADVRB	ArgM_LOC	null	null	junk
near	8041	كنار	N	POSDEP	null	نزديک$n-10189	جا$n-12733	location
sanctuary	8042	محراب	N	MOZ	null	null	null	location
her	8043	او	PR	MOZ	null	null	null	junk
notice	8044	مشاهده	N	NVE	null	null	null	role_state
noticed	8045	مي‌كرد	V	ROOT	null	رويت‌كردن$v-8581	رخداد$n-13136	role_action
which	8046	كه	SUBR	VCL	null	null	null	junk
caused	8047	باعث	ADJ	MOS	Arg2	سبب$n-13414	مشخصه$n-12756	no
wonder	8048	شگفتي	N	NEZ	null	شگفت‌زدگي$n-20348	خصوصيت_رواني$n-12725	role_state
his	8049	ش	PR	MOZ	null	null	null	junk
be	8050	مي‌شد	V	PRD	null	موجب شدن$v-7980	رخداد$n-13136	no
.	8051	.	PUNC	PUNC	null	null	null	junk

Figure 4. The part of the dataset produced based on the part of the completed corpus shown in Table 7.

to STATIC-OBJECT-STATE, it is very probable that the token before it in the sentence has been mapped to STATIC-OBJECT. This relation holds in reverse order in English sentences. The last token (excluding the punctuation mark) of most sentences in Persian is the action that has occurred in that sentence. Depending on the presence of a ROLE or an ANIMATED-OBJECT in the sentence, it will be denoted as a ROLE-ACTION or OBJECT-ACTION. A sample of these interdependent scene elements is shown in Table 8. Ignoring these sequential relations among the tokens of a sentence can cause some important information about the tokens to be discarded. As a sequential labeling graphical model, CRF can find the globally best label sequence for all tokens of a sentence, which is better than labeling each token of a sentence one at a time (Lafferty *et al.* 2001).

Another reason for choosing sequential modeling to solve the problem at hand is the imbalanced nature of the collected data and the ability of the CRF model to handle this imbalanced data (Sutton and McCallum 2012). The proportion of different elements of the conceptual scene model in the prepared dataset was diverse, as shown in Table 9. The proposed scene model consisted of 13 elements (including JUNK and NO labels). The proportion of the JUNK label was 50%, the ROLE and NO elements together comprised approximately 25% of the dataset, and the

https://doi.org/10.1017/S1351324922000390 Published online by Cambridge University Press

Table 9. The distribution of all scene elements across the completed corpus, produced dataset, training, and test sets. 4th column: The proportion percentage of each scene element in the dataset. Last column: The percentage of each scene element in the test set in proportion to its total number in the dataset

Scene element	No. in corpusin	No. in dataset	Percentage dataset	No. in Training set	No. in Test set	Percentage in Test set
ROLE	691	860	10.82%	738	122	14.19%
ROLE-ACTION	476	476	5.99%	405	71	14.92%
ROLE-STATE	526	538	6.77%	475	63	11.71%
ROLE-INTENT	329	331	4.17%	265	66	19.94%
ANIMATED-OBJECT	58	70	0.88%	48	22	31.43%
ANIMATED-OBJECT-STATE	26	27	0.34%	21	6	22.22%
OBJECT-ACTION	36	36	0.45%	20	16	44.44%
STATIC-OBJECT	122	132	1.66%	108	24	18.18%
STATIC-OBJECT-STATE	72	72	0.91%	60	12	16.67%
LOCATION	271	272	3.42%	213	59	21.69%
TIME	117	117	1.47%	106	11	9.40%
NO	1130	1145	14.41%	935	210	18.34%
JUNK	3802	3870	48.70%	3284	586	15.14%
Total	7656	7946	100%	6678	1268	

other ten elements comprised the remaining 25% of the dataset. To preserve the natural sequence of words in a sentence, none of the tokens belonging to the common stop word lists were eliminated from the dataset. These stop word tokens comprised half of the dataset and redoubled the imbalance. These imbalanced statistics were, to some extent, due to the domain of the selected stories, which was real stories about encounters by each prophet with his tribe. ROLE was a frequent scene element in comparison with STATIC/ANIMATED-OBJECT or other scene elements in those stories. The CRF model can handle this imbalanced data and learn the elements with a small number of samples, as in NER (Finkel, Grenager, and Manning 2005) and POS tagging problems (Pandian and Geetha 2009). The increase in the average accuracy of the mapping task in the third attempt (sequential modeling), that is, 85.7% compared with the second attempt (non-sequential modeling), that is, 76.58%, confirmed the ability of CRF to model sequential and imbalanced data.

CRF is a graphical model that uses traditional features of each token of a sentence in a dataset to predict the sequence of labels and also uses contextual information of neighbor tokens, such as their features or labels. It uses this contextual information to model the interdependencies among variables. Considering this property of CRF, its feature set can be very large and include the features and labels of one, two, three, or more tokens before and after the current token and their combinations. Nevertheless, an inefficient large feature set can extend the time required for training and can decrease accuracy. Selecting an efficient feature set is a critical success factor in CRF. The selection process of the final feature set for the CRF for the mapping problem at hand is discussed in Section 6.2. CRFsuite software (Okazaki 2007) was used to learn the CRF model in the current study.

https://doi.org/10.1017/S1351324922000390 Published online by Cambridge University Press

Table 10. The collected dataset statistics

	Token No.	Sentence No.	Scene No.	Story No.
Dataset	7946	451	232	12
Training set	6678	385	200	10
Test set	1268	66	32	2

6. Results and evaluation

6.1 Dataset statistics

The collected dataset had 7946 tokens in 451 sentences. These 451 sentences formed 232 scenes of 12 stories. The story lengths varied from 350 to 1500 tokens. Meeting the challenge of incorrect SRL tagging of sentences in the ITRC corpus eliminated some sentences and decreased the story lengths. The use of CRF as a sequential labeling model raises the expectation of a decrease in accuracy for stories with incomplete sentences (stories from which some sentences have been deleted). 15.96% of the 7946 tokens of the dataset (1268 tokens in 66 sentences, 32 scenes, and 2 stories) was separated for the test set. The distribution of each scene element across the training and test sets is shown in Table 9. The 3rd, 5th, and 6th columns show the number of each type of scene element in the collected dataset, training, and test sets, respectively. The last column shows the percentage of each scene element in the test set in proportion to its total number in the dataset. This was done to guarantee that the minimum proportion of the less frequent scene elements in the test set was approximately 12% (except one). 10% of training set was assigned to validation set. Table 10 summarizes the dataset statistics.

Precision, recall, and F1-score accuracy measures were calculated for evaluation. These accuracy measures are used publicly in text classification (Alpaydin 2014). The formulas used to calculate these measures are shown above.

$$precision = \frac{TP}{TP + FP} \quad recall = \frac{TP}{TP + FN}$$

$$F_1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 * precision * recall}{precision + recall}$$

TP = No. of correct classifications of positive tokens.
FP = No. of incorrect classifications of positive tokens.
TN = number of correct classifications of negative tokens.
FN = number of incorrect classification of negative tokens.

6.2 CRF feature set

The final selected feature set of CRF in this study contains 15 features, as shown in Table 11. As stated in Section 5, CRF can use the features of the neighbor tokens as contextual features of a token to learn the interdependencies between tokens of a sentence. To select the best feature set for the problem at hand, the collected dataset feature set of the token itself comprising FORM, GPOS, DEPREL, ARG, WSD, Super-WSD, and Scene-Element was selected as the base. Then, the features of neighbor tokens were temporarily added to the feature set so that the average F1-score accuracy measure for all scene elements could be monitored. Whenever adding a neighbor feature increased the average F1-score, that neighbor feature was added to the selected feature set permanently. Otherwise, it was removed. The different features tested during this process and the resulting average F1-score for the feature set are listed in Table 12.

https://doi.org/10.1017/S1351324922000390 Published online by Cambridge University Press

Table 11. The final selected feature set of CRF in this study

No.	Feature	Description
1-7	FORM, GPOS, DEPREL, ARG, WSD, Super-WSD, and Scene-Element	Features of the token itself
8	FORM[−1]	The FORM feature of the token before
9	GPOS[−1]	the GPOS feature of the token before
10	DEPREL[−1]	The DEPREL feature of the token before
11	ARG[−1]	The ARG feature of the token before
12	Super-WSD[−1]	The Super-WSD feature of the token before
13	WSD[+2]	the WSD feature of the two token after
14	Scene Element[−1] | Scene Element[−2] | Scene Element[−3]	The combination of the Scene Element feature of the one, two and three token before
15	Scene Element[+1] | Scene Element[+3]	The combination of the Scene Element feature of the one and three token after

6.3 Evaluation and discussion

The detailed accuracy measures with the selected feature set in Table 11 are provided in Table 13. As listed in Table 9, the ANIMATED-OBJECT-STATE and OBJECT-ACTION scene elements each had nearly 30 tokens in the total dataset; thus, the CRF could not learn them, so they were excluded from the evaluation. The average F1-score for the other ten scene elements (excluding NO as the unwanted label) was 85.7%, which is satisfactory.

Precision, recall, and F1-score measures for some scene elements, including ROLE, ROLE-ACTION, ANIMATED-OBJECT, LOCATION, TIME and JUNK, were satisfactory. The accuracy measures for ROLE-STATE and ROLE-INTENT were lower than others because of the complex distinction of states from intents for the ROLE samples. The confusion matrix shown in Figure 5 confirms this. These two (and also ROLE-ACTION) classes had the most tokens misclassified as NO. Nearly one-third of the ROLE-INTENT samples were misclassified as ROLE-STATE; hence, decreases occurred in ROLE-INTENT's recall and ROLE-STATE's precision. The misclassification of the ANIMATED-OBJECT-ACTIONs as ROLE-ACTION decreased ROLE-ACTION's precision significantly.

The lowest F1-score was for the STATIC-OBJECT-STATE scene element. The STATIC-OBJECT-STATE and ANIMATED-OBJECT were both low-frequency scene elements that had nearly 70 tokens in the total dataset (nearly 0.9% of the dataset), but the ANIMATED-OBJECT scene element was learned very well. This means that the selected feature set contained enough information to recognize the ANIMATED-OBJECT scene elements from others. For the STATIC-OBJECT-STATE, the few tokens in the training set prevented CRF from recognizing it from others. Only 2 of 12 actual STATIC-OBJECT-STATE tokens were correctly predicted, and nearly half of them were misrecognized as ROLE-STATE. However, CRF did not significantly confuse other scene elements as being STATIC-OBJECT-STATE. In the case of STATIC-OBJECT, the tokens were wrongly misclassified as STATIC-OBJECT-STATE, LOCATION, and NO, which resulted in a reduced recall. The presence of STATIC-OBJECTs in the description of LOCATIONs is a reason for confusion, as in "at the bottom of the sea," in which "sea" is annotated as a STATIC-OBJECT and "bottom" is annotated as LOCATION. Despite the low number of TIME tokens in the test set, only one of its tokens was misclassified as NO, and some NO tokens were wrongly predicted as TIME, which slightly reduced the precision.

https://doi.org/10.1017/S1351324922000390 Published online by Cambridge University Press

Table 12. The different features which were tested during the feature selection process and the resulting average F1-score for all scene elements using that feature set

No.	The feature set to be tested	Accept	Avg. F1-score
0	Base feature set: FORM, GPOS, DEPREL, ARG, WSD, and Super-WSD	base	84.1
1	Removing the declaration of "NO" as the unwanted label (the tokens with no mapping in the visual scene) and learning it as a scene element.	–	82.8
2	**FORM[–1]**	+	**84.2**
3	FORM[–2]	–	83.9
4	FORM|FORM[–1]	–	83.7
5	**GPOS–1**	+	**84.4**
6	GPOS|GPOS[–1]	–	83.8
7	GPOS[–2]	–	84.7
8	GPOS|GPOS[–1]|GPOS[–2]	–	83.6
9	**DEPREL[–1]**	+	**84.5**
10	DEPREL|DEPREL[–1]	–	83.7
11	DEPREL[–2]	–	84.0
12	DEPREL|DEPREL[–1]|DEPREL[–2]	–	84.2
13	**ARG[–1]**	+	**84.6**
14	ARG|ARG[–1]	–	83.6
15	ARG[–2]	–	85.0
16	ARG|ARG[–1]|ARG[–2]	–	84.4
17	WSD[–1]	–	84.2
18	WSD|WSD[–1]	–	84.3
19	WSD[–2]	–	84.7
20	WSD|WSD[–1]|WSD[–2]	–	85.0
21	**Super-WSD[–1]**	+	**85.0**
22	Super-WSD|Super-WSD[–1]	–	84.9
23	Super-WSD[–2]	–	84.5
24	Super-WSD|Super-WSD[–1]|Super-WSD[–2]	–	84.2
25	label[–1]	–	84.7
26	label[–2]	–	84.5
27	label[–1]|label[–2]	–	84.8
28	label[–3]	–	84.3
29	label[–1]|label[–3]	–	83.7
30	label[–2]|label[–3]	–	84.8

https://doi.org/10.1017/S1351324922000390 Published online by Cambridge University Press

Table 12. Continued

No.	The feature set to be tested	Accept	Avg. F1-score
31	**label[−1]|label[−2]|label[−3]**	+	**85.4**
32	FORM[+1]	−	84.9
33	FORM|FORM[+1]	−	84.6
34	FORM[+2]	−	84.5
35	FORM|FORM[+1]|FORM[+2]	−	84.5
36	GPOS[+1]	−	85.1
37	GPOS|GPOS[+1]	−	85.1
38	GPOS[+2]	−	85.1
39	GPOS|GPOS[+1]|GPOS[+2]	−	84.4
40	DEPREL[+1]	−	85.0
41	DEPREL|DEPREL[+1]	−	85.4
42	DEPREL[+2]	−	85.1
43	DEPREL|DEPREL[+1]|DEPREL[+2]	−	84.5
44	ARG[+1]	−	85.0
45	ARG|ARG[+1]	−	85.0
46	ARG[+2]	−	85.2
47	ARG|ARG[+1]|ARG[+2]	−	85.3
48	WSD[+1]	−	84.8
49	WSD|WSD[+1]	−	84.8
50	**WSD[+2]**	+	**85.6**
51	WSD|WSD[+1]|WSD[+2]	−	84.7
52	Super-WSD[+1],	−	85.0
53	Super-WSD|Super-WSD[+1]	−	85.0
54	Super-WSD[+2]	−	84.9
55	Super-WSD|Super-WSD[+1]|Super-WSD[+2]	−	84.7
56	label[+1]	−	85.4
57	label[+2]	−	84.9
58	label[+1]|label[+2]	−	85.0
59	label[+3]	−	84.9
60	**label+1]|label+3]**	+	**85.7**
61	label+2|label+3]	−	84.9
62	label+1|label+2]|label+3]	−	85.3

https://doi.org/10.1017/S1351324922000390 Published online by Cambridge University Press

Table 13. The precision, recall, and F1-score of the 10 scene elements of the conceptual scene model learned by the CRF model

Scene element	Precision	Recall	F1-score
ROLE	89.0	92.6	90.8
ROLE-ACTION	64.2	73.2	68.4
ROLE-STATE	44.3	61.9	51.7
ROLE-INTENT	66.7	54.5	60.0
ANIMATED-OBJECT	100	81.8	90
STATIC-OBJECT	66.7	41.7	51.3
STATIC-OBJECT-STATE	40.0	16.7	23.5
LOCATION	81.7	83.1	82.4
TIME	76.9	90.9	83.3
JUNK	94.7	97.8	96.2
Average accuracy	85.0	87.1	85.7

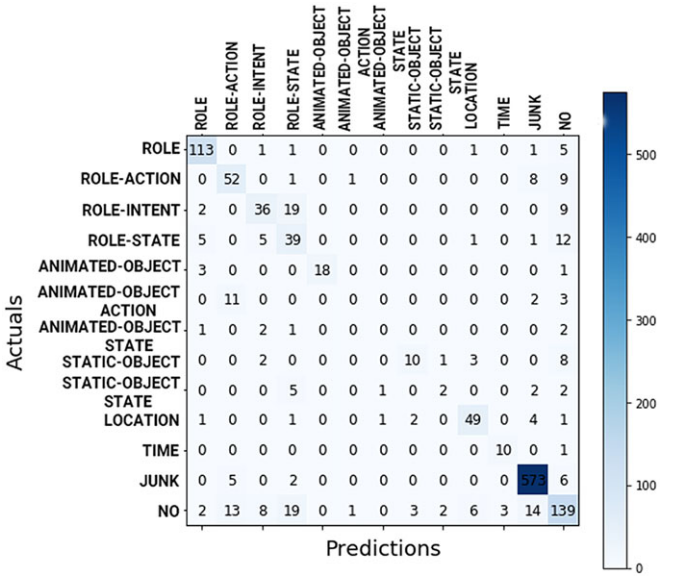


Figure 5. The heat map confusion matrix of all scene elements learned by the CRF model.

6.4 Feature ablation study to expand the dataset

The dataset collected in this study was limited to 7946 tokens because of the limited number of tokens in the SRL annotated corpus. WSD and Super-WSD tags also were annotated manually. The lack of more SRL annotated sentences and the costs of manual annotation of additional data are two barriers to expand the dataset in the future steps of the study. A feature ablation study was done to understand the proportion of these two (and other) annotation types to system performance. Each of the main annotation types, FORM, GPOS, DEPREL, ARG, WSD, Super-WSD, and their related features were removed one at a time, and then, the system performance was

https://doi.org/10.1017/S1351324922000390 Published online by Cambridge University Press

Table 14. A feature ablation study to understand the proportion of each annotation type to system performance. 3rd column: Average F1-score while features of Second column removed, 4th column: The difference of F1-score while some features are removed in comparison with the base feature set

Annotation type	Removed features	Avg. F1-score	Difference
FORM	FORM, FORM[-1]	**82.1**	**–3.6**
GPOS	GPOS, GPOS[-1]	84.5	–1.2
DEPREL	DEPREL, DEPREL[-1]	84.4	–1.3
ARG	ARG, ARG[-1]	**85.5**	**–0.2**
WSD	WSD, WSD[+2]	84.1	–1.6
Super-WSD	Super-WSD, Super-WSD[-1]	**82.2**	**–3.5**

measured. It is shown in Table 14 that removing the token itself (presented as FORM feature) and the Super-WSD feature reduced the average accuracy significantly. Conversely, removing the SRL feature (presented as ARG) caused the minimum accuracy reduction. These results showed that to expand the dataset for the future steps of the study, the existence of SRL tagged sentences is not critical, but manual annotation of WSD and Super-WSD of tokens could not be ignored.

6.5 Comparison with non-sequential learning

As stated in Section 1, in the second attempt to map Persian natural language story text to conceptual scene model elements, non-sequential machine learning models were applied. These models predict the scene element mapped to each token of the story one at a time. In this phase, Decision Table (Kohavi 1995) and ZeroR models were learned as rule-based classifiers and Decision Stump and J48 (Quinlan 1993) models were learned as decision trees using Weka 3 software (Frank, Hall, and Witten 2016). The SVM model was also tested. The dataset used in this phase was produced based on the completed corpus and consisted of FORM, GPOS, DEPREL, ARG, WSD, Super-WSD, and Scene-Element, as the dataset used for learning the CRF model, with this difference that tokens of all the sentences were concatenated. The division of the dataset into training and test sets was similar to that was used for the CRF model. The accuracy measures of these non-sequential models and the corresponding accuracy measure of the CRF model are shown in Table 15. The best average accuracy with this feature set was 76.58%, which was for the J48 model and was lower than the average accuracy of the CRF model (85.7%). The tokens with the JUNK label (nearly half of the dataset) were not removed in this dataset, but their removal is common in non-sequential modeling to prevent fake accuracies. As a second test case, the tokens with a JUNK label were removed from the dataset, which significantly reduced the average accuracy of the best non-sequential model to 60.70%. Preserving the natural order of tokens in sentences prevented application of this change for the CRF model. The imbalanced nature of the scene elements in the collected dataset, shown in Table 9, was one reason for the low performance of these non-sequential models. A filter was applied to the dataset in the first test case (the most comparable case to the CRF dataset; dataset which includes JUNK tokens) to artificially balance the number of tokens of each scene element to approximately 600. Applying this filter reduced the accuracy measure of each scene element (excluding JUNK and NO) minimally 6% and maximally 30%. The results are shown in the fifth column of Table 15. A significant decrease in the accuracy of all models confirms the natural imbalance of the scene elements mapped from the story tokens, which could not be handled by the artificial balancing of the dataset.

https://doi.org/10.1017/S1351324922000390 Published online by Cambridge University Press

Table 15. The accuracy measures of non-sequential models in the second attempt to solve the mapping problem. Columns 3–5 are average F1-score for all scene elements with a different feature set. 3rd column: The dataset similar to the CRF dataset. 4th column: Tokens with JUNK label were removed from the dataset. 5th column: A filter was applied to artificially balance the dataset as preprocessing

No.	Model	Avg. F1-score dataset1	Avg. F1-score dataset2	Avg. F1-score dataset1 class balanced
1	ZeroR	46.21	34.02	46.21
2	Decision Table	73.11	55.28	64.04
3	J48	**76.58**	**60.70**	**68.38**
4	Decision stump	59.38	36.36	7.41
5	SVM	53.71	40.27	47.63
	CRF	**85.7**	–	–

Table 16. Detailed accuracy measures of the J48 model with feature set: FROM, GPOS, DEPREL, ARG, WSD, Super-WSD, and Scene-Element, based on the completed corpus. 5th, 6th, and Last column: precision, recall, and F1-score of CRF model with the same feature set

Scene element	Precision of J48	Recall of J48	F1-score of J48	Precision of CRF	Recall of CRF	F1-score of CRF
ROLE	88.7	90.2	89.4	89.0	92.6	90.8
ROLE-ACTION	46.8	**81.7**	59.5	64.2	73.2	**68.4**
ROLE-STATE	34.5	60.3	43.9	44.3	61.9	**51.7**
ROLE-INTENT	58.6	25.8	35.8	66.7	54.5	**60.0**
ANIMATED-BJECT	58.6	77.3	66.7	100	81.8	**90**
STATIC-OBJECT	24.1	29.2	26.4	66.7	41.7	**51.3**
STATIC-OBJECT-STATE	**100**	8.3	15.4	40.0	16.7	**23.5**
LOCATION	84.6	55.9	67.3	81.7	83.1	**82.4**
TIME	76.9	90.9	83.3	76.9	90.9	83.3
JUNK	88.2	99.1	93.3	94.7	97.8	**96.2**
Weighted average	80	76.6	75.3	85.0	87.1	**85.7**

The detailed accuracy measures of the non-sequential model with the best results, J48, with the first test case dataset are shown in Table 16. The strengths and weaknesses of this model for the prediction of mapping between tokens and scene elements are similar to CRF. ROLE, ROLE-ACTION, ANIMATED-OBJECT, LOCATION, TIME, and JUNK were the best scene elements learned by J48. All scene elements were learned better by CRF. The precision of STATIC-OBJECT-STATE and recall of ROLE-ACTION increased significantly for J48. Nearly all accuracy measures of STATIC-OBJECT were decreased in comparison with CRF. This means that discarding sequential information about tokens, such as their position in the sentence and the scene elements of the neighboring tokens, caused misclassification of other scene elements as STATIC-OBJECT. Non-sequential learning did not recognize the tokens with STATIC-OBJECT-STATE and ROLE-INTENT labels and misclassified them as other scene elements.

https://doi.org/10.1017/S1351324922000390 Published online by Cambridge University Press

7. Conclusion

The present study is part of PERSIS MEANS and maps Persian natural language text to a conceptual scene model aimed at generating meaningful animation based on an input story. The data-driven approach of learning a CRF model was used for the mapping task. To the best of the authors' knowledge, the required dataset does not exist; thus, as the headmost TTSC system, which converts Persian text, a dataset was collected for this task. This process faced challenges such as a lack of required off-the-shelf NLP modules and a significant error rate in the output of available modules or corpora. Some of the required information was available in a corpus with a limited number of tokens. Human annotators manually annotated the available corpus with the required information for the intended dataset. Evaluation of the results showed acceptable accuracy for the mapping task. The next stage of the research will enrich the conceptual scene model mapped from the input text, in this stage, with conceptual factors to enable the production of meaningful animation.

Acknowledgments. We acknowledge the Iran Telecommunication Research Center (ITRC), specially the Qur'anic Question and Answer Project team for their great help and providing the NLP modules required for this research. We would also like to express our gratitude for Mojgan Farhoodi, Ehsan Darrudi, Maryam Mesgar and Meisam Ahmadi for their invaluable guidance and consultation.

References

Adorni G., **Di Manzo M.** and **Giunchiglia F.** (1984). Natural language driven image generation. In *Proceedings of the 10th International Conference on Computational Linguistics*, COLING 1984, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 495–500.

Alpaydin E. (2014). *Introduction to Machine Learning*, 3rd Edn. Cambridge, MA: The MIT Press.

Arian N. and **Sabbagh M.** (2017). Semantic labeling of sentences in Persian language with supervised method. In *Proceedings of the 22nd National CSI Computer Conference*, CSICC 2017, Tehran, Iran. Computer Society of Iran, pp. 1–8.

Chang A.X., **Eric M.**, **Savva M.** and **Manning C.D.** (2017). SceneSeer: 3D Scene Design with Natural Language. CoRR, pp. 1–10.

Chang A.X., **Monroe W.**, **Savva M.**, **Potts C.** and **Manning C.D.** (2015). Text to 3D scene generation with rich lexical grounding. In *The 53rd Annual Meeting of the Association for Computational Linguistics and The 7th International Joint Conference of the Asian Federation of Natural Language Processing*, Beijing, China. Association for Computational Linguistics, pp. 1–10.

Chang A.X., **Savva M.** and **Manning C.D.** (2014a). Learning spatial knowledge for text to 3D scene generation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar. Association for Computational Linguistics, pp. 2028–2038.

Chang A.X., **Savva M.** and **Manning C.D.** (2014b). Semantic parsing for text to 3D scene generation. In *Workshop on Semantic Parsing*, Baltimore, Maryland, USA. Association for Computational Linguistics, pp. 17–21.

Coyne B., **Rambow O.**, **Hirschberg J.** and **Sproat R.** (2010). Frame semantics in text-to-scene generation. In *Knowledge-Based and Intelligent Information and Engineering Systems*, Lecture Notes in Computer Science, vol. 6279. Springer Berlin Heidelberg, pp. 375–384.

Coyne B. and **Sproat R.** (2001). WordsEye: an automatic text-to-scene conversion system. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH 2001, New York, NY, USA. ACM, pp. 487–496.

Fillmore C. (1982). Frame semantics. *Linguistics in the Morning Calm.* Hanshin Publishing Company, pp. 111–137.

Finkel J.R., **Grenager T.** and **Manning C.** (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL 2005, Stroudsburg, PA, USA. Association for Computational Linguistics, pp. 363–370.

Fort K., **Adda G.** and **Cohen K.B.** (2011). Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics* **37**(2), 413–420.

Frank E., **Hall M.A.** and **Witten I.H.** (2016). *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, 4th Edn. Morgan Kaufmann.

Glass K. and **Bangay S.** (2008). Automating the creation of 3D animation from annotated fiction text. In *IADIS 2008: Proceedings of the International Conference on Computer Graphics and Visualization 2008*, MM'10, Amsterdam, The Netherlands. IADIS Press, pp. 3–10. 00006.

Glass K. and **Bangay S.** (2009). A method for automatically creating 3D animated scenes from annotated fiction text. *International Journal on Computer Science and Information System* **4**(2), 103–119.

https://doi.org/10.1017/S1351324922000390 Published online by Cambridge University Press

Hassani K. and **Lee W.-S.** (2016). Visualizing natural language descriptions: A survey. *ACM Computing Surveys (CSUR)* **49**(1), 1–34.

Helfiandri M.A., **Zakhralativa Ruskanda F.** and **Khodra M.L.** (2020). Generating Scene Descriptor from Indonesian Narrative *Text*. vol. CFP2013V-ART, Bandung, Indonesia. IEEE, pp. 1–6.

Hong J.-H., **Cho S.-H.**, **Jeon J.-U.** and **Park S.-Y.** (2018). Development and evaluation of text-to-scene model for Korean language writing education as a Foreign language. *Journal of The Korean Society for Computer Game* **31**(3), 63–70.

Iran Telecommunication Research Center (2014). Qur'anic Question and Answer Project. http://quranjooy.itrc.ac.ir.

Jackendoff R. (1990). *Semantic Structures*. *Current Studies in Linguistics Series*, vol. 18. Cambridge, MA: MIT Press.

Jain P., **Bhavsar R.**, **Kumar A.**, **Pawar B.V.**, **Darbari H.** and **Bhavsar V.C.** (2018). Tree adjoining grammar based parser for a Hindi text-to-scene conversion system. In *3rd International Conference for Convergence in Technology*, I2CT, Pune, India. IEEE, pp. 1–7.

Johansson R., **Nugues P.** and **Williams D.** (2004). Carsim: A system to convert written accident reports into animated 3D scenes. In *Proceedings of the 2nd Joint SAIS/SSLS Workshop Artificial Intelligence and Learning Systems, AILS-04*. Department of Computer Science, Lund University, pp. 76–86.

Kayser D. and **Nouioua F.** (2009). From the textual description of an accident to its causes. *Artificial Intelligence* **173**(12), 1154–1193.

Kohavi R. (1995). The power of decision tables. In *Proceedings of the 8th European Conference on Machine Learning*, ECML 95, Berlin, Heidelberg. Springer Berlin Heidelberg, pp. 174–189.

Lafferty J., **McCallum A.** and **Pereira F.C.** (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning (ICML)*, MA, USA. Morgan Kaufmann, pp. 282–289.

Landis J.R. and **Koch G.G.** (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174.

Lu R.-Q. and **Zhang S.-M.** (2002). From story to animation–full life cycle computer aided animation generation. *Acta Automatica Sinica* **28**, 321–348.

Ma M. (2006). *Automatic Conversion of Natural Language to 3D Animation*. PhD Thesis, University of Ulster.

Mesgar M., **Hajizade M.**, **Darrudi E.**, **Farhoodi M.**, **Mohamadzade M.**, **Alavi T.**, **Davoudi M.**, **Sarabi Z.** and **Khalash M.** (2014). Semantic role labeling of Persian language based on dependency tree. Technical report, Iran Telecommunication Research Center, Tehran, Iran. sent to get published.

Miaoulis G. and **Plemenos D.** (2009). *Intelligent Scene Modelling Information Systems*. *Studies in Computational Intelligence*, vol. **181**. Berlin, London: Springer. 00000.

Miller G.A. (1995). WordNet: A lexical database for English. *Communications of the ACM* **38**(11), 39–41.

Nazari M. (2006). Film production and play.

Okazaki N. (2007). CRFsuite: A fast implementation of Conditional Random Fields (CRFs).

Palmer M., **Gildea D.** and **Kingsbury P.** (2005). The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics Journal* **31**, 1.

Pandian S.L. and **Geetha T.V.** (2009). CRF models for tamil part of speech tagging and chunking. In *Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy*, Berlin, Heidelberg. Springer Berlin Heidelberg, pp. 11–22.

Pardhi V., **Shah K.**, **Vaghasiya J.** and **Hole V.** (2021). Generating a scene from text for smart education. In *ICCICT*, Mumbai, India. IEEE, pp. 1–6.

Quinlan J.R. (1993). *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.

Qur'anic Question and **Answer Project** (2014a). Semantic role labeling manual of style. Technical report, Iran Telecommunication Research Center, Tehran, Iran.

Qur'anic Question and **Answer Project** (2014b). Syntactic labeling manual of style on the basis of dependency grammar in Persian. Technical report, Iran Telecommunication Research Center, Tehran, Iran.

Rouhizadeh M. (2013). *Collecting Semantic Information for Locations in the Knowledge Resource of a Text-to-Scene Conversion System*. *Master of Science*, Oregon Health & Science University, Oregon, USA.

Ruppenhofer J., **Ellsworth M.**, **Petruck M.R.**, **Johnson C.R.** and **Scheffczyk J.** (2016). *FrameNet II: Extended Theory and Practice*. Berkeley, CA: International Computer Science Institute.

Shamsfard M. (2011). Challenges and open problems in Persian text processing. In *5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. Lecture Notes in Artificial Intelligence, vol. 8387. Poznan, Poland: Springer, pp. 65–69.

Shamsfard M., **Hesabi A.**, **Fadaei H.**, **Mansoory N.**, **Famian A.**, **Bagherbeigi S.**, **Fekri E.**, **Monshizadeh M.** and **Assi S.M.** (2010a). Semi automatic development of farsnet; the persian wordnet. In *Proceedings of 5th Global WordNet Conference*, GWA2010, vol. 29, Mumbai, India. Indian Institute of Technology.

Shamsfard M., **Jafari H.S.** and **Ilbeygi M.** (2010b). STeP-1: A set of fundamental tools for Persian text processing. In *7th Language Resources and Evaluation Conference*, LREC 2010, Valletta, Malta. European Language Resources Association, pp. 859–865.

Surdeanu M., **Johansson R.**, **Meyers A.**, **Marquez L.** and **Nivre J.** (2008). The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning (CoNLL 2008)*, Manchester, UK. Association for Computational Linguistics, pp. 159–177.

Sutton C. and **McCallum A.** (2012). An introduction to conditional random fields. *Foundations and Trends in Machine Learning* **4**(4), 267–373.

Tabibzadeh O. (2006). *Verb Capacity and Fundamental Structure of Sentence in Current Persian*. Tehran, Iran: Markaz Publishing.

Takahashi N., **Ramamonjisoa D.** and **Ogata T.** (2007). A tool for supporting an animated movie making based on writing stories in xml. In *Proceedings of IADIS International Conference Applied Computing*, Salamanca, Spain. International Association for Development of the Information Society, pp. 405–409.

Ustalov D. and **Kudryavtsev A.** (2012). An ontology-based approach to text-to-picture synthesis systems. In *Proceedings of the Second International Workshop on Concept Discovery in Unstructured Data (CDUD 2012) In Conjunction with the Tenth International Conference on Formal Concept Analysis (ICFCA 2012)*, vol. 871, Leuven, Belgium. Katholieke Universiteit Leuven, pp. 94–101.

Yadav P., **Sathe K.** and **Chandak M.** (2020). Generating animations from instructional text. *International Journal of Advanced Trends in Computer Science and Engineering* **9**(3), 3023–3027.

Zeng X., **Tan M.-l.** and **Ren S.** (2016). The implementation of graphic constraints for automatic text to scene conversion. In *International Conference on Artificial Intelligence and Computer Science*, AICS 2016, Guilin, China. World Scientific Pubilshing Company, pp. 364–367.

Cite this article: Hashemi-Namin M, Jahed-Motlagh MR and Torkaman Rahmani A (2023). Recognition of visual scene elements from a story text in Persian natural language. *Natural Language Engineering* **29**, 693–719. https://doi.org/10.1017/S1351324922000390