RESEARCH ARTICLE OPEN ACCESS

A Model for Employing Semantic Role Labeling To Extract Predicate Argument Structure

Hadia Abbas Mohammed Elsied [1], Naomie Salim [2], Atif Khan [3]
Department of Computer Science [1], Sudan University of Science & Technology Khartoum, Sudan
Faculty of Computing [2], Universiti Teknologi Malaysia, 81310, Skudai, Johor, Malaysia
Faculty of Computing [3], Universiti Teknologi Malaysia, 81310, Skudai, Johor, Malaysia

ABSTRACT
In this study we aim to represent a dataset from sentence level form to predicate argument structure form, which is considered as a higher- level of abstraction. This new representation can be processed further in various applications such as text summarization and plagiarism detection .We use SRL (Semantic Role labeling) to identify sentence constituents then we implement a model to extract the predicate argument structure from the sentences that undergo SRL automatically , we compare our results with a manual predicate argument structure extraction, we got a good result according to precision and recall values.
Keywords :— SRL , NLP , predicate Argument Structure

I. INTRODUCTION

Semantic Role Labeling (SRL) has been widely applied in text content analysis tasks such as text retrieval[1], information extraction[2], text categorization [3] and sentiment analysis [4]. In the area of text summarization, [5]introduced a work that combined semantic role labeling with general statistic method (GSM) to determine important sentences for single document extractive summary, also [6] introduce a work of abstractive summarization uses SRL.

The sentence-level semantic analysis of text is related with the characterization of events, such as figuring out "who" did "what" to "whom," "where," "when," and "how." "what" took place is established by the predicate of the statement or the clause , and the other remaining sentence constituents expresses the contributors in the event (such as "who" and "where"), as well as further event properties (such as "when" and "how").The main task of semantic role labeling (SRL) is to indicate exactly what semantic relations hold among a predicate and its associated participants and properties. Exemplary roles used in SRL are labels such as Agent, Patient, and Location for the entities participating in an event, and Temporal and Manner for the characterization of other aspects of the event or participant relations. This type of role labeling thus yields a first- level semantic representation of the text that indicates the basic event properties and relations among relevant entities that are expressed in the sentence [7].

II. SEMANTIC ROLE LABELLING (SRL)

SRL is a task in natural language processing (NLP) consisting of detection of the semantic arguments associated with the predicate or verb of a sentence and their classification to their specific roles , more over it is the underlying relationship that a participant has with the main verb in the clause [4], also known as semantic case, thematic role, theta role (generative grammar), and deep case (case grammar). The goal of SRL is to discover the predicate argument structure of each predicate in a given input sentence[5] . According to [6] the task of SRL is to find all arguments for a given predicate in a sentence and label them with semantic roles.

Semantic role labeling (SRL) is a process to identify and label arguments in a text. SRL can be extended for the events characterization task that answer simple questions such as "who" did "what" to "whom", "where", "when", and "how".The main task of SRL is to show what specific relations hold among a predicate with respect to its associated participants . As the definition of the PropBank and CoNLL-2004 shared task [10] there are six different types of arguments labeled as A0-A5 and AA. These labels have different semantics for each verb as specified in the PropBank Frame files. In addition, there are also 13 types of adjuncts labeled as AM-adj where adj specifies the adjunct type. SRL aims to identify the constituents of a sentence, with their roles such as Agent, Patient, Instrument etc., and the adjunctive arguments of the predicate such as Locative, Temporal, with respect to the sentence predicates [3]. This type of role labeling thus produce a first level semantic representation of

the text that indicates the basic event properties and relations among relevant entities that are expressed in the sentence [7].

In this study we employ SRL to extract the Predicate Argument Structure (PAS) to be as a representation for our dataset, we use DUC 2002 dataset , for the SRL we use SENNA toolkit[9] .SENNA is a software distributed under a non-commercial license, which produces a host of Natural Language Processing (NLP) predictions: semantic role labeling (SRL) ,part-of-speech (POS) tags, chunking (CHK) and name entity recognition (NER). As a preprocess for our dataset we extract the text only out of other document's data and tags , next we employ the SRL to parse each sentence and label the semantic phrases /words in each sentence properly , we referred to these phrases as semantic arguments . Semantic arguments are accumulated in tow groups : core arguments (Arg) and adjunctive arguments (ArgM) as illustrated in table 1.

In this study, we consider A0 for subject, A1 for object, A2 for indirect object as core arguments, and ArgM-LOC for location, ArgM-TMP for time , ArgM-DIR, ArgM-MNR as adjunctive arguments , V for predicate (Verb). We put into account all the complete predicates associated with the single sentence structure so as to avoid loss of important terms/words that participate to the meaning of a sentence and its predicates. We suppose that predicates are complete if they have at least two semantic arguments. The predicate argument structure which is extracted used as semantic representation for each sentence in the document collection. We represent the sentence which contains one predicate by simple predicate argument structure where the sentence which contains more than one predicate will be represented by composite predicate argument structure that is the number of predicates in a sentence is equal to the number of extracted predicate argument structure extracted from the same sentence.

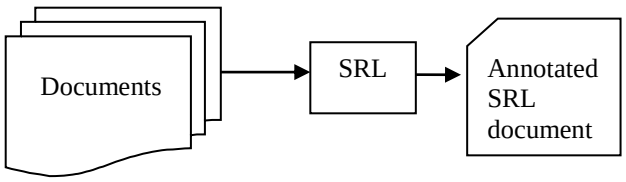

Fig. 1 The general process of SRL

Hurricane	NNP	0	-	B-A1	0	0	0
Gilbert	NNP	S-PER	-	E-A1	0	0	0
swept	VBD	0	swept	S-V	0	0	0
toward	IN	0	-	B-A2	0	0	0
the	DT	0	-	I-A2	0	0	0
Dominican	NNP	B-LOC	-	I-A2	0	0	0
Republic	NNP	E-LOC	-	I-A2	0	0	0
Sunday	NNP	0	-	E-A2	0	0	0
,	,	0	-	0	0	0	0
and	CC	0	-	0	0	0	0
the	DT	0	-	0	B-A0	0	0
Civil	NNP	B-ORG	-	0	I-A0	0	0
Defense	NNP	E-ORG	-	0	E-A0	0	0
alerted	VBD	0	alerted	0	S-V	0	0
its	PRP$	0	-	0	B-A1	0	B-A0
heavily	RB	0	-	0	I-A1	S-AM-MNR	I-A0
populated	VBN	0	populated	0	I-A1	S-V	I-A0
south	JJ	0	-	0	I-A1	B-A1	I-A0
coast	NN	0	-	0	I-A1	E-A1	E-A0
to	TO	0	-	0	I-A1	0	0
prepare	VB	0	prepare	0	I-A1	0	S-V
for	IN	0	-	0	I-A1	0	B-A2
high	JJ	0	-	0	I-A1	0	I-A2
winds	NNS	0	-	0	I-A1	0	I-A2
,	,	0	-	0	I-A1	0	I-A2
heavy	JJ	0	-	0	I-A1	0	I-A2
rains	NNS	0	-	0	I-A1	0	I-A2
and	CC	0	-	0	I-A1	0	I-A2
high	JJ	0	-	0	I-A1	0	I-A2
seas	NNS	0	-	0	E-A1	0	E-A2
.	.	0	-	0	0	0	0

Fig. 2 An example of an annotated sentence, in columns. Input consists of words (1st), PoS tags (2nd), named entities (3rd). The 4th column marks target verbs, and their propositions are found in remaining columns.

Figure 2 shows an excerpt of a document after undergo SRL , the results are in columns where 1st column consists of the words tokens , 2nd column contain the Part of speech tagging , 3rd column is name entity recognition , 4th column is targeted verbs and remaining columns contains the role labeling for each targeted verb(predicate), as shown in figure2 we have 4 targeted verb and this implies 4 columns for role labeling , each corresponds to one verb.

TABLE I
SEMANTIC ROLE LABELING(SRL)ARGUMENTS

Arg_labelling		Arg_modifier	
rel	verb	*ArgM-ADV*	adverbial modification
A0	subject	*ArgM-DIR*	direction
A1	object	*ArgM-DIS*	discourse marker
A2	Indirect object	*ArgM-EXT*	extent marker
A3	Start point	*ArgM-LOC*	location
A4	End point	*ArgM-MNR*	manner
A5	Direction	*ArgM-MOD*	general modification
		ArgM-NEG	negation
		ArgM-PRD	secondary predicate
		ArgM-PRP	purpose
		ArgM-REC	reciprocal
		ArgM-TMP	temporal marker

III. PREDICATE ARGUMENT STRUCTURES

The form of apredicate(verb) along with its Arguments is called predicate argument structure .

In this study we consider two types of predicate argument structures (PAS) simple predicate argument structure and composite argument structure , the simple one is considered if we have one verb in a sentence and the composite one if we have more than one verb in the sentence , and we consider a PAS as a PAS if at least contains a verb and one other argument A0 or A1 .

Example

Consider the following sentence represented by composite predicate argument structures.

S:Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas.

The corresponding composite predicate argument structures PAS1 and PAS2 are obtained after applying semantic role labeling to sentence *S*

Precision	Recall	F_measure
0.905598	0.932293	0.918455

PAS1: [A1: Hurricane Gilbert] [V: swept] [A2: toward the Dominican Republic Sunday].

PAS2: [A0: the Civil Defense][V: alerted][A1: its heavily populated south coast to prepare for high winds , heavy rains and high seas]

IV. EXPERIMENT

This section illustrates the practical part of our proposed predicate argument structure extraction scheme .

in this experiment we use the DUC 2002 [13] dataset ,first we did preprocess for the dataset such as removing the html tags and segmenting the text in to separate sentences then we employ the Semantic Role Labeling (SRL) , we use SENNA toolkit to achieve the SRL , Then the SRL files are used as input to our system , the system extract PAS's from each sentence in the SRL file, the extracted PAS's are equal to the number of verbs associated to each sentence , which can be processed further for many other tasks such as summarization , categorization and classification .

V. EVALUATION

We evaluate our results with a manual ones , we got high precision and recall which asserts that our model can be characterized as excellent , as shown in the Figure 2.

we use the formula shown in figure 3 to calculate the precision and recall.

$$Precision = \frac{SPP \cap HPP}{HPP}$$

$$Recall = \frac{SPP \cap HPP}{SPP}$$

$$F\_Measure = \frac{2 \times Precision \times Recall}{Recall + precision}$$

Fig. 3 Recall ,Precision formula's

Where SPP is System produced PAS's and HPP is Human produced PAS's .The average precision and recall for the tested DUC 2002 documents are shown in table 2.

TABLE2
RECALL,PRECISION AND F_MEASURE FOR SYSTEM

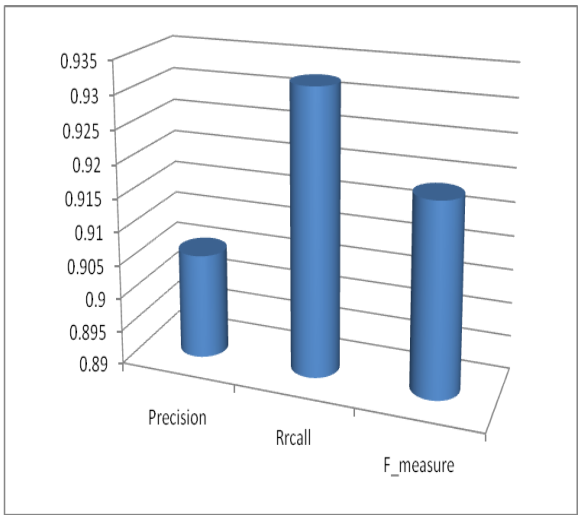

Fig. 4 Recall ,Precision and F_measure

VI. CONCLUSIONS

SRL can be employed for each sentence in a document in the intended DUC 2002 dataset to extract predicate argument structure which is considered as semantic representation of sentence to be used further for other applications such as summarization , categorization and classification to extract PAS's we use SENNA toolkit to employ SRL then we use our

system to extract PAS's from those SRL files .We evaluate our results with a manual ones , we got high precision and recall which asserts that our model can be characterized as exellent.

REFERENCES

[1] Y. Liu, H. T. Loh, Y. T. Kamal, and S. B. Tor, Natural Language Processing and Text Mining. 2007.

[2] S. Shehata, F. Karray, and M. S. Kamel, "An efficient concept-based retrieval model for enhancing text retrieval quality," Knowl. Inf. Syst., vol. 35, no. 2, pp. 411–434, 2013.

[3] C. K. Emani, C. Ferreira, D. Silva, B. Es, and P. Ghodous, "Improving Open Information Extraction using Domain Knowledge," pp. 1–7,vol. 34, pp. 257–287, 2008.

[4] R. Johansson and J. Persson, "Text Categorization Using Predicate – Argument Structures Department of Computer Science," vol. I, pp. 142–149, 2009.

[5] P. Teknikal, U. Pendidikan, S. Idris, T. Malim, and T. M. Perak, "Jurnal Teknologi," vol. 4, pp. 185–190, 2015.

[6] L. Suanmali, N. Salim, and M. S. Binwahlan, "Fuzzy Genetic Semantic Based Text Summarization," 2011 IEEE Ninth Int. Conf. Dependable, Auton. Secur. Comput., pp. 1184–1191, Dec. 2011.

[7] A. Khan, N. Salim, and Y. Jaya Kumar, "A framework for multi-document abstractive summarization based on semantic role labelling," Appl. Soft Comput., vol. 30, pp. 737–747, 2015.

[8] L. Màrquez, X. Carreras, K. C. Litkowski, and S. Stevenson, "Semantic Role Labeling: An Introduction to the Special Issue," Comput. Linguist., vol. 34, pp. 145–159, 2008.

[9] X. Carreras, L. Màrques, and L. Màrquez, "Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling," Proc. CoNLL-2004, pp. 89–97, 2004.

[10] C. Aksoy, A. Bugdayci, T. Gur, I. Uysal, and F. Can, "Semantic Argument Frequency-Based Multi-Document Summarization," pp. 470–474, 2009.

[11] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural Language Processing (almost) from Scratch," J. Mach. Learn. Res., vol. 12, no. Aug, pp. 2493–2537, 2011.

[12] V. Punyakanok, D. Roth, and W. Yih, "The Importance of Syntactic Parsing and Inference in Semantic Role Labeling," Comput. Linguist.,

[13] P. Over, "Introduction to DUC-2002 : an Intrinsic Evaluation of Generic News Text Summarization Systems Document Understanding Conferences (DUC)…," 2002.