Middle-East Journal of Scientific Research 23 (6): 1222-1227, 2015
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2015.23.06.22276

A Comparative Study on Arabic Grammatical Relation Extraction Based on Machine Learning Classification

Mohanaed Ajmi Falih and Nazlia Omar

Center for AI Technology, Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia (UKM) Bangi, Selangor, Malaysia

Abstract: Grammatical Relation (GR) can be defined as a linguistic relation established by grammar, in which the linguistic relation is an association between linguistic forms or constituents. Fundamentally, GRs determine grammatical behavior, such as the placement of a word in a clause, verb agreement and passivity behavior. The GR of Arabic is aprerequisite for many natural language processing applications, such as machine translation and information retrieval. This study focuses on Arabic GR-related problems. The main difficulty of determining grammatical relations in Arabic sentences is ambiguity. Such grammatical ambiguity is caused by the large and complex nature of Arabic sentences. This study primarily aims to develop an efficient GR extraction technique to analyze modern standard Arabic sentences and address these issues with an optimum solution. This paper proposes a machine learning classification method to recognize subject, object and verb. To extract the correct subject, object and verb from sentence structure, the proposed technique enhances the basic representations of Arabic using Support Vector Machines (SVM), k-Nearest Neighbor (KNN) and a combination between SVM and KNN algorithms. The system used 80 Arabic sentences as a training and test data set, with the length of each sentence ranging from 3 to 20 words. The results obtained by combination classification between SVM and KNN algorithms achieved 94.44% recall, 93.33% precision and 93.48% F-measure. This result proves the viability of this approach for GR extraction of Arabic sentences.

Key words: Arabic language processing・Feature extraction・Machine learning classification

INTRODUCTION

In linguistics, a grammatical relation (GR) is defined as the correlation and connection between the constituents in a clause. Common examples of GRs in conventional grammar are the direct object, indirect object and subject. GRs are also referred to as syntactic functions. These functions are usually the typical classes of object and subject and are crucial in linguistic theory, involving a variety of approaches ranging from functional and cognitive theories to generative grammar.

Numerous modern grammar theories likely recognize many other types of grammatical relations, which are complementary, predicative and specific. The most important role of GRs within grammar theories involves dependency grammars, which are accompanied by several distinct grammatical relations. Each individual dependency grammar performs a grammatical function. More often than not, experts and researchers in linguistics and grammar are able to identify the subject and object within a particular clause or sentence. However, their attempts to theoretically propose appropriate definitions for these concepts are usually quite vague and, therefore, arguable.

These arguments arise in cases where many grammar theories confirm the grammatical relations and rely heavily on them for describing the concepts of grammar, while steering clear of providing credible definitions. However, many values can be verified to describe grammatical relations.The precision and recall of bracketed constituents are frequently implemented in parser assessment metrics and the structure of the syntactic constituents of sentences is typically viewed as the output of a parser. Alternatively, sentences are analyzed for various reasons by many types of parsers via different methods. A diagram to depict the structures of constituents is usually not the most appropriate kind of output.

Correponding Author: Mohanaed Ajmi Falih, Center for AI Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM) Bangi, Selangor, Malaysia.

Both the precision and recall of GRs can be executed to evaluate parsers and several advantages of implementing GRs compared to other types of evaluation metrics have been discussed in the literature [1]. The use of GRs is prompted by importance of this information in the analysis of the syntactic complexity in various situations in linguistics.

A grammatical relation is defined as a form of linguistic connection based on grammar, which can usually be found among several constituents and linguistic forms [2]. The extraction of GRs essentially determines grammatical actions, such as the placing of a certain term in a sentence or clause, verb-based agreement and passive behavior. The Arabic language in general requires the extraction of GRs as a condition for many natural language processing (NLP) programs and applications, including machine translation and information retrieval. This chapter providesa description of the methods employed by previous studies, namely machine learning clustering and classification, to resolve this issue and the various GRs that have been generated as a result.

Numerous studies have employed different methods to propose a language parser in several different languages, but only a few works have focused primarily on GR extraction. Most methods for a full parser do not focus specifically on the extraction of grammatical relations. Several applications are available, such as the creation of an Arabic-based parser, Arabic parsing via Grammar Transforms, a machine learning-based classification for the GR of Arabic terms and the POLA-based grammar approach for GR extraction in the Malay language [3].

The machine learning method of general classification may help to resolve the current issues, including morphology [4, 5] and syntactic parsing [6]. Importantly, precision and recall are the most common methods used to assess GR extraction models, because both methods for the bracketed constituents are usually implemented as assessment-based metrics for parsers. This implementation often describes the constituent syntactic structure of the sentences or phrases as the output of a particular parser. On the other hand, sentences are evaluated by different types of parsers using various methods and for various purposes. Depicting constituent structures via diagrams is not always appropriate. The aim of this paper is Arabic GR extractions based on machine learning classification.

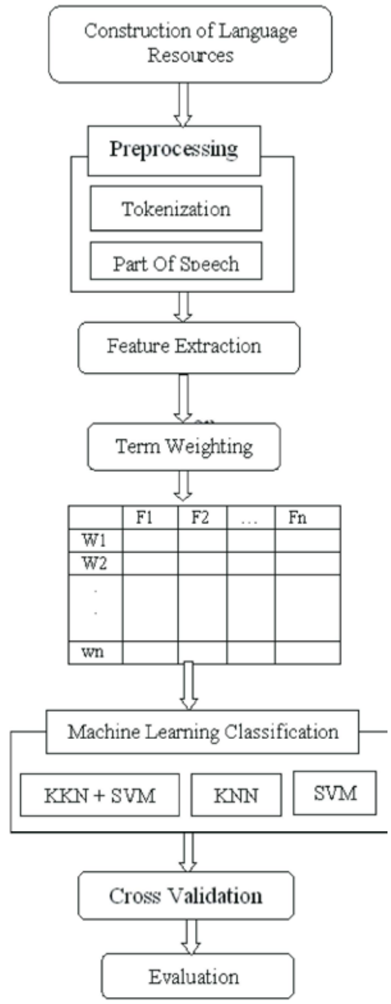

Fig. 1: Architecture of machine learning classification for GR extraction.

METHODSAND MATERIALS

This section presents the method used in Arabic GR extraction models, which consists of several phases. Figure 1 shows the overall architecture of the method, which involves the following phases:

Construction of Language Resources: Given thatan Arabic corpus of new sentences annotated with GRs was not available for training a data-driven system, a manually-constructed corpus was prepared for this study. The corpus consisted of 80 sentences from Othman [7]. Each sentence in the corpus was manually annotated with the GRs, such as subjects, objects and predicates. Table 1 shows a sample of the Arabic sentences from the corpus annotated with the Grs.

Table 1: Sample of Arabic sentences from the corpus annotated with GRs

English sentence	Arabic sentence	POS	Relations extraction
Proved	أثبت	V	
Student	الطالب	N	I-SUB
and	و	CONJ	
teacher	المعلم	N	E-SUB
yesterday	أمس	ADV	
that	ان	PR	
the experience	التجربة	N	E-OBJ
wrong	خطأ	ADJ	

Pre-Processing: New Arabic sentences must undergo a pre-processing phase before the grammatical relations in these sentences can be extracted and classified using machine learning methods. In addition, the sentences should be divided into clauses or phrases to facilitate the extraction and classification of the grammatical relations.In this system, new Arabic sentences are passed through pre-processing steps detailed below.

Tokenization is very important in natural language processing, which can be seen as a preparation stage for all other natural language processing tasks. Tokenization is the process of breaking up words in a continuous text to form units, which can be characters, words, numbers, sentences, or any other suitable form [8].

The disambiguation of a part of speech (POS) can be defined as an operation in whicha computational reorganization of the active POS is established based on its usage in a certain context [9]. In this step, each word is tagged to its unique POS. For example:

Features Extraction: The aim of this phase is to convert each word into a feature vector. Features have been introduced in this work for the classification and foundation of grammatical relations. Three different kinds of features from the sliding windows have been optimized from the previous works carried out by [10-12].

Term Weighting:A pre-processing method used for the enhancement of the presentation of a word as a feature vector. Term weighting aids in the finding of vital terms in a collection of documents to perform ranking [13]. Several term weighting systems are available, with the popular ones being Term Frequency (TF), Inverse Document Frequency (IDF) and Term Frequency-Inverse Document Frequency (TF-IDF).

Machine Learning Classification: The grammatical relations extraction and classification approach in this work is primarily a machine learning approach, in which one of the machine learning classification methods is employed to classify each word based on one of the grammatical relations.

The K-Nearest Neighbor classifier is a renowned occurrence-based classifier, which is known to be a powerful tool for solving various text classification issues [14]. However, the k-NN is known as lazy learning because it postpones the decision to generalize outside the training data until every new query occurrence has been experienced [15].

Traditional texts are very accurately categorized by support vector machines (SVMs), which usually perform better than the K-Nearest Neighbor classifier. Unlike the K-Nearest Neighbor and Maximum Entropy classifiers, SVM function is based on the large-margin concept instead of on the theory of probability [16].

Classifier models can be implemented by combining different classification algorithms and by using different combination techniques. Various subsets of features can be used to construct combining classifiers. Feature extraction is conducted to attain more efficient computation, with greater accuracy. As such, different feature selection methods will be assessed in the experiments for this research, which will use a combination of k-NN and SVM algorithms, in which the SVM algorithm for classification exploits the k-NN algorithm as regards the distribution of test samples in a feature space [17].

Cross Validation: A validation technique model used to evaluate how the results of a statistical analysis are generalized into an independent dataset. This model is used primarily in settings meant for prediction. Furthermore, the model is used to compute the accuracy of a predictive model in practice [18]. In a prediction problem, the model is usually fed with a dataset comprising known data on which training is conducted (training dataset) and a dataset comprising unknown data, against which the model is tested (testing dataset).

Evaluation: The function of the GR extraction and classification operation may be represented by the reclamation R, precision P and the micro-average. However, a standard system will show a minimized time response and the permitted space. Table 3 presents a comparison between the word results of a human and a computer.

The number of words that have been assigned via human prudence and the designator and which possess the appropriate GR, is considered TP (true positive). The number of related words that have been assigned via human prudence but inconsequentially with as regards

Table 2: Examples of POS structures

English word	Arabic Word	POS	Description
Ignores	يتجاهل	V	Verb
Human	الانسان	NN	Singular Noun
Advice	النصيحة	NN	Singular Noun
When	عندما	CC	Particle Conjunction
Be	يكون	V	Verb
In	في	IN	Preposition
Yesterday	امس	ADV	Adverb
Need	الحاجة	NN	Singular Noun
It	اليها	P	Pronoun

Table 3: Assignment processing

Classifier Assigned g	Yes (g)	No (g)
	TP	FP
	FN	TN

the classifier is denoted by FN (false negative). Furthermore, FP (false positive) denotes the designated words that are unrelated as regards human prudence but have been correctly classified as regards the categorizer. Finally, TN (true negative) is considered the total number of words that have been wrongly classified by human prudence as well as by the classifier.

However, to calculate the accuracy metric (precision measure), which is best able to recover the words (where these words are assigned by the end-user as being appropriate), the following mathematical formula can be used:

$$P_r = \frac{TP}{(TP+FP)} \tag{1}$$

Meanwhile, the metric that shows the ability to recover the related words can be expressed as:

$$R_e = \frac{TP}{(TP+FN)} \tag{2}$$

The most common measure for evaluating GR extraction and classification systems is the F-measure, which is a combination of the precision and recall functions:

$$F1 = \frac{2\Pr \times \mathrm{Re}}{\Pr + \mathrm{Re}} \tag{3}$$

RESULTS AND DISCUSSION

Data Description: This experiment employs a manually assembled corpus for Arabic GR extraction, because an Arabic corpus of new sentences annotated with GRs is currently unavailable to traina data-propelled set-up. The 80 Arabic sentences in the corpus, which are derived from [7] are annotated by hand with GRs that include subjects, objects and predicates. An illustration of sentences in Arabic annotated with GRs is displayed in Table 1.

Experimental Results: This study focused on 80 sentences in Arabic from [7]. K-Nearest Neighbor (KNN) and Support Vector Machines (SVM) were the two algorithms employed for this undertaking. Fourteen features comprising the part of speech for specific words were analyzed on a dataset. These include five word features, three POS, three prefixes and threesuffixes. The features employed for this study are elaborated in Table 4.

Table 4: The feature extraction layout utilized for this study

Name Feature	Feature Symbol	Feature Extraction	Details
	F1	s_1	Initial char of the word
Prefixes	F2	s_1s_2	First two chars of the word
and	F3	$s_1s_2s_3$	First three chars of the word
Suffixes	F4	s_n	Last char of the word
	F5	$s_{n-1}\,s_n$	Last two chars of the word
	F6	$s_{n-2}\,s_{n-1}\,s_n$	Last three chars of the word
	F7	w_0	Existing word
Word	F8	w_{+1}	Word following the existing word
Features	F9	w_{+2}	Two words following the existing word
	F10	w_{-1}	Word prior to the existing word
	F11	w_{-2}	Two words prior to the existing word
	F12	p_0	Part of speech of the existing word
Part	F13	p_{-1}	Part of speech of the word prior to the existing word Of
Speech	F14	p_{+1}	Part of speech of the word following the existing word

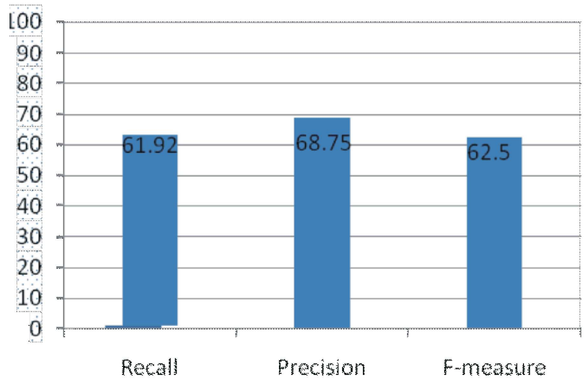

Fig. 2: The accuracy percentage (%) achieved by KNN

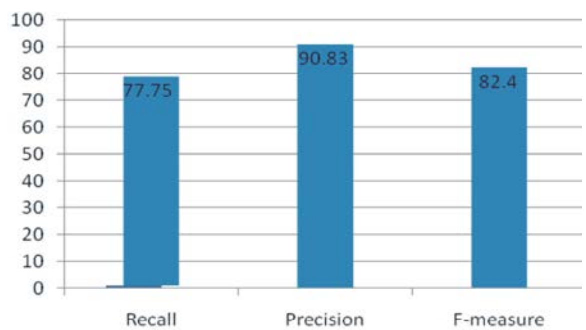

Fig. 3: The accuracy percentage (%) achieved by SVM

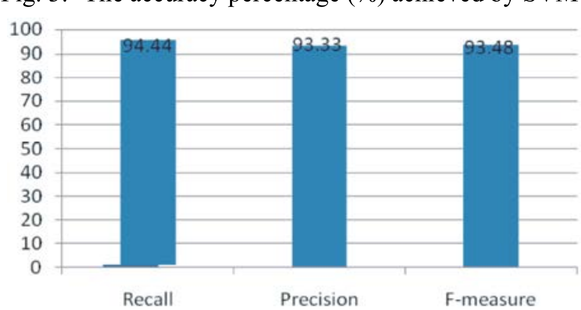

Fig. 5: The accuracy percentage (%) achieved by KNN and SVM

Individual Classifier Method: In this approach, a significant number of runs are executed for every recognized classifier with the inclusion of the feature array for every run.

As portrayed in Table 4,the application of the KNN classifier together with the array of features resulted in an optimum outcome of 61.92% for recall, 68.75% for precision and 62.5% for F-measure. The horizontal alignment in Figure 2 defines the assessment of findings derived from tests, whereas the vertical alignment portrays the degree of accuracy achieved.

The outcomes generated via SVMs with the feature array were recorded as 77.75% for recall, 90.83% for precision and 82.40% for F-measure. These outcomes are detailed in Table 4. Figure 3 displays the outcomes achieved via SVMs together with the array of feature extractions. The horizontal alignment represents an assessment of outcomes obtained from the test and the vertical alignment registers the degree of accuracy.

Combinedclassifiersapproach: Via classifier merging, a variety of combinations can be utilized together with the feature array using voting algorithm. The merging of SVMs and KNN resulted in recall achieving an accuracy level of 94.44%, precision 93.33% and F-measure 93.48% (Table 4).

The levels achieved via a KNN and SVM merger together with the array of feature extractions is exhibited in Figure 5. The horizontal alignment displays the assessment of outcomes obtained from the test and the vertical alignment records the degree of accuracy achieved.

Comparison of Results with Previous Studies: The outcome of the experimental results were compared with [7]. Acknowledged as the most relevant study in this sphere, [7] arrived at an optimum outcome by employing a rule-based procedure. With a selection of 80 sentences in Arabic (3 to 20 words per sentence), an F-measure of 89.60% was realized. However, an F-measure of 93.48% was accomplished with the combination of KNN and SVMs for classification in this research, which agrees well with [7].

CONCLUSIONS

In this study, the information gathered from test results was scrutinized and assessed for the proposal of a working model. An appraisal of the outcomes acquired by the application of a classifier with an array of features revealed that the merging of KNN and SVMs resulted in a 93.48% F-measure. Thus, in comparison to previous classification research on the subjects, objects and verbs in Arabic script, the degree of accuracy achieved by this study isproven superior.

ACKNOWLEDGMENTS

Our sincere thanks to Universiti Kebangsaan Malaysia (UKM) which have supported this study. Also, we would like to acknowledge and thank the General Directorate of Education in Babylon, Ministry of Education, Republic of Iraq and Babylon University for contributed in this research.

REFERENCES

1. Briscoe, E. And J. Carroll, 2002. `Robust accurate statistical annotation of general text'. In Proceedings of the 3rd International Conference on Language Resources and Evaluation, Las Palmas, Gran Canaria, pp: 1499-1504
2. Attia, M., 2009. Parsing Arabic using Treebank-Based LFG Resources. Proceedings of the LFG09 Conference.Miriam Butt and Tracy Holloway King (Editors).
3. Ab Aziz, M., 2006. pola grammar technique for grammatical relation extraction in malay language. Malaysian Journal of Computer Science, 19(1):
4. Neumann, G., C. Braun and J. Piskorski, 2000. A Divide and Conquer Strategy for Shallow Parsing of German Free Texts. In Proceedings of ANLP 2000:
5. Attia, M., 2008. Handling Arabic Morphological and Syntactic Ambiguities within the LFG Framework with a View to Machine Translation. PhD Dissertation, UK: University of Manchester.
6. Ahmed, A.F., 1999. Developing an Arabic Parser in a Multilingual Machine Translation System. Tesis M. Sc. thesis, 1999, Cairo University ISSR.
7. Othman, H., 2012. Rule-Based Approach for Arabic Grammatical Relation Extraction.
8. Huang, C., P. Simon, S. Hsieh and L. Prevot, 2007. Rethinking Chinese Word Segmentation: Tokenization, Character Classification, or Word break Identification.
9. Kroeger Paul, 2005. Analyzing Grammar: An Introduction. Cambridge: Cambridge University Press. pp: 35. ISBN 978-0-521-01653-7.
10. Jesus, G. and M. Lluis, 2004. SVM Tool: A general POS tagger generator based on Support Vector Machines. Proceeding of the 4the International Conference on Language Resources and Evaluation.
11. Albared, M., N. Omar and M. Ab Aziz, 2009. Classifiers combination to Arabic Morphosyntatic disambiguation, Proceeding of International Conference on Electrical Engineering and Informatics, Selangor, Malaysia.
12. Dehdari, J., L. Tounsi and J. Van Genabith, 2011. Morphological features for parsing morphologically-rich languages: a case of Arabic. In Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages, pp: 12-21.
13. Raghavan, P., C.D. Manning and H. Schutze, 2008. Scoring, term weighting and the vector space model. Introduction to Information Retrieval, pp: 100.
14. Dua, R., P. Hart and D. Stork, 2001. Pttern Classification. New York: Wiley.
15. Altman, N.S., 1992. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46(3): 175-185.
16. William, H., A. Teukolsky Saul, T. Vetterling William and B.P. Flannery, 2007. Section 16.5. Support Vector Machines. Numerical Recipes: The Art of Scientific Computing (3rd ed.). New York: Cambridge University Press.
17. Rong Li, Hua-Ning Wang, Han He, Yan-Mei Cui and Zhan-Le Du, 2006. Support Vector Machine combined with K-Nearest Neighbors for Solar Flare Forecasting.
18. Mosteller Frederick, 1948. A k-sample slippage test for an extreme population. Annals of Mathematical Statistics, 19(1): 58-65.