Global Journal on Technology

Vol 7 (2015) 00-00

Selected Paper of 3nd Global Conference on Computer Science, Software, Networks and Engineering, (COMENG-2015)
19-21 November 2015, İstanbul Aydın Üniversitesi – İstanbul, Turkey

SEMANTIC ROLE LABELING WITH RELATIVE CLAUSES

Metin Bilgin*, Department of Computer Engineer-Phd Student, Yıldız Technical University,34220,Turkey.
Mehmet Fatih Amasyalı, Department of Computer Engineer, Yıldız Technical University,34220,Turkey.

Suggested Citation:
Bilgin, M. & Amasyalı, M.F. (2015). Semantic Role Labeling With Relative Clauses,*Global Journal on Technology*[Online]. 2015, 07, pp 000-000.Available from: www.awer-center.org/pitcs

Received date:Oct 26, 2015; Revised date:Nov 07,2015; accepted date.
Selection and peer review under responsibility of Assoc. Prof. Dr. Özcan Asilkan.

Abstract

One of the main tool of the computerized linguistic studies is to find semantic role labeling automatically. Clauses that contain many conclusion are harder than simple clauses as semantic role labeling. This study proposes that a clause should semantic role labeling the dependent clauses, and these clauses should be semantic role labeling rather than separate the items of whole clause. This approach seems to separate a hard problem into simple sub-parts and it gives a higher success than semantic role labeling. to separate the clauses to the dependent clauses and finding its items; Condition Random Fields (CRF) algorithm was used.

Keywords: Natural Language Processing, Semantic Role Labeling, Condition Random Fields

1. Introduction

Studies of natural language process, it is needed to have many application to semantic role labeling automatically. To semantic role labeling; it leads to work more correctly in language natural process problems as data extraction, dialogue systems, text classification, text comprehension.

A clause is a syntax that mentions a feeling, a thought, a wish or a conclusion. In this syntax, there can be one or more than conclusion. If we examine the clauses as a structure, we can see it separates to four groups [1].

*ADDRESS FOR CORRESPONDENCE: Metin Bilgin, Department of Computer Engineer-Phd Student, Yıldız Technical University, İstanbul 34220, Turkey. *E-mail address*:metin_bilgin@hotmail.com /Tel.: +0-505 806 03 62

If the clause contains a conclusion, it is a simple clause; but if it contains more than one conclusion, it is a complex sentence. But one of this conclusion is main clause, and the other/s are dependent clauses that mention other conclusion. If there is more than one dependent clause in a narration, it is a sequential clause. The clauses in sequential clauses depend each other by comma or semicolon connects. Dependent clauses connect by conjunctions. There are examples of four groups below;

"You have to study hard." and "Every person has a tree interest." Simple Clause
"The game that we watched yesterday (Dependent clause) / nobody liked it. (Main Clause)" Complex Sentence
"It snows outside, it must be cold." Sequential Clause
"You have been old when you are familiar to your environment." Dependent Clause

In the literature for automatic discovery of the elements have been made of Turkish clauses. Özköse and Amasyalı in a survey conducted by simple (gerunds and infinitives) without the extraction of the duo found the elements of Turkish sentences of life sciences and the item has been performed [2]. The items to find hand-made, a rule based method is used. Again the study by Çoşkun was prepared by hand has been used in a rule-based structure [3].

Conditional random fields (CRF) Aygül et al. 2000 uses to find the elements of simple clauses in Turkish and automatically , using a data set on Turkish simple clauses, sentence elements are allocated to CRF [4].

Zafer's study conducted by context-independent rules of grammar, morphological analysis, and has developed a parser based on the validation rule. The developed system of independent and grammar rules contain validation rules that is working for all the Turkic languages. The study is implemented for Turkish and Turkmen [5].

Studies usually used simple clauses and rule sets were created manually. Yet, the texts that are encountered in everyday life mainly in the form of a compound clause.

Apart from this study, for the English, despite many instances [6] There is no study that uses the elements of the Turkish sentences CRF to find another. However CRF name recognition of the Turkish entity (Name Entity Recognition) there are a few studies for that use. One of them is Şeker and Eryigit [7] by a study in news texts. Another study, Özkaya and Diri [8], made by e-mails and texts is working on it. In both surveys, 3-4 different entity types (person name, location name, Institution Name, etc.) and have been trying to find 90% of near success has been achieved. Singla et al. the commitment made by Hin parse (Dependency Parsing of words in the sentence and connecting words they connect in the work of the carried out a study to determine the labels) [9].

In the present study, non-simple (depending on the compound and ordered) have been focused on the elements of the clause separation. Unlike the existing literature, non-simple clauses instead of one of the first side of the element as a whole and then broken up into clauses, the separation of the elements have been proposed. This can be seen as a complex problem into its sub parts separate it is an approach and that is simple. CRF was used instead of manual allocation of items to the production rules. This second declaration. Conditional random fields are given information about section. Third is

information about the data sets used in the chapter are given. The fourth chapter was informed about the experiments carried out. in part fifth, the results obtained were interpreted.

2. Conditional Random Fields

The possibility of statistical inference algorithms based on the algorithms that are used in labeling an array. The most common statistical array that are used in labeling the graph(Graph-based models, hidden Markov Model (hidden Markov model-HMM), maximum entropy Markov model maximum entropy Markov Model MEMM), and CRF systems. Studies CRF' s graph-based models have shown that it is best for you. We use the graph in Figure 1 shows of CRF.

CRF machine learning and pattern recognition, structured data that are used in statistical classification method. CRF, problems in natural language processing, the input sequence is used to predict label sequences [10]. CRF, statistical machine learning classification based on the alignment proposed by lafferty and his colleagues method. The sequence within a sequence classifier to assign a label to each unit they try. Calculate a probability distribution over the possible labels and elect the most probable label sequence [10].

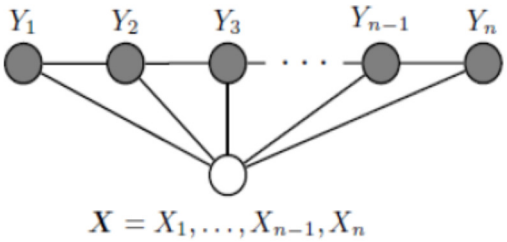

Figure 1. CRF Model Graph

$$P(y, x)= \frac{1}{Z_\theta(x)}\exp\left(\sum_j \lambda_j t_j (y_{t-1}, y_i, x, i)+ \sum_k \mu_k s_k (y_i, x, i)\right) \quad (1)$$

Equation (1) As can be seen in the parameters attribute of the function t. label (y_t) and T-1. label (y_(t-1)), as a function of X and a word sequence. To use the attribute functions are functions that determine the qualities desired in machine learning. In this equation j, k if the number of transitions between sequences in the sequence refers to the number of words. $\sum_j \lambda_j t_j (y_{t-1}, y_i, x, i)$, the transitions between sequences, $\sum_k \mu_k s_k (y_i, x, i)$, refers to the corresponding inputs and outputs.

3. Used Data Set

To separate the elements of a sentence phrases, clauses, clauses before the separation of the whole items to the side then the side with the separation of the elements of the novels in order to compare the approaches of various news sites and simple, non-obtained a sentence of 1278 pieces. Items separated for the measurement of the success of the first approach sentences are required. For the second approach, both elements must become dedicated to both dependent clauses and this clause of the sentence. Firstly, Fatih Parser was using natural language processing parser library [11]

was carried out by using the program the analysis of words in a sentence. Fatih Parser, a syntactic parser designed for Turkish and other Turkic languages.

Conducted sentences word analysis writing your own program with the help of our system are labeled according to the labeling of our choosing. Sentence analysis of sentences the sentences provided to the system is the key to being inserted into the relevant place and thus the system to take note of the phrases while labeling side is provided. Ready to be handed off to the system were analyzed and the attached example sentences in Table 1 can be seen. The average of sentences side-the number Claude 2.25%.

Table 1. Keyword analysis is made example sentences

"yazar/isim" "da/2conj" "ısrar/isim" "et/verb mek/+fiil_mastar_mek ten/isim_çıkma" "vazgeç/fiil er/fiil_genişzaman_ır"
"canan/Özel_isim" "kadın/Özel_isim" "ağla/fiil mak/+fiil_mastar_mek tan/isim_çıkma" "perişan/sıfat" "hale/isim" "gelir/isim"
"heyecan/isim ımız/isim_sahiplik_biz_imiz" "git/fiil erek/+fiil_süreklilik_erek" "art/fiil ıyor/fiil_şimdikizaman_iyor"
"böylece/isim" "çok/adv" "çalış/fiil an/+fiil_dönüşüm_en ın/isim_tamlama-ın" "yüksek/sıfat" "emekli/isim" "maaş/isim ı/isim_belirtme" "alacak/isim"
"doğru/sıfat yu/isim_belirtme" "söyle/fiil yin/fiil_emşr_siz_in" "yan/isim ınız/isim_sahiplik_siz_iniz da/isim_kalma yım/fiil_kişi_ben_im"

4. Experimental Results

The dataset for the experiments introduced in the previous section was used. In the results, studies will be automatically allocated to the sentence of the first sentence, and then to be allocated to elements of the studies are presented. The work that is performed for training and testing phase of the CRF-based C# Program which is written in CRFSHARP program has been used[12].

4.1. Detection of the dependent clauses with the CRF

1278 sentence has been given to the system and automatically labeled. Through the program developed automatically tagged sentences, to convert it into a format that can be given to the CRF system is provided. Descriptions of terms used in tagging are shown in Table 2. CRF and given to the system that can be automatically tagged examples of sentences that can be made can be seen in Table 3.

Table 2. Labeling Definitions

Label	Description
Start	Dependent/Basic clause mentions start
Continue	Dependent/Basic clause mentions continue
Finished	Dependent/Basic clause mentions finish
Empty	Mention the blanks in the sentence

Punctuation	Mention punctuation marks

Table 3. Example of Automatically Labelled

Example Clauses		
Introduction 1	**Introduction 2**	**Exit**
kaymakam	isim	Start
ın	isim_tamlama-ın	Continue
bos	bos	Empty
karı	isim	Continue
sı	isim_sahiplik_o_ı	Continue
bos	bos	Empty
ol	verb	Continue
an	+fiil_dönüşüm_en	Continue
bos	bos	Empty
canan	Özel_isim	Continue
ın	isim_tamlama-ın	Continue
bos	bos	Empty
yusuf	Özel_isim	Continue
u	isim_belirtme	Continue
bos	bos	Empty
aşağıla	fiil	Continue
ma	fiil_dönüşüm_me	Continue
sı	isim_sahiplik_o_ı	Continue
bos	bos	Empty
bile	fiil	Finished
o	pron	Start
nu	acc	Continue
bos	Bos	Empty
etki	Fiil	Continue
le	fiil_olumsuz_me	Continue
mez	fiil_geniszaman_ır	Continue
.	Nokta	Finished

Our 1278 sentence, 250 for test, training them is reserved for 1028. Separation of clauses and sentences in the training set automatically base side of the experiment presented in Table 4 success rates on the test set and the number sentence.

Table 4. Training Set Success Rate

Training Set Clause	Exit Function	Test Success Rate
100	21025	98.49
250	37355	98.46
500	59015	99.3
1028	104080	99.59

As seen in Table 4 The process for the automatic determination of the sentences side could be performed with very great accuracy. In addition, increasing the number of sentences in the training set can be seen to make a positive impact on the success.

4.2. Effects of Dependent Clauses to separate items

In Section 4.1 of the sentences, sentence elements to a successful on the big side of the process of separation of the allocation process as a whole, not divided into clauses in the sentence can be applied to the side it has been seen that the idea of using.

Sentences and sentence elements divide elements are deal located as a whole side devoted to two separate systems were prepared for comparison. First, the system as a whole to distinguish them from elements of the sentence, to distinguish them from the second system to the side the side lines you the first sentence then separates the elements of a sentence.

Both systems in education Aygül et al. created by 2000-simple sentences 1000 sentences randomly from the data set were used [4]. The compound generated in the test sentence randomly selected up to 100 in study 4.1 were used. The first system training and test sentences is given to the system as a whole. 2. The same training system, the training set is used. Compound sentences in the test set of 100 each side-Claude set up a test to be a new sentence. Thus, our district 225 test-sentence test set was created. The sentences that are used when tagging the tags shown in Table 5. Also the side was divided into clauses and sentences of a test set are presented in Table 6 given in every case.

Table 5. Used Labels

Label	Definition
o	Subject
bn	Direct Object
bsn	Indirect Object
dt	Indirect Component
zt	Adverb
y	Verb
Punctuation	Punctuation marks (.-, etc.)

Table 6. Example of Labeling

With Dependent Clause			
People	half	naked	walk
o	zt	zt	y
meal	Do not find		
bsn	y		
situation	come		
bsn	y		

Compound Clause								
İnsanlar	yarı	çıplak	dolaşıp	yemek	bulamayacak	hale	gelirler	.
o	zt	zt	y	bsn	y	bsn	y	Punctuation

The sample sentence in Table 7 is seen in a training set.

Table 7. Example of Train Sentence

Introduction 1	**Introduction 2**	**Exit**
bugün	Zaman	zt
ler	İsim_çoğul_eki_ler	zt
de	İsim_kalma	zt
hava	İsim	o
lar	İsim_çoğul_eki_ler	o
çok	zarf	zt
sıcak	Sıfat	y
.	Nokta	Punctuation

The size was taken as 3 in the training of CRF. in other words, "hava" when calculating a probability value for the word preceding the word "de" that specifies the type previous "isim_kalma",the next word "lar" that specifies the type and the next "isim_çoğul_eki_ler" for them are being taken into account and a probabilistic model is being created. Side effect of the clauses separate the elements it is seen in Table 8 The results of experiments.

Table 8. Results

Test Set	**Success Rate (%)**
Compound Sentence (Test Seti-1)	43.91
Dependent Clause (Test Seti-2)	59.58

As seen in Table 8, the clauses of the sentence by dividing it into the elements of the allocation to the success of the system increased.

4.Conclusion and Discussion

Separate the sentence elements is an important issue in linguistic terms. We conducted two studies to be able to do it automatically.

As a result of the first study, the sequence labeling algorithm CRF Sentence often preferred to the elements in the process that can be used in the allocation process has been proven. The CRF trained with manually labeled data to separate the clauses and basic sentence high side of the system has been shown to be a success. Here is a definite correlation between the size of the training set is confirmed by experiments that the increase of success.

In the second study, in differentiating rather than being directly given to the system of sentences, sentence elements, clauses, divided into base and side of the provision, concluded that success increases significantly. Here, in sequence labeling, each verified in its own right that contains the prediction values of consistency and commitment. Trained with simple sentences, compound sentences with clauses of a system to increase success on a test set that can be used from the side confirmed the thesis.

Yet the Turkish system is sized to be able to express the training set. However, as the amount of labeled data increases the reliability and success of the system are projected to increase. Also a deficiency of the system as the side to which the label has been connected with the basic values of clauses. Future studies aimed at overcoming these shortcomings is as planned.

To access the data set used in the study metin_bilgin@hotmail.com to the address by e-mail can be requested.

7. Acknowledgment

Collecting when creating our data, the sentence values for the identification of the elements of the sentences and leave the main and dependent back, Thank you for Serdar Düz, language and literature teachers for their support, Yeliz İnci, Ayşe Kalkan.

References

[1]Türkçe Kaynak Sitesi, Yapılarına Göre Cümleler ,(2015, 10 June), Retrieved from: http://www.turkcesinifi.com/yapilarina-gore-cumleler-t781.html

[2] Özköse, C., & Amasyalı, M.F., "Cümle Öğelerinden Hayat Bilgisi Çıkarımı". Türkiye Bilişim Vakfı, Bilgisayar Bilimleri ve Mühendisliği Dergisi, Aralık 2012, 6.

[3] Coşkun,N., "Türkçe Tümcelerin Öğelerinin Bulunması", Master Thesis, Istanbul Technical University, Science Institute, 2013.

[4] Aygül, M., &, Karaalioğlu, G., & Amasyalı, M.F., "Koşullu Rastgele Alanlarla Basit Türkçe Cümlelerin Öğelerine Ayrılması", Sigma-YTU Journal of Engineering and Natural Sciences,2014, 32(1),pp 23-30.

[5] Zafer,H.R., "Türki Diller İçin Uyarlanabilir Sözdizimsel Ayrıştırıcı", Master Thesis, Fatih University, Science Institute, 2011.

[6] Sutton, C., & McCallum,A., “An Introduction to Conditional Random Fields”, Foundations and Trends in Machine Learning,2011, 4 (4),pp 267-273.

[7] Şeker, G.A.,& Eryiğit, G., “Initial explorations on using CRFs for Turkish Named Entity Recognition”, 24th International Conference on Computational Linguistics, COLING 2012, Mumbai, India, 2012.

[8] Özkaya, S.,& Diri, B., “Named Entity Recognition by Conditional Random Fields from Turkish informal texts”, Signal Processing and Communications Applications (SIU),Antalya,2011.

[9] Singla, K., Tammewar, A., Jain, N. , Jain, S. ," Two-stage Approach for Hindi Dependency Parsing Using MaltParser", Workshop on Machine Translation and Parsing in Indian Languages,Mumbai,2012.

[10] Lafferty, J.,& McCallum,A., &Pereira,F., "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data International Conference on Machine Learning (ICML),San Francisco,2001.