*Semantic Role Labeling With Relative Clauses**

***Metin BİLGİN*[1]**

***Mehmet Fatih AMASYALI*[1]**

Abstract

One of the main goals of computerized linguistic studies is automatically finding the elements of a sentence. Separation of sentences with too many judgements into their elements is a more complicated process when compared with simple sentences. In this study, instead of separating the whole sentence into its elements, the separation of sub clauses into their own elements is suggested. This approach can be considered as dividing a hard problem into sub parts; and this had higher rate of achievement when compared with dividing the whole sentence into its elements. Condition Random Fields (CRF) algorithm was used for dividing the sentences automatically into sub-clauses and finding its elements.

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

1. Introduction

In the studies about natural language processing, division of sentences into their components automatically is necessary for many applications. Dividing a sentence automatically into its components makes it possible for various natural language processing problems such as information inference, dialogue systems, text classification and text understanding.

Sentence is a syntax that indicates a feeling, a thought, a request and a judgement. One or more judgements might exist in a syntax. When we examine the sentences in terms of

**This paper is presented in 3rd GLOBAL CONFERENCE ON COMPUTER SCIENCE, SOFTWARE, NETWORKS AND ENGINEERING*

[1] *Department of Computer Engineer-Phd Student, Yıldız Technical University, metin_bilgin@hotmail.com*

[2] *Department of Computer Engineer, Yıldız Technical University*

structure, we see that they are separated into 4 groups [1].

Sentences with a single judgement are called simple sentences, and sentences with more than one judgements are called compound sentence. But one these judgements is the main sentence, and other(s) are the sub-clauses that define the side sentences. If there are more than one sentences in an expression which are connected to each other in terms of meaning, this sentence is called tiered sentence. Sentences inside a tiered sentence are connected to each other with a comma or semicolon. Bound sentence is connected by conjunctions. Examples about these 4 groups are below.

"You have to study hard." and "Everyone should love trees." Simple Sentence
"No one liked (Main sentence) / the game we watched together yesterday (sub-clause)" Compound Sentence
"It snows outside, weather must be cold." Tiered Sentence
"You are old whenever you are get used to your environment." Bound Sentence

There are studies in the literature conducted in order to find the elements of Turkish sentences automatically. In the study conducted by Özkose and Amasyali, the elements of simple (without verbal) Turkish sentences were found and life science inference was made for element pairs [2]. Manually generated rule based method was used to find the elements. Another study conducted by Coşkun has also used a manually prepared rule based structure [3].

Aygül et. al. have used the CRF to find the elements of Turkish simple sentences and they have used CRF on a Turkish data set consisting of 2000 simple sentences; dividing the sentence into its elements [4].

Study conducted by Zafer has developed an analyser relying on grammatical rules independent from the context, morphological analysis and validity rules. Developed system works for all Turkic languages with independent grammatical rules that include validity rules. The study was made for Turkish and for Turkoman [5].

Simple sentences and manually established rule sets were used in the studies. But, the texts faced in daily life are mostly in the form of compound sentences.

Except for these studies, although there are many examples for English [6], there is no other study using CFR in order to find the Turkish sentence elements. There are some studies that use CRF for Turkish Name Entity Recognition. One of them is a study conducted by Şeker and Eryiğit [7] on the texts of news. Another study was conducted by Özkaya and Diri {8] and it is on the e-mail texts. In both studies, 3-4 different entity name types (name of individual, name of places, name of institution etc.) and the success rate was around 90 %. The Dependency Parsing study conducted by Singla et.al. for the Indian language is about the determination of a word other words are dependent in a sentence and the dependency labels [9].

This study focuses on dividing the non-trivial sentences (compound, bound and tiered) into their elements. Different from the existing literature, instead of dividing the non-trivial sentences into their elements; it was suggested to divide them into sub-clauses first and then into the elements. This is an approach that could also be seen as dividing a complex problem into its simple sub-divisions. Instead of manually creating the rules of dividing into elements, CRF was used. The second part of this paper gives information about Condition Random Fields. Third part gives information about the used data set. Fourth part gives information about the experiments conducted. Fifth part interprets the obtained results.

2. Conditional Random Fields

CRF proposed by Lafferty et al is a method of sequencing a machine learning based on the statistical classification [10]. Array classifiers tried to throw a label to each unit in an array. They calculate a probability distribution on the possible tags and they select the best possible label combination. Accordingly CRF model can be defined as a probability model was developed to calculate the p(o*|s*) probability. Here while specifying the $o^* = o_1,...,o_n$ possible outcomes tags, it specifies the $s^* = s_1,...,s_n$ input data.

It is a frequently used method in the problems such as CRF, NER, POS labelling, SP and so on. Formula for CRF is given in Equation 1 and its form in Figure 1.

$$P(s|o) = \frac{1}{Z_{(\vec{o})}} \prod_{t=1}^{(\vec{o})} \exp\left(\sum_j \alpha_j f_j(s_t, s_{t-1}) + \sum_k \beta_k g_k(s_t, o_t)\right) \quad (1)$$

$Z_{(\vec{o})}$ is a normalization factor for all possible label sequences.

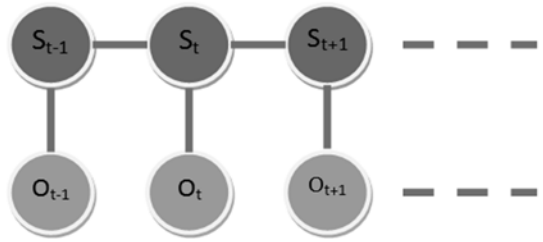

Figure 1 Condition Random Fields [3]
(S-State, O-Observe)

Quality functions are determined for each word in the training corpus. In the training set quality functions, also label information in the designated word is available. According to this benefiting quality functions and action sequences of the specified word weight value of each attribute is calculated. Some attributes may be high weight to throw that word that label type, some qualifications may lower the weight to assign a label. Thanks to educate the system a CRF model which you can find weight values for each feature is created. CRF model created through training is can be used to label previously unlabelled words. After determining the nature of each word, thanks to the CRF model became apparent that the weight of each character, the calculated probability of each word is assigned to each label [11].

As a result, if we consider the most likely label combination as Y*. Each word sequence (o) can be found as given in equation 2 by selecting the most likely.

$$Y^*=\arg\max(P(s|o)) \quad (2)$$

3. Used Data Set

1278 non-trivial sentences were gathered from various news sites and novels in order to compare the two approaches of dividing a sentence into its elements as a whole and then dividing the sentence into sub-clauses and then into its elements. In order to measure the success of first approach, we need sentences divided into their elements. And for the second approach, we need sub-clauses and the divided sub-clauses.

By using the FatihParser program [11] in the study which operates with Zemberek Natural Language Process library; the analysis of the words were done. FatihParser is a syntax analyser designed for Turkish and other Turkic languages. Table 1 indicates the word analyses in the sample sentences.

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 data cluster suggested in the previous chapter were used in trials. The results present the work conducted for dividing the sentences into sub-clauses automatically, followed by studies dividing them into their elements. For the Training and Test phase of the studies; CRFSHARP program written by CRF based C# language was used [12].

4.1. Determination of Sub-Clauses via CRF

1278 sentences were installed into the system, providing automated labelling. Through the developed program, sentences were labelled automatically and were transformed into a format for CRF system. The definitions of the terms used in labelling are indicated on Table 2. Some examples of the sentences that were automatically labelled and transformed for CRF system 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

Out of 1278 sentences we have; 250 were used for test and 1028 were kept for training. The number of sentences for the experiment of dividing the sentences automatically into main and sub-clauses and the success rates on test cluster are seen on Table 4.

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 it can be seen on Table 4, the automated determination of sub-clauses was actualized with a very high success rate. Also, it can be seen that the increase of number of sentences in training cluster has made a positive effect on success.

4.2. **Impact of Dividing the Sub-Clusters into Its Elements**

As the division of sentences into sub-clauses was achieved on Chapter 4.1.; it was seen that the idea of using the sub-clause divided version of a sentence instead of dividing the whole sentence into its elements was an applicable idea. 2 different systems were prepared in order to compare the division of sentences as a whole and dividing the sub-clauses into its elements. First system divides the sentence into its elements as a whole; and he second system divides the sentence into its sub-clauses first and then divides the sub-clauses into their elements. Randomly selected 1000 sentences from the 2000 sentence data set created by Aygül et.al were used for the training of both systems [4]. And for the test, 100 compound sentences created in our study were randomly selected and used. At the training and test of first system, sentences are inserted into the system as a whole. Same education set is used for the training of second system. And for the test set; 100 compound sentences becoming a new sentence for each sub-clause was developed. So, a test set consisting of 225 sentences was created in our test set. Labels used for labelling the sentences are indicated on Table 5. Also, single test set sentences provided as a whole and divided into sub-clauses are seen on Table 6.

Table 5. Used Labels

Label	Definition
o	Subject
bn	Direct Object
bsn	Indirect Object
dt	Indirect Component
zt	Adverb
y	predicate
Punctuati	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

Table 7 indicates a sample training set sentence.

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 window dimension for CFR training was set as 3. In other words, when the probability rate is calculated for the word "air" (hava); previous word "de", previous type definer "isim_kalma"i next word "lar" and next type definer "isim_cogul_eki_ler" is taken into account and a probability model covering all of these is created.

The flow chart diagram about the study conducted is seen at Figure 2.

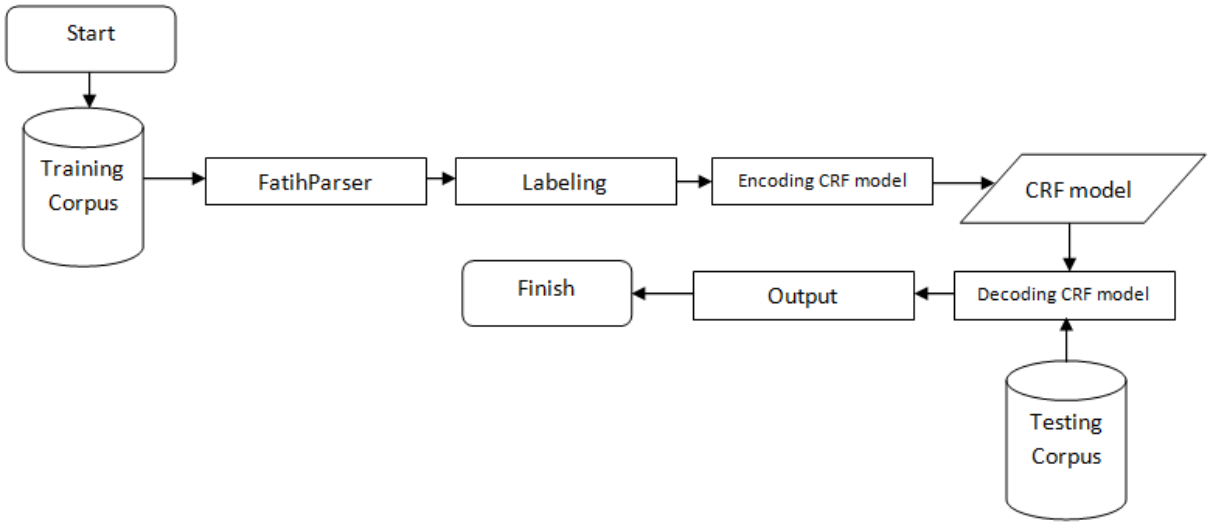

Figure 2 Flow Chart Diagram for application

The results of the experiment of dividing the sub-clauses into elements are seen on Table 8.

Table 8. Results

Test Set	Accuracy (%)
Compound Sentence (Test	43.91
Dependent Clause (Test	59.58

As Table 8 indicates, as the sentences are divided into sub-clauses; system had a higher rate of success in terms of dividing into elements.

4. Conclusion and Discussion

Dividing the sentences into their elements is an important issue for linguistics. We have conducted two studies in order to do this automatically.

As a result of the first study, it has been proved that CFR algorithm which is frequently preferred for sequence labelling transactions is also applicable in Dividing the Sentence into Its Elements. It was seen that the CRF system trained by manually labelled data had a great success rate in dividing the sentence into main and sub-clauses. The experiment has proved that there is a direct proportion

between the size of training set and the increase of success.

Second study has reached the conclusion that instead of giving the sentences as a whole, dividing them into main and sub-clauses significantly increase the success. When the sequence labelling is done, it is proved that each sub-clause has a unique consistency and dependency. The thesis that sub-clauses can be used in order to increase the success of a system trained by simple sentences on a test set with compound sentences is hereby proved.

The training set of the system is not yet in the size of expressing Turkish. Along with this, as the amount of labelled data increases; it is assumed that the reliability and success of the system will also increase. As a defect of the system, it was not found how the sub-clauses are bound to the main clause with a label. Forthcoming studies are planned in order to overcome such deficiencies.

To access the data sets used in this study, an e-mail request can be sent to metin_bilgin@hotmail.com

7. Acknowledgment

The authors would like to thank Serdar Düz, Yeliz İnci and Ayşe Kalkan for his help in the preparation of the manuscript

REFERENCES

[1] Turkish Source Site, Sentences according to structure, (2015, 10 June), Retrieved from: http://www.turkcesinifi.com/yapilarina-gore-cumleler-t781.html

[2] Özköse, C., Amasyalı, M.F., "Common sense Knowledge Acquisition by Sentence Analysis ", Turkey IT Foundation, Computer Science and Engineering Journal, December 2012, 6.

[3] Coşkun,N., “Finding constituents of Turkish sentences", Master Thesis, Istanbul Technical University, Science Institute, 2013.

[4] Aygül, M., Karaalioğlu, G., Amasyalı, M.F. , " Prediction of Function Tags of the Simple Turkish Sentences by Condition Random Fields", 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.