Information Processing and Management xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

On constituent chunking for Turkish

Ozkan Aslan[a,*], Serkan Gunal[b], Bekir Taner Dincer[a]

[a] *Department of Computer Engineering, Mugla Sitki Kocman University, Mugla, Turkiye*
[b] *Department of Computer Engineering, Anadolu University, Eskisehir, Turkiye*

ARTICLE INFO

Keywords:
Chunking
Shallow parsing
Turkish
Constituent
Conditional random fields
Natural language processing

ABSTRACT

Chunking is a task which divides a sentence into non-recursive structures. The primary aim is to specify chunk boundaries and classes. Although chunking generally refers to simple chunks, it is possible to customize the concept. A simple chunk is a small structure, such as a noun phrase, while constituent chunk is a structure that functions as a single unit in a sentence, such as a subject. For an agglutinative language with a rich morphology, constituent chunking is a significant problem in comparison to simple chunking. Most of Turkish studies on this issue use the IOB tagging schema to mark the boundaries.

In this study, we proposed a new simpler tagging schema, namely OE, in constituent chunking for Turkish. “E” represents the rightmost token of a chunk, while “O” stands for all other items. In reference to OE, we also used a schema called OB, where “B” represents the leftmost token of a chunk. We aimed to identify both chunk boundaries and chunk classes using the conditional random fields (CRF) method. The initial motivation was to employ the fact that Turkish phrases are head-final for chunking. In this context, we assumed that marking the end of a chunk (OE) would be more advantageous than marking the beginning of a chunk (OB). In support of the assumption, the test results reveal that OB has the worst performance and OE is significantly a more successful schema in many cases. Especially in long sentences, this contrast is more obvious. Indeed, using OE means simply marking the head of the phrase (chunk). Since the head and the distinctive label “E” are aligned, CRF finds the chunk class more easily by using the information contained in the head. OE also produced more successful results than the schemas available in the literature.

In addition to comparing tagging schemas, we performed four analyses. Along with the examination of window size, which is a parameter of CRF, it is adequate to select and accept this value as 3. A comparison of the evaluation measures for chunking revealed that F-score was a more balanced measure in contrast to token accuracy and sentence accuracy. As a result of the feature analysis, syntactic features improves chunking performance significantly under all conditions. Yet when withdrawing these features, a pronounced difference between OB and OE is forthcoming. In addition, flexibility analysis shows that OE is more successful in different data.

1. Introduction

Chunking, also called shallow parsing or light parsing, is a task that divides a sentence into non-recursive structures. The primary aim is to specify chunk boundaries and classes. Although chunking generally refers to simple chunks, it is possible to customize the concept. A simple chunk is a small structure, such as a noun phrase (NP), a verb phrase (VP), or a prepositional phrase (PP), while a

[*] Corresponding author.
E-mail address: ozkanaslan@mu.edu.tr (O. Aslan).

https://doi.org/10.1016/j.ipm.2018.05.004
Received 22 December 2016; Received in revised form 11 April 2018; Accepted 15 May 2018

constituent chunk is a structure that functions as a single unit in a sentence, such as a subject, or an object. Some natural language processing (NLP) tasks utilize the chunking because it is easier than full parsing. Besides, more complicated processes such as named entity recognition (Wibawa & Purwarianti, 2016), text contextualization (Bellot, Moriceau, Mothe, SanJuan, & Tannier, 2016), text summarization (Tayal, Raghuwanshi, & Malik, 2017), sentiment analysis (Chan & Chong, 2017), and natural language understanding (Lopez-Gazpio et al., 2016) can use the data obtained from this process.

Chunking is a kind of classification especially in machine learning. In this sense, "*deciding where a chunk ends*" is a key research question (Abney, 1991). Various statistical machine learning methods have been used for chunking in the literature. Several examples to those methods are support vector machines (Wu & Chang, 2007), the memory-based approach (Rekha Raj & Reghu Raj, 2015), the hidden Markov model (Ibrahim & Assabie, 2013) and conditional random fields (Khoufi, Aloulou, & Hadrich Belguith, 2015). Additionally, there are rule-based studies (Ariaratnam, Weerasinghe, & Liyanage, 2014; Karad & Joshi, 2015).

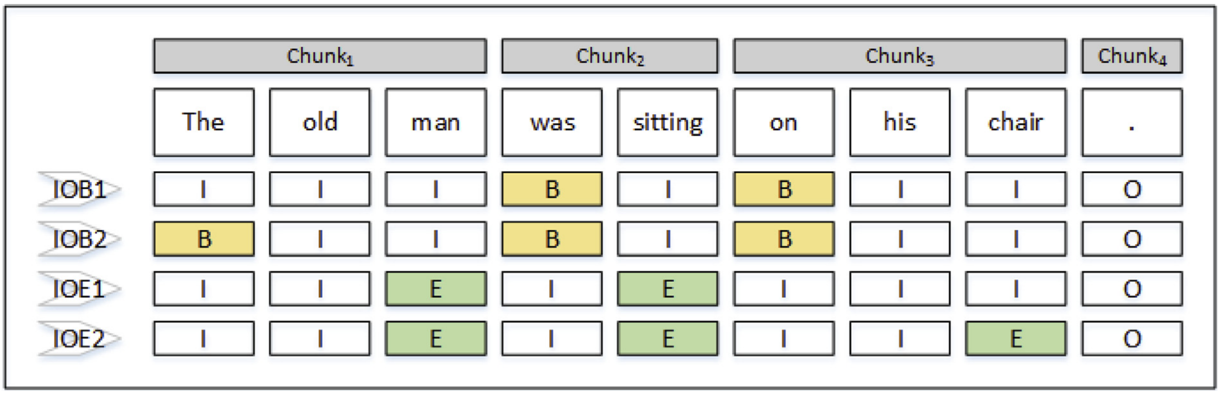
In recent years, there has been a tendency towards neural network based methods for chunking. Collobert et al. (2011) used a multilayer neural network architecture and obtained a performance close to the ones reported by several works based on support vector machine and conditional random fields. Søgaard and Goldberg (2016) applied deep bi-directional recurrent neural networks and secured a higher performance than the one proposed in Collobert et al. (2011). Wang, Qian, Soong, He, and Zhao (2016) used bidirectional long short-term memory recurrent neural network and accomplished a promising result.

Most of the studies on chunking are specific to a certain language, such as Arabic (Khoufi, Aloulou, & Hadrich Belguith, 2015), Basque (Aduriz & İlarraza, 2003), Bengali (Sarkar & Gayen, 2014), Burmese (Aung & Moe, 2015), Czech (Radziszewski & Grác, 2013), Hindi (Gahlot, Krishnarao, & Kushwaha, 2009), Kazakh (Wu & Altenbek, 2016), Russian (Anisimov, Makarova, & Polyakov, 2016), and Tamil (Ariaratnam et al., 2014).

Considering the types of languages, Turkish is an agglutinative one with a rich morphology. Unlike English, this feature of Turkish makes it difficult to decompose a sentence into simple chunks such as NP, VP, and PP (El-Kahlout & Akın, 2013). First, major structures of the sentence should be detected to obtain the simple chunks. They are constituents that are members of the main verb of the sentence. In the literature, several works have focused on obtaining the simple chunks for Turkish such as Adalı and Tantuğ (2015), Kutlu and Cicekli (2016), Yıldız, Solak, Ehsani and Görgün (2015), and Kutlu (2010). Many constraints caused by the features of Turkish are discussed in those works: morphological and syntactical structures interact with one another since Turkish is an agglutinative language, and it is more challenging to establish chunking rules because it is a free-order and pro-drop language. El-Kahlout and Akın (2013) and El-Kahlout, Akın, and Yılmaz (2014) covered Turkish constituent chunking.

Several representations have been identified in the literature for the chunking task. The IOB1 tagging schema is one of the representations proposed by Ramshaw and Marcus (1999). According to this schema, while inside-chunk words acquire label *I*, outside-chunk words adopt label *O*. If two chunks are adjacent, label *B* is used for differentiating the end of the first chunk and the beginning of the second chunk. In contrast the IOB2 tagging schema proposed by Ratnaparkhi (1998), uses label *B* for the first word of the chunk, label *I* for the other words of the chunk and label *O* for the ones that are out of the chunk. In addition to the IOB1 and IOB2 tagging schemas, Sang and Veenstra (1999) compared self-produced IOE1 and IOE2 with other tagging schemas on English texts in terms of performance criteria obtained for NP chunking. These researchers reported that there is no significant difference between the tagging schemas. Fig. 1 exemplify the four schemas mentioned above.

In this work, we proposed a new tagging schema named OE for constituent chunking of Turkish sentences. In reference for OE, we also used a schema called OB. These can be converted into existing schemas in the literature with a slight loss. We also performed five analyses. The first analysis is related to the window size parameter of conditional random fields (CRF) that is a notable statistical machine learning method used for chunking. According to the results of the examination, the optimum value is 3. In the second analysis, we employed three measures to evaluate the performance of chunking: F-score (chunk-based), token accuracy and sentence accuracy (token-based). In conclusion, we can easily argue that F-score is a more balanced measure. In the third analysis, we processed 13 features used in the literature for Turkish constituent chunking. Among them, DLabel, a syntactic feature, improves chunking performance significantly under all conditions. By removing the DLabel and other syntactic feature, namely distance, we obtained the same result in terms of the difference between OB and OE when compared to all features. The fourth analysis is a



	Chunk$_1$			Chunk$_2$		Chunk$_3$			Chunk$_4$
	The	old	man	was	sitting	on	his	chair	.
IOB1	I	I	I	B	I	B	I	I	O
IOB2	B	I	I	B	I	B	I	I	O
IOE1	I	I	E	I	E	I	I	I	O
IOE2	I	I	E	I	E	I	I	E	O

Fig. 1. An example for tagging schemas.