Utilizing Annotated Wikipedia Article Titles
to Improve a Rule-Based
Named Entity Recognizer for Turkish

Dilek Küçük

Electrical Power Technologies Group
TÜBİTAK Energy Institute
06800 Ankara, Turkey
dilek.kucuk@tubitak.gov.tr

Abstract. Named entity recognition is one of the information extraction tasks which aims to identify named entities such as person/location/organization names along with some numeric and temporal expressions in free natural language texts. In this study, we target at named entity recognition from Turkish texts on which information extraction research is considerably rare compared to other well-studied languages. The effects of utilizing annotated Wikipedia article titles to enrich the lexical resources of a rule-based named entity recognizer for Turkish are discussed after evaluating the enriched named entity recognizer against its initial version. The evaluation results demonstrate that the presented extension improves the recognition performance on different text genres, particularly on historical and financial news text sets for which the initial recognizer has not been engineered for. The current study is significant as it is the first study to address the utilization of Wikipedia articles as an information source to improve named entity recognition on Turkish texts.

1 Introduction

Information extraction (IE) from textual data is an important natural language processing task which gains increasing research attention mainly due to the need to automatically process and analyze textual documents available on the Web in several different languages. Named entity recognition (NER) is a well studied IE subtask targeting at the extraction of the names of people, organizations, and locations, along with some numeric and temporal expressions from free natural language texts [1].

NER studies generally range from manually engineered rule-based systems to learning-based and statistical systems [2,3]. While the former systems, which make use of gazetteers and other lexical resources, do not require annotated corpora, they achieve low success rates when tested on diverse text genres, particularly on genres different from the initial target domain of these systems. The learning/statistical systems are freed from this problem since they can be trained
on new domains provided that the required annotated training corpora are available [2]. Nevertheless, building annotated corpora manually is a highly costly task and the latter systems need an additional training phase on such corpora in order to be applicable to new domains. Decision trees [4], Bayesian learning [5], hidden Markov models (HMMs) [6], relational learning (such as inductive logic programming) [5], support vector machines (SVM) [7,8], and conditional random fields (CRF) [9] are among the most prominent learning/statistical techniques used for IE tasks including NER.

On well-studied languages such as English, several different NER studies with high performance rates have been conducted and reported in the literature while NER research on other languages including Turkish is quite rare. Considering the NER research conducted on Turkish texts, to the best of our knowledge, the first study on the topic is the language-independent named entity recognizer reported in [10] where the recognizer is evaluated on Turkish texts along with other texts in Romanian, English, Greek, and Hindi. The statistical name tagger proposed in [11] is an HMM based statistical system [11] to extract person, location, and organization names in Turkish texts. In [12], the first rule-based NER system for Turkish is described. This system utilizes a set of lexical resources and patterns bases to extract person, location, and organization names as well as money/percentage and date/time expressions from Turkish news texts [12]. This system is turned into a hybrid recognizer with a capability to extend its lexical resources by extracting high-confidence named entities from annotated corpora through rote learning and the ultimate hybrid system [13] is shown to outperform its rule-based predecessor. The aforementioned rule-based recognizer has also been successfully integrated into automatic and semi-automatic semantic video annotation systems for Turkish news videos [14,15] where named entities are automatically extracted from video texts to be used as semantic annotations for the videos. A system utilizing CRF and a set of morphological features is presented in [16] where the author argues that CRF provides advantages over HMMs during NER in Turkish. Finally, a rule learning system for the NER task in Turkish texts is presented in [17].

In this study, we annotate Wikipedia article titles with named entity tags and utilize this annotated data to enrich the resources employed by the rule-based NER system presented in [12]. After this extension to the recognizer, we have evaluated the performance of this extended recognizer against its predecessor on diverse text genres to observe the effects of this annotated data to named entity recognition performance and discussed the evaluation results. As Wikipedia articles are known to constitute a plausible information source for different knowledge-based applications, the effects of their utilization to improve a rule-based NER system is a significant contribution to the literature especially for Turkish texts for which NER studies are still limited in number and scope. The rest of the paper is organized as follows: In Section 2 the process of annotating Wikipedia article titles in Turkish is presented, Section 3 is devoted to the evaluation results of the ultimate NER system, corresponding to the original rule-based system enriched with the Wikipedia annotations, against its