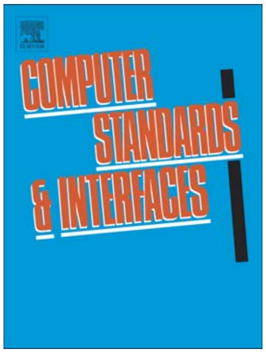

Automatic Compilation of Language Resources for Named Entity Recognition in Turkish by Utilizing Wikipedia Article Titles

Dilek Küçük

PII: S0920-5489(15)00019-7
DOI: doi: 10.1016/j.csi.2015.02.003
Reference: CSI 3007

To appear in: *Computer Standards & Interfaces*

Received date: 5 October 2013
Revised date: 23 October 2014
Accepted date: 8 February 2015

Automatic Compilation of Language Resources for Named Entity Recognition in Turkish by Utilizing Wikipedia Article Titles

Dilek Küçük[a,*]

[a] *Electrical Power Technologies Group*
TÜBİTAK Energy Institute
Ankara, Turkey

Abstract

Named entity recognition is one of the well-defined information extraction tasks and is crucial for several automatic text processing applications. In this paper, we present an automatic approach to compile language resources for named entity recognition in Turkish by utilizing Wikipedia article titles. In the first phase of the proposed approach, a subset corresponding to about one twentieth of all Wikipedia article titles in Turkish is annotated with the named entity tags corresponding to the basic types of person, location, and organization names. This annotated subset is then utilized as a training data set to automatically categorize the remaining Wikipedia titles into one of the basic named entity types by employing the k-nearest neighbor algorithm. This classification approach has led to the construction of a significant lexical resource set for Turkish named entity recognition with a favorable overall precision rate. In order to observe the contribution of the ultimate resource set, several experiments on different text genres are conducted after extending an existing named entity recognizer for Turkish with the resource set. The results of these experiments are quite promising and serve to confirm that the resource set can contribute to the named entity recognition task on distinct text genres. The automatically generated resource set and the extended named entity recognizer (with this resource set) stand as significant contributions to information extraction research on Turkish texts, as related studies and language resources are currently far from being sufficient, especially compared to those of the well-studied languages like English.

Keywords: information extraction, named entity recognition, language resources, k-nearest neighbor algorithm, Wikipedia, text mining

[*] Corresponding author. Tel: +90 312 2101830 Fax: +90 312 2101033
Email address: dilek.kucuk@tubitak.gov.tr (Dilek Küçük)