Computers, Environment and Urban Systems xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

Address standardization using the natural language process for improving geocoding results

Dilek Küçük Matci*, Uğur Avdan

Research Institute of Earth and Space Sciences, Anadolu University, Iki Eylul Capmus, 26470 Eskisehir, Turkey

ARTICLE INFO

Keywords:
The natural language processing
Geocoding
Address matching
Levenshtein distance
Match rating compute

ABSTRACT

Geocoding is a tool that can be used in many areas such as the development of disaster prevention systems, crime mapping and the monitoring of communicable diseases, and which has gradually gained importance. However, the use of geocoding is not yet possible in some areas where it could serve as an effective tool, for various reasons such as inconsistencies in address formats, including inaccurate numbering systems, misspellings, the use of abbreviations and a lack of data that refers to the geocoding process. This study seeks to address these problems by way of a standardization process. To that end, it employs a method that decomposes addresses used as input data in geocoding, identifies spelling mistakes and abbreviations, and reorganizes the addresses through the Natural Language Process (NLP). As test data, the addresses of primary schools in the district of Eskisehir are taken. First the geocoding process is performed on the data set, using both Google geocoding API and ArcGIS geocoding API. Then, the addresses are reformatted into three address formats by applying standardization processes. Geocoding is performed on the re-formatted addresses and the results compared to the non-standardized results. The standardization used is shown to make a significant improvement in the accuracy of the geocoding results. The method used in this study is significant not only in increasing the accuracy of the geocoding process, but also in sustaining its wider use.

1. Introduction

An address is information that allows a specific location to be reached by predetermined directives. It is essential that addresses are converted into coordinates when analyzing incidents such as epidemics, crimes, or accidents (Davis & Fonseca, 2007). The conversion of text-based postal addresses to geographic coordinates is known as geocoding.

Address geocoding is a tool which can be used to digitally locate places on a map based only on addresses. It allows us to locate addresses downloaded from various computer systems on various types of digital maps in various situations, such as the investigation of epidemics (Tassinari et al., 2008; Wey, Griesse, Kightlinger, & Wimberly, 2009), crime mapping or analysis systems (Ratcliffe, 2004; Scribner, Cohen, Kaplan, & Allen, 1999), disaster and risk management systems (Johnson, Stanforth, Lulla, & Luber, 2012), and political science (Haspel & Knotts, 2005). It, however, depends on the accuracy and format of the addresses.

As the accuracy of the geocoding process depends on the format of the addresses, it is of vital importance to have not only correct and precise input addresses, but also to have them written in a certain standard way. The metrics commonly employed in evaluating the quality of geocoding results are completeness, positional accuracy, and repeatability (Zandbergen, 2008).

The accuracy of the geocoding process can be influenced by various factors, such as the geographical areas of the addresses, the quality of the reference database, match scores and geocoding algorithms. Previous research has shown that geocoding results are more accurate in urban areas than rural areas ((Bonner et al., 2003), (Cayo & Talbot, 2003),(Ward et al., 2005)), and geocoding results are more accurate for community addresses than commercial addresses (Zandbergen, 2008).

In the case of Turkey, the biggest challenge in achieving accurate geocoding results is inaccuracies in the input address data. Errors in enumeration work, the use of too many address components, and the lack of a specific address standard can all lead to confusion (Cbsgm, C. B. S. G. M., 2009). Research examining address data in Turkey and performing geocoding, has indicated that mismatched addresses result from errors in door numbers, the incompatibility of street names, incomplete addresses, misspellings, incomplete offset data, typographical errors, and improper formats (35, 28, 15, 9, 5, 4 and 4%, respectively) (Yildirim, Yomralioglu, Nisanci, & Inan, 2014).

The scale of the problem becomes evident when one compares the

* Corresponding author.
E-mail addresses: dkmatci@anadolu.edu.tr (D. Küçük Matci), uavdan@anadolu.edu.tr (U. Avdan).

https://doi.org/10.1016/j.compenvurbsys.2018.01.009
Received 17 January 2017; Received in revised form 8 January 2018; Accepted 15 January 2018

Table 1
Geocoding results obtained from the data.

	Geocoding results obtained from addresses of schools in New Jersey(%)	Geocoding results obtained from addresses of schools in Eskishehir (%)
≤100	82.1	46.7
100–200	9.4	19.3
200–400	3.3	14.7
400–600	1.2	6.0
600–800	0.0	2.7
≥800	4.1	10.7

geocoding outcomes based on school addresses in New Jersey, US, with those of school addresses in Eskisehir (Eskişehir), Turkey, which is the area used as the case study in this research. Geocoding has been performed by using 264 schools addresses in New Jersey and 233 schools addresses in Eskishehir, all provided by the official web pages of the respective Ministries of Education (Milli Eğitim Bakanlığı, 2016; Schools, 2016). The results of the calculation of the distances between the obtained coordinates and the actual coordinates of the schools is shown in Table 1.

When the geocoding process is applied to the 264 school addresses in New Jersey, 246 (93.2% of all addresses) coordinates can be obtained. 82.1% of these coordinates are located at a distance less than 100 m from the actual location. On the other hand, when the geocoding process is performed on the 233 school addresses in Eskishehir, only 150 coordinates can be obtained (65% of coordinates). 46.7% of these coordinates are located at a distance less than 100 m from the actual location. According to these results, the success rate of geocoding performed with US addresses is 82.1%, while this ratio is 46.7% for the addresses in Turkey. Considering the fact that address standardization is part of the address matching process, the latter success rate is extremely low. For this reason, a customized standardization process based on specific regions is a necessity.

The process of translating manually written addresses into a certain digital format is known as address standardization. Unlike human written addresses, in the standardized format every section of the address, e.g. road, city, etc., can be separately identified (Abbasi, 2005). This study seeks to increase the success rate by the standardization of addresses through an innovative use of Natural Language Processing, which uses artificial intelligence methods to communicate with the computer in natural language.

Chomsky's study of Natural Language Processing (Chomsky, 1986) is highly influential, and Natural Language Processing has been used in a number of pieces of research for various purposes, including correcting misspellings (Kukich, 1992), summarizing text (Hu & Liu, 2004), voice interaction with a computer (Weber, 2003), and translating between natural languages (Nitta, Okajima, & Yamano, 1987).

Improving geocoding process results is not an unknown topic. Research which examines concerns over accuracy (Dickson et al., 2017), has undertaken geocoding with three tools and reached match rate results between 82% and 88%. This research suggests that parsing/standardizing tools can greatly improve the results of geocoding and seeks to improve the results by using different tools in different parts of the geocoding process. For example, while one tool is used to determine the input addresses, another is used to geocode addresses at street level. The same strategy is used in another study (Yang, Bilaver, Hayes, & Goerge, 2004) to test three geocoding tools and demonstrate that each tool provides better results in different parts of the geocoding process. As a result, the research concludes that the best strategy would be to combine the three methods.

In another study (Tian et al., 2016), an optimized address matching method for Chinese geocoding with three components, address modeling, address standardization and address matching, is proposed. The suggested model is structured around a standardization process based on an address tree model proposed in a previous study (Mengjun, Qingyun, & Mingjun, 2015). In this model the input address string is parsed and organized as a collection of address elements X and a semantic collection S. The root node is created and address element X1 is extracted. Finally, all the semantic elements in S1 associated with X1 are traversed to create address semantic nodes and connected to the root node. It is concluded that 60.4% of the addresses are matched accurately for company data, 86% of the addresses are matched successfully at a matching degree > 60%, and the corresponding matching rates for disease data are 49.3% and 98.6%, respectively, at the same matching degrees. These results are obtained by using the geocoding service created by them and the local data in their study area. It is pointed out that the reasons for there being different results for different data, include the company data being prepared in a more regular manner, and many of the addresses in the disease data not being located in their study area.

Advanced probabilistic methods such as Hidden Markov Models (HMM) are used to deal with misspellings and misplacements in some studies (Peter Christen, Churches, & Willmore, 2004; Christen, Willmore, & Churches, 2006; Churches, Christen, Lim, & Zhu, 2002). (Christen et al., 2004) propose a technique based on HMM to clean and standardize addresses and reach a 72.87% address level matching using Geocoded National Address File (GNAF) data.

Our study does not perform a simple standardization process. The model developed in this research enables the correction of spelling mistakes in input data, completion of the missing data in the addresses, and most importantly, by examining which address structure gives the higher success rate in the geocoding process, it offers a new address format. In order to improve the results of the geocoding process, the test dataset is standardized in three formats (PTT, Google, ArcGIS) using an NLP-based approach. For this purpose, parser, semantic analyzer and generator modules contained in the NLP systems are established. With these modules, raw address texts are parsed into address components; misspellings, abbreviations and incorrect address formats are corrected; and address data is created in the new format. In order to examine the impact of the address formats on the results of the geocoding process, the geocoding process is applied to the reproduced address data in the various formats and the raw address data. The online world geocoding services of google geocoding API and ArcGIS geocoding API are used to perform the geocoding operations.

2. Materials and methods

2.1. Dataset

In the study, the addresses of 233 primary schools located in the Eskishehir (Eskişehir) region are selected as a test dataset. These addresses are obtained from the official websites of the schools. Examples of the data are given in Table 2.

Addresses in Turkey include words that indicate areas, such as street (sokak), neighborhood (mahalle), and avenue (cadde or bulvar). However, there is neither a single address format for the input address nor a specific standard for the abbreviations. For example, while one uses the abbreviation “Sk.” to denote the word sokak, the other uses “Sok”. More importantly, some of the addresses omit the information regarding the province. In the study, the actual coordinates of the schools are obtained using Google Earth on top of the school buildings, to check the accuracy of the coordinates found through the geocoding process.

This study is conducted based on school addresses because in Turkey schools are also used for purposes other than education, for instance, in the nationwide central examinations run by the Ministry of Education. In many of these examinations, candidates are assigned to a school buildings located in provinces other than where they live, which poses a problem in finding the location of the examination centre. With the method proposed in this study the addresses of schools, hotels or