Computer Science • 16 (4) 2015 http://dx.doi.org/10.7494/csci.2015.16.4.395

Wojciech Korczyński

RETRIEVAL AND INTERPRETATION OF TEXTUAL GEOLOCALIZED INFORMATION BASED ON SEMANTIC GEOLOCALIZED RELATIONS

Abstract *This paper describes a method for geolocalized information retrieval from natural language text and its interpretation by assigning it geographic coordinates. Proof-of-concept implementation is discussed, along with a geolocalized dictionary stored in a PostGIS/PostgreSQL spatial relational database. The discussed research focuses on the strongly inflectional Polish language; hence, additional complexity had to be taken into account. The presented method has been evaluated with the use of diverse metrics.*

Keywords geolocalization, geolocalized dictionary, geolocalized relations, natural language processing

Citation Computer Science 16 (4) 2015: 395–414

1. Introduction

Each year, human beings generate more and more data. Many computer science domains set themselves the crucial task of comprehending and interpreting this vast amount of information. One such field is natural language processing, which has gained more and more popularity over the last several years. Nevertheless, there still contains a lot of unexplored and challenging areas, especially in research focused on languages with complex inflectional system (such as Polish). Many solutions, which obtained high evaluation results for languages with simple inflectional system (e.g., English), have much worse efficiency when an additional complexity layer is added in the form of strong inflection.

Some natural language processing tasks are designed for retrieving useful data from text. The mechanism of retrieval and interpretation of geolocalized information is the main topic of the current paper, which describes a method for the automatic determination of geographical coordinates of the locations described in natural language text.

The aforementioned geolocalized information retrieval and coordinate determination technique is based on the concept of semantic geolocalized relations, which this paper also introduces.

As an example of a practical application of the introduced theory, a system able to determine the coordinates of places described in text has been created. This system determines semantic geolocalized relations and matches them against a geolocalized dictionary created with OpenStreetMap data and stored in the PostgreSQL database extended with the PostGIS module (which enables the creation and use of spatial databases). Because of the availability of test data, our research was focused on the city of Krakow (in southern Poland).

The test data came from a real-life application, a part of the IBM Smart Cities project. Since the texts used in these experiments were in Polish, the specificity of the language (in particular inflectional system) had to be taken into consideration.

The effectiveness of the discussed implementation was deeply analyzed and evaluated with the use of diverse metrics. The results look promising and may be a valuable basis of future work and development.

Taking into account the test data features and cognitive research characteristics, this paper focuses on determining locations in the city of Krakow. Therefore, an appropriate spatial reference system (SRS) was used. However, the presented technique and implementation may be applied to any other place in the world and any SRS.

The presented research undertakes the problem of coping with a complex inflectional language such as Polish. However, it has to be emphasized that the discussed theory and its implementation are applicable in any other language.

This paper is an extended version of the work published in [17]. The presented theory of geolocalized relations has been expanded, and the system construction (along with descriptions of the technologies and processing strategy used) is analyzed

more profoundly. In addition, a related-work review and the possibilities of further development of the introduced idea have been included.

The paper starts with a short overview of works related to the discussed subject (Section 2). Section 3 introduces the concept of a geolocalized dictionary. In Section 4, the notion of geolocalized expressions (along with the geolocalized relations and their taxonomy) is explained. The aforementioned implementation and obtained experimental results are presented and discussed in Sections 5 and 6, respectively. Section 7 concludes the paper and suggests future work.

2. Related works

The field of Geographical Information Retrieval (GIR) and georeferencing has been researched for some time, especially in the subjects of geographic and spatial queries [19] or mapping of Internet resources to geographic locations [28]. Although studies on assigning coordinates to places described in text had been already conducted [39], the first TREC (Text Retrieval Conference)-style forum for GIR systems evaluation, GeoCLEF, took place in 2005 [8]. However, as far back as 1994, Woodruff and Plaunt presented a system called GIPSY in [39], which extracted location names from a text and constructed with spatial information in order to support the georeferenced document indexing. Over the next years, other studies in the field of textual geographical information retrieval and georeferencing were carried out; i.a. by Pouliquen et al., who created a system that analyzed natural language documents in search of place names and then produced maps depicting the geographical coverage of texts about a given topic [32].

Currently, tasks related to geographic data – geographic information retrieval from natural language texts, recognition of spatial qualifiers, textual geographic information processing (e.g., disambiguation of place names, interpretation of vague names and qualifiers), etc. – are widely researched. Notable popularity was gained by a subject of geographic information retrieval from web resources such as common web content, user logs [36], or social media content [3, 11]. Work presented by Han et al. in [11] is particularly interesting, as it focuses on text-based geolocation prediction for Tweets, which are examples of user-generated content (UGC). The discussed approach based on various feature selection methods in order to recognize location-indicative words. Moreover, the influence of different languages on the obtained results was examined. In [34], Roller et al. described a supervised method for geolocating documents by comparing them to previously constructed pseudo-documents, created by concatenating training documents labeled with geographical coordinates. A similar technique was applied by Wing and Baldridge in [38]. Karimzadeh et al. introduced a web tool capable of recognizing place names in short texts and then associating them with actual points on a map [14]. Many papers focus on geolocalization in a large scale (i.e., geolocalization of geographical objects such as cities, rivers, lakes, mountains, etc.) and discusses the methods of relatively rough estimation.

The problem of geo-coding (i.e., the task of mapping textual geographical data to specific coordinates) was also widely discussed in [20], where emphasis was put on resolving names of populated places. An algorithm for geoparsing place names of fine granularity was introduced by Derungs and Purves in [5].

The notion of spatial relations was also presented by many researchers. Klien and Lutz discussed these relations in [15] and proposed an automated method for the semantic annotation of geographic data. Egenhofer and Franzosa focused in [6] on spatial relations between sets. Spatial relations in natural language were also widely talked over in [35] and the problem of their semantics was undertaken (e.g., in [12]).

It has to be noticed that a vast amount of works in the field of textual geographical information retrieval and georeferencing was focused on English and was not concerned with strongly inflectional languages such as Polish. A problem of geolocalization of places described in Polish texts was touched upon in [13]. This paper describes an algorithm for estimating coordinates of cities and villages mentioned in the Polish geographical encyclopedia.

The novelty of the presented paper has to be emphasized. Not only does it deal with strongly inflectional languages, but also aims at geolocation in smaller scale with the use of the novel approach to the problem based on the idea of *semantic geolocalized relations*. The introduced method applies particular taxonomy of these relations and serves for precise geo-coding of objects of fine granularity.

3. Geolocalized dictionary

Processing languages with complex and difficult inflectional system (such as Polish) is a very challenging task. The vast majority of Polish words are inflectable [26, 40].Referring to location and infrastructural objects (which this paper focuses on), people use inflected names of streets or places every day. For example, in Polish, one cannot say *Idę ulicą* ***Kijowska*** (*I am going down Kijowska street*) or *Czekam na ulicy* ***Mazowiecka*** (*I am waiting on Mazowiecka street*) but *Idę ulicą* ***Kijowską*** and *Czekam na ulicy* ***Mazowieckiej***. Thus, in order to be able to correctly recognize street or place names included in Polish texts, it was necessary to create a dictionary of inflections of Krakow street names as well as a list of places in the city such as churches, universities, squares, bus and tram stops, etc. It should be emphasized that no such dictionary has been found.

Data exported from the OpenStreetMap project (this project is discussed in Section 4) was the basis of this dictionary. The proper XML file describing a map of Krakow was downloaded, and information tagged as *highways* was the matter of interest. This data was extracted with the use of the Imposm Python module[1].

Moreover, CLP was used. CLP is a library capable of determining the part of speech of a given word and its set of inflective forms [7, 24]. It is worth noting that information about part of speech is also relevant, since there is no need to look for

[1] http://imposm.org/

inflective forms of nouns – if a street name is a noun, it maintains the same form in any case, e.g.: *To jest ulica* ***Lea*** (*This is Lea street*), *Idę ulicą* ***Lea*** (*I am going down Lea street*), *Czekam na ulicy* ***Lea*** (*I am waiting on Lea street*).

Unfortunately, the CLP library was not able to find forms of street names that were uncommon words (i.e., not contained in the Polish dictionary – e.g., *Altanowa street*, *Soboniowicka street*, etc.). Therefore, the transducer described in [18] was used to perform stemming (word's core extraction [22]) and assign to each word its inflective label that specifies the unambiguous set of inflective forms [18, 23].

Eventually, the dictionary of inflections of Krakow street names consisted of 2988 names (out of 3322 exported from OpenStreetMap – some names were rejected because their inflections were not determined in the aforementioned process). This amount was satisfactory enough to continue work. It did not contain all of the names of streets in Krakow, but the creation of the discussed dictionary was not the main aim of the research, as the dictionary itself was only an important auxiliary element.

Apart from the dictionary of inflections of street names, a list of locations in Krakow was made. This list was also created based on data exported from OpenStreetMap. There were 2685 locations such as: churches, universities, schools, shops, galleries, hospitals, restaurants, squares, bus and tram stops, monuments, etc.

A dictionary of inflections of location names was not created, as it was not as essential as the previous one was. Localization is specified with the use of street names more frequently. Moreover, some inflected location names could be recognized by an error-correction mechanism, such as Levenshtein edit distance [21]. However, it is obvious that this method is not as effective as if a dictionary of inflections of location names was applied. Therefore, research with the use of such a dictionary would be an interesting subject of future work.

4. Geolocalized expressions

In the beginning of discussing geolocalized expressions and the taxonomy of geolocalized relations, a definition of *preposition* has to be reintroduced. According to the Merriam-Webster Dictionary, a preposition is "a word or group of words that is used with a noun, pronoun, or noun phrase to show direction, location, or time, or to introduce an object"[2]. So it is an inseparable word that helps define relations between objects. Exemplary prepositions are: *on, over, by, above, in, between, in front of.*

Moreover, a definition of a *prepositional phrase* would also be helpful. According to the Macmillan Dictionary, a prepositional phrase is "a phrase consisting of a preposition and the noun or pronoun that comes after it"[3]. Some exemplary prepositional phrases are: *in the building, on the table, through the tunnel, behind the door, between the trees.*

[2]http://www.merriam-webster.com/dictionary/preposition
[3]http://www.macmillandictionary.com/dictionary/british/prepositional-phrase

A geolocalized expression is a kind of prepositional phrase that describes some place. Geolocalized expressions provide information about the localization of given object. It should be precise so one can interpret it and determine geographical coordinates of the described object's localization. Hence, these expressions should be richer. Some exemplary geolocalized expressions are presented below:

- *in* the main building of University of Science and Technology in Krakow → `(50.06457° N, 19.92341° E)`,
- *by* the Adam Mickiewicz monument on the Main Square → ca. `(50.06145° N, 19.93794° E)`,
- *on* the Main Square on the right side of the entrance to St. Mary's Church → ca. `(50.06145° N, 19.93875° E)`.

The geolocalized expressions (as well as an attempt to interpret them and determine geographical coordinates based on the analysis of such expressions) are the main topic of this paper.

In order to determine coordinates of the localization described in text, a concept of *geolocalized relation* [25] has to be introduced. According to the Macmillan Dictionary, a relation is "a connection between two or more people or things"[4]. Thus, a geolocalized relation used to determine a given object's position is the relation that occurs in a space between two or more elements of that space. Such a relation may be defined according to the schema:

$$xRy \iff x + \ (predicate) + preposition \ + y,$$

where x and y represent objects between which the given relation occurs. It is worth emphasizing that, in some cases (discussed later), y may be a pair of objects. *Predicate* (e.g. *is*, *lies*) is an optional element because of the fact that geolocalized relations may appear in elliptical sentences, for example. In some cases, there is a prepositional phrase instead of the *preposition*.

4.1. Taxonomy of geolocalized relations

The classification of geolocalized relations divides them into 4 types, which are discussed below in detail. For each type of relation, a set of characteristic prepositions can be defined.

4.1.1. The I type of geolocalized relation: membership

The I type of geolocalized relation describes the case of a relation between two objects that enables us to determine the precise and exact location of the first object. This case is illustrated in Figure 1 – one can state that the red object lays **exactly** *on* the black road and **exactly** *in* the black object.

Prepositions characteristic for the I type of geolocalized relations are: *in*, *on*, *inside*.

[4]http://www.macmillandictionary.com/dictionary/british/relation

Examples of the usage of the I type of geolocalized relation: *an apple lies **in** the box*; *I am waiting **inside** the building.*

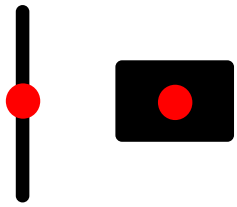

Figure 1. Illustration of the I type of geolocalized relation.

4.1.2. The II type of geolocalized relation: adherence

The II type of geolocalized relation is also a kind of relation between two objects, and it is used when one is not able to determine the location of a given object precisely, as it is only possible to state that this object is localized somewhere within a small distance from the second object. This case is illustrated in Figure 2 – the red object lays *by* the black one, **somewhere** in the area marked by arrows.

Prepositions characteristic for the II type of geolocalized relation are: *behind*, *along*, *round*, *next to*, *by*, *in front of*, *over*, *under*, *above*.

Examples of the usage of the II type of geolocalized relation: *a cat is standing **behind** the couch*; *I left the car **by** the city hall*; *new lights were installed **along** the path.*

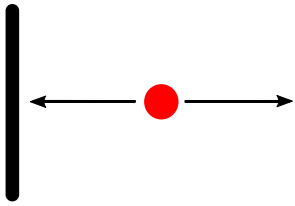

Figure 2. Illustration of the II type of geolocalized relation.

4.1.3. The III type of geolocalized relation: neighborhood

The exact location of an object cannot be determined in the case of the III type of geolocalized relation either. One only knows that this object is located somewhere in the area limited by pair (or more) of other objects. It can also be located in the area set by many objects; e.g., *among* trees, *among* buildings. Figure 3 is an illustration of

the III type of geolocalized relation. It presents the red object located **somewhere** *between* two black ones.

Prepositions characteristic for the III type of geolocalized relation are: *between*, *among*.

Examples of the usage of the III type of geolocalized relation: *there is a new shop* ***between*** *the university and the train station*; *I left my bike somewhere* ***among*** *the trees*.

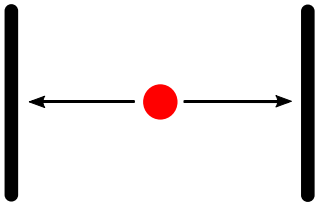

Figure 3. Illustration of the III type of geolocalized relation.

4.1.4. The IV type of geolocalized relation: intersection

The IV type of geolocalized relation is a relation that occurs between one object and the pair of other objects (usually roads, streets, etc.). In the case of this type of relation, a preposition is replaced by a prepositional phrase. Using the IV type of geolocalized relation, one is able to specify exact and precise localization of a given object, as it is located at the intersection of the other two objects.

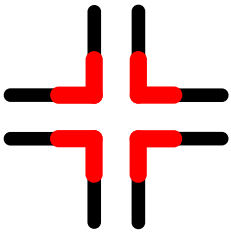

Figure 4. Illustration of the IV type of geolocalized relation.

Prepositional phrases characteristic for the IV type of geolocalized relation are: *on the corner*, *at the crossroads*, *on the junction*, *at the intersection*.

Examples of the usage of the IV type of geolocalized relation: *there was an accident* ***at the intersection*** *of X St. and Y St.*; *there is a restaurant* ***on the corner*** *of the Z St.*

However, it is worth noting that, if some object is located *on the corner* of two streets, it is necessary to provide additional information to determine which corner

is in question. This case is presented by Figure 4 – the possible locations of a given object are marked in red. The number of corners taken into consideration may be lower: if one of these streets ends at this junction, there are only two corners. And if both streets end in this place, there is just one corner.

In the presented work, intersections are described with the help of just one point, so all corners are indistinguishable.

4.2. Representation of geolocalized relations

In order to be able to find geolocalized relations in text, an extension for the **PostgreSQL** database management system, named **PostGIS**, was used. It allows us to handle an object-relational database as a spatial database that supplies functions that operate on geographic data. The data used in the presented work was created within the framework of the **OpenStreetMap** project.

4.2.1. OpenStreetMap

OpenStreetMap[5] is a collaborative project created with the motivation of providing free and unconstrained access to precise geographical data (whereas such information is very expensive and mostly unavailable). Its main aim is to create a world map that would be free, editable, and open for general use [4, 10].

Maps are made from data that can be divided into 3 types of objects: nodes, ways, and relations. A node is a basic unit: it contains information about longitude and latitude and represents a key point on a map (junctions, bus and tram stops, buildings, etc.). A way consists of a list of ordered nodes, while a relation is a group composed of nodes, ways, and other relations [1]. Relations are used to represent (for example) bus or tram routes, squares, etc.

Information about each object is stored in *tags* in the form of `key=value`. Values of some keys may be selected only from a strictly defined set (e.g., among permissible values for *building* key are among others: *house*, *church*, *hospital*, *university*), whereas values of the other keys are not restricted in any way; e.g., *addr:street* (street next to which a given object is located), *phone* (phone number connected with a given location) [1, 10].

OpenStreetMap data for the given area may be exported to the *OSM XML* (*OpenStreetMap Extensible Markup Language*) file [10] e.g. with the use of **imposm.parser** Python library[6].

Unfortunately, there were some problematic issues that had to be resolved while using OpenStreetMap data:

- There is no easy method for exporting data of big areas such as the city of Krakow. Therefore, there is the need to use a website with previously prepared data[7]. Unfortunately, this data contains some smaller towns and villages that

[5] http://www.openstreetmap.org
[6] http://imposm.org/docs/imposm.parser/latest
[7] https://mapzen.com/metro-extracts/

add unnecessary information and causes problems with the ambiguity of street names.

- OpenStreetMap does not provide any verification and validation of the data created [1], and as a consequence, it may be noisy. Such invalid data needed to be filtered out.
- Some parts of the data were inconsistent. Inconsistency revealed in using different names for the same street; e.g., *aleja Adama Mickiewicza*, *aleja Mickiewicza*, *aleja A. Mickiewicza* are all the names of the same street. Removal of inconsistency is a complex problem and required the creation of groups of each object's different names, from which the most complete one could be chosen.

4.2.2. PostGIS

Geographical data exported from OpenStreetMap was stored in an object-relational PostgreSQL database management system[8] extended with PostGIS module[9], which adds the possibility of spatial database creation. Such a database allows us to process spatial, geometrical, and geographical objects using SQL (Structured Query Language) queries [2, 30].

PostGIS provides over 300 function and operators [30], all of which follow *Simple Features for SQL*[10] specification acknowledged by the OCG (*Open Geospatial Consortium*) [30, 33]. As it may be used freely, it gains more and more popularity as a database management system for applications operating on spatial data [29].

Integration of the OpenStreetMap data and PostgreSQL/PostGIS spatial database is ensured on a high level [41]. There are many programs that help with importing this data to the database; e.g., *OSMOSIS*, which was used for the purpose of this research. Not only does it import spatial data, but it also creates a complete database schema, taking characteristics of OpenStreetMap data under consideration [9, 37].

While using the PostgreSQL/PostGIS spatial database, the currently used *spatial reference system* (SRS) has to be kept in mind. SRS specifies i.a. how the coordinate system and data on plane are presented and which measure units (meters, feet, degrees, etc.) are used. For instance, a PostgreSQL/PostGIS user has to pay attention if the distance between two places is calculated as the length of a segment that connects them in a straight line or along the Earth's zone; if necessary, one should apply a planar projection [30].

In the presented research *ETRS89/Poland CS92*, the SRID (*Spatial Reference System Identifier*): 2180[11], was used, as it was created especially for the zone of Poland. It uses a flat rectangular coordinate system that was built by assigning points of the Earth's surface to the points on a plane according to the Gauss-Krüger projection. The basic measure unit is the meter [31]. However, there are other systems that might

[8] http://www.postgresql.org
[9] http://www.postgis.org
[10] http://www.opengeospatial.org/standards/sfa and http://www.opengeospatial.org/standards/s
[11] http://spatialreference.org/ref/epsg/2180

be used, such as the very popular *NAD83/Massachusetts Mainland* (SRID 26986)[12]. Moreover, it has to be emphasized that any other SRS could be used, depending on the geographical area being considered.

Table 1 contains information about the number of particular objects imported to the PostgreSQL/PostGIS database for the purpose of our research.

Table 1

Number of objects imported to the PostgreSQL/PostGIS database

Type of object	Number of elements
Nodes	699725
Ways	76726
Relations	948

5. System construction

A proof-of-concept implementation has been created as an example of the practical application of the presented theory. A system that determines coordinates of places described in text was created in the **Python**[13] programming language. This technology was chosen with regard to its efficiency and simplicity of programming, very good text processing tools, and trouble-free integration with the PostgreSQL database management system. In order to connect the program written in Python to a database, the **psycopg2**[14] adapter was used.

The system may be divided into two parts:

- the main part that processes input texts and returns found coordinates of described places,
- `relations` module, which contains implementations of relationhips introduced in Section 4.1.

`Relations` module consists of three parts:

- implementation of geolocalized relations that describe places/points only (e.g., *in* some point, *next to* the some place, *between* two points),
- implementation of geolocalized relations that describe places/points as well as streets (e.g., *between* some street and some point),
- implementation of geolocalized relations that describe streets only (e.g., *on* some street, *next to* the some street, *between* two streets, *on the junction* of two streets).

[12]http://spatialreference.org/ref/epsg/26986
[13]http://www.python.org
[14]http://initd.org/psycopg

5.1. The processing strategy

Before implementing the system, the proper data processing strategy had to be chosen depending on their characteristics. In the following points, the consecutive processing steps are discussed.

- First of all, the standard action of filtering out the most frequent words had to be performed. These words do not have any meaning for the context of the main information. They are usually contained in the so-called *stop-list* that consists of the most common words in a given language or corpus (an exemplary stop-list for Polish language is available on Wikipedia[15] – it was used in the experiments after removing all prepositions, which are crucial in the presented research).
- Since the system is designed to work with informal texts that contain many mistakes (such as typographical or spelling errors), the proper correction needs to be performed. In order to correctly recognize wrongly typed words, the Levenshtein edit distance was applied.
- Moreover, as this work focuses on texts that are in Polish (which is an inflective language), each recognized inflected street name had to be replaced with its basic form. The geolocalized dictionary introduced in Section 3 was an essential part of this step, as it contained basic forms of inflected street names.
- In the next step, the system tried to recognize a type of geolocalized relation used in a text by looking for keywords – prepositions characteristic for each type of relation.
- At the very end, the system attempted to interpret the retrieved geolocalized information and determine the coordinates of the described places by applying PostGIS spatial query appropriate for a given type of relation.

These successive stages of data processing are illustrated in Figure 5.

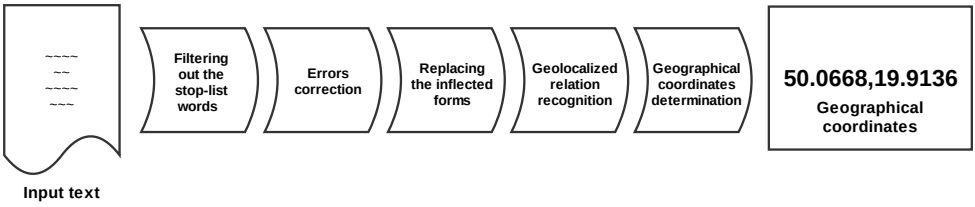

Figure 5. Illustration of the successive stages of data processing performed by the system.

Figure 6 presents consecutive processing steps for an exemplary sentence *Proszę o usunięcie dziury na skrzyżowaniu Królwskiej i Kijoskiej, bo ciężko się jeździ* (*I ask for repairing the hole at the intersection of Królwska St. and Kijoska St. because it's hard to drive*). The names of both streets are typed incorrectly.

[15]http://pl.wikipedia.org/wiki/Wikipedia:Stopwords

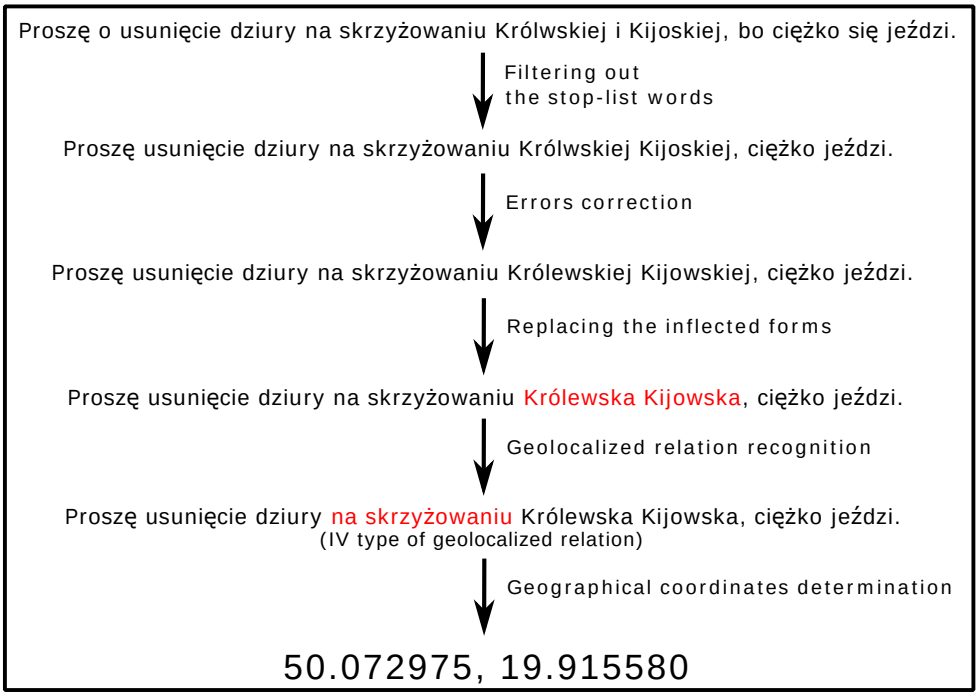

Figure 6. Successive stages of data processing for an exemplary sentence.

6. Experimental results

6.1. Corpus

A corpus with the test data used in our experiments came from a real-life application from the IBM Smart Cities project and was a set of 210 complaints sent to ZIKiT (*Polish: Zarzd Infrastruktury Komunalnej i Transportu w Krakowie* – Board of Municipal Infrastructure and Transport in Krakow). Each complaint described a single location and contained real remarks about failures, problems with infrastructure, bad states of streets, etc.

This corpus was exceptionally valuable, as there were coordinates of the described place added to each complaint. This data was useful as a model, to which coordinates determined by the system were compared.

Finally, in order to prepare the corpus for experiments, it was indispensable to add a piece of information to each complaint about the used geolocalized relation to evaluate the effectiveness of the geolocalized relation recognition.

Unfortunately, this corpus had some major drawbacks, as some of texts did not carry any information about any place – either it was not included by the complaint's

author, or the received complaint could be classified as a SPAM message. Such input texts negatively influenced the obtained results.

6.2. Metrics

The effectiveness of the implementation has been evaluated using two kinds of metrics – effectiveness measures and the simple geolocalized measure.

The first group consists of well-known measures: precision (percentage of retrieved elements that are relevant), recall (percentage of relevant elements that are retrieved), and F1 measure (harmonic means of precision and recall) [27]. All of them have been used for effectiveness evaluation of recognizing the geolocalized relation type used in an input text.

The second metric is based on geographical information, as it measures the distance between two points on a map. For each point X determined by the system, an error has been defined as the distance between this point and a model point Y to which the input text refers.

6.3. Results analysis

The main aim of performed experiments was to verify the effectiveness of determining the geographical coordinates of places described in natural language texts. These texts came from the corpus discussed in Section 6.1.

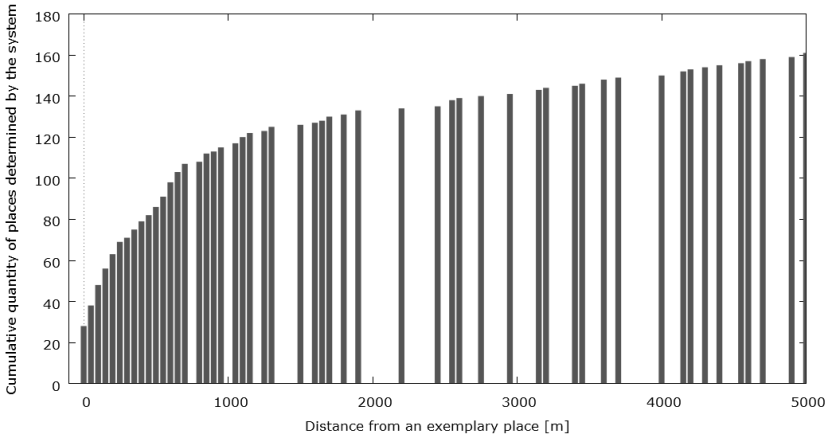

Figure 7. Error cumulative distribution.

Firstly, an error computed as a distance between point determined by the system and a model point defined by geographical coordinates of place described in a text was studied. Figure 7 illustrates a cumulative distribution (i.e., each bar represents the number of cases with errors less than the given value) of such distances (errors), calculated for each text in a corpus. A fifty-meter range of distances may be found

on the x-axis, which has been limited to 5000 meters. The cumulative number of distances included in the particular range has been specified on the y-axis.

The results show that, in 28 cases, the distance between the determined point and the place described in the text was smaller than 50 meters (and smaller than 100 meters in 38 cases). Generally, these results seem to be satisfactory, since an error was smaller than 1000 meters in the significant number of input texts. It also has to be noticed that larger errors were caused by the meaningless texts mentioned in Section 6.1. Moreover, some of these texts lacked precision in their place description.

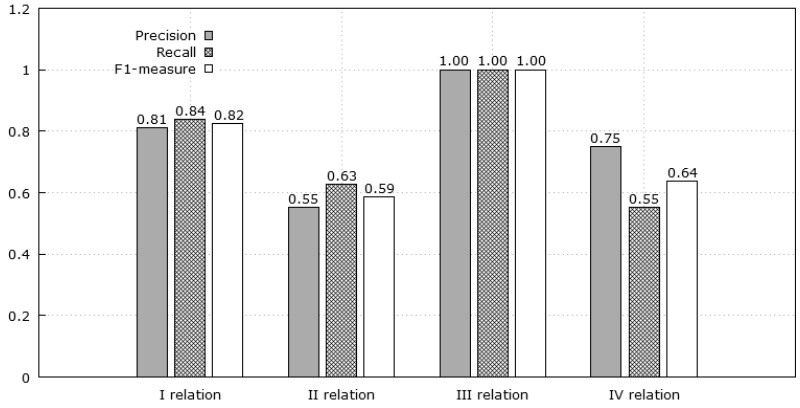

Figure 8. Effectiveness of geolocalized relation type recognition.

In the second experiment, the effectiveness of the geolocalized relation type recognition was examined with the use of the three aforementioned metrics: precision, recall, and F1 measure. Figure 8 presents the obtained results. Besides, Table 2 contains information about the number of correctly recognized relations compared to the total number of relations used in the corpus.

Table 2

Number of correctly recognized relations

Type of relation	Correctly recognized	Total
I relation	103	123
II relation	27	43
III relation	6	6
IV relation	21	38

One can see that these numbers are also quite good. The III type of geolocalized relation was recognized with no mistakes. In the case of the I type of geolocalized relation, the effectiveness was also satisfactory, while the results of recognition of the remaining relations might be acknowledged as acceptable. It should be kept in mind that this paper describes research in a quite-unexplored area, and the presented

system should be treated as a proof-of-concept (which may be the basis of further development).

7. Conclusions

This paper discussed the theory of geolocalized relations. The main part of the work was focused on the mechanism of geolocalized relation retrieval from natural language text and afterwards interpreting them by determining geographical coordinates of places described in an input text. Finally, an implementation based on OpenStreetMap data and the PostgreSQL/PostGIS database was introduced as a practical application of the presented theory. The obtained results were evaluated with the use of effectiveness and geolocalized measures.

The presented research was focused on processing Polish texts describing places located in the area of the city of Krakow. Nevertheless, the introduced theory and created system might be applied to any language and any area of the world.

Taking into consideration the novelty and cognitive character of the research based on particular semantic geolocalized relations (as well as the fact that the implementation was only a proof-of-concept), the obtained results look very promising and prove that further work and development of the discussed idea could be very beneficial.

Future work should focus on the improvement of accuracy of geographical coordinate determination. First of all, since the current implementation iterates over each word in a text and looks upon each word as a separate unit, the Named Entity Recognition system would be particularly helpful in the identification of multiword names of streets, places, or other objects. Moreover, the process of name recognition should be aware of objects that can be called different names but should have the same coordinates assigned (e.g., words like *campus*, *dormitories*, *student residences* can describe the same place). Besides, more precise text analysis would be useful: such mechanisms as exact address recognition (that is, composed of both street names and numbers) or taking into consideration additional descriptions (which could restrict the search area) would increase accuracy and efficiency. Furthermore, interpretations of the II type of geolocalized relation could be improved by taking into account expressions in which a distance is mentioned (e.g., *5 metres from...*, *10 meters behind...*, etc.), as it enables us to set a precise object neighborhood radius (see: Sec. 4.1.2). Finally, system accuracy would be certainly increased if a dictionary of inflections of location names similar to the one created for street names (see: Sec. 3) was introduced. It could be also interesting to test how another error-correction mechanism (e.g., n-gram similarity [16] instead of the Levenshtein edit distance) would influence the obtained results. Last but not least, the problem of preposition ambiguity should be taken into account as well as the issue of more complex cases of sentences (i.e., with many prepositions), where a straightforward search for keywords – prepositions characteristic for each type of geolocalized relation – could not provide satisfactory results.

Currently, research focused on an application of a Named Entity Recognition system to discussed implementation is being conducted, and valuable results are expected to be obtained.

References

[1] Ballatore A., Bertolotto M., Wilson D.C.: Geographic Knowledge Extraction and Semantic Similarity in OpenStreetMap. *Knowledge and Information Systems*, pp. 1–21, 2012.

[2] Blasby D.: Building a Spatial Database in PostgreSQL. Refractions Research, 2001. `http://wiki.postgis.org/files/OSDB2_PostGIS_Presentation.pdf`.

[3] Cheng Z., Caverlee J., Lee K.: You are where you tweet: a content-based approach to geo-locating twitter users. In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pp. 759–768. ACM, 2010.

[4] Corcoran P., Mooney P.: Characterising the Metric and Topological Evolution of OpenStreetMap Network Representations. *The European Physical Journal Special Topics*, vol. 215(1), pp. 109–122, 2013.

[5] Derungs C., Purves R.S.: From text to landscape: locating, identifying and mapping the use of landscape features in a Swiss Alpine corpus. *International Journal of Geographical Information Science*, vol. 28(6), pp. 1272–1293, 2014.

[6] Egenhofer M.J., Franzosa R.D.: Point-set topological spatial relations. *International Journal of Geographical Information System*, vol. 5(2), pp. 161–174, 1991.

[7] Gajęcki M.: *Słownik fleksyjny języka polskiego CLP – opis użytkowy.* Katedra Lingwistyki Komputerowej, Katedra Informatyki, Akademia Górniczo-Hutnicza, Kraków, 2008.

[8] Gey F., Larson R., Sanderson M., Joho H., Clough P., Petras V.: *GeoCLEF: the CLEF 2005 cross-language geographic information retrieval track overview.* Springer-Verlag, Berlin – Heidelberg, 2006.

[9] Goetz M., Lauer J., Auer M.: An Algorithm Based Methodology for the Creation of a Regularly Updated Global Online Map Derived from Volunteered Geographic Information. In: *Proceedings of the Fourth International Conference on Advanced Geographic Information Systems, Applications, and Services*, C.P. Rückemann, B. Resch, eds, pp. 50–58. Valencia, 2012.

[10] Haklay M., Weber P.: OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing*, vol. 7(4), pp. 12–18, 2008.

[11] Han B., Cook P., Baldwin T.: Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, pp. 451–500, 2014.

[12] Herskovits A.: *Language and spatial cognition.* Cambridge University Press, 1987.

[13] Jaśkiewicz G.: Geolocalization of 19th-century villages and cities mentioned in geographical dictionary of the kingdom of Poland. *Computer Science*, vol. 14(3), pp. 423–442, 2013.

[14] Karimzadeh M., Huang W., Banerjee S., Wallgrün J.O., Hardisty F., Pezanowski S., Mitra P., MacEachren A.M.: GeoTxt: A Web API to Leverage Place References in Text. In: *Proceedings of the 7th Workshop on Geographic Information Retrieval*, GIR '13, pp. 72–73. ACM, New York, NY, USA, 2013.

[15] Klien E., Lutz M.: The role of spatial relations in automating the semantic annotation of geodata. In: *Spatial Information Theory*, pp. 133–148. Springer-Verlag, Berlin – Heidelberg, 2005.

[16] Kondrak G.: N-Gram Similarity and Distance. In: *String Processing and Information Retrieval*, M. Consens, G. Navarro, eds, *Lecture Notes in Computer Science*, vol. 3772, pp. 115–126. Springer-Verlag, Berlin – Heidelberg, 2005.

[17] Korczyński W., Korzycki M.: Extraction and application of geolocalized dictionaries, vol. 1, pp. 593–600. STEF92 Technology, 2014.

[18] Korzycki M.: *Transducer skończenie stanowy jako narzędzie rozpoznawania form tekstowych wyrazów polskich.* Ph.D. thesis, Katedra Informatyki, Wydział Elektrotechniki, Automatyki, Informatyki i Elektroniki, Akademia Górniczo-Hutnicza, Kraków, 2008.

[19] Larson R.R.: Geographic information retrieval and spatial browsing. *GIS and Libraries: Patrons, Maps and Spatial Information*, pp. 81–124, 1996.

[20] Leidner J.L.: *Toponym resolution in text.* Ph.D. thesis, University of Edinburgh, 2007.

[21] Levenshtein V.: Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics-Doklady*, vol. 10(8), pp. 707–710, 1966.

[22] Lovins J.B.: Development of a Stemming Algorithm. *Mechanical Translation and Computational Linguistics*, vol. 11, pp. 22–31, 1968.

[23] Lubaszewski W.: *Gramatyka leksykalna w maszynowym słowniku języka polskiego.* Prace Instytutu Języka Polskiego. Polska Akademia Nauk, Instytut Języka Polskiego, Kraków, 1997.

[24] Lubaszewski W.: *Słowniki komputerowe i automatyczna ekstrakcja informacji z tekstu.* AGH Uczelniane Wydawnictwa Naukowo-Dydaktyczne, Kraków, 2009.

[25] Lubaszewski W., Gatkowska I.: Struktura semantyczna języka naturalnego. In: *Interfejs dla osób z dysfunkcją wzroku. Model kognitywny i przykład dobrej praktyki*, I. Gatkowska, W. Lubaszewski, eds, pp. 49–106. Wydawnictwo Uniwersytetu Jagiellońskiego, Kraków, 2013.

[26] Łuczyński E.: Fleksja języka polskiego z punktu widzenia ontogenezy mowy. *Biuletyn Polskiego Towarzystwa Językoznawczego*, vol. LVIII, pp. 157–165, 2002.

[27] Manning C.D., Raghavan P., Schütze H.: *Introduction to Information Retrieval.* Cambridge University Press, New York, 2008.

[28] Markowetz A., Brinkhoff T., Seeger B.: Geographic information retrieval. In: *Next generation geospatial information: from digital image analysis to spatio temporal databases*, P. Agouris, A. Croitoru, eds, pp. 5–17, A.A. Balkema Publishers, Leiden – London – New York – Philadelphia – Singapore, 2005.

[29] McArdle G., Ballatore A., Tahir A., Bertolotto M.: An Open-Source Web Architecture for Adaptive Location Based Services. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 38(2), pp. 296–301, 2010.

[30] Obe R.O., Hsu L.S.: *PostGIS in Action*. Manning Publications Company, 2011.

[31] Pażus R.: *Państwowy system odniesień przestrzennych – systemy i układy odniesienia w Polsce*. Główny Urząd Geodezji i Kartografii, Departament Geodezji Kartografii i Systemów Informacji Geograficznej, 2009.

[32] Pouliquen B., Steinberger R., Ignat C., De Groeve T.: Geographical Information Recognition and Visualization in Texts Written in Various Languages. In: *Proceedings of the 2004 ACM Symposium on Applied Computing*, SAC '04, pp. 1051–1058. ACM, New York, NY, USA, 2004.

[33] Ramsey P.: *Introduction to PostGIS*. Refractions Research, Victoria, 2007.

[34] Roller S., Speriosu M., Rallapalli S., Wing B., Baldridge J.: Supervised text-based geolocation using language models on an adaptive grid. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1500–1510. Association for Computational Linguistics, 2012.

[35] Shariff A.R.B., Egenhofer M.J., Mark D.M.: Natural-language spatial relations between linear and areal objects: the topology and metric of English-language terms. *International Journal of Geographical Information Science*, vol. 12(3), pp. 215–245, 1998.

[36] Wang C., Xie X., Wang L., Lu Y., Ma W.Y.: Detecting Geographic Locations from Web Resources. In: *Proceedings of the 2005 Workshop on Geographic Information Retrieval*, GIR '05, pp. 17–24. ACM, New York, NY, USA, 2005.

[37] Whitelegg N.: Using OpenStreetMap Data. Southampton Solent University, 2011. `http://www.bcs.org/upload/pdf/open-street-map-data-180313.pdf`.

[38] Wing B.P., Baldridge J.: Simple Supervised Document Geolocation with Geodesic Grids. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies – Volume 1*, HLT '11, pp. 955–964. Association for Computational Linguistics, Stroudsburg, PA, USA, 2011.

[39] Woodruff A.G., Plaunt C.: GIPSY: Automated geographic indexing of text documents. *Journal of the American Society for Information Science*, vol. 45(9), pp. 645–655, 1994.

[40] Wróbel H.: *Gramatyka języka polskiego*. Spółka Wydawnicza „Od Nowa", Kraków, 2001.

[41] Zheng J., Chen X., Ciepłuch B., Winstanley A.C., Mooney P., Jacob R.: Mobile Routing Services for Small Towns Using CloudMade API and OpenStreetMap. In: *Proceedings of the 14th Joint International Conference on Theory, Data Handling and Modelling in GeoSpatial Information Science*, pp. 149–154. Hong Kong, 2010.

Affiliations

Wojciech Korczyński
AGH University of Science and Technology, Department of Computer Science, Krakow, Poland, wojciech.korczynski@agh.edu.pl

Received: 20.02.2015
Revised: 9.08.2015
Accepted: 25.08.2015