A Secured Preposition-Enabled Natural Language Parser for Extracting Spatial Context from Unstructured Data

Patience U. Usip(✉), Moses E. Ekpenyong, and James Nwachukwu

Computer Science Department, University of Uyo, Uyo, Nigeria
{patiencebassey,mosesekpenyong}@uniuyo.edu.ng,
nwachukwujames7@gmail.com

Abstract. Acquiring data within the health domain is generally intractable due to privacy or confidentiality concerns. Given the spatial nature of health information, and coupled with the accompanying large and unstructured dataset, research in this area is yet to flourish. Further, obtaining spatial information from unstructured data is very challenging and requires spatial reasoning. Hence, this paper proposes a secure Preposition-enabled Natural Language Parser (PeNLP), sufficient for mining unstructured data to extract suitable spatial reference with geographic locations. The proposed PeNLP is a subcomponent of a larger framework: the Preposition-enabled Spatial ONTology (PeSONT) – an ongoing project. The short term impact of PeNLP is its availability as a reliable information extractor for spatial data analysis of health records. In the long run, PeSONT shall aid quality decision making and drive robust policy enactment that will greatly impact the health sector and the populace.

Keywords: Knowledge representation · Ontology · Spatial reasoning
Unstructured data

1 Introduction

The emergence of massive patient databases of electronic health records has provided new opportunities to test clinical hypotheses, inform clinical decision making, and optimize healthcare services. Improved medical decision making requires accurate predictive models, but the spatial nature of observational health data presents unique challenges, and requires an understanding of the impact of data representation on prediction. In [1], a sparse coding representation of medical records was implemented, and interfaced with the Observational Health Data Sciences and Informatics (OHDSI) software tools for predictive modeling. An empirical evaluation of the performance of traditional predictive models with and without sparse coding was also demonstrated to prove the importance of data representation as a step in building predictive models.

Traditional analytic methods are often ill-suited to the evolving world of healthcare big data usually characterized by massive volume, complexity, and velocity [2]. Numerous machine learning methods have effectively addressed such limitations, but they are still subject to the usual sources of bias that commonly arise in observational

V. Odumuyiwa et al. (Eds.): AFRICOMM 2017, LNICST 250, pp. 163–168, 2018.
https://doi.org/10.1007/978-3-319-98827-6_14

studies. Consequently, novel methods for developing good model estimates to efficiently represent, predict, and evaluate datasets containing healthcare utilization, clinical, personal devices, and many other sources, are required. Spatial ontology is therefore concerned with the application of human intelligence to spatial reasoning, hence, simplifying the complexities involved in scanning through numerous pages of textual data or listening to a multimedia files in search of spatial (location-based) concepts, which are necessary requirements for decision making. Decision making towards the formulation of strong public health policies in support of good health and well-being [3] (for instance) depends on real-time location-based data. These data are sometimes unavailable and unstructured and may lead to inconsistent data manipulation because of their geographic spread.

Health data requires ethical approval and hence are intractable to obtain due privacy or confidentiality concerns. The ability to link data in a manner that protects patient privacy has improved dramatically through the use of salting and hashing methodologies [2]. The sparse health data including spatial or location-based information has generally hindered research progress owing to the volume of data, alongside their characteristics (such as the unevenness of data completeness), which raises questions about the potential for using new methods to analyze areas such as treatment effectiveness, health care value, strengths and weaknesses of alternative care organization models, and policy interventions. Although statistical methods are mostly used to provide answers to these questions, they are either too generic or insufficient to handle spatial data analysis [4]. The categorization of locational data (place terms) using Linguistic and Logical perception [5] plays a key role in the classification and engineering in PeSONT.

The focus of this paper is limited to public health, with an attempt to answer basic research questions posed as follows:

(i) Can spatial concepts be obtained from a poll of unstructured public health data?
(ii) Are the extracted data, spatial?
(iii) Is the confidentiality of health data compromised?
(iv) Do resulting decisions share equal accuracy with all spatial data?
(v) If research progress in public health is hindered, how can researchers, decision makers and the public, enjoy a strong public health policy?

To answer the above research questions, a spatial tool that accepts unstructured data such as health data (as input) to produce geographic context-aware reference in spatial form (as output), for decision making and policy enactment, is proposed. The extracted spatial information shall be evaluated using three metrics namely, Precision, Recall and F-measure. The resulting locational concepts and attributes shall be linked to the GIS tool, and identified geographic locations ported to Google Maps.

2 Formal Theory/Concept

Within the medical domain, spatial ontology should integrate the principles of mereology (the study of parts and the wholes they form) and spatial reasoning [6]. Whereas mereology has been explored in various ways as applications of predicate logic to formal ontology, spatial reasoning on the other hand forms a central component of medical research and practice, and must be incorporated into any successful medical informatics programme. The spatial concepts most often utilized in this field are not the quantitative, point-based concepts of classical geometry, but rather qualitative relations among extended objects. Hence, this paper pursues formalism for qualitative spatial relations – patients/healthcare services and location relations. The proposed PeNLP is a location based parser that rests on specific application subcomponents of typical location-based service architecture [6]. Although, several location-based systems exist, none of these systems integrate ontology formalism, as most existing location-based systems are those created from environmental information, trained and mapped to an area of interest [7].

This paper gained insights and motivation from the inconsistency in data obtained for reasoning with the spatial qualification logic [8]. Spatial qualification problem is well-known in artificial intelligence (AI), and is concerned with the non-recognition of agent's presence at a specific location at a particular time as a qualification for carrying out an action or participate in an event, given its known location antecedents [9]. The implementation of SQL however depends greatly on spatial geographical information system (GIS) data, and requires expensive GIS software and device to access, hence, introducing greater reasoning problems for big datasets [8].

3 Methodology

The Context-based Preposition-enabled Spatial ONTology (PeSONT), is an on-going project that provides spatial ontology – a classified repository of spatial (locational and temporal) concepts – based on some given contexts (textual or multimedia), which are to a great extent unstructured. The components of the proposed PeNLP are shown as subset components of the context-based PeSONT framework (see Fig. 1).

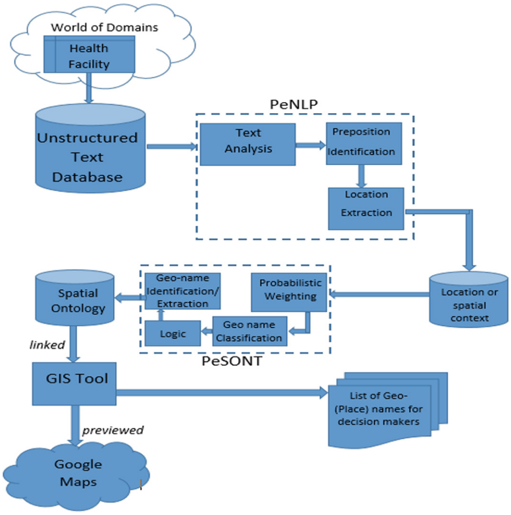


Fig. 1 The context-based PeSONT framework

The activity flow of the proposed PeNLP system includes the collection of unstructured public health data from several health facilities, be it textual or multimedia. The PeNLP process begins with text analysis, from where the relevant prepositions are identified, and the spatial contexts are extracted and storage in a spatial context repository. This repository then feeds the PeSONT component, which comprises a set of geographically-enabled tools for name identification and classification using formal representation logic. The extractions are finally stored in a spatial ontology database and linked to GIS tools for effective visualization and decision making purposes.

4 PeNLP Implementation and Evaluation

4.1 The PeNLP Algorithm

The PeNLP algorithm in Fig. 2 describes the human intelligent approach to placial noun identification. Using spatial reasoning, prepositions are first identified, before identifying words or phrases after the preposition. The algorithm splits the statements into sentences using a boundary marker (the period or full stop (.)), then the existence of verbs in each of the sentences are searched for (mostly the main verbs). In the absence of a main verb the auxiliary verb is retrieved before breaking the sentence into its subject and predicate. The algorithm then checks for prepositions in both the subject and the predicate. On finding any preposition, the word or phrase is extracted as the placial noun. This process is repeated for all sentences, and the placial nouns stored as spatial contexts in the repository. A formal representation of the part of speech (POS) and the classification of placial noun/spatial context based on Agarwal's structure is given using the following regular expressions:

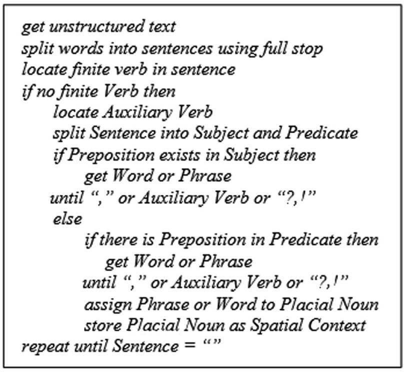
<Sentence> = <Subject><Verb><Predicate>

<Subject> = <Noun Phrase>

<Predicate> = <Noun Phrase>

<Noun Phrase>= <Article><Preposition><Noun>

<Noun Phrase> = <GeoLocation><Location Description><Generic PlaceNames> ....



```
get unstructured text
split words into sentences using full stop
locate finite verb in sentence
if no finite Verb then
    locate Auxiliary Verb
    split Sentence into Subject and Predicate
    if Preposition exists in Subject then
        get Word or Phrase
    until "," or Auxiliary Verb or "?,!"
    else
        if there is Preposition in Predicate then
            get Word or Phrase
        until "," or Auxiliary Verb or "?,!"
        assign Phrase or Word to Placial Noun
        store Placial Noun as Spatial Context
repeat until Sentence = ""
```

Fig. 2 PeNLP algorithm

The Agile software development lifecycle was followed during the development of the system. The process involved the building of the system prototype, testing, modification and re-building as need arises. A phase-wise approach was adopted for building the prototypes, with continuous update made to the existing prototype until the new system was obtained. The system programming tools used include the Java programming language, JXbrowser library, JavaScript, HTML, Cascading Style Sheet and Microsoft Access for spatial context database.

4.2 Evaluation Mechanism

The experimental plan features sample unstructured data collected directly from the temporary database, to accuracy. The unstructured data are textual data and the resulting geo-locations or placial nouns from the PeNLP is also textual but structured. The evaluation metrics we shall use to measure the correctness of the resulting spatial concept for the PeNLP and PeSONT include Precision and Recall, and are given in Eqs. (1) and (2) respectively [10]:

$$\text{Presision} = \frac{\text{correct} + 0.5 * \text{partial}}{\text{correct} + \text{spurious} + \text{partial}} \tag{1}$$

$$\text{Recall} = \frac{\text{correct} + 0.5 * \text{partial}}{\text{correct} + \text{missing} + \text{partial}} \tag{2}$$

The F-measure is used in conjunction with Precision and Recall, as a weighted average of the two. With the weight set to 0.5, both Precision and Recall are deemed equally important.

5 Conclusion

Empowering a community to collaboratively generate evidence that promotes better health decisions and better care is important and mission critical to improving health and healthcare services. This paper has proposed an ontology-based approach to formal characterization of unstructured data. Precise formal characterizations of all spatial relations assumed by PeNLP and PeSONT are necessary to ensure that the information embodied in the ontology can be fully and coherently utilized in a computational environment. The paper therefore serves as a springboard toward actualizing this goal, but more rigorous research along this line is required.

References

1. Gill, M.S., Ryan, P.B., Madigan, D.: Sparse coding for predictive modeling of observational health outcomes. In: Proceedings of OHDSI Symposium, Washington Hilton, pp. 1–2 (2016)
2. Crown, W.H.: Potential application of machine learning in health outcomes research and some statistical cautions. Value Health **18**(2), 137–140 (2015)
3. UN: United Nations general assembly draft outcome document of the United Nations summit for the adoption of the post-2015 development agenda. http://srsg.violenceagainstchildren.org/sites/default/files/documents/docs/A_69_L.85_EN.pdf. Accessed Sept 2017
4. Waller, L.A., Gotway, C.A.: Applied Spatial Statistics for Public Health Data. Wiley Series in Probability and Statistics. A Wiley Interscience. Wiley, Hoboken (2004)
5. Bennett, B., Agarwal, P.: Semantic categories underlying the meaning of 'place'. In: Winter, S., Duckham, M., Kulik, L., Kuipers, B. (eds.) COSIT 2007. LNCS, vol. 4736, pp. 78–95. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74788-8_6
6. Donnelly, M., Bittner, T., Rosse, C.: A formal theory for spatial representation and reasoning in biomedical ontologies. Artif. Intell. Med. **36**(1), 1–27 (2006)
7. Wang, C., Shi, Z., Wu, F.: Intelligent RFID indoor localization system using Gaussian filtering based extreme learning machine. Symmetry **9**(30), 1–16 (2017)
8. Bassey, P.C., Akinkunmi, B.O.: An Alibi Reasoner based on the Spatial Qualification Model. In: Proceedings of ISKO International Conference on Transition from Observation to Knowledge to Intelligence, France, pp. 261–270 (2014)
9. Bassey, P.C., Akinkunmi, B.O.: Introducing the spatial qualification problem and its qualitative model. Afr. J. Comput. ICTs **6**(1), 191–196 (2013)
10. Maynard, D., Peters, W., Li, Y.: metrics for evaluation of ontology-based information extraction. In: Proceedings of the EON Workshop (2006)