Expert Systems With Applications 95 (2018) 351–367

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Spatial Information Extraction from Short Messages

Sarah Zenasni[a,*], Eric Kergosien[c], Mathieu Roche[a,b], Maguelonne Teisseire[a]

[a] TETIS, APT, Cirad, CNRS, Irstea, Montpellier University, Montpellier, France
[b] Cirad, TETIS, Montpellier, France
[c] GERiiCO, Lille 3 University, Lille, France

ARTICLE INFO

Article history:
Received 18 July 2017
Revised 4 October 2017
Accepted 10 November 2017
Available online 16 November 2017

Keywords:
Spatial entities
Spatial relations
Similarity measures
Short message corpora
Text mining

ABSTRACT

Texts in addition to maps and satellite images, have become an important spatial data resource in recent years. Electronic written texts used in mediated interactions, especially short messages, have triggered the emergence of new ways of writing. Extracting information from such short messages, which represent a rich source of information, is highly important in order to discover domain-relevant information in the text and facilitate information retrieval. However, short messages are hard to analyse because of their brief, unstructured and informal nature. This paper focuses on the kinds of special or unique spatial entities and relations are contained in short messages. A new entity extraction method specifically dedicated to French short messages (SMS and tweets) is outlined to address this issue. The method is then tested on more traditional sources, like newspaper texts. This work is crucial in order to take advantage of the vast amount of geographical knowledge expressed in heterogeneous unstructured data. Firstly, we propose a process in which new spatial entities are extracted (e.g. *motpellier, montpelier, Montpel* are associated with *Montpellier*). Secondly, we identify new spatial relations that precede spatial entities (e.g. *sur, par*). Finally, we propose general patterns for the extraction of spatial relations. The task is very challenging and complex due to the specificity of short message language, which is based on weakly standardized modes of writing. The experiments were carried out on the three French corpora (i.e. 88milSMS, tweets, and Midi Libre) and highlight the efficiency of our proposal for identifying new kinds of spatial entities and relations.

1. Introduction

Much current research in text analysis and information retrieval fields is driven by the growing need to browse and scrutinize great quantities of information. However, many studies related to spatial information analysis conducted in recent years have focused on various types of large datasets (satellite images, GPS data and textual data).

Concerning textual data, many studies have described the particular way of expressing spatial information in natural language. Some of them define spatial information, expressed in textual documents, as a spatial entity (SE) and one or more spatial relation(s) (SR) (Lesbegueries, Sallaberry, & Gaio, 2006). The identification of such spatial information is important to improve the efficiency of information retrieval systems.

The SE has been extracted from various types of corpora such as newspaper articles, tweets, and many other types of textual data. Previous studies were focused mainly on the identification and extraction of simple SE, defined as toponymic names or place names (e.g. *Paris, Montpellier, gare Montparnasse*). However, more recent studies have been focused on the recognition of more complex SE integrating SR (e.g. *au nord de Paris, près de l'église Saint-Paul*). The aim of identifying these complex SE is to resolve issues in many areas, including reduction of ambiguity, geographic information retrieval, navigation, traffic management, spatial reasoning, query answering, etc.

The recent development of communication technologies has contributed to the emergence of new forms of written text. In particular, communication via short messages has become a social phenomenon. Millions of messages are exchanged daily to communicate, participate in contests, obtain information (place names, locations, themes, opinion, etc.), and countless other services. The various technological advances have made the creation, acquisition, storage, and sharing of digital documents simpler and easier. Extracting information from such corpora, which represent a rich source of information, is highly important in order to discover

* Corresponding author. Tel.: 0033638940317.
E-mail addresses: sarah.zenasni@teledetection.fr (S. Zenasni), eric.kergosien@univ-lille3.fr (E. Kergosien), mathieu.roche@cirad.fr (M. Roche), maguelonne.teisseire@irstea.fr (M. Teisseire).

https://doi.org/10.1016/j.eswa.2017.11.025
0957-4174/

domain-relevant information in the text and facilitate information retrieval. However, the multitude and variety of these corpora and the regular emergence of new formulations of messages and terms (new vocabulary, spelling mistakes, etc.) make it difficult to automatically recognize and extract this spatial information.

In fact, in addition to the volume of data to be processed, such corpora are characterized by the use of a multitude of different expressions to express the same named entity (NE), and especially the same SE (e.g. *motpellie, montpel* for *Montpellier*). This task is essential for analyzing and providing smart navigation methods. For example, users look for specific information but do not know what form they take. Societal issues linked to a better automatic understanding of short messages are important, especially to improve systems for detecting events (cultural events, natural disasters, etc.), emergency rescue systems (Monteiro, 2015) or to enhance the accuracy of system recommendations of places of interest (Rikitianskii, Harvey, & Crestani, 2014). Spatial information extraction from short messages, especially those in the French language, is a challenging problem since no annotated datasets for supervised learning are available.

In this global context, we propose a new method combining several Natural Language Processing (NLP) approaches, including statistical analysis (similarity measures), lexical analysis (presence or absence of accents (Figuerola, Rodríguez, & Berrocal, 2001; Kobus, Yvon, & Damnati, 2008; Savoy, 1999), similar prefixes (Youngja & J., 2001; Yu, Hripcsak, & Friedman, 2002)), grammatical analysis (part-of-speech (POS) tagging), and a text mining approach (n-grams of words) for the identification and extraction of simple and complex SE from large text corpora in SMS and tweets.

Indeed, basic similarity measures produce interesting results, but they are not adapted to short messages because of their lexical and syntactic specificities. Hence, we combine the similarity measure with the unaccented method and similar prefixes for SE extraction to provide a more robust approach. It is essential to benefit from the vast amount of geographical knowledge expressed in diverse natural language texts.

The remainder of the paper is structured as follows. Section 2 presents a brief introduction to existing spatial information extraction studies. In Section 3, the proposed approach is described. In Section 4, we outline the experimental analysis and results of the application to short message corpora (SMS and tweets). A discussion on the application to different types of text (short message corpus and standard corpus) is presented in Section 5. Finally, Section 6 presents the conclusion and prospects of our work.

2. Related work

Compared with research in standard corpora, very limited resources and NLP methods have been developed for processing short message corpora in French language texts. Sections 2.1 and 2.2 present related work on standard and short message corpora, respectively.

2.1. Spatial information extraction from texts

Several studies have described the particular ways of expressing spatial information in natural language. Some have focused on the identification of simple SE, also known as Absolute Spatial Entities (ASE), whereas complex SE integrating Spatial Relations (SR) are qualified as Relative Spatial Entities (RSE) (Lesbegueries et al., 2006). A first family of studies was focused on the identification of simple SE. More specifically, Named Entity Recognition (NER) is the process of identifying and classifying NEs in text, such as persons, locations[1], organizations. These methods can be categorized in three classes (Wu, Fan, Lee, & Yen, 2006): Rule-based NER, machine learning NER, and hybrid NER.

A) *Rule-based methods* focus on NE extraction using a set of rules (e.g. grammatical, syntactic, orthographic features) in combination with a list of dictionaries (e.g. list of countries, cities, etc.) that are manually pre-defined by experts (Mansouri, Affendey, & Mamat, 2008). Among the rule-based approaches, Wakao, Gaizauskas, and Wilks (1996) described an information extraction system in which four classes of entities (organization, person, location name, and time expression) are recognized and classified. They took advantage of graphological, syntactic, semantic knowledge, and discourse level information for proper name recognition and classification. The approach was tested on 30 unseen Wall Street Journal texts. The global precision and recall scores were 93% and 91%, respectively.
Stern and Sagot (2010) described NP (from the French Noms Propres, i.e. proper nouns) using a system for NER which works in two steps: one for detection and typing, and the other one for disambiguation and resolution of NE. A context-free grammar consisting of 130 rules was developed to detect and type NEs. Then disambiguation heuristics based on quantitative and qualitative information were applied to reduce ambiguity. The NP achieved an F-measure score of 77% on a manually annotated French newswire corpus from Agence France-Presse.
Alfred, Leong, On, and Anthony (2014) proposed a NER algorithm for Malay articles. First, the proposed method tokenizes sentences. The tokenized words were then evaluated using a POS tag dictionary containing 8700 words that were retrieved from the Thesaurus Bahasa Melayu. Finally, specific rules based on POS-tagging contexts were applied to detect three major types of NEs (person, organization, and location). The proposed method achieved an F-measure score of 89%.
These types of approach can obtain relevant results in specific domains, although they are often domain and language dependent (Mansouri et al., 2008).

B) *Machine learning methods* employ a classification model to identify NEs (supervised or unsupervised). Unsupervised learning approaches draw inferences from datasets consisting of unlabeled input data. In this type of approach in Kim, Kang, and Choi (2002), the authors proposed a model based on an unlabeled corpus and a training set automatically built from a small-scale NE dictionary. Firstly, the authors automatically built an NE tagged corpus. Secondly, they extracted syntactic relations from the training set and target documents. Thirdly, a classification is progressed by three different learners independently and those results were combined into one result. Finally, using a rule, the system predicted the category for yet to be labeled test examples. The experimental results showed 73% precision and recall for Korean news articles.
Lample, Ballesteros, Subramanian, Kawakami, and Dyer (2016) proposed a semi-supervised learning approach that combines character-based word representations learned from an annotated corpus and unsupervised word representations induced from unannotated corpora. They tested their approach on the CoNLL-2002[2] and CoNLL-2003[3] datasets for the English, Spanish, German, and Dutch languages.
Munro and Manning (2012) presented a new approach that jointly learns to identify NEs in parallel text. The proposed method: (1) generates seeds by calculating the edit likelihood deviation between candidate entities across both languages

[1] In our case, location is called Spatial Entity (SE).
[2] http://www.cnts.ua.ac.be/conll2002/ner/.
[3] http://www.cnts.ua.ac.be/conll2003/ner/.