Entity Recognition and Resolution in Semi-structured Data

Nuno Freire
IST / INESC-ID
Apartado 13069,
1000-029 Lisboa, Portugal
nuno.freire@ist.utl.pt

ABSTRACT

Potentially usable business information exists in unstructured form. This information, although machine readable, resides in unstructured human language texts that are difficult to process by computers. Within this information are references to real world entities, which are the focus of this paper. More specifically, we address the recognition of references to entities and their resolution, in the context of semi-structured data. This kind of data is structured according to a model which defines only generic semantics for its data elements, or includes data elements that contain natural language text. Semi-structured data presents new challenges and opportunities. In this kind of data, grammatical evidence is very often insufficient for entity recognition, since short sentences and simple expressions are predominant. However, the contextual information given by the structure of the data opens new possibilities for innovative techniques. We propose an approach to integrating the support for structured data throughout the complete process. It will be evaluated by studying three entity types in two scenarios with different semi-structured data formats, each one with distinct characteristics.

Categories and Subject Descriptors

E.1 [Data] Data Structures – Records; I.2.7 [Artificial Intelligence]: Natural Language Processing - Text analysis

General Terms

Algorithms. Documentation Experimentation.

Keywords

Entity recognition; Entity resolution; Semi-structured data.

1. INTRODUCTION

Potentially usable business information exists in unstructured form. This information, although machine readable, follows a data model with generic semantics, or may not follow any data model, that is, it consists in human language texts. Although its exact percentage cannot be determined, it is widely accepted that from 80% to 90% of business information may exist in unstructured forms [1].

As society becomes more data oriented, much interest arisen in the last decades for these unstructured sources of information. This interest gave origin to the research field of information extraction, which looks for automatic ways to create structured data from these unstructured data sources [2].

A particular scenario of information extraction deals with the references to entities (such as persons, organizations, places, etc.) that exist within unstructured data. This scenario deals with two particular problems: how to locate these references (entity recognition) and how to resolve the references to their real world entity (entity resolution). The typical output of this scenario is the exact location of references to entities within the data, and their resolution to a set of entities known to be disambiguated (such as geographic gazetteers, persons' databases, etc.).

This paper addresses the techniques for performing entity recognition and resolution on semi-structured data sources. This kind of data has different characteristics from those of unstructured texts, which have been the focus of previous research, and our preliminary experiences already identified that existing techniques underperform when applied to semi-structured data. We will address the problem by studying how references to entities exist in semi-structured data sets, and how exiting recognition and resolution techniques can be adapted to provide better results.

This document follows in Section 2, with a description of the problem. Section 3 presents related work and our approach to entity recognition and resolution in semi-structured data is provided in Section 4. The main scientific challenges are described in Section 5, and the plan for evaluation of our approach is presented in Section 6. The main contributions of this work are described in Section 7.

2. The problem

This section describes the problem to be approached in our work. It will describe and characterize semi-structured data, with particular focus on how references to entities typically appear. It also addresses the differences between this kind of data and textual data commonly addressed in information extraction research.

We consider semi-structured data to be any data that is encoded according to a model which contains data elements with at least one of the following characteristics:

- The data elements are modeled with general semantics. For example, a data element for the subject of a movie may contain many potential data attributes, such as a theme, a location, a time period, etc.
- The data elements are modeled to contain unstructured data, that is, natural language text (which can contain references to entities). For example, the title of a book, abstracts, or general notes elements.

JCDL'11, June 12–17, 2011, Ottawa, Ontario, Canada.

Data with these characteristics can be found everywhere: in organizations' primary data sources; in data on the Web; on the personal desktop; etc. With the rise of interoperability between systems and organizations, some data models are designed precisely to have general semantics. It is a way to enable the encoding of data originating from heterogeneous sources in a standard way. Examples of such cases are the models based on the general Dublin Core[1] terms for resource description, which are widely used in the Web for describing millions of resources.

Within these data elements, references to entities can often be found. However, due to the lack of structure, to be efficiently processed by computers, they need to be identified by means of information extraction techniques.

Previous research on information extraction has focused mainly in natural language processing. Information extraction processes are composed of subtasks, where each task simplifies the text by transforming it into more machine processable structures. For example, before entities can be recognized it may be necessary to identify the language of the text, to tokenized the text into paragraphs, sentences and words, and to classify the words for their part-of-speech category (that is, whether they are a verb, a noun, etc.). It is by reasoning on the output of this process that references to entities are recognized. This process is therefore dependent of evidence given by the natural language text to identify a reference to an entity, and also its type.

The resolution of the reference to a known entity is usually addressed as a separate problem to entity recognition. The techniques used are largely independent of the entity recognition task. Techniques used for resolution typically need to be adapted to the domain and data set being addressed.

Semi-structured data presents a new setting for entity recognition and resolution. It presents new challenges and opportunities for the application of information extraction. The possible lack of natural language text in semi-structured data is a challenge for entity recognition, but the contextual information given by structure of the data opens new possibilities for improvement.

3. Related work

This section describes the state of the art of two particular sub problems of information extraction: entity recognition, and entity resolution. Although entity resolution is sometimes seen as a sub problem of entity recognition, much research on the topic exists from other areas of computer science. For this reason both problems are addressed in this section separately.

3.1 Entity Recognition

Entity recognition is a subtask of information extraction, with the purpose to recognize information units such as names (for example, of persons, organizations) and numeric expressions [3]. Current state-of-the-art solutions can achieve near-human performance, effectively handling the ambiguous case, and achieving F-score accuracy around 90%. A recent survey of this area is available in [4].

Initial approaches, which are nonetheless still commonly used, were based on manually constructed finite state patterns and/or collections of entity names [4]. In general, the pattern-based approaches attempt to match against a sequence of words in much the same way as a general regular expression matcher. However, named entity recognition is considered as a typical scenario for the application of machine learning algorithms, because of the potential availability of many types of evidence, which form the input variables for the algorithms [5].

A major factor supporting the use of machine learning algorithms for entity recognition reasoning is their capacity to adapt to each case. Thus, they can be deployed with greater flexibility on distinct corpus from different domains, languages, etc. Different types of text analysis methods make available several types of evidence on which to base the named entity recognition reasoning. But not all evidences will be present in every corpus, and not all text analysis techniques will be able to identify the same types of evidence, so the capacity of machine learning algorithms to adapt to each case seams to make it a very good solution for entity recognition, which is supported by the rising trend in usage of machine leaning in this research area [4].

Two particular types of supervised machine learning algorithms have been successfully used for entity recognition. Early applications applied classification algorithms, which basically classify words, or groups of words, according to their entity type. Nevertheless, in entity recognition, as in other natural language related tasks, the problem was shown to be better solved with sequential classification algorithms. The earliest sequence classification techniques applied to entity recognition were Hidden Markov Models [5]. However this technique does not allow the learning algorithm to incorporate into the predictive model the wide range of evidence that is available for entity recognition. This limitation has lead to the application of other algorithms such as the Maximum Entropy Markov Model [5] and Conditional Random Fields [6]. Conditional Random Fields is currently the technique that provides the best results for entity recognition. It has sequence classification learning capabilities together with the flexibility to use all the types of evidences that entity recognition systems can gather.

3.2 Entity Resolution

Entity resolution is a problem common to many areas of computer science. Data created and processed in information systems often represents real world entities. The way these entities are represented in the data is specific to the context where it is manipulated. Typically, these representations are focused in fulfilling the requirements of the business processes that create and use them.

Entities are represented by a set of attributes that are specific to the context of the business. When these attributes do not comprise a unique identifier, the representation may be ambiguous. An entity may have multiple different representations, and each representation might match the description of multiple entities.

Entity resolution refers to the general problem of determining, within a data set, which representations of entities actually represent the same real world entity. It is a common problem to many research communities, although the term used is not always the same. Common terms used include record linkage, record matching, merge-purge, de-duplication, instance identification,

[1] http://dublincore.org/

database hardening, reference reconciliation, reference disambiguation, and object consolidation [7].

In the context of information extraction we can find the entity resolution problem as part of the entity recognition problem. The resolution of the entities appears as a final phase of the process, where the recognized entities are to be matched against a disambiguated data set, such as an ontology, relational data, etc. This particular form of entity resolution is typically called disambiguation [8], since the main problem is to choose the right entity, when more than one matches the name.

In its most basic form, the resolution of entities is performed by comparing the recognized name against the collection of names, where the names have to be compared. In more complex scenarios, where the resolution is to be done against a structured dataset, the inherent semantics of structured data can be explored to provide evidence for supporting the resolution. An example can be found in the resolution of place names against a geographic gazetteer, where the detailed relations between places can help to disambiguate [9]. Another example can be seen in the disambiguation of person names in a social network [10].

The final step of the resolution step is a reasoning process, where the evidence gathered about the recognized entity is compared with the resolution candidates. Similarly to entity recognition, the reasoning may be implemented as a set of manual rules, or by using machine learning techniques for classification [7]. Many machine learning techniques for classification have been used in entity resolution, and no technique seems to outperform all others across all data sets [7].

4. Proposed Approach

This section describes the main opportunities for improvement when applying existing entity recognition techniques on semi-structured data, and a proposal for an approach to address them.

In previous work, we observed that current entity recognition techniques underperform in semi-structured data. According to our analysis, we can point out the following three limitations:

- Semi-structured data may not contain enough grammatical evidence to support the recognition and resolution of entities.
- The language of the text within semi-structured data may be hard to determine, either because it is not clearly stated or automatic language guessing techniques have poor performance on very small texts.
- Pragmatic approaches consist in extracting the text from the structured data, and provide it as input to the system. This approach does not take advantage of the data structure, and any evidence that it can provide.

The structure of the data may provide relevant contextual evidence. Therefore it may compensate for the lack of grammatical evidence, or improve results in a similar way as previous work used corpus level evidence in entity recognition.

Entity resolution should also benefit from the structure of the data, particularly when it is done on data sets with rich semantics, such as those that nowadays exist as open linked data.

In order to exploit the structure of the data, it is proposed an approach that will integrate the support for structured data throughout the complete recognition and resolution process. Figure 1 describes in more detail the process of the traditional black box approach.

The first task consists in selecting the text within the record, where entity names are to be recognized and resolved. The rest of the recognition and resolution process is performed only on the text, ignoring any other information from the record.

The process follows with basic text processing to transform the text into a series of paragraphs, sentences and finally tokens. The text tokens are then analyzed and associated with any evidence that may support the recognition and resolution. Typically it includes grammatical analysis of sentences, analysis of words features (e.g. capitalization), and checking the existence of the tokens in collections of entity names.

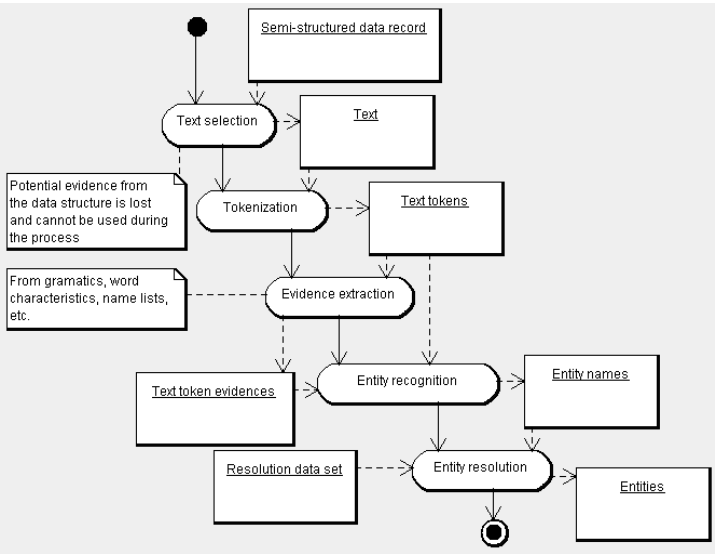

Figure 1 – The process of black box application of entity recognition and resolution to semi-structured data

Table 1 - Comparison of the black box and the proposed approaches

Activity	Black box approach	Proposed approach
Text selection	The text is selected from the relevant data fields, and passed forward. The context provided by the data structure is not available for the activities that follow.	The relevant data fields are selected and passed forward. The fields and respective records are available throughout the process.
Tokenization	Performed similarly in both approaches.	
Evidence extraction	Evidence is extracted from the text tokens, and the target resolution data set.	The data field itself provides extra evidence. Also extra evidence can be extracted from structured fields on the same record.
Entity recognition	Performed by machine learning, dictionary based, or rule based techniques.	Performed with the same techniques, but with more evidence available.
Entity resolution	Performed by machine learning or rule based techniques.	Performed with the same techniques, but with more evidence available.

The following task is the recognition of the entity names. It consists in reasoning on the sequence of tokens and their evidence. It results in the recognized names and their position in the source text.

Finally, a reasoning process resolves the recognized names, to specific entities described in the target resolution data set.

In our approach, the same general tasks are executed. However, they are executed with the record always available, from where further evidence may be used. We also propose that the choice of techniques should be appropriate for the characteristics of the text in these records. Table 1 highlights the main changes of executing the process according to the proposed approach.

In order to test our approach, its implementation will be carried out as a general framework for executing entity recognition and resolution on semi-structured data. This framework will be designed to allow the comparison of results from several state of the art techniques, across different domains and data formats.

5. Scientific Challenges

Our work addresses the general scientific challenge of matching unstructured against structured data, in order to recognize and link similar entities. To achieve the general goal, some specific challenges are faced by the tasks of entity recognition and resolution.

The main challenges faced by entity recognition are:

- How to perform entity recognition with limited grammatical evidence. What other sources of evidence can be used.
- How to cope with language uncertainty of the text.
- How to generate machine-learned predictive models.
- How to achieve a flexible system, adaptable to requirements for either high precision or high recall.

The main challenges faced by entity resolution are:

- How to perform the resolution of simple references to entities. What other sources of evidence can be used?
- How to cope with computational performance.
- How to generate machine-learned predictive models.

6. Evaluation

The evaluation of our approach will be carried out by studying the recognition and resolution of three entity types in two scenarios, which have semi-structured data formats with different characteristics. The entity types to be studied are the *geographical place*, the *person*, and the *corporate body*.

The target data set for resolution of the *geographical places* will be Geonames[2]. This geographical dataset covers all countries and contains over eight million place names.

The target data sets for resolution of *persons* and *corporate bodies* will be the *Virtual International Authority File*[3], a joint project of several national libraries. It is a consolidated data set containing data gathered for many years about the authors of the bibliographic resources.

For both scenarios, a sample of the data sets will be manually annotated, with the position of the entities and their resolution. A description of both scenarios is provided next.

6.1 Bibliographic Data

Our first scenario will address the recognition and resolution of entities in the data sets from Europeana[4], with descriptions of digital objects of cultural interest. This dataset follows a data model using mainly Dublin Core attributes, structured in a schema named Europeana Semantic Elements[5]. It contains data elements which have general semantics. In this data set, entity names appear in data fields for titles, textual descriptions, tables of contents, and subjects.

The dataset contains records originating from several European institutions from the cultural sector. It therefore, contains very

[2] http://www.geonames.org/

[3] http://viaf.org

[4] http://www.europeana.eu/

[5] http://group.europeana.eu/web/guest/technical-requirements/

heterogeneous data. Several European languages are present, even within the description of the same object. Institutions, from where this data originates, follow different practices for describing the digital objects, which also contributes for having highly heterogeneous data. Grammatical evidence will be very limited in this data set, so it will provide a good scenario for the evaluation of the evidence made available by the structure of the data. We believe this is an example of the emerging general scenario of heterogeneous information systems interoperability, either internally to a specific company or in collaborative business networks (such as supply chains, etc.).

6.2 On-line Semi-structured Documents

Wikipedia[6] is a web-based, collaborative, multilingual encyclopaedia project, containing over 16 million articles. A part of Wikipedia's contents is openly available as semi-structured data, in the context of another community effort, DBpedia[7]. This initiative aims to extract structured information from Wikipedia and to make it available on the Web. This dataset allows the usage of data from Wikipedia in flexible ways for a variety of purposes, including research for several topics of computer science, including entity recognition and resolution.

For the evaluation, we will use DBpedia as the semi-structured data set where the recognition and resolution of references to entities will be performed. This dataset is centred on the articles of Wikipedia, and therefore they can describe any concept.

DBpedia's data is mainly structured. Associated with the articles we can find categories, languages, geographical coordinates and links to other articles. Our primary focus will be the unstructured text within the abstracts. These will consist in well-formed texts, so evidence will be available from linguistic analysis. This data set will allows us to measure the contribution of evidence given by the structured data to improve the quality of entity recognition and resolution on well-formed text.

7. Expected Results

The main contribution of this work will be a methodology for applying entity recognition and resolution in semi-structured data, but during the course to the main goal other contributions to some more specific research topics will be achieved.

We expect to achieve improved entity recognition results on unstructured data and without well-formed text, which therefore lacks grammatical evidence for recognition of entities. But our work also will study and evaluate the use of other types of evidence. For example, this work may also have impact on other areas, such as the entity recognition on damaged texts resulting from optical character recognition.

By studying entity recognition together with entity resolution we also expect to be able to use evidence in the recognition phase, normally used only for entity resolution, therefore, enabling the recognition process on unstructured data without well-formed text, and improving the results on well-formed text.

Currently, no manually annotated data sets are known for research in entity recognition and resolution on semi-structured data. Two manually annotated data sets will be created, and will be a contribution for future research.

[6] http://www.wikipedia.org

[7] http://dbpedia.org

8. REFERENCES

[1] Seth, G., "Unstructured Data and the 80 Percent Rule", Clarabridge Bridgepoints, 3rd quarter 2008.

[2] S. Sarawagi, "Information Extraction," Found. Trends databases, 2008.

[3] R. Grishman, B. Sundheim, "Message Understanding Conference - 6: A Brief History", International Conference on Computational Linguistics, 1996.

[4] D. Nadeau, S. Sekine, "A survey of named entity recognition and classification". Linguisticae Investigationes, 2007.

[5] A. McCallum, D. Freitag, and F. C. N. Pereira, "Maximum Entropy Markov Models for Information Extraction and Segmentation," 17th International Conference on Machine Learning, 2000.

[6] J.D. Lafferty, A. McCallum, and F.C.N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labelling Sequence Data," 18th International Conference on Machine Learning, 2001.

[7] A. Elmagarmid, P. Ipeirotis, and V. Verykios, "Duplicate Record Detection: A Survey," Knowledge and Data Engineering, IEEE Transactions, 2007.

[8] E. Amitay, N. Har'El, R. Sivan, A. Soffer, "Web-a-where geotagging web content" Conference on research and development in information retrieval, 2004.

[9] B. Martins, H. Manguinhas, and J. Borbinha, "Extracting and Exploring the Geo-Temporal Semantics of Textual Resources," International Conference on Semantic Computing, IEEE, 2008.

[10] R. Hölzer, B. Malin, and L. Sweeney, "Email alias detection using social network analysis," 3rd international workshop on Link discovery, ACM, 2005.