Adopting Ontologies for Multisource Identity Resolution

Milena Yankova
Department of Computer,
Science University of Sheffield
Regent Court, 211 Portobello
Street
United Kingdom
milena@dcs.shef.ac.uk

Horacio Saggion
Department of Computer
Science, University of
Sheffield
Regent Court, 211 Portobello
Street
United Kingdom
saggion@dcs.shef.ac.uk

Hamish Cunningham
Department of Computer
Science, University of
Sheffield
Regent Court, 211 Portobello
Street
United Kingdom
hamish@dcs.shef.ac.uk

ABSTRACT

Identity resolution aims at identifying the newly presented facts and linking them to their previous mentions. Our main hypothesis is that variations of one and the same fact can be recognised, duplications removed and their aggregation actually increases the correctness of fact extraction. Our approach to the identity problem has been implemented as Identity Resolution Framework (IdRF). The framework provides a general solution identifying known and new facts in specific domains, and it can be used in different applications for processing of different types of entity. It uses an ontology for internal and resulting knowledge representational formalism. The ontology not only contains the representation of the domain, but also known entities and properties. Apart from extracting information from textual sources, we also exploit structured information available in databases mapping the database schema to the ontology and populating the ontology with existing knowledge. Our main goal is not to advocate one criterion among the others, but to introduce widely applicable solution of the identity resolution problem, we present a set of customisable criteria as well as a mechanism new criteria to be added. We have carried two series of experiments in two different business intelligence domains - company profiling and recruitment - achieving rather encouraging result.

1. INTRODUCTION

The question of consolidating redundant information arises when one uses multiple sources to obtain information. Formally the consolidation procedure aims at identifying the newly presented facts and linking them to their previous mentions. Once the information coming from different sources (textual documents, web pages, database records, ontologies or other knowledge representations) is interlinked and the noise of redundancy and incompleteness is removed, it can be then easily queried and used for various tasks. This problem is also known as “object consolidation” or “record linkage” in the Database community, or “cross document coreference resolution” in NLP community, “ontology population” in Semantic Web community, etc.

Recognising identical or similar information across different sources is of paramount importance and in particular can lead to improved extraction performance from single sources [25]. The main goal of such applications is to extract new information from a stream of documents. Thus, the decision about which information is new and which has been already extracted is crucial for the application success. Aggregation of extracted information has many advantages such as: complementing partial information from one source, increasing extraction confidence. It also enables further aggregation of details about the known objects, so each of the identified mentions may provide additional information to the initial object description.

Our main hypothesis is that variations of one and the same fact can be recognised, duplications removed and their aggregation actually increases the correctness of fact extraction. Fact variations, presented in different ways, will improve the ability of a system to recognise at least one mention of the fact. Once aggregated, the information can be easier analysed and searched, retrieving more accurate and relevant result. For example, a multisource information extraction system in the domain of corporative profiling will extract information composing company profiles from different web sources; this may include identification of the company names, head-office address and phone, sector and industry, executive team, number of employees, profit and revenue of the companies as well as earnings per shares. The system will have to identify all these details and compile a single profile, although not all of them can be easily found presented by a single source. In addition to syntactic differences on how information may be expressed (e.g. the names of the company “Marks & Spencer Group Plc” has been also seen as “M&S”, “Marks and Sparks” and “Marks’ ”) , one source may report information which is missing in the other, or one source may report information which is an update of information reported in the other source. Bringing the company called “Marks and Spencer” into focus (see Fig 1) beside complete contact details, its web site[1] lists short executive biographies, however financial information is missing. On the other site Yahoo!Finance[2] provides the

[1] http://www.marksandspencer.com
[2] http://uk.finance.yahoo.com

missing figures, although company contacts and executive's list is very brief. Other sources of information with different complexity of details also exist (e.g. Wikipedia[3] etc.)

This work is focused on providing a general solution to the identity problem and recognising different mentions of one and the same fact coming from different sources. We are interested in the problem of consolidation or merging of information found in different sources by ontology-based information extraction (OBIE) – the process of identifying in text concepts, instances, and relations expressed in an ontology. Traditionally, information extraction (see [12]) only classifies information chunks as belonging to pre-defined types, in an OBIE system, identity resolution aims at establishing a *reference link* between an object residing in the system's knowledge base and its mention in context (e.g. text) or asserting the new instance if not present in the knowledge base. Hence our solution uses ontologies as internal knowledge representation formalism, the benefits of which are described later on this paper.

Various approaches have been applied in the past to merging information from various sources [25, 23], however our approach is unique because it provides a general and adaptable framework to address the problem. Because we are interested in multi-source extraction, apart from extracting information from textual sources, we exploit structured information available in databases mapping the database schema to the ontology and populating the ontology with existing knowledge. Our main goal is not to advocate one criterion among the others, but to introduce widely applicable solution of the identity resolution problem, we present a set of customisable criteria as well as a mechanism new criteria to be added. The final decision about the identity may also depend on the domain and entity type; therefore it is presented as a semi automatically tuneable machine. As it will be shown later on, our results of the application of the framework to two different but related problems are very promising.

The work to be described here has been carried out in the context of the business intelligence (BI) Musing Project. The Musing projects is developing tools and modules based on Natural Language Processing (NLP) technology and reasoning to mitigate the efforts involved in gathering, merging, and analysing/annotating multisource information for BI applications. Here multisource information plays an important role since it is unlikely that a single (textual) source will contain all up-to-date information required by our target applications.

2. KNOWLEDGE REPRESENTATION VIA ONTOLOGIES

One of the applications we are working with is a system that keeps information on companies updated for competitive intelligence. Such information is available in many different sources such as web pages, financial news, however it is in non-structured textual form. It is usually the task of a human analyst to search in thousands of documents information on companies in order to insert the information in a data base or knowledge repository.

In our application we use information extraction techniques to transform unstructured and semi-structured documents into structured representations. After extraction, the information is further processed populating the ontology knowledge base. Ontologies are widespread and they are already successfully used for IE in systems like SemTag [6] , h-TechSight [20] and KIM [22] as in other areas (e.g. in biology and chemistry). Several ontology description languages have already been standardised.

The ontologies are already used for approaching the identity resolution problem. [9] present the advantages of semantically enhanced annotation for resolving co-references from different sources. Another example of using ontologies in this domain is the innovative work of [15] for extending standartised ontology description languages to enable approximation of instances. The authors introduce new "Rough Description Language" to represent and reason about similarity of instances.

In this we also adopt ontology as our main knowledge representation formalism. It has been chosen because of its detailed entity description that is complemented with semantic information. The expected benefit from using semantic representation is the opportunity to associate general type objects with concrete instances of ontology classes; in this way we will be able to recognise not only the type of the objects, or the class they belong to, but also the individual instances they refers to. For example, different appearances of "M&S" on different sources (e.g. different web pages) are extracted and collected as a single instance which all mentions point to. Such a semantic linkup of the identified objects guaranties more detailed description as opposed to a simple syntactic representation. In this way it provides more details, which serving as evidence can improve the accuracy of object comparison.

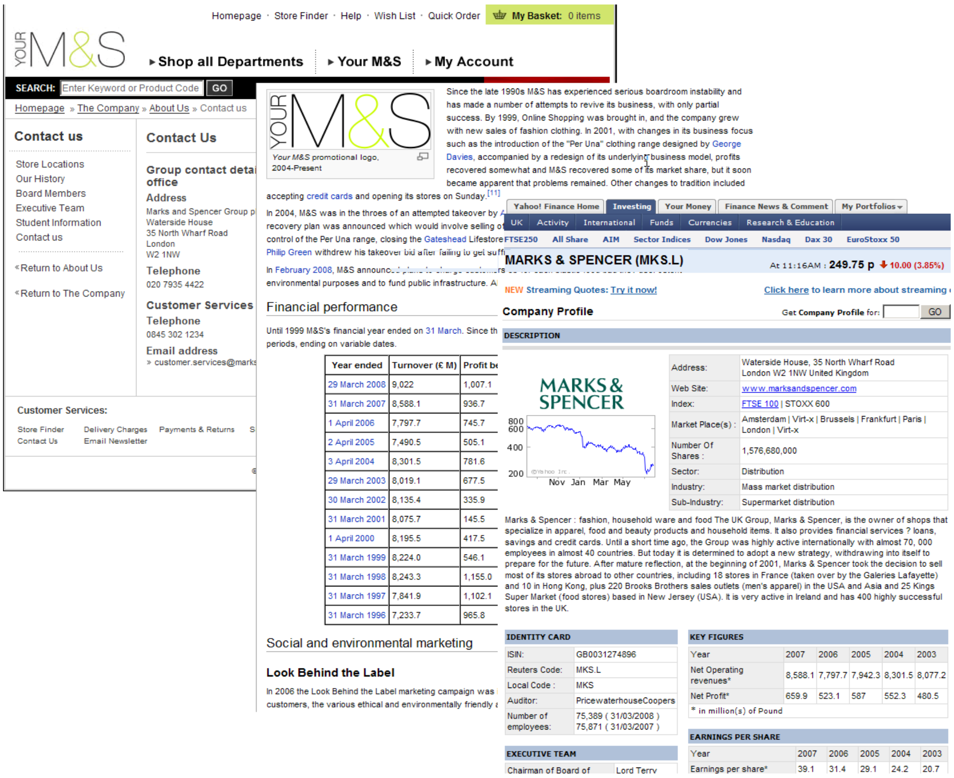
The ontological representation also provides a standard mechanism for introducing relations between concepts that can be used for further comparison enrichment Ontologies contain concepts arranged in class/sub-class hierarchies (e.g. a *company* is a type of *organization*), relations between concepts (e.g., a *company* has employees), and properties (e.g., a company has one (and only one) foundation date). The concepts targeted by this application are the company name, its main activities, its number of employees, its board of directors, turnover, etc. In order to represent the target concepts we are working with *domain ontologies* which represent the domain of application and which capture the experts' knowledge and allow us to encode detailed entity description.

Our domain ontology is based on PROTON [Terziev et al, 05] ontology, which is designed to be easy extendable for different domains or specific tasks. The knowledge base that actually contains the ontology and the instances associated with it is stored in the semantic repository provided by KIM [Popov et al, 03] that is based on OWLIM [Kiryakov et al, 05] and Sesame .

2.1 Mapping Databases to Ontologies

Many databases already contain information relevant for our application domain. The database schema is the data description that holds the meaning of the data, although the

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



Homepage · Store Finder · Help · Wish List · Quick Order
My Basket: 0 items
YOUR M&S
▸ Shop all Departments
▸ Your M&S
▸ My Account
SEARCH: Enter Keyword or Product Code GO
Homepage » The Company » About Us » Contact us

Contact us

Store Locations
Our History
Board Members
Executive Team
Student Information
Contact us

«Return to About Us

«Return to The Company

Contact Us

Group contact detai office

Address

Marks and Spencer Group p
Waterside House
35 North Wharf Road
London
W2 1NW

Telephone

020 7935 4422

Customer Services

Telephone

0845 302 1234

Email address

» customer.services@marks

Customer Services:

Store Finder · Delivery Charges · Payments & Returns · S
Contact Us · Email Newsletter

Your M&S promotional logo, 2004-Present

Since the late 1990s M&S has experienced serious boardroom instability and has made a number of attempts to revive its business, with only partial success. By 1999, Online Shopping was brought in, and the company grew with new sales of fashion clothing. In 2001, with changes in its business focus such as the introduction of the "Per Una" clothing range designed by George Davies, accompanied by a redesign of its underlying business model, profits recovered somewhat and M&S recovered some of its market share, but it soon became apparent that problems remained. Other changes to tradition included accepting credit cards and opening its stores on Sunday.[11]

In 2004, M&S was in the throes of an attempted takeover by A recovery plan was announced which would involve selling o control of the Per Una range, closing the Gateshead Lifestore Philip Green withdrew his takeover bid after failing to get suffi

In February 2008, M&S announc environmental purposes and to fund public infrastructure. A

Financial performance

Until 1999 M&S's financial year ended on 31 March. Since th periods, ending on variable dates.

Year ended	Turnover (£ M)	Profit be
29 March 2008	9,022	1,007.1
31 March 2007	8,588.1	936.7
1 April 2006	7,797.7	745.7
2 April 2005	7,490.5	505.1
3 April 2004	8,301.5	781.6
29 March 2003	8,019.1	677.5
30 March 2002	8,135.4	335.9
31 March 2001	8,075.7	145.5
1 April 2000	8,195.5	417.5
31 March 1999	8,224.0	546.1
31 March 1998	8,243.3	1,155.0
31 March 1997	7,841.9	1,102.1
31 March 1996	7,233.7	965.8

Social and environmental marketing

Look Behind the Label

In 2006 the Look Behind the Label marketing campaign was customers, the various ethical and environmentally friendly

Yahoo! Finance Home · Investing · Your Money · Finance News & Comment · My Portfolios
UK · Activity · International · Funds · Currencies · Research & Education
FTSE250 · All Share · AIM · Sector Indices · Dow Jones · Nasdaq · Dax 30 · EuroStoxx 50

MARKS & SPENCER (MKS.L) At 11:16AM : **249.75 p** 10.00 (3.85%)

NEW Streaming Quotes: Try it now! Click here to learn more about streaming

Company Profile Get Company Profile for: GO

DESCRIPTION

Address:	Waterside House, 35 North Wharf Road London W2 1NW United Kingdom
Web Site:	www.marksandspencer.com
Index:	FTSE 100 | STOXX 600
Market Place(s) :	Amsterdam | Virt-x | Brussels | Frankfurt | Paris | London | Virt-x
Number Of Shares :	1,576,680,000
Sector:	Distribution
Industry:	Mass market distribution
Sub-Industry:	Supermarket distribution

Marks & Spencer : fashion, household ware and food The UK Group, Marks & Spencer, is the owner of shops that specialize in apparel, food and beauty products and household items. It also provides financial services ? loans, savings and credit cards. Until a short time ago, the Group was highly active internationally with almost 70, 000 employees in almost 40 countries. But today it is determined to adopt a new strategy, withdrawing into itself to prepare for the future. After mature reflection, at the beginning of 2001, Marks & Spencer took the decision to sell most of its stores abroad to other countries, including 18 stores in France (taken over by the Galeries Lafayette) and 10 in Hong Kong, plus 220 Brooks Brothers sales outlets (men's apparel) in the USA and Asia and 25 Kings Super Market (food stores) based in New Jersey (USA). It is very active in Ireland and has 400 highly successful stores in the UK.

IDENTITY CARD

ISIN:	GB0031274896
Reuters Code:	MKS.L
Local Code :	MKS
Auditor:	PricewaterhouseCoopers
Number of employees:	75,389 (31/03/2008) 75,871 (31/03/2007)

KEY FIGURES

Year	2007	2006	2005	2004	2003
Net Operating revenues*	8,588.1	7,797.7	7,942.3	8,301.5	8,077.2
Net Profit*	659.9	523.1	587	552.3	480.5

* in million(s) of Pound

EXECUTIVE TEAM

Chairman of Board of	Lord Terry

EARNINGS PER SHARE

Year	2007	2006	2005	2004	2003
Earnings per share*	39.1	31.4	29.1	24.2	20.7

Figure 1: Example of a company profile across different sources

relations and the semantics of table elements are limited and often hard to interpret. This limitation also affects identity resolution process and prevents the system from correct data interpretation. As a consequence, binging databases to other knowledge representational formalism e.g. ontologies requires deep understanding and domain expertise and is usually done manually producing mapping between the particular database schema and given ontology. In this work, we use company profiles stored in a MySql Relational Database Management System which has been manually mapped to the Musing ontology using scripts. Examples of the record fields in the data base are: organization, section, url, name, address, etc. (see Figure 2). The scripts shown on Figure 3 map for example a record field such as "organization name" into the attribute "hasAlias" in our domain ontology.

Once the mapping between the database and the ontology is produced, the records of interest are easily transformed to formal entity descriptions with respect to the ontology. Before inserting a new record in the database, the identity resolution process is called upon to find the referent instance: if the reference is found, only the new details are inserted, if there is no existing instance to refer to, a new one is created. In this way, the resulting knowledge base (formally stored in a database) is a reference point to all different mentions of known entities.

```
+--------------+        +--------------------+
| ml_orgs      |        | ml_org_details     |
+--------------+        +--------------------+
| org_id       |        | org_id             |
| type_id      |        | office_employeed   |
| sect_id      |        | total_employees    |
| url          |        | empband            |
| parent_id    |        | empind             |
| name         |        | activity_desc      |
+--------------+        | num_branches       |
                        | profit             |
+--------------+        | sme                |
| ml_address   |        | turnover           |
+--------------+        | turngrade          |
| org_id       |        | turnmod            |
| address      |        | yearest            |
| postal       |        +--------------------+
| phone1       |
| phone2       |
| fax          |
+--------------+
```

Figure 2: A sample of the RDBMS schema realated to company profiling

2.2 Ontology-based Information Extraction

Information extraction is the process of extracting from text specific facts in a given target domain [13]. For example, in extracting information about companies key elements to be

```
_columnToURI.putForward("ml_orgs.org_id",
                               "musing:hasOrgId");
_columnToURI.putForward("ml_orgs.type_id",
                               "musing:hasType");
_columnToURI.putForward("ml_orgs.name",
                               "protonu:hasAlias"
_columnToURI.putForward("ml_orgs.sect_id",
                               "musing:hasSector");
_columnToURI.putForward("ml_orgs.url",
                               "musing:hasWebsite");
_columnToURI.putForward("ml_orgs.parent_id",
                               "musing:hasParentID");
_columnToURI.putForward("ml_address.postal",
                               "musing:hasPostal");
_columnToURI.putForward("ml_address.phone",
                               "musing:hasPhoneNumber");
_columnToURI.putForward("ml_address.phone2",
                               "musing:hasPhoneNumber");
_columnToURI.putForward("ml_address.fax",
                               "musing:hasFaxNumber");
_columnToURI.putForward("ml_address.address",
                               "musing:hasAddress");
_columnToURI.putForward("ml_org_details.total_employees",
                               "musing:hasNumEmp");
_columnToURI.putForward("ml_org_details.turnover",
                               "musing:hasTurnover");
```

Figure 3: Mapping between RDBMS and an Ontology for Company Information

extracted are the company address, contact phone, fax numbers, and e-mail address, products and services, members of the board of directors and so on. The info ration to be extracted is pre-specified and the system is tailored to extract those specific elements. The field of information extraction has been fuelled by two major US international evaluations efforts, from 1987 until 1997 the Message Understanding Conferences [17, 5] and since 2000 the Automatic Content Extraction Evaluation.

Tasks carried out during information extraction are named entity recognition, which consists on the identification and classification of different types of names in text; coreference resolution, which is the task of deciding if two linguistic expressions refer to the same entity in the discourse; semantic role recognition, which deals with the recognition of semantic roles to sentence constituents (e.g. agent, goal); and relation extraction, which identifies relations between entities in text. Information extraction usually employs the following natural language processing technologies: parts-of-speech taggers, morphological analyser, named entity recognisers, full (or partial) parsing, and semantic interpretation including nominal and verb coreference. These linguistic processors are generally available although some may require domain adaptation, for example while a parts-of-speech tagger for English could be used without major need for adaptation, a named entity recogniser should usually need adaptation to a new application domain.

There are two main approaches to the development of IE systems: (i) Hand-crafted systems which rely on language engineers to design lexicons and rules for extraction, and (ii) machine learning systems which can be trained to perform one or more of the IE tasks. Learning systems are given either an annotated corpus for training or a corpus of relevant and irrelevant documents together with only a few annotated examples of the extraction task, in this case some non-supervised techniques such as clustering can also be applied.

Rule-based systems can be based on gazetteer lists - lists of keywords which can be used to identify known names (e.g. New York) or give contextual information for recognition of complex names (e.g. Corporation is a common postfix for a company name).- and cascades of finite state transducers which implement pattern matching algorithms over linguistic annotations (produced by various linguistic processors) [22].

In our work we are mainly interested in Ontology-based information extraction which aims at identifying in text concepts and instances from an underlying domain model specified in an ontology. Our ontology-based information extraction system has been developed with the GATE platform which provides a set of tools for development of information extraction applications. In particular GATE provides support to work with ontologies.

The extraction prototype uses some default linguistic processors from GATE, however the core of the system, the concept identification program was developed specifically for this application. In addition to specific processes such as phrase chunking, lexicons and gazetteer lists have been created to perform gazetteer lookup processes. Rules for concept identification have been specified in regular grammars implemented in the JAPE language. A key element in the annotations created by the system is the encoding of ontological information - our applications create *Mention* annotations which make reference to the target ontology as well as the ontological concept a string of text refers to.

As an example of extraction rule, Figure 4 shows rules for the recognition of products (similar rules were developed for the recognition of services, etc). The rule for the identification of information in company profiles is implemented in the JAPE language. Jape rules have a left-hand side and a right-hand side - the left hand side of the rule is a regular pattern over linguistic information; the right hand side of the rule is Java code which creates an annotation enriched with feature/values. The rule in the Figure uses syntactic information - such as noun phrases (NP) identified by a process of rule-based chunking (e.g. a determinant followed by a sequence of nouns is a noun phrase) as well as lexical information (e.g. verbs used to express product information such as produce, manufacture, make). The left-hand side of the rule is a regular pattern over annotations produced by GATE, the right-hand side in this case is a Java program which creates for each noun phrase in the context a *Mention* annotation tied to the domain ontology. Other, much simpler rules not shown here are used to exploit the regularity of semi-structured documents, i.e., specific markup signaling the presence of a concept (e.g. “Telephone: +1-333-999-2222”).

The result of the automatic annotation is further analysed by (i) a module which produces RDF triples associating different pieces of information together (e.g. a company with its number of employees, a company with its CEO) – see Fig-

```
Rule: MainProduct1
// manufactures X equipment
(
 {Lookup.majorType == produce}
 (KIND)?
) (
  ({NP}|(LIST))
 ({Lookup.majorType == equipment})?
):mention --> {


//get the mention annotations in a list
List annList = new
ArrayList((AnnotationSet)bindings.get("mention"));
//sort the list by offset
Collections.sort(annList, new OffsetComparator());
//iterate through the matched annotations
for(int i = 0; i < annList.size(); i++)
   {
    Annotation anAnn = (Annotation)annList.get(i);
    if (anAnn.getType().equals("NP"))
   {
      // add features and values to annotaction:
      // link to the ontology
      FeatureMap features = Factory.newFeatureMap();
      features.put("class", "Product");
      features.put("note", "Main_Product");
      features.put("xbrl_id", "genInfo.company");
      features.put("rule", "MainProduct1");
      // create the annotation
      annotations.add(anAnn.getStartNode(),
                   anAnn.getEndNode(), "Mention",
      features);
    }
  }
}
```

Figure 4: A JAPE rule for identification of information in company profiles

```
<?xml version="1.0" encoding="UTF-8" ?> - <root
xmlns:indicator="http://musing.deri.at/ontologies/v0.5/int/indicator#"
xmlns:time="http://musing.deri.at/ontologies/v0.5/general/time#"
xmlns:protonu="http://musing.deri.at/ontologies/v0.6/protonu/protonu#"
xmlns:protont="http://musing.deri.at/ontologies/v0.6/protont/protont#"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">



- <entry id="company+address">
  <feature name="company_name" />
  <feature name="address_property" />
  <feature name="content_text" />
  <generated name="address_id" />
  <generated name="company_id" />
- <rdf>
    - <protonu:Address rdf:ID="${address_id}">
        <protont:description rdf:datatype=
          "http://www.w3.org/2001/XMLSchema#string">$
            {content_text}</protont:description>
     </protonu:Address>
    - <protonu:Company rdf:ID="${company_id}">
        <protonu:hasAlias rdf:datatype=
          "http://www.w3.org/2001/XMLSchema#string">$
           {company_name}</protonu:hasAlias>
      <protonu:hasAddress rdf:resource="${address_id}" />
        </protonu:Company>
  </rdf>
 </entry>

- <entry id="company+executive">
  <feature name="company_name" />
  <feature name="employee_property" />
  <feature name="job_title" />
  <feature name="name" />
  <generated name="person_id" />
  <generated name="job_id" />
  <generated name="company_id" />
- <rdf> - <protont:Person rdf:ID="${person_id}">
    <protonu:hasAlias rdf:datatype=
        "http://www.w3.org/2001/XMLSchema#string">$
        {name}</protonu:hasAlias>
    <protonu:hasPosition rdf:resource="${job_id}" />
    </protont:Person>
  - <protont:JobPosition rdf:ID="${job_id}">
      <protont:description rdf:datatype=
        "http://www.w3.org/2001/XMLSchema#string">$
        {job_title}</protont:description>
      <protont:holder rdf:resource="${person_id}" />
      <protont:withinOrganization rdf:resource="${company_id}" />
    </protont:JobPosition>
  - <protonu:Company rdf:ID="${company_id}">
      <protonu:hasAlias rdf:datatype=
        "http://www.w3.org/2001/XMLSchema#string">$
        {company_name}</protonu:hasAlias>
      <protonu:hasEmployee rdf:resource="${person_id}" />
    </protonu:Company>
  </rdf>
  </entry>

</root>

<!--  RDFTemplate, generated at 20080205-111054 from file
company-templates.xml template company+executive  -->
```

Figure 5: RDF templates produced from the analysis of a company profile. It includes a company instance, a person instance, and a job position instance all related according to what is specified in the Musing ontology

ure 5, and (ii) the ontology population module responsible for knowledge base population. An evaluation of the performance of the extraction system indicates good results with over 84% F-score [18]. Evaluation of ontology population (e.g., resolution and merging) is presented in the following sections.

3. USAGE OF ONTOLOGIES IN IDRF

Our approach to the identity problem has been implemented as Identity Resolution Framework (IdRF). The framework provides a general solution identifying known and new facts in specific domains, and it can be used in different applications for processing of different types of entity. The IdRF uses an ontology for internal and resulting knowledge representational formalism. The ontology not only contains the representation of the domain, but also known entities and properties. Therefore the expected input to IdRF is an entity and its associated properties and values – as specified in the ontology, the output is an integrated representation of the entity which will have new properties and values.

IdRF is based on the PROTON [26] ontology, which has been extended for our particular domain of company profiling following specific task carried for the MUSING project. The knowledge base that actually contains the ontology and the instances associated with it is stored in the semantic repository provided by KIM [22] that is based on OWLIM [14] and Sesame.

Each class the extended ontology is described by a set of customisable identity criteria. So formed class models, described in details later in this section, are used by the framework to measure the similarity between two instances of a given class. These criteria use ontological operations and similarity computation between extracted and stored values. The weighting criteria are specified according to the type of entity and the application domain.

The framework processes the entities in consequence. Thus, after identity resolution of an entity, the ontology Knowledge Base (KB) is updated and will contain all entities processed so far with their full semantic description aggregated during the resolution process. The effect of this continuous updating of the KB is that the identity criteria are constantly refined, thus improving the identity resolution by both refining the evidence calculation and introducing new entities serving as identity goals. Details about the two effects are given below:

- The evidence calculation is refined when a new value of an attribute, property or relation is added to an existing instance description. Then, the identity criteria for this instance is changed in order to reflect the newly available data adding new comparison restrictions. For example if the person age is added to his/her description, the age restriction will be added a new identity criterion.
- New entities added to the knowledge base represent potentially new goals for resolution. They are created by insertion of entirely new objects to the KB. When entities are processed in a later stage, they have to be compared not only two the previously available entities but also to the newly added instances.

3.1 Identity Class Models

Execution of the IdRF is based on what we call Class Models - that handle the differences of the entity types represented as ontology classes. They are configured as formulas and express different conditions for candidates *retrieval and comparison* during the identification process.

The formulas are valid for entities of the same type or class. However when two instances are from different classes then, the formula that is attached to the most specific common class is used. In case if one of the classes is subclass of the other e.g.

$$\text{class(C1)\&class(C2)\&subClassOf(C2,C1)} \qquad (1)$$

Each class model is expressed by a single formula based on first order probabilistic logic; in this way we are able to encode something specific about each class in the corresponding formula. Rule inheritance between classes is also supported allowing the set of formulas to be easily expanded for a new class. This is especially useful when the ontology is extended and refined or the focus of a particular IdRF application changes.

The formulas are described by predicates from a common pool of predicates which are implemented as Java classes making them extensible. Each formula is manually composed by combining predicates by the usual logical connectives like like "&", "|", "not" and "⇒". Predicates can be weighted in the formulas using real values from 0 to 1 - that can be attached to each of the predicates in the formulas using the logical connective "&".

```
namespace: rdf: "http://www.w3.org/1999/02/22-rdf-syntax-ns"
rdfs:"http://www.w3.org/2000/01/rdf-schema"
protons:"http://proton.semanticweb.org/2005/04/protons"
protonu:"http://proton.semanticweb.org/2005/04/protonu"
musing:"http://www.ontotext.com/2007/07/musing"


"protons:Entity":
    SameAlias()

"protont:Organization":
      let parentCond = Super()
          sectorCond =
          StrictSameAttribute(<protonu:hasURL>)
          aliasCond = OrganizationLD(<protonu:hasAlias>)
      in  parentCond | aliasCond | sectorCond

"musing:Company":
      Super()|
      OrganizationCombine(<protonu:hasAlias>) \&
      StrictSameAttribute(<musing:hasPostal>) \&
      StrictSameAttribute(<musing:hasSector>)
```

Figure 6: Example of SeRQL query for a "musing:Company" class model

The example on Figure 6 shows the corresponding formulas for two classes the main class *protonu:Company* and its subclass - application specific extension *musing:Comapany.* It is essential that several formulas can use one and the same predicate as part of their definitions (e.g *StrictSameAttribute()* on Figure 6) and each primitive predicate is implemented as Java class. The idea of defining a number of simple predicates instead of a single complex one follows the library like or "code reuse" approach in software development. This allows us to support an extendable set of reusable primitive predicates from which someone can compose complex formulas in a declarative way.

Class models are used in two stages of the framework pipeline: (i) during the retrieval of potential matching candidates from the ontology - applying a strict criteria; and (ii) during actual comparison of entities potential matching pairs using a soft criteria. They are also evaluated differently depending on which component use them.

3.2 Pre-filtering

It filters out the irrelevant part of the ontology and forms a set of instances similar to the source entity. It is intended to restrict the whole amount of ontology instances to a reasonable number, to which the source entity will be compared. The selected instances are potential target instances that might be identical to the source object; they already appear in the knowledge base and are somehow similar to the source object.

In this case the engine does not formally evaluate the class model/formula but composes a SeRQL or SQL query. The query embodies the model strong equivalency criteria or

```
  select DISTINCT
  V1
from
  {V1} <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
             {<http://ontotext.com/2007/07/musing#Company>};
       [<http://ontotext.com/2007/07/joci#hasURL> {V2}];
       [<http://www.w3.org/2000/01/rdf-schema#label> {V3}];
       [<http://www.w3.org/2000/01/rdf-schema#label> {V4}];
       [<http://ontotext.com/2007/07/musing#hasPostal> {V5}]
where
  (V2 = "http://www.marksandspencer.com") or
  ( (   ( (V3 like "*marks" IGNORE CASE )
            or (V3 like "*marks *" IGNORE CASE )
        ) and
        ( (V4 like "*spencer" IGNORE CASE )
            or (V4 like "*spencer *" IGNORE CASE )
        )
     )
     and (V5 = "W2 1NW")
   )
```

Figure 7: Example of SeRQL query for a "musing:Company" class model

restrictions with concrete values from the currently processed object e.g. some of the extracted values found by the OBIE system are used to retrieve entities from the knowledge repository.

For example retrieving instances from the system knowledge base, which are similar to a company called “MARKS & SPENCER” according to the class model for "musing:Company" on Figure 6, will result in automatically composed SeRQL query on Figue 7.

Once the query is prepared it is sent to the semantic repository and the retrieved objects are returned to the pre-filtering component.

3.3 Evidence Collection

It collects as much as possible evidence about the similarity between the source entity and each of the targets in the ontology. A set of similarity criteria is computed by comparing corresponding attributes in the entity descriptions. Different comparison criteria are possible: some are based on string representation e.g. text edit distance, inverted frequency based matching; others can be web appearance, context similarity, etc. All this is specified in the formulas described above.

This component calculates the similarity between two objects based on their class model, which is expressed by a probabilistic logic formula resulting in a real number from 0 to 1. Value 0 means that the given entities are totally different and value 1 means that they are absolutely equivalent. Any value between 0 and 1 mean that these entities are equivalent but only with a specific confidence.

In is important to note that each formula is evaluated for each pair of instances - the new coming entity and the matching candidate. The same is true also for the predicates calculation. Predicates have access to all details of both instances, although they can also take instance attributes, properties and relations as attributes.

Formally, a predicate value is calculated according to its algorithm, which reflects the specificity of the predicate and its attributes. As an example, *OrganizationLD()* predicate on Figure 6 normalises organisation names (e.g. preprocessing suffixes and abbreviations) and computes Levenshtein Distance between them.

Once calculated the values of different predicates are combined according to the logical connectives in the corresponding formula. In this setting the the usual logical connectives are expressed as arithmetic expressions, e.g. $a \vee b \equiv a + b - ab$, and so on for the other possible combinations of binary operators.

3.4 Data Integration

Once all the evidences for different identity possibilities are collected, the IdRF decides which is the best identity match. It is this third stage of identification process that encodes the strength of the presented evidence for choosing the candidate favored by the class model natively stored as part of the Class Model. The successful candidate must pass a certain preset threshold which meaning is to balance precision and recall according to the specific application requirements.

After the decision is made the incoming entity is registered to the ontology. Thus, the resulted knowledge base contains aggregated data from the incoming entity. Each of the new coming entities can be either not matched or successfully identified with an existing instance. If the system is not able to find a reliable matching candidate, the incoming object is inserted as a new instance in the KB. More interesting case is when the source entity is successfully associated with an existing instance, then the object description is added to the description of the identified KB instance. Hence the details of the entity are enriched. The results from the current identification are stored back in the ontology knowledge base and are used for the identity resolution of the next incoming objects.

4. CASE-STUDY EVALUATION

Our case-study is focused on company profiling. We use ontology based information extraction system implemented as GATE[4] application to collect information about UK companies, e.g. name, web-site and contact details including phone and fax number, from various web pages. Then, the extracted templates are compared and updated with manually collected data about 1,801,868 different companies provided by a company called "Market Location"[5].

In order to enable merging of those two sources we use the identity resolution framework. The process has targeted a set of 310 UK companies and attempted unification against an already populated knowledge base. The input to the framework is a set of RDF statements generated with respect to our domain ontology and following the system templates as shown on Figure 5. Each of these statements is either a new fact or a known fact. In the first step, the system selects all possible matches from the set of already collected company profiles. Formally the preexisting com-

[4] http:
gate.ac.uk

[5] http://www.marketlocation.com

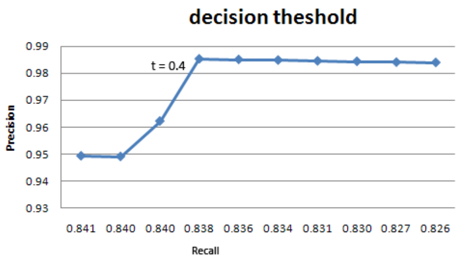


Figure 8: Threshold ROC curve analysis

pany records are stored and retrieved from a RDBM - MySql repository - using manually created mapping between the DB schema and the Ontology as explained earlier. Then the system calculates the candidate identity evidence using the corresponding formula for the particular ontology class - here "protont:Company" - that is shown on Figure 6.

Finally the framework makes a decision about the identity of the candidates based on a pre-set threshold and registers the successful matches. We have used ROC curve analysis to calculate the best value for our company profiling threshold. As shown on Figure 8 the threshold of *0.4.* gives the best performance balancing the system's Precision and Recall.

As our evaluation shows (see Table 1) the accuracy of identity resolution process is very promising giving 89% F-measure. In order to prove portability that similar results can also be achieve in other domains, we have performed another experiment on merging job-offers extracted from corporate web-sites. The set of job offers collected in this experiment consists off automatically extracted vacancies. This set has 45% rate of redundancy, e.g. two of each fife vacancies are redundant and usually described with less details. The f-measure for identity resolution in the requitement domain is 85% which is rather encouraging and indicates potential of the proposed approach (also presented on Table 1).

	Precision	Recall	F-measure
Company profiles	0.86	0.84	0.89
Vacancies	0.82	0.89	0.85

Table 1: Evaluation of Information Merging

There are two aspects of data improvement while merging different datasets - improve are both correctness and completeness of the collected data. In our company profiling application we emphasize on completeness, retrieving as much as possible company details. As shown on Table 2 the Recall of the OBIE data or how many of the company attributes have been extracted is 92%. After integration however it increase to 97%, which is rather acceptable result. Unfortunately the Precision or the correctness of the obtained results is initially only 70% and does not show significant growth after integration reaching 73%.

The main causes for this is that the chosen sources either fail to provide desired details or give us inaccurate information. Although our pre-existing database records have been manually collected there are a lot of empty fields. Table 3 gives examples of company details (address, website and phones) that have been missing in the company records before integration. Therefore, we plan next series of experiments using up to five instead of two sources in order to explore how possible inter-source agreement on company details can improve the correctness of collected data.

	Precision	Recall	F-measure
OBIE templates	0.70	0.92	0.80
Integrated profiles	0.73	0.97	0.83

Table 2: Evaluation of company profiles before and after consolidation

profile element	Update statistics
postal and address	20%
website	23%
phone	24%

Table 3: Statistics on updating details in the pre-existing data

5. RELATED WORK

Previous experiments in multi-source information extraction have been taken mainly in the area of Text Summarization, Databases and Co-reference Analysis. Saggion [24], studies the effect of different document contexts (e.g. full document, summary) and term representations (e.g. words, named entities) for entity clustering. An approach which uses named entities of type organisation to disambiguate person names proved to be very competitive. Although not deciding about the identity, Mann and Yarowsky [19] use semantic information that is extracted from documents to inform a hierarchical agglomerative clustering algorithm. Semantic information here refers to factual information about a person such as the date of birth, professional career or education.

In database identity is addresses as "record linkage" or "deduplication" problem of discovery identical record in different databases during their merging. A leading position in this field is hold by Bilenko and Mooney who present a framework for duplicate detection using trainable measure of textual similarity (a learnable text distance function) [4]. A comprehensive survey about different methods used for de-duplication in database field is given by [7]. In contrast to our work, all the approaches mentioned above are based and limited to string comparison of the corresponding field, not exploring the relations between record fields and hardly use even the fields' interdependencies.

Co-reference analysis is closely related to the extraction from multiple sources. It refers to the process of determining whether or not multiple mentions of entities refer to the same object and enables the extraction of relations among entities as well as complex propositions. Cross-document co-reference analysis pushes the task into considering whether mentions of a name in different documents are the same. There is little published work on cross-document co-reference analysis, and it has generally been evaluated on a small corpus of documents [11]. Significant results have been presented by Bagga and Baldwin [3, 2], as well as Gooi and Allan [11]. They use a Vector Space Model to compute similarities between personal summaries (sentences extracted) for each pair of documents. Hence, the co-reference problem has been addressed as language phenomena and solved only in textual context. On the other hand the identity resolution approach presented here is not bound to the source text, therefore handles wider range of objects.

In contrast to co-reference resolution which identifies different names that correspond to one and the same object, entity identification aims at disambiguation of identical names referring to different objects often addressed as author's name disambiguation. Aswani et al.[1] base their approach on web searches while looking for the author home pages, as well as on papers titles and abstracts. Other approaches based on web searching hits, author names distances etc. are given in [27]. Kousha and Thellwall [16] also explore different web sources e.g. Google Scholar[6] for tracking citations. Identification of names in news articles is another popular domain. Fernandez et al. [8] work on identification of names in news articles and base their approach on the categories of the news articles and "news trends" - news phenomena where one and the same entity appears in several consequent articles, usually connected with daily news. They use ontologies for internal representation and adopt their page ranking algorithm for this task. All these approaches, although successive in their particular domain - person names disambiguation, do not provide a general solution for similar problems in other domains, which our identity resolution framework does.

The IdRF proposed knowledge representation – ontologies – are already used for approaching the identity resolution problem. Funk et al.[9] present the advantages of semantically enhanced annotation for resolving co-references from different sources. Another example of using ontologies in this domain is the innovative work of Klein et al.[15] for extending standardised ontology description languages to enable approximation of instances. Although not being focused on the identity resolution but on duplicated data representation, the authors introduce new "Rough Description Language" as well as reasoning formalism on similarity of instances. Another notable aspect of using semantics for matching knowledge representation structures is presented by [10]. The authors define Match as an operator that takes two graph-line structures and produces mappings among the nodes that correspond semantically to each other. However, the processing is based mainly on the node labels, even if their comparison is based on WordNet [21] and the graph structure is restricted to a tree.

6. CONCLUSIONS AND FUTURE WORK

We have presented a general framework for identity resolution which is based on ontologies and can be adapted to different ontology-based information extraction and ontology-population applications. We have also demonstrated and evaluated the application of the framework in the context of two ontology-based information extraction systems where we have obtained positive results.

This work does not pretend to be complete but rather lays the foundations of series of experiments and further research of the problem of data aggregation and identification. Both experiments we have used illustrating the proposed approach aim at integration of rather well defined and rich of details entities e.g. company profiles. Obviously this is not true for a wide range of other object e.g. people mentioned in documents only with their names.

Researching how uniqueness of the details and their number influence the process of identification is targeted for our future work. We will look into adapting the framework in different applications in the Musing project where in addition to companies, we are extracting from multiple sources information for ontology population for entities such as persons, and locations.

Acknowledgements

This work was partially supported by the EU-funded projects MUSING (IST-2004-027097).

7. REFERENCES

[1] Niraj Aswani, Kalina Bontcheva, and Hamish Cunningham. Mining information for instance unification. In *International Semantic Web Conference*, 2006.

[2] A. Bagga and A. Biermann. A methodology for cross-document coreference. In *Proceedings of the Fifth Joint Conference on Information Sciences*, pages 207–210, 2000.

[3] Amit Bagga and Breck Baldwin. Entity-based cross-document coreferencing using the vector space model. In Christian Boitet and Pete Whitelock, editors, *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, pages 79–85, San Francisco, California, 1998. Morgan Kaufmann Publishers.

[4] Mikhail Bilenko and Raymond J. Mooney. Employing trainable string similarity metrics for information integration. In *Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web*, pages 67–72, Acapulco, Mexico, August 2003.

[5] N. Chinchor. Overview of muc-7. In *In Proceedings of MUC-7*, 1998.

[6] S. Dill, N. Eiron, D. Gibson, D. Gruhl, and R. Guha. Semtag and seeker: Bootstrapping the semantic web via automated semantic annotation. In *Proceedings of the The Twelfth International World Wide Web Conference (WWW 2003)*, Budapest, Hungary, 2003.

[7] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, and Vassilios S. Verykios. Duplicate record detection: A survey. Technical report, TKDE, January 2007.

[8] Norberto Fernandez, Jose M. Blazquez, Jesus A. Fisteus, Luis Sanchez, Michael Sintek, Ansgar Bernardi, Manuel Fuentes, Angelo Marrara, and Zohar Ben-Ashe. News: Bringing semantic web technologies into news agencies. In *International Semantic Web Conference*, 2006.

[9] Adam Funk, Diana Maynard, Horacio Saggion, and Kalina Bontcheva. Ontological integration of information extraction from multiple sources. In *International Workshop on Multi-source, Multi-lingual Information Extraction and Summarisaton*, 2007.

[10] Fausto Giunchiglia, Pavel Shvaiko, and Mikalai Yatskevich. S-match: an algorithm and an

[6] http://scholar.google.com/

implementation of semantic matching. In *ESWS*, pages 61–75, 2004.

[11] Chong Jeong Gooi and James Allan. Cross-document coreference on a large scale corpus. In *Proceedings of the Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting*, Boston, 2004.

[12] R. Grishman. Information Extraction: Techniques and Challenges. In *Information Extraction: a Multidisciplinary Approach to an Emerging Information Technology*, 1997.

[13] R. Grishman and B. Sundheim. Message understanding conference Ű 6. a brief history. In *In Proceedings of COLING-96*, pages 466–471, 1996.

[14] Atanas Kiryakov, Damyan Ognyanov, and Dimitar Mano. Owlim Ű a pragmatic semantic repository for owl. In *SSWS 2005, WISE*, USA, 2005.

[15] Michal C.A. Klein, Peter Mika, and Stefan Schlobach. Approximate instance unification using roughowl. 2007. submitted.

[16] Kayvan Kousha and Mike Thelwall. Google scholar citations and google web-url citations. a multi-discipline exploratory analysis. *Journal of the American Society for Information Science and Technology*, 58(7):1055–1065, 2007.

[17] W. Lehnert, C. Cardie, D. Fisher, E. Riloff, and R. Williams. Òuniversity of massachusetts: Muc-3 test results and analysis. In *in Proceedings of MUC-3*, pages 116–119. Morgan Kaufmann, 1991.

[18] W. Lehnert, D. Fisher, J. McCarthy, E. Riloff, and S. Soderland. Òuniversity of massachusetts: Muc-4 test results and analysis. In *in Proceedings of MUC-4*, pages 151–158, 1992.

[19] G. S. Mann and D. Yarowsky. Unsupervised personal name disambiguation. In W. Daelemans and M. Osborne, editors, *Proceedings of the* 7^{th} *Conference on Natural Language Learning (CoNLL-2003)*, pages 33–40. Edmonton, Canada, May 2003.

[20] D. Maynard, M. Yankova, A. Kourakis, and A. Kokossis. Ontology-based information extraction for market monitoring and technology watch. In *ESWC Workshop "End User Apects of the Semantic Web"*, Heraklion, Crete, 2005.

[21] George A. Miller. Wordnet: a lexical database for english. In *Communications of the ACM 38 (11)*, pages 39 – 41, November 1995.

[22] Borislav Popov, Atanas Kiryakov, Damyan Ognyanoff, Dimitar Manov, and Angel Kirilov. Kim - a semantic platform for information extaction and retrieval. In *Journal of Natural Language Engineering*. Cambridge University Press, 2004.

[23] Dragomir R. Radev. A common theory of information fusion form multiple text sources step one: Cross-document structure. In *Proceedings of the First (SIGdial) Workshop on Discourse and Dialogue*, pages 74–83, Somerset, NJ, May 2000.

[24] H. Saggion. Experiments on semantic-based clustering for cross-document coreference. In *International Joint Conference on Natural Language Processing*, Hyderabad, India, January 2008. AFNLP.

[25] H. Saggion, J. Kuper, T. Declerck, D. Reidsma, and H. Cunningham. Intelligent multimedia indexing and retrieval through multi-source information extraction and merging. In *IJCAI 2003*, Acapulco, Mexico, 2003.

[26] Ivan Terziev, Atanas Kiryakov, and Dimitar Mano. Base upper-level ontology (bulo) guidance. Technical Report Deliverable 1.8.1, SEKT project, UK, July 2005.

[27] K. Yang, J. Jiang, H. Lee, and J. Ho. Extracting citation relationships from web documents for author disambiguation. Technical Report TR-IIS-06-017, Institute of Information Science, Academia Sinica Taipei Taiwan, December 2006.