Proceedings of the 2nd International Workshop on Exploiting Large Knowledge Repositories *and the* 1st International Workshop on Automatic Text Summarization for the Future

Satellite events of the 28th edition of the SEPLN Conference 2012
September 7th, Castellón de la Plana, Spain

Volume 882 of CEUR-WS.org: `http://ceur-ws.org/Vol-882/`

Edited by: Ernesto Jiménez-Ruiz, Horacio Saggion, María José Aramburu Cabo,
Roxana Danger, Antonio Jimeno-Yepes, Elena Lloret, Manuel Palomar.

Preface

There were 13 papers submitted (9 long papers, 2 short papers and 2 statements of interest) each of which was reviewed by at least three members of the program committee. The program committee selected 11 papers for oral presentation.

Exploiting Large Knowledge Repositories workshop. Large knowledge repositories (LKR) are being created, published and exploited in a wide range of fields, including Bioinformatics, Biomedicine, Geography, e-Government, and many others. Some well known examples of LKRs include the Wikipedia, large scale Bioinformatics databases and ontologies such as those published by the EBI or the NIH (e.g. UMLS, GO), and government data repositories such as data.gov. These repositories are publicly available and can be used openly. Their exploitation offers many possibilities for improving current information systems, and opens new challenges and research opportunities to the information processing, databases and semantic web areas.

The main goal of this workshop is to bring together researchers that are working on the creation of new LKRs on any domain, or on their exploitation for specific information processing tasks such as data analysis, text mining, natural language processing and visualization, as well as for knowledge engineering issues, like knowledge acquisition, validation and personalization.

Automatic Text Summarization for the Future workshop. Due to the great proliferation of online documents and information, it becomes necessary to develop automatic tools capable of filtering redundant and irrelevant information, thus presenting the most important one in an efficient and effective manner. This is the goal of Automatic Summarization, which aims at producing a concise document, keeping the essential information.

Research into Automatic Summarization began in the 50s with the purpose of summarizing scientific texts. Recently, new challenges have appeared in this research area. In the context of the Internet, not only is information being constantly updated, but there is also a lack of quality control of what is being published on the Web. Social networks, blogs, reviews, etc. are non-traditional texts of informal nature, and they therefore constitute a big challenge for the new generation of summaries.

Another challenge for automatic summarization is the generation of abstracts, where it is necessary to take into consideration natural language generation techniques and be able to adapt them from one domain to another. In addition to these, efforts are needed to produce summaries in languages other than English and in multiple languages.

The main goal of this workshop is to bring together researchers working on Automatic Summarization, encouraging research into little explored areas such as new textual gentres as well as old, forgotten ones, or summarization in languages other than English.

Acknowledgements

We thank all members of the program committee, additional reviewers, authors and local organizers for their efforts. We would also like to acknowledge that the work of the workshop organisers was greatly simplified by using the EasyChair conference management system (`http://www.easychair.org`) and the CEUR Workshop Proceedings service (`http://ceur-ws.org/`).

Program Committee

Laura Alonso	Universidad Nacional de Crdoba
Ahmet Aker	University of Sheffield
María José Aramburu Cabo	University Jaume I
Ana Armas	University of Oxford
Yassine Benajiba	Columbia University
Rafael Berlanga Llavori	Universitat Jaume I
Ester Boldrini	Universitat d'Alicante
Hakan Ceylan	University of North Texas
Iria Da Cunha	Universitat Pompeu Fabra
Roxana Danger	Imperial College London
Manuel De La Villa Cordero	Universidad de Huelva
Alberto Díaz	Universidad Complutense de Madrid
Atefeh Farzindar	NLP Technologies Inc.
Maria Fuentes	Universitat Politècnica de Catalunya
Robert Gaizauskas	University of Sheffield
George Giannakopoulos	NCSR Demokritos
Jorge Gracia	Universidad Politécnica de Madrid
Ramon Granell	University of Oxford
Nicolas Hernandez	Université de Nantes
Ernesto Jimenez-Ruiz	University of Oxford
Antonio Jimeno-Yepes	NLM, NIH
Senay Kafkas	EMBL Outstation Hinxton The European Bioinformatics
Evgeny Kharlamov	Free University of Bozen-Bolzano
Leila Kosseim	Concordia University
Guy Lapalme	Université de Montréal
Maria Liakata	University of Wales, Aberystwyth
Elena Lloret	Universitat d'Alacant
Dolores María Llidó	Universitat Jaume I
Despoina Magka	Oxford University Computing Laboratory
Marco Mesiti	DICO - University of Milano
Jean-Luc Minel	Universit Paris Ouest Nanterre La Défense
Shamima Mithun	Concordia University
Jose Mora	Universidad Politécnica de Madrid
Paloma Moreda	Universitat d'Alacant
Rafael Muñoz	Universitat d'Alacant
Victoria Nebot	Universitat Jaume I
Ani Nenkova	University of Pennsylvania
Manuel Palomar	Universitat d'Alacant
Thiago Pardo	Universidade de São Paulo
María Pérez	Universitat Jaume I
Laura Plaza	Universidad Complutense de Madrid
Bastien Rance	NLM, NIH

Dietrich Rebholz-Schuhmann	European Bioinformatics Institute
Horacio Rodriguez	Universitat Politècnica de Catalunya
Paolo Rosso	POlytechnic University Valencia
Horacio Saggion	Universitat Pompeu Fabra
Ismael Sanz	Universitat Jaume I
Giorgio Stefanoni	University of Oxford
Juan Manuel Torres-Moreno	Laboratoire Informatique d'Avignon
Jorge Vivaldi	Universitat Pompeu Fabra
René Witte	Concordia University
Dina Wonsever	UdelaR - Fing
Dmitriy Zheleznyakov	Free University of Bozen-Bolzano
Yujiao Zhou	University of Oxford

Table of Contents

Invited talk

Long papers

Short papers

Statements of interest

Generation of Patent Abstracts: A Challenge for Automatic Text Summarization

Leo Wanner, ICREA and DTIC, UPF

It is well known that patents drive the modern economies. But they do even more: patents also serve as a valuable and unique source of up-to-date scientific and technological information. It is assumed that only 10% to 15% of the content presented in patents are described in other publications as well. The worldwide stock of patents thus comprises about 85% to 90% of scientific knowledge. Given that central parts of patents are authored in an idiosyncratic and complex language which is difficult to read and comprehend, and since author-written patent abstracts have the goal to obfuscate the precise nature and the real scope of the inventions rather than to clarify them, an efficient access to this knowledge, for instance, via concise and transparent summaries, appears crucial. However, partially due to the aforementioned language idiosyncrasy, which implies extremely long sentences with complex repetitive linguistic constructions, common extraction-oriented automatic text summarization techniques cannot be expected to show an acceptable performance when applied to patents. Other, more content-oriented (or abstractive) summarization techniques are needed. In my talk, I will present the recent and ongoing research on patent summarization carried out by the Natural Language Processing Group of the Department of Information and Communication Technologies, UPF as member European consortia. I will first describe the techniques for the summarization of patent claims developed in the scope of the PATExpert project and outline then how these techniques are about to be improved in the TOPAS project by considering information from other sections of a patent, notably the description of the invention. In the last part of my presentation, I will summarize the remaining challenges and suggest some lines of future research which are crucial if we want automatic patent summarization to be a real alternative to (semi-)manual abstracting, which still dominates the patent domain.

Towards an ontology based large repository for managing heterogeneous knowledge resources

Nizar Ghoula and Gilles Falquet

ICLE, Centre Universitaire d'Informatique, University of Geneva, Switzerland
`{Nizar.Ghoula,Gilles.Falquet}@unige.ch`

Abstract. Knowledge based applications require linguistic, terminological and ontological resources. These applications are used to fulfill a set of tasks such as semantic indexing, knowledge extraction from text, information retrieval, etc. Using these resources and combining them for the same application is a tedious task with different levels of complexity. This requires their representation in a common language, extracting the required knowledge and designing effective large scale storage structures offering operators for resources management. For instance, ontology repositories were created to address these issues by collecting heterogeneous ontologies. They generally offer a more effective indexing of these resources than general search engines by generating alignments and annotations to ensure their interoperability. However, these repositories treat a single category of resources and do not provide operations for reusing them. The aim of this research is building a large repository of knowledge resources. This repository is a collection of heterogenous resources represented in different languages and offers a set of operations to generate new resources based on the existing ones.

Key words: Resources repository, Operations, Ontology of resources, Knowledge representation

1 Introduction

Knowledge extraction and representation is a widely explored research problem. Most of the proposed solutions to this problem are based on the usage of auxiliary knowledge resources [1]. This knowledge currently exists in resources of different types such as terminologies, glossaries, ontologies, multilingual dictionaries or aligned text corpora. These resources are represented using various formalisms and languages such as predicate logic, description logic, semantic networks and conceptual graphs, etc. As part of an application that requires the use of external resources, a designer is often required to perform painstaking research and pre-treatment in order to collect and build adequate resources to his application needs. Resolving this problem relies on finding at first the right resources before extracting the required knowledge and then representing it in a common formalism. It is then important to have repositories offering access to more diverse resources in different formalisms. Moreover, the right knowledge resource for an application must be constructed and adapted to the application. This adaptation may involve operations such as selecting a part of a resource, composing it with another one, translating it to another language or representing it in a different formalism [2] [3] [4].

In this paper, we present a model and a taxonomy of abstract operations for managing and extracting knowledge from resources. We consider the possibility of combining these operators to perform complex processes such as semantic enrichment or generating a new resource by merging some other resources.

2 Methodology

A central point of our approach is to build a repository of knowledge resources. This repository should offer the possibility to store and integrate heterogenous knowledge resources and organize their usage in common context. It should also offer operators for managing and combining these resources. For this we have proposed a three steps methodology:

- propose a method and a formalism allowing to represent heterogeneous terminological, linguistic and ontological knowledge resources;
- define the major representation languages by means of the repository's concepts (Resource, Entity, Relation, etc.);
- define a set of operations performed on these resources to generate new resources bases on some criteria;
- propose multiple implementations per operator depending on the resource type and the representation language;
- implement a resources repository to study and resolve scalability problems that arise by evaluating the usability of such a system.

Our approach is not focused on a particular domain, it aims to represent different resources from diverse domains and manipulate them using different operations. We distinguish two categories of resources. The first category is about autonomous resources like ontologies, corpora or terminologies. These resources are widely used in multiple applications of knowledge management. The second one represents enrichment resources like annotations or alignments. They link two or more autonomous resources and they result from the application of a process on autonomous resources.

3 State of the art

For managing heterogeneous resources in large knowledge repositories we need to resolve the problem of resources representation and storage at first and then address the problem of defining and implementing resources management operators (collected from existing approaches and classified by type such as alignment operators, annotation services, translation mechanisms, etc.).

3.1 Knowledge resources repositories

Some large repositories have been created to offer a more effective indexing for knowledge resources than common search engines. For example, Swoogle[1] indexes more than 10 000 ontologies; DAML repository[2] provides search based on ontology components (classes, properties, ...) or metadata (URI, funding source, ...); BioPortal[3] has similar searching and browsing tools [5] and offers the possibility to annotate and align different ontologies. Many other portals [6] [7] offer access to linguistic or ontological resources. However, these portals are dedicated each for a specific category of resources (Swoogle is focused on ontologies, ACL[4], CLARIN[5] or META-NET[6] are focused on corpora and linguistic resources).

A repository containing heterogeneous types of knowledge resources is needed. Hence, multiple languages for representing these resources are required. For this purpose, it is necessary to develop a set of knowledge resources operators that can import, export and process these resources while keeping a trace of their origin (the provenance of the resources, for example externally imported or generated from the combination of multiple ones).

3.2 Resources representation models

There are many models for knowledge representation, but they usually focus on one or two aspects only: ontological, terminological, lexical, textual, documentary, etc. It is more difficult to find models representing various aspects of knowledge or resources of different kinds. For the integration of heterogeneous resources, [8] have proposed a model of terminologies and ontologies. This remains faithful to the representation of each resource model without using common abstract entities. For example, instead of considering a term or a concept as an abstract entity these classes have different representations depending on the resource, which creates redundancy in the instances. A model of the multilingual aspect in ontology has been proposed by [9], its development is an association between a meta-model of ontologies and a linguistic model. Another model to unify the management of linguistic resources in multilingual environment has been developed to centralize the management of linguistic resources within a platform called Intuition [10]. This model is characterized by its exploration of the structure of linguistic forms. The application of this model allows to represent ontological entities and identify lexical units by taking into account the syntactic and semantic multilingual relations. This model cannot represent pure linguistic resources. [11] proposed a Linguistic Meta-Model (LMM) allowing a

---

[1] http://swoogle.umbc.edu

[2] http://www.daml.org/ontologies

[3] http://bioportal.bioontology.org

[4] http://www.aclweb.org

[5] http://www.clarin.eu/external/

[6] http://www.meta-net.eu

semiotic-cognitive representation of knowledge and linguistic resources. It represents individuals and facts in an open domain perspective.

In our case, we need to preserve the originality of all resources and treat them within their original context and representation language. This is why we propose a meta-model treating a resource as an entity in the repository. Each resource can have different derivations which are also resources represented in different languages.

3.3 Resources re-engineering

In the context of mapping linguistic and ontological resources, [12] have proposed an approach to integrate and merge Wikipedia and WordNet to enrich an ontology (YAGO[7]). The ontology is extracted from these two resources by adding new facts[8] extracted from Wikipedia as individuals, classes from the conceptual categories in Wikipedia and each "synset" of WordNet. This approach shows that the combination of multiple resources makes possible building or extending existing resources. Another methodology [13] focuses on a pattern based approach for re-engineering non-ontological resources into ontologies. This type of approach is a perfect component or a framework to add in the repository. It offer a comparative study of re-engineering methods of non-ontological resources. By means of this framework we can design a decision support algorithm for choosing the best reuse method based on the type of the resource since all reuse methods are supposed to be implemented by means of services or operators in the repository.

4 A meta-model for integrating heterogeneous resources

Since there exist many different (and incompatible) ways to express knowledge in resources (from formal logic to semi-formal or natural languages). Moreover, the same resource may be involved in processes that can only handle specific representation formalisms. For instance, an ontology alignment algorithm might be implemented for OWL ontologies, while another algorithm might be about resources in a WordNet-like model. It can be the same for other processes like automated text annotation, multilingual text alignment, word sense disambiguation, etc.

We have proposed a MOF-based model[9] to unify the representation of heterogeneous resources in a common formalism [14]. This model allows to describe the metadata of any kind of knowledge resource and then associate different representations (derivations) of the resource's content in many languages (formalisms) which are by them selves represented in the repository by means of a

---

[7] Yet Another Great Ontology

[8] relative to all existing data in a knowledge base

[9] MOF is an acronym for Meta-Object Facility: http://www.omg.org/mof/

common terminology (namespace of the repository). The implementation of this model includes an ontology, called TOK_Onto[10].

Depending on the user's needs, a resource in the repository can be represented differently using multiple languages, each language uses a subset of the resource's entities and link them in a different way compared to another language (for example, a class hierarchy representation links the concepts of an ontology using the subClassOf relation which leads to a different derivation of this resource, otherwise a semantic network representation of that resource will lead to the use of another set of relations). Table 1 shows some example of languages that have been described in the current version of the repository.

Table 1. Examples of resource content models (languages) and their principal components

Model	Components
Concept hierarchy	Concept, ISA_Relation, ...
WordNet Like	Concept, Term, Lexical_Form, Hypernym_Relation, Meronym_Relation, Term_Form_Relation, ...
Graph ontology	Class, Taxonomic_Relation, Relation, Relation_Label, etc.
Translation memory	Text_Segment, Language, Translation_Relation, Language_Relation
Ontology Alignment	Concept, Correspondence_Relation, ...

For example, to represent an ontology we can focus on the hierarchy of classes if we need it in a task of classification. We can also represent the same ontology by focusing on axioms and complex expressions using logics if we need it for a reasoning task.

5 Taxonomy of operations on knowledge resources

The aim of a resources repository is not only to collect heterogenous knowledge resources but especially to offer instruments for reusing them. In order to formalize the definition of processes over these resources, we have defined a set of generic primitive operations. We represented then an abstract class of operators in the repository's ontology in order to manage multiple implementations for each operator and to represent restrictions about each implementation. We define a process as a sequence of operators applied on resources' derivations. By means of processes descriptions we managed to construct a process dictionary that stores each instance of a process and apply it each time there is an

---

[10] http://cui.unige.ch/isi/onto/tok/OWL_Doc/

evolution in the involved resources. Therefore, we must develop a subsequent meta-operators. The definition of these operators depends on the treatment of the resources.

5.1 Representation operators

These are the basic construction operators for representations. The abstraction and reification operations create the resources in the repository and map them to their original derivation in the repository (representation of the resource in its original language). Language mapping operations creates new derivations in other languages.

Importation or abstraction We denote by i_{RL} the import operation that produces an instance of a resource R in the resources repository and by creating the content of the resource in its original language$L$. This operation can be followed by a derivation which produces a derivation of the resource in a representation language.

Exportation or reification We denote by e_{RL} the export operation that transforms a derivation of a resource R expressed in a language L and its metadata into an external file in a certain formalism related to the derivation's language. Reification is generally used at the end of a process (sequence of operations) to produce the new resource. Consequently this operator can have as much instances as the possible combinations from the representation languages implemented in the repository (for example OWL, UML, DL, Graphs, etc.) to the possible required formats (txt, xml, rdf, ttl, n3, etc.).

Derivation This abstract operator is used to create new representations of a resource in different languages (represented already in the repository). For instance, an UML class diagram could be derived into a *Class diagram* representation, then mapped to WordNet-like lexical ontology model (by dropping all the associations except *part-of* and *subclass*). Since a derivation may "forget" information, in general $\mu_{L_2L_1}$ is not the inverse of $\mu_{L_1L_2}$. It is not always necessary to preserve the entire contents of a resource when deriving a new representation of its content (this can be compared to generating a view in the relational approach). In particular, if the representation language is less expressive than the original language it is obvious that some knowledge will be lost.

5.2 Enrichment operators

The enrichment operations generate new alignments or annotations on existing resources. They are generally based on sophisticated algorithms (more precisely heuristics) and use auxiliary resources like lexical ontologies.

Alignement Alignment allows to express explicitly the correspondences between resources [15]. An alignment method consists of defining a distance between the entities of a resource and calculating the best match between them by minimizing the distance measure or maximizing the similarity measure [16]. An alignment operator takes as input two resources R_i and R_j represented in a language L_1 and a set of auxiliary resources represented in other languages $L_2, \ldots$ to produce an alignment resource represented in a language L_{al}.

The signature of this operator is :

$$\mathsf{Op}_{\mathsf{Align}} : L_1, L_1, [L_2, \ldots] \to (L_1, L_{al})$$

L_{al} is a language that includes the alignment relations used to represent the correspondences ($\sqsubseteq, \equiv$, etc.), $\mathsf{Op}_{\mathsf{ALIGN}}$ is the operator used for the alignment.

A typical example of the need for simplified languages is the ontology alignment task. Most of the current alignment algorithms can align ontologies represented in OWL language, but they do not take advantage of all the semantics expressed in such ontologies [17]. They are based on the textual labels attached to each class and the structure of the ontology. The structure of a used resource is generally a graph representing the class hierarchy and a set of properties relating two classes, e.g. there is an axiom of the form $Class_1 \sqsubseteq$ *property* **only/some** $Class_2$. In this case, it is much more appropriate to represent an OWL ontology by its graph instead of the full description logic model. This will adapt the resources for the alignment algorithms that are able to align any type of ontology expressed as a labelled graph.

Annotation The annotation operator is used to describe elements of a resource R_1 in terms of a resource R_2, this description is through adding a set of relationships between entities of these resources according an annotation language.

The signature of this operator is:

$$\mathsf{OP}_{\mathsf{Ann}} : L_1, L_2 \to L_1, L_2, L_{ann}$$

where L_1 is the language of the resource's derivation to annotate and $L_2, \ldots$ are the languages of the resources' derivations that serve as reference in the annotation. L_{ann} is the annotation language. For example, *word sense disambiguation* is a kind of annotation operation. Starting from a natural language text and a reference lexical ontology (and possibly other resources), it produces a set of correspondences between the text words and their meanings (the concepts of the ontology).

5.3 Selection and combination operations

These operations are intended to produce new resources' derivations by selecting and combining entities of one or more resources.

Selection This type of operation selects entities from a resource's derivation to generate a new resource's derivation in the same language. This filtering is specified by a boolean function applied on each entity. The computation of the filtering function for a resource entity may depend on other entities from the same resource or others entities associated to it by means of annotations or alignments. In addition, the selection may generate a natural alignment between entities of the original and new resource's derivations. Each selected entity is associated to its original entity.

The signature of a selection operation is of the form

$$\mathsf{Op}_{\mathsf{Sel}} : L_1 \rightarrow L_1$$

where L_1 is the language of the input resource and the resulting selection.

For instance, in a description logic ontology, this operator can select individuals in the ABox (Assertional Box), leaving the TBox (Terminological Box) untouched (as in a database **selection**) or it can select a subset of the TBox, and hence drop the ABox entities that depend on unselected TBox concepts or roles (as in a database **projection**).

Composition Composition operations may be applied on alignments and annotations. It is an operator that generates new derivation of the composed resources in the same language.

The composition of two alignment resources (from S_1 to S_2 and from S_2 to S_3 results in a new alignment resource from S_1 to S_3. The semantics (relation type) of the resulting alignment depends on the relation types of the given alignments. If A_1 and A_2 have the same relation type R and R is transitive, then $A_1 \circ A_2$ has type R.

Merge The idea of the merge operation is to build a new resource by taking all the entities of two given resources [18] [3]. Depending on the representation language, the operation can take different forms. For example, using the merge operator on two ontologies in the language DL (description logic) is reduced to perform the union operation of their vocabularies and axioms:

- (merge) disjoint union of the vocabularies and axioms plus equivalence and subsumption axioms corresponding to the given alignment;
- (replace) if named concept C of an ontology O_1 is aligned (equivalence) with the named concept D of an ontology O_2 then the operators drops every axiom that defines C ($C \equiv \ldots$ and $C \sqsubseteq \ldots$), keeps the axioms that define D and add the axiom $C \equiv D$. This is a way to replace the definitions given in O_1 by those in O_2 (used, for instance, when O_2 is considered as more reliable than O_1).

The signature of the merge operator has the form:

$$\mathsf{Op}_{\mathsf{Merge}} : L_1, L_1, [L_{al}] \rightarrow (L_1)[L_{al}]$$

This operator takes as parameters a list of resources represented in the same language and uses auxiliary resources such alignments between them. Merging two alignments or annotations can occur only if they are about a common resource. First, for each resource R_i to merge, we must consolidate and merge all correspondences whose source is R_i and represented in the same alignment language L_{al}. A multiple inputs and outputs alignment resource is constructed and represented within the language L_{al}. Both the set of resources to merge and the constructed alignment provide required ingredients for the merge.

6 Conclusion and Further work

Our main objective is to build a large repository for integrating heterogeneous resources represented in different languages. We have identified three major steps for implementing this repository. First we have defined an upper level model for representing knowledge resources and dealing with different representation languages. Then we have defined a set of abstract operators having multiple implementations in order to combine the content of the repository and generate new resources from existing ones. We will focus on defining examples and a set of use cases in order to validate this approach and finally address the scalability issues. To ensure the usage of the repository by means of knowledge representation and resources management operators we are currently focusing on the following issues: (1) define a model for each processing task using resources, these tasks models should be the result of a reflection on a set of use cases; (2) define and implement a set of heuristics for the automatic detection of entity mappings to construct alignments between resources during the execution of any task.

For the third part of this research we will focus on the experimentation and the implementation of the repository. An implementation of a prototype is intended to prove the research results and define software requirements by studying the available technologies and APIs that can be used. For instance, we should address the following issues:

- evaluation and study of RDF storage approaches must be driven to select the best storage API to use for storing knowledge resources especially focus on the scalability issues;
- for the sake of generality we should investigate the possibilities for providing resources management operators using web services;
- define the interface that should be used for the repository's portal and the define the criteria of accessibility and user profiles.

References

1. Hendler, J., Golbeck, J.: Metcalfe's law, web 2.0, and the semantic web. Web Semant. **6** (February 2008) 14–20
2. D'Aquin, M., Schlicht, A., Stuckenschmidt, H., Sabou, M.: Criteria and evaluation for ontology modularization techniques, Berlin, Heidelberg, Springer-Verlag (2009) 67–89

3. Pinto, H.S., Martins, J.P.: A methodology for ontology integration. In: K-CAP'01:Proceedings of the 1st international conference on Knowledge capture, New York, NY, USA, ACM (2001) 131–138
4. Sabou, M., Lopez, V., Motta, E.: Ontology selection: Ontology evaluation on the real semantic web. In: In Workshop: Evaluation of Ontologies for the Web (EON 2006), 15th International World Wide Web Conference, Edinburgh (2006)
5. Noy, N.F., Shah, N., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Montegut, M., Rubin, D.L., Youn, C., Musen, M.A.: Bioportal: A web repository for biomedical ontologies and data resources. In: International Semantic Web Conference (Posters & Demos). (2008)
6. Sabou, M., Dzbor, M., Baldassarre, C., Angeletou, S., Motta, E.: Watson: A gateway for the semantic web. In: Poster session of the European Semantic Web Conference, ESWC. (2007)
7. Kiryakov, A., Ognyanov, D., Manov, D.: Owlim - a pragmatic semantic repository for owl. In: WISE Workshops. (2005) 182–192
8. Vandenbussche, P.Y., Charlet, J.: Méta-modèle général de description de ressources terminologiques et ontologiques. In Gandon, F.L., ed.: Actes d'IC, PUG (2009) 193–204
9. Montiel-Ponsoda, E., Aguado de Cea, G., Gómez-Pérez, A., Peters, W.: Modelling multilinguality in ontologies. In: Coling 2008: Companion volume: Posters, Manchester, UK, Coling 2008 Organizing Committee (August 2008) 67–70
10. Cailliau, F.: Un modèle pour unifier la gestion de ressources linguistiques en contexte multilingue. In Mertens, P., ed.: Verbum ex machina: actes de la 13e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2006) : Leuven., Presses univ. de Louvain, 2006 (2006) 454–461
11. Picca, D., Gliozzo, A.M., Gangemi, A.: Lmm: an owl-dl metamodel to represent heterogeneous lexical knowledge. In: LREC, European Language Resources Association (2008)
12. Suchanek, F., Kasneci, G., Weikum, G.: YAGO: A core of semantic knowledge - unifying WordNet and Wikipedia. In Williamson, C.L., Zurko, M.E., Patel-Schneider, Peter F. Shenoy, P.J., eds.: 16th International World Wide Web Conference (WWW 2007), Banff, Canada, ACM (2007) 697–706
13. García-Silva, A., Gómez-Pérez, A., Suárez-Figueroa, M.C., Villazón-Terrazas, B.: A pattern based approach for re-engineering non-ontological resources into ontologies. In: Proceedings of the 3rd Asian Semantic Web Conference on The Semantic Web. ASWC '08, Berlin, Heidelberg, Springer-Verlag (2008) 167–181
14. Ghoula, N., Falquet, G., Guyot, J.: Tok: A meta-model and ontology for heterogeneous terminological, linguistic and ontological knowledge resources. In Huang, J.X., King, I., Raghavan, V.V., Rueger, S., eds.: Web Intelligence, IEEE (2010) 297–301
15. Kalfoglou, Y., Schorlemmer, M.: Ontology mapping: the state of the art. Knowl. Eng. Rev. **18**(1) (2003) 1–31
16. Euzenat, J., Meilicke, C., Stuckenschmidt, H., Shvaiko, P., Trojahn, C.: Journal on data semantics xv. Springer-Verlag, Berlin, Heidelberg (2011) 158–192
17. Shvaiko, P., Euzenat, J.: Ontology matching: State of the art and future challenges. IEEE Transactions on Knowledge and Data Engineering **99**(PrePrints) (2011)
18. Noy, N.F., Musen, M.A.: Anchor-PROMPT: Using Non-Local Context for Semantic Matching. In: Workshop on Ontologies and Information Sharing at the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-2001), Seattle, WA (2001)

Enhancing the expressiveness of linguistic structures

J. Mora, J. A. Ramos, G. Aguado de Cea

Ontology Engineering Group – Universidad Politécnica de Madrid. Spain
{jmora, jarg, lupe}@fi.upm.es

Abstract. In the information society large amounts of information are being generated and transmitted constantly, especially in the most natural way for humans, i.e., natural language. Social networks, blogs, forums, and Q&A sites are a dynamic Large Knowledge Repository. So, Web 2.0 contains structured data but still the largest amount of information is expressed in natural language. Linguistic structures for text recognition enable the extraction of structured information from texts. However, the expressiveness of the current structures is limited as they have been designed with a strict order in their phrases, limiting their applicability to other languages and making them more sensible to grammatical errors. To overcome these limitations, in this paper we present a linguistic structure named "linguistic schema", with a richer expressiveness that introduces less implicit constraints over annotations.

Keywords: Pattern Matching, Pattern Recognition.

1 Introduction

Text understanding covers a series of tasks such as document classification [13], machine learning [9], information retrieval [3], etc. To perform these tasks, two processes are generally carried out: the recognition of structures and the interpretation of them. In the first one, the aim is to find some specific structures (for example, the pattern *AGENT buys OBJECT* in the text of a web page). Depending on the results found in the search (for example, *AGENT=Pepe* and *OBJECT=flores*, *AGENT=Paco* and *OBJECT=bombones*) the interpretation process triggers the action corresponding to the task performed (learning task, classification task, etc.). In other words, during the interpretation process, the document is classified (for example, G*oods Transactions*), something is learnt (for instance, *Pepe* and *Paco* are instances of *Person*), some information is retrieved (for example, *flores* and *bombones* are goods sold in the Web), etc. Generally speaking, the process of structure recognition is common and independent of the interpretation process, although this process can be instantiated in a battery of structures that might be needed for a later specific interpretation. However, the recognition process itself does not vary. It is in the above mentioned structures on which this work is focused: studying and upgrading their representations and the expressiveness of these representations. This expressiveness will determine the searches: the greater the expressiveness, the more searches can be conducted and the more complex these searches can be. Large scale corpora present greater opportunities in terms of quantity and variety. On a par with these possibilities they present new challenges with respect to the variety of grammatical constructions used, freedom of language (as opposed to controlled vocabularies), and diversity in topics for the interpretation process. However, these factors increase the ambiguity in the recognition of structures in the text. Therefore, the language of representation for these structures and its components, operators and hypotheses is of paramount importance.

Although recognition structures are widely used, and many examples with different interpretations can be found, it is not so easy to find a specification of the language in which these linguistic structures are expressed, nor the formalization used to express the restrictions involved. Furthermore, these representations of structures have been focused more on human legibility than on machine interpretation, although computational systems need a formal form of representations to work. In fact, these systems use a formal representation, but this is implicit and has not been fully explained. For that reason, sharing the structures, defined following a specific representation, is not a trivial issue.

In this paper we present a well defined proposal of formal representation to express linguistic structures of recognition. For the purpose of this work, we have named them "linguistic schemas", in which the meaning of all the elements appearing in the structures is made explicit. Moreover, a formal representation of these linguistic schemas, which is also interpretable by a computational application, is specified. The main aim is to provide these recognition structures with the capability of being reused and shared by different tools and systems, and to allow this formal representation to be explicit, well defined and computationally interpretable. This proposal aims at solving the complex problem of expressiveness in linguistic structures for NLP.

Thus, section 2 presents the representation specifications of linguistic structures. Section 3 offers a view of the linguistic scenario in which we can find the need for these new linguistic structures. The representation of the linguistic schemas is presented in section 4 and they are exemplified. Section 5 analyzes the expressiveness of the existing recognition structures comparing them with the new one developed and presents the results and future work. Finally references are also included.

2 Linguistic structures in use

Linguistic patterns are used in Computational Linguistics to understand natural language texts. Among the most outstanding projects it is worth noting the program PHRAN (PHRasal Analysis) [2, 16], which tackles the implementation of an approach based on knowledge. PHRAN deals with pattern-concept pairs (PCPs), whose linguistic components are phrasal patterns that may present different abstraction levels. This means that the pattern may be composed by a word, a literal string, as "Digital Equipment Corporation" or a general phrase as "<component> <send> <data> to <component>", enabling any object with the semantic category "component" to appear in the first and last position, any verbal form of "send" to appear in the second position, the word "to", in the fourth position, etc. There is also a conceptual template associated to each phrasal pattern, in which the meaning of the phrasal pattern is described.

In the field of information acquisition from machine readable dictionaries (MRDs), Hearst [5] developed a set of lexical-syntactic patterns restricted to identifying hyponymy relations in texts. Kim and Moldovan [7] created the *FP-structures* (*Frame-Phrasal pattern structure*), which are pairs composed by a frame of meaning and a phrasal pattern, as the one used in PALKA (*Parallel Automatic Linguistic Knowledge Acquisition System*).

More recently, the development of systems for automatic knowledge extraction has generated a substantial amount of works focused both on representations and systems. A detailed analysis can be found in the compilatory study by [17].

All in all, the lexical-syntactic patterns are generally expressed by means of operators in the *Backus-Naus Form* (BNF) in order to compose regular expressions in context-free grammars. Jacobs *et al.* [6] make this explicit when they take the following operators to express lexical-semantic patterns:

Lexical features that can be tested in a pattern: token "name" (ej. “*AK-4T*”), root (ej. “*shoot*”), lexical category (ej. “*adj.*”)
Variable assignment from pattern components: ?X =
Logical combination of lexical feature tests: OR, AND, NOT
Wild cards: $ **-** 0 or 1 token, *** -** 0 or more tokens, **+ -** 1 or more tokens
Grouping operators: <> for grouping, [] for disjunctive grouping
Repetition: * - 0 or more, + - 1 or more
Range: *N - 0 to N, +N - 1 to N
Optional constituents: {} - optional

Linguistic patterns, be they lexical-syntactic, semantic or, as in the case of PALKA, structures of phrase frames, are always ordered sets of components that express characteristics or constraints on the phrase elements. In every case, the phrase element order and the pattern component order will be the same, even if not explicitly indicated, as all of them are patterns for English, a language with a strict phrase order [12] compared to other Romance languages, for instance. However, when the texts processed by the system are written in a natural language without these constraints, these patterns, which are equivalent to regular expressions, do not fulfill the objectives; then, a wider representation enabling not to specify the order in which the phrase elements should appear, is required. Therefore one of these wider patterns will match the same phrases as a set of ordered patterns, which correspond to different permutations of the same pattern components.

This problem is partially solved by Hazez [4], as he takes morphemes, words, grammatical categories or a syntactic pattern as linguistic patterns. These linguistic patterns are managed as segments to which certain set operators, such as union and intersection, and other operators that express position and content are applied.

Linguistic patterns based on annotations can be found in other cases, as in Specia and Motta’s work [14], but the annotations used are always simplified. Thus, in the following example taken from Specia and Motta, based on the relation extraction between phrasal components, and performed by the system Minipar [8], everything is simplified to a triplet over which the patterns are established: <noun_phrase, verbal_exp., noun_phrase>. In this same line, syntactic patterns are applied to disambiguate [11]. Table 1 contains a comparison of the pattern features in these approaches.

Table 1. Comparison of pattern features

Phrasal Pattern	Lexical-syntactic pattern	(Hazez)	(Specia and Motta)
Elements			
literal string, general phrase, semantic category identifier	token "name", root, lexical category, conceptual category, variable	variables, morphemes, words, grammatical categories, linguistic pattern	triplet of token annotations
Operators			
order (in general phrase)	OR, AND, NOT, $, *, +, <>, [], *N, +N, { }, order (in sintaxis)	set operators (∪,∩, etc.), position, content (⊃,⊂, etc.)	
Hypothesis			
	closed world		search triplets

3 Linguistic scenario

The purpose of this work is the understanding of Spanish texts annotated electronically by software tools. In order to enable the automatic application of these patterns to large scale corpora, we have established some constraints over the phrases in several levels, specifically in orthographical, morpho-syntactic, and syntactic levels. However, the possibility of including annotations of any other level (such as semantic, pragmatic or discursive) remains open.

As for the works about annotation and creation of linguistic patterns to extract information from texts in Spanish, the initiatives grow in number and importance as the multilinguality significance increases in the Internet. In the framework of the European project SEKT[1], one of the use cases was focused on the Spanish legal terminology for the creation of ontologies in the legal domain. For this task Hearst's taxonomic relation patterns [5] were translated into Spanish and new patterns were created with the purpose of using the knowledge obtained to enrich ontologies [15].

Related with knowledge extraction for ontology enrichment and population in Spanish we can find another classification attempt in Álvarez de Mon y Rego y Aguado de Cea [1]. These authors extended Hearst's patterns by focusing on certain patterns with classification verbs such as *clasificar*, *figurar*, *distinguir* or *dividir*, that allow a more complete extraction of concepts hierarchically related.

Nica's *et al.* [11] work about desambiguation has been also applied to Spanish for extracting syntactical-semantic patterns (formalizations of the argument-predicate structure related with a verb) from an annotated corpus [10].

We decided to represent linguistic structures in XML format to work computationally with these structures in an easier way as XML is the language most widely used for knowledge representation and many tools can process it. However, files in XML cannot be easily read by humans because of the verbosity of its syntax.

---

[1] http://www.sekt-project.com/

4 Linguistic schemas

A linguistic schema is a set of constraints over the tokens of a phrase (token contraints) and over the relations between these tokens (phrase constraints). Token constraints are expressed as a set of values of characteristics of annotations of a token. Phrase constraints are expressed using operators (optimality, grouping, etc.) over token constraints or other phrase constraints.

As previously stated, the complete representation of the schemas is stored in XML files for an easier computational processing. Although these files can be read by a person, this task is rather tedious and can be untractable if the number or size of the schemas grows significantly.

For this reason a shortened and user-friendly annotation is defined. This annotation may serve as a mnemonic of the schemas that appear in a file. It does not comply with the XML conventions and may not contain all the information available in the schema. However, the annotation is much easier to read, and, if used correctly, it may identify the schema that is referred to without any short of ambiguity.

Furthermore, this notation has been extended with additional operators, which are not present in the XML notation, to increase the expressiveness and improve the shortness. These operators are replaced by combinations of the operators available in the XML notation. As an example, the optionality operator (see section 4.2) applied to a token would be replaced with a disjunction between this token and the negation of the same token.

A brief summary of the notation proposed (for a friendly representation) is:

Token constraints (Elements): constant (ej. "*shirt*"), identifier (ej. "*ANIMAL*")

Phrase constraints (Operators):

Order operators: A ⊕ B – A appears before B, A + B – A appears immediately before B

Disjunction operators: A | B – A and B can appears, A / B – A or B can appear

Grouping operator: ( ) – group

Repetition operator: * - 1 or more times

Negation operator: ¬ A – A doesn't appear

Optionality operator: [ ] – optional

General hypothesis: Open world

4.1 Terms

For the purposes of this work, a term in a linguistic schema is the set of constraints, in other words, the set of elements applied to one single token. In the user-friendly syntax, these terms may be displayed with two different types of symbols: constants and identifiers.

- Constants are words written as they appear in the text, for example *clasifica*.
- Identifiers are used to retrieve values instead of restricting them, and they appear as strings in uppercase, for example "ACTOR".

Terms with identifier and/or lemma

In those cases in which an identifier ("ACTOR") or a lemma (*clasificar*) is specified this will be shown in the set of constraints of a token.. For example, when a token has as a constraint the lemma *clasificar*, and its morpho-syntactic value is “main verb”, only the lemma will be shown. If both data about the same term are used in the information, then the identifier will be shown. For example, when a token has as a constraint the lemma “*clasificar*” and as text the identifier “CONJUGATED_FORM”, then only the identifier “CONJUGATED_FORM” will be shown. If a set of identifiers is specified for a token, then the identifier whose value has previously appeared will be used, according to the annotation standard used. For instance, if two identifiers are assigned to a token, such as the values of gender (“GENDER”) and syntactic function (“FUNCTION”), only the former will be shown, i.e., “GENDER”.

It is possible to use the identifier to refer to any of the non constant terms. For instance, the next schema can be written using the identifiers A, B, C and Z:

```
A + come + B + y + C + en + Z
```

This schema would match a phrase such as “*Pepe come pan y chocolate en el patio de la escuela*”, and in this matching the identifiers will take the values corresponding to this specific phrase: *A=Pepe*, *B=pan*, *C=chocolate*, *Z=patio*.

Terms with the category specified in any annotation level

For those terms for which no identifier or lemma are specified, but the value of, at least, a category in an annotation level is defined, the name of that category will be shown. Taking as reference the previous example, the values “verb” or “direct object” will appear instead of “come” and “B”.

If an abbreviated form is specified for any category in the standard used[2] and possibly with information about additional attributes (for example “Fused_Prep-At” for “Fused Preposition-Article”), then the most specific abbreviated form will be shown for each annotation level, being the most specific form the one that includes more information about the additional attributes. Thus, when we want to identify a token that is an ordinal pronoun, but of which we do not want to obtain any other information, its lemma or value for any other category (as in the phrase “el primero es el grande”) is described as “Ordinal_pronoun” as we only want to restrict this word to this type of pronoun.

4.2 Operators

As previously mentioned, operators define the relations among the different parts of a schema. It is necessary to point out that the order in which the parts of the phrase must appear is not specified by the element appearance order in the schema; therefore, if it is necessary to set this order, then it must be specified explicitly. This can be done with two symbols:

- With the symbol ‘+’: the expression “*symbol1* + *symbol2*” means that “*symbol2*” must appear immediately after “*symbol1*”.

---

[2] http://pln.oeg-upm.net/annotation/ontotag

- With the symbol ‘⊕’: the expression “*symbol1* ⊕ *symbol2*” means that “*symbol2*” must appear after “*symbol1*”, immediately or not.

There are other symbols besides the previous ones which express different relations. These symbols are the following:

- **‘*’**: expresses repetition.
  For example, “*symbol**” means that “*symbol*” may appear more than once.
- **‘(’** and **‘)’**: groups several symbols.
  For example, “(*symbol1* + *symbol2*)*” means that “*symbol1*” may appear several times, all of them followed by “*symbol2*”.
- **‘[’** and **‘]’**: means that whatever is between both square brackets is optional.
  For example, “[*symbol*]” means that “*symbol*” may appear or not in the phrase.
- **‘|’**: means that either what is in the left side or what is in the right side must appear.
  For example “*a*|*an*” means that “*a*” or “*an*” must appear.
- **‘/’**: means that either what is on the left side or what is on the right side must appear, but not both of them.
  For example “*a/an*” means that “*a*” or “*an*” must appear, but not “*a*” and “*an*” at the same time.
- **‘¬’**: means that the next element must not appear in the phrase. When combined with the symbols + and ⊕, it may indicate that the said symbol must not appear in some specific positions of the phrase.
  For example, “¬*symbol*” means that “*symbol*” may not appear in the phrase.

Examples of linguistic schemas

To show the versatility and possibilities of the linguistic schemas we include some examples, expressed in the user-friendly notation.

We want to identify who buys things to María, and which those things are. Hence, we express these constraints in a linguistic schema setting the main verb (“*compra*”) and the indirect object (“*a María*”). The rest of the phrase and the order of appearance are not restricted. These constraints may be expressed with the next linguistic schema:

```
X compra Y a + María
```

This schema would match phrases like “*Pepe compra a María flores*”, “*Pepe a María flores compra*”, “*Pepe compra flores a María en domingo*”, “*A María Pepe le compra flores*” and “*Juan a María compra bombones de licor en Santander*”.

It is worth mentioning that the previous schema would be equivalent to the next one, since linguistic schemas have no implicit order, as it happens in the case of lexical-syntactic patterns. Also, the name assigned to the identifier does not change the recognition capabilities of a linguistic schema:

```
SOMEONE a + María SOMETHING compra
```

This schema would match with exactly the same phrases as the previous one. However, it would take 24 lexical-syntactic patterns (P(4,4) = 4! = 24) to match the same phrases using patterns, as there are four pattern components in the previous example, (1) SOMEONE, (2) a+María, (3) SOMETHING and (4) compra. Moreover, these patterns could also have additional elements in the phrase, resulting in a larger list of lexical-syntactic patterns.

Because of this combinatorial explosion and the open world assumption, processing a schema requires more computational power than a pattern. However, our proposal for a schema represents a set of patterns in a more compact way, enabling a further optimization and more efficient algorithms.

The application of these linguistic schemas to Spanish does not mean that they cannot be used for other languages. For the lexical-syntactic pattern

```
X buys Y for María
```

the equivalent linguistic schema would be:

```
X + buys + Y + for + María
```

An example can be seen in pln.oeg-upm.net/process/linguisticschemas.

5 Comparison and discussion

Once we have described and exemplified the notation proposed, we will compare the expressiveness of our notation with the lexical-syntactic pattern notation, accepted by Jacobs *et al.*[6].

The first point is that the notation we propose assumes the open world assumption. This assumption means that everything that is not described in the schema is not restricted, thus, it can appear or not.

Table 2. Comparison of Lexical-syntactic patterns and Linguistic schemas

Lexical-syntactic patterns	**Linguistic schemas**
Lexical features:	
token "name"	text value
Root	lemma value
lexical category	values of these categories
conceptual category	
Combination of lexical features:	
OR	operator '|'
AND	implicit
NOT	operator '¬'
Wild cards:	
$, *, +	These operators are unnecessary taking into account the open world hypothesis
Variable assignment from pattern components:	
?X =	identifiers
Grouping operators:	
<>	'( )'
[]	combination of '( )' and '/'
Repetition:	
*	combination of '[ ]' and '*' inside

+	‘*’
Range:	
*N, +N	Extensional representation
Optional constituents:	
`{}`	‘[ ]’

In the Table 2, we can see the notation with the regular expressions used by Jacobs *et al.* (left column) and the correspondences to our notation (right column).

In the case of range, the term “Extensional representation” involves iterating n times the term optionally. That is, it can be represented, but it does not have an operator or a sign that compresses this expression.

Besides covering completely the expressiveness of the previous notation, the new notation contributes the following functionalities:

- It takes into account the values of all the characteristics of the annotations (not only lexical and conceptual categories).
- It includes identifiers to be used in any value of the annotation (not only the four lexical characters provided by Jacobs).
- It includes operators of exclusive disjointness ‘/’ in and out of the group.
- It includes operators of order ‘+’ and ‘⊕’.
- It allows applying these operators to sub-schemas (not only to the four lexical characters dealt in Jacobs’).
- The open world assumption allows ignoring which tokens can appear or not in phrases. For this reason wild cards are not necessary.

We consider that this comparison shows that any lexical-syntactic pattern expressed in traditional notation can be expressed in terms of the notation proposed as a linguistic schema.

This representation permits, first, describing constraints about text annotations and, second, dealing with other previously created linguistic structures.

With these characteristics we have designed linguistic structures which describe phrases in languages that do not have a rigid morpho-syntactic order, such as Spanish.

As future work, the implementation of an assistant (already designed) for editing schemas will make linguists work easier and will contribute to a greater automatization.

The assistant should allow the definition of many schemas comfortably. Presumably, this combination of quality and quantity should allow a greater automation of NLP tasks, improving the results when processing large scale corpora.

7 Acknowledgments

This work has been supported by the project *BabeLData* (TIN-2010-17550), funded by *INIA*, under *Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica* (I+D+i) of *Ministerio de Ciencia e Innovación.*

8 References

[1] Álvarez de Mon y Rego I, Aguado de Cea G (2006) The phraseology of classification in Spanish: integrating corpus linguistics and ontological approaches for knowledge extraction. BAAL/IRAAL Joint Int. Conf., Ireland.

[2] Arens Y (1986) *CLUSTER: An approach to Contextual Language Understanding*. Ph.D thesis, Univ. of California at Berkley, 1986.

[3] Baeza-Yates R, Ribiero-Neto B (1999) *Modern information retrieval*. Addison Wesley Longman, Essex, England.

[4] Hazez SB (2001) Linguistic pattern-matching with contextual constraint rules. *IEEE International Conference on Systems, Man, and Cybernetics*. Vol. 2. Pages: 971-976. Tucson, AZ, USA, October 7th-10th, 2001.

[5] Hearst MA (1992) Automatic Acquisition of Hyponyms from Large Text Corpora. *COLING-92.* Nantes, 23-28 August, 1992.

[6] Jacobs PS, Krupka GR, Rau LF (1991) Lexico-semantic pattern matching as a companion to parsing in text understanding. *In Fourth DARPA Speech and Natural Language Workshop*, pp. 337-342, 1991.

[7] Kim JT, Moldovan DI (1993) Acquisition os Semantic Patterns for Information Extraction from Corpora. *Ninth Conf. AI applications*, 1993.

[8] Lin D (1993) Principle based parsing without overgeneration. *31st ACL*, Columbus, pp. 112-120. 1993.

[9] Mann T (1993) *Library research models*. Oxford University Press, NY.

[10] Navarro B, Moreno-Monteagudo L, Martínez-barco P (2006) Extracción de relaciones sintagmáticas de corpus anotados. *Procesamiento de Lenguaje Natural*, ed. SEPLN, nº 37, septiembre 2006, pp: 59-66

[11] Nica I, Martí NA, Montoyo A, Vázquez S (2004) Intensive Use of Lexicon and Corpus for WSD. *Procesamiento de Lenguaje Natural*, ed. SEPLN, nº 33, septiembre 2004, pp: 147-154

[12] Quirk R, Greenbaum S (1977) *A University Grammar of English,* London, Longman.

[13] Sebastiani F (2002) Machine learning in automated text categorization. ACM Computing Surveys, Vol. 34, No. 1, pp. 1–47. 2002.

[14] Specia L, Motta E (2006) A hybrid approach for extracting semantic relations from texts. *2nd Workshop on Ontology Learning and Population en COLING/AC*L 2006. Sydney, Australia. July 22nd, 2006.

[15] Völker J, Vrandečići D, Sure Y (2006) SEKT Project *D3.3.3 Data-driven Change Discovery.* SEKT Project.

[16] Wilensky R, Arens Y (1980) PHRAN: A Knowledge-Based Natural Language Understander. *18th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA. June 1980.

[17] Zhou N, Zhou X (2004) Automatic Acquisition of Linguistic Patterns for Conceptual Modeling. *Course "INFO629: Concepts in Artificial Intelligence*". Drexel University, Fall 2004.

Integrating Large Knowledge Repositories in Multiagent Ontologies

Herlina Jayadianti, Carlos Sousa Pinto, Lukito Edi Nugroho, Paulus Insap Santosa

herlinajayadianti@gmail.com
csp@dsi.uminho.pt
lukito@mti.ugm.ac.id
Insap@mti.ugm.ac.id

Abstract. Knowledge is people's personal map of the world. According to the knowledge differences, it is possible different groups of people have different perceptions about the same reality. Each perception can be represented by using ontologies. In the research underlying this paper we are dealing with a multiple ontologies. In that context, each agent explores its own ontology. The goal of this research is to generate a common ontology including a common set of terms, based on the several ontologies available, in order to make possible to share the common terminology (set of terms) that it implements, between different communities. In this paper we are presenting a real implementation of a system using those concepts. The paper provides a case study involving groups of people in different communities, managing data using different perceptions (terminologies), and different semantics to represent the same reality. Each user – belonging to a different community - uses different terminologies in collecting data and as a consequence they also get different results of that exercise. It is not a problem if the different results are used inside each community. The problem occurs if people need to take data from other communities, sharing, collaborating and using it to get a more global solution.

Keywords. Heterogeneity, Agents, SPARQL, Ontology alignment, Common ontology.

1 Introduction

In information technology, a repository is a central place in which an aggregation of data is kept and maintained in an organized way. Repository is a place where things are collected. Depending on how the term is used, a repository may be directly accessible to users or may be a place from which specific databases, files, or documents are obtained for further relocation or distribution in a network. As an example scenario, institution A, institution B and institution C are working in a domain D. Repositories which contain information about that domain can be scattered in different places. One of the main problems that we can find in such a scenario is related to the existence of different perceptions and to the use of different representations and terms in each repository in each institution. Our problem is how to combine different repositories from different institutions and how to manage knowledge between these different repositories. Heterogeneity in data, in semantic and in perception between each institution is the major problem we need to solve. We use ontologies to solve those problems. Using ontologies we can shared different conceptualizations, different terminologies, and different meanings between systems [18]. However, tasks on distributed and heterogeneous systems demands support from more than one ontology.

We can distingue four types of heterogeneity [1]: (1) Paradigm heterogeneity that occurs if distinct agents express their knowledge using different modelling paradigms; (2) Language heterogeneity which occurs if distinct agents express their knowledge in different representation languages; (3) Ontology heterogeneity that occurs if distinct agents make different ontological assumptions about their domain of knowledge; (4) Content heterogeneity which occurs if distinct agents express different knowledge the same reality. Ontology integration [4], [9-12] is one way to solve the problem of heterogeneity and it can be done using several approaches. For example, ontology merging, ontology matching or ontology alignment. The integration of ontologies creates a new ontology by reusing other available ontologies through assembling, extending, or specializing operations. In integration processes the source ontologies and the resultant ontology can have different amounts of information [2]. We need to map ontologies in order to make compatible different terminologies (sets of terms). While having some common ground, either within an application area or for some high-level general concepts, this could alleviate the problem of data and semantic heterogeneity [5].

Ontology alignment or ontology matching [3], [13], [14] is the process of determining correspondence between concepts. Given two ontologies $i = (C_i, R_i, I_i, A_i)$ and $j = (C_j, R_j, I_j, A_j)$, we can define different types of (inter ontology) relationships among their terms. If two ontologies have at least one common component (relation, hierarchy, type, etc.) then they may be compared. Since the characteristics (attributes) of concepts capture the details of those concepts, they provide a good opportunity to find similarities [1].

In this paper we describe an approach to solve the problem of data and semantic heterogeneity using a common ontology derived from several different ontologies, using an ontology alignment process. This paper is organized as follows: (1) Introduction; (2) In this section we present several definitions of the terms used in operations involving ontologies, in order to avoid possible misunderstandings; (3) In this section we present the case study that underlies the work described in the paper; (4) This section describes the implementation of the proposed solution; (5) In this section we refer the used technologies and preliminary results of our work ; and (6) the paper ends with the Conclusions.

2 Operations Involving Ontologies – Used Terminology

To avoid potential misunderstandings, we present the definitions of the terms used throughout this paper.

- **Ontology Combination** is the process of using two or more ontologies and can be used to implement alignment, merge or integration of different ontologies. The combined ontologies usually hold data which is relevant to all ontologies involved.[6], [7]
- **Ontology Merging** is the process of building a single ontology through the merging of several source ontologies. Usually the source ontologies cover similar or overlapping domains. [8]
- **Ontology Alignment** is the process of determining correspondence between concepts and the process of creating a new ontology from two or more ontologies by overlapping the common parts. The domains of the source ontologies are different from the domain of the resulting ontology, but there is a relation between these domains. [3], [13], [14]
- **Ontology Matching** is the process of reaching global compatibility between two or more ontologies so that the resulting ontology is consistent and coherent. [3]
- **Ontology Mapping** is the process of relating similar concepts or relations from different sources through some equivalence relation. Mapping allows finding correspondences between the concepts of two ontologies. If two concepts correspond, then they mean the same thing or closely related things. Currently, the mapping process is regarded as a promise to solve the problem between ontologies since it attempts to find correspondences between semantically related entities that belong to different ontologies. It takes as input two ontologies, each consisting of a set of components (classes, instances, properties, rules and axioms). [15], [16], [17]

3 Heterogeneity And Interoperability Problems

In this section, we describe the problem we are trying to solve and an approach to solve it. Considering some reality, different groups of people (different communities) have different opinions, use different sets of data about it and have diverse perceptions about that reality. Figure 1 represents several communities that faced reality with different perceptions *(Perception_1, Perception_2, and Perception_N).* Perceptions are converted into data that is saved into separate storage devices not interconnected. Repositories *db1, db2, and dbN* contain different data, different concepts, different terms, and different semantics. It depends on people in the group who look at reality (policy makers) and people who create and store data (users that use technology). Users who deal with computers has a very important role in controlling and changing the terminology and semantic of the data. Each group (community) uses technology to find data. It is very difficult for those different groups to get similar results and the problem happens if people need to use data from another group in order to share, collaborate and use it to get a more global solution.

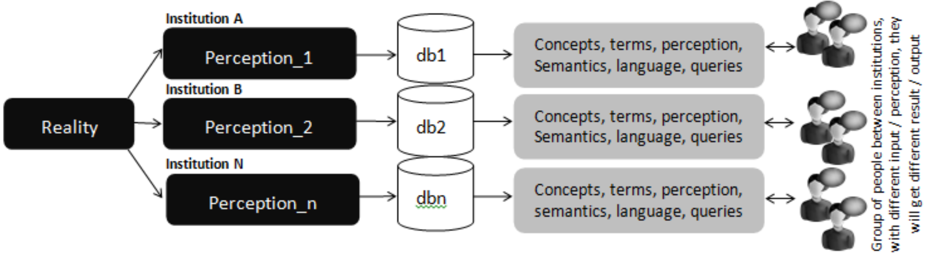

Institution A
Perception_1
db1
Concepts, terms, perception, Semantics, language, queries
Institution B
Reality
Perception_2
db2
Concepts, terms, perception, Semantics, language, queries
Institution N
Perception_n
dbn
Concepts, terms, perception, semantics, language, queries
Group of people between institutions, with different input / perception, they will get different result / output

Fig. 1. The Problem of Different Perceptions

The solution presented in this paper is based on different knowledge about the same reality based on different perceptions and uses a mechanism that works with a set of common concepts, common terms, common semantics, common languages, and a set of common queries (See Figure 2). Users in each community still can use their different concepts, terms, and perceptions as inputs for querying the system. According to the proposed solution, we aim to get similar answers (output) from such a common layer that acts like an interface between the different systems and the users.

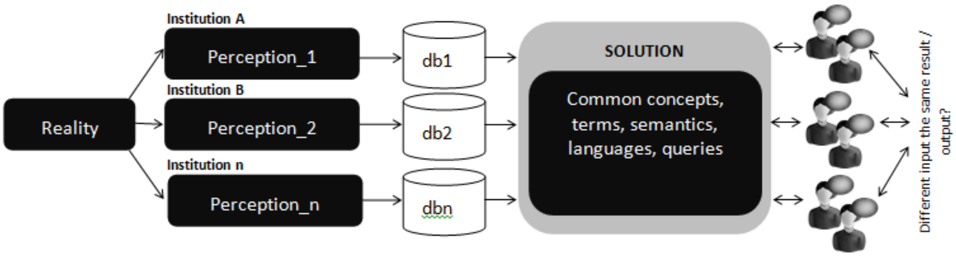


Fig. 2. Towards a Solution of Different Perceptions

4 Using Ontologies to implement the solution

Ontology is defined as a formal, explicit specification of a shared conceptualization [18]. Tasks on distributed and heterogeneous systems demand support from more than one ontology. Multiple ontologies need to be accessed by different systems. Different perceptions about the same reality led to dissimilar ontologies for the same domain. Thus, various organisms with different ontologies do not fully understand each other. To solve this problem, it is necessary to use ontology alignment geared for interoperability.

4.1 Ontology Alignment

Ontology Alignment [13], [14] is the process of creating a new ontology from two or more ontologies by overlapping common parts and determining correspondences between ontology entities. Entities of the source ontologies are different from entities of the resulting ontology, but there is a relation between these entities. Based on the fundamental concepts above and on Figure 2, the solution for solve the problem is to use ontology alignment (see Figure 3) to create a new ontology (a common ontology) by overlapping the common parts of the original ontologies. Common part is a common word recognized and used with the same meaning by different communities. CO (Common Ontology) is expected to overcome the differences that exist in the different source ontologies. In Figure 3 we use ontology UV_1 from institution A, UV_2 from institution B, and ontology UV_n from institution $_N$. CO will contain terms that will be equated with each term in the source UVs.

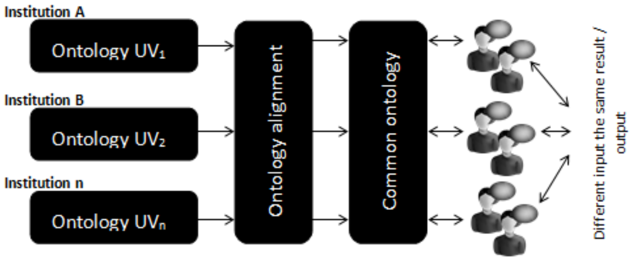


Fig. 3. Ontology Alignment

4.2 Dictionary and Search engine analysis

To get the CO terms we analyzed several dictionary such as WordNet[1] and Thesaurus[2] (See Table 1).

---

[1] Wordnet is a large lexical database or electronic dictionary for English. WordNet implements measure of similarity and relatedness among terms. Measures of similarity use information found in an *is–a* hierarchy of concepts, and quantify how much concept A is similar to concept B. http://wordnet.princeton.edu/

[2] Thesaurus is a reference work that lists words grouped together according to similarity of meaning (Synonym or antonym). http://en.wikipedia.org/wiki/Thesaurus/ and http://thesaurus.com/

Search string	*Synonym*	
	Wordnet 2.1	Thesaurus
People	Group, Family, Masses, Mass, Family Line	Citizens Community, Family, Folk, Folks, General Public, Heads, Persons, Population, Society
Person	Individual, Someone, Somebody	Human, Identity, Individual, Individuality

Table 1. Synonym results found by Wordnet and Thesaurus using “People“ and “Person“ as search string

There are four senses for the term people in Wordnet (version 2.1).

Sense 1 *people -- ((plural) any group of human beings (men or women or children) collectively) => group, grouping*

Sense 2 *citizenry, people -- (the body of citizens of a state or country) => group, grouping*

Sense 3 *people -- (members of a family line; "his people have been farmers for generations) => family, family line, folk*

Semantic Web Search Engines such as Swoogle[3], Watson[4], and Sindice[5] (See Table 2) accept queries in a format that varies from one tool to another.

Search string	Semantic Search Engine					
	Swoogle		Watson		Sindice	
	Number of references	Time	Number of Terms	Time	Number of references	Time
People	1,818	0.456	**12,348**	0	**12,709,732**	2.19
Person	**16,320**	0.237	3,046	0	77,724,899	0.04
Group	3,812	0,381	3,742	0	690,570	0.68
Family	2,209	0,565	7,326	0	7,081,244	2.24
Individual	1,010	0,469	854	0	237,692	2.05

Table 2. Different results found by several search engines using “People“, “Person“, “Group“, “Family“ and “Individual“ as search strings

Different from other types of platforms that can be used to find suitable ontologies, which usually only provide browse functionalities, Semantic Web Search Engines (SWSE) permit another degree of automation. For instance, a query on Sindice for ontologies including the term “People”, returned more than 12.699.661 results in 2.72 second, where near 4.568.172 documents (0.03 second) of them were RDF files. Data from Table 2 was taken on June 20, 2012.

4.3 A Case Study

To demonstrate the capabilities of the described mechanisms we implemented an alignment process between original ontologies using data about poverty. Poverty is not the focus of our research. We just use that case as a real scenario that allows us to demonstrate the validity of our approach. We combine different existing terminolo-

---

[3] Swoogle is the first Web search engine dedicated to online semantic data. Its development was partially supported by DARPA and NFS (National Science Foundation). http://swoogle.umbc.edu/

[4] Watson development was partially supported by the NeOn (http://www.neon-project.org) and the OpenKnowledge (http://www.openk.org) project. http://kmi-web05.open.ac.uk/WatsonWUI/

[5] http://sindice.com/

gies about the same reality (poverty in this case) used by different communities in order to get a common set of terms that can be transparently used by those communities, while maintaining the original terms in the data sources. We use Indonesia as the country for the example because in that country there are several institutions in charge of dealing with poverty data, generating problems due to differences in the criteria used by them to make their surveys, even considering that the semantics of these different criteria are the same. For example, let's consider the two institutions, BKKBN[6] (institution A) and BPS[7] (institution B), that are responsible for collecting data on poverty. Each institution has a different system and use different sets of terms to describe the same domain and different criteria to classify people as poor or not. In fact, institution A uses 24 criteria and institution B has 14 criteria to define poverty.

Institution A: *"Normally all family members have* ***meal*** *two or more times a day"*

Institution B: "*Minimum two times per day the family have* ***food****"*

Meal and *food* have the same meaning, as well as *suit* and *clothes* or *clinic* and *hospital.* To be similar *(≅)* or not equal *(≠)* depend on several factors, such as the programmer's interpretation, the needs of the system itself, and last but not least the domain/area that we are talking about. One term has always a strong relationship with the domain. In this research, we focus on poverty domain, identifying terms that are most commonly used by users.

Table 3 shows some examples of criteria and terms in the domain of poverty from two different institutions. Currently, both institutions are working separately to collect and manage data on poverty. Each institution sends data to the government based on its perception. Institution A (BKKBN) is more focused on family welfare and institution B (BPS) is more concerned with basic needs. The major problem of this situation is the great impact on aid distribution.

	Criteria from Institution A	Criteria from Institution B
Classes	Area, Assets, Contraceptive, Education, FoodConsume, GovernmentAid, Hospital, HealthProblem, HouseCondition, Person	Asset, BirthControlMethod, EducationLevel, Food, GeographicArea, GovernmentHelp, HealthCondition, Clinic, HouseParameter, JobArea, Person.
Object Properties	isComposedBy, hasFrequentlyEat, PassTheStudyFrom, hasRarelyEat, has Assets, hasChildren, hasfamily, hasHouseCondition, hasJobPositionAs	EnergyUsedForCooking, hasEduBackground, hasFrequentlyEaten, hasLargestFloorMadeFrom, hasRarelyEaten,
Data properties	Address, has Age, FrequentlyEatenADay, hasMarriageStatus, hasSalary, hasaGoodHouseCondition	hasAge, DistrictCode, FloorArea, FullName, HouseCondition, JobsArea, NameOfFood, FloorArea, Salary ≈ hasWage, hasStatus.

Table 3. Example of Classes, Object Properties, and Data properties From Two Institutions

---

[6] Badan Keluarga Berencana Nasional (BKKBN) or National Population and Family Planning Board is a governmental agencies that appointed to conduct a survey of poverty in Indonesia. http://www.**bkkbn**.go.id

[7] Badan Pusat Statistik (BPS) or Central Berau of Statistic is a non departmental government institution directly responsible to the President of Indonesia. http://www.**bps**.go.id

Based on the criteria of both institution (see Table 1), we identify an example of Classes, ObjectProperties, and DataProperties to be used by institutions A and B (see Table 3). We can see that:

- **Terms (classes) in Ontology UV$_1$** = {Area, Assets, Contraceptive, Education, FoodConsume, GovernmentAid, Hospital , HealthProblem, HouseCondition, Person}
- **Terms (classes) in Ontology UV$_2$** = {Asset, BirthControlMethod, EducationLevel, Food, GeographicArea, GovernmentHelp, HealthCondition, Clinic, HouseParameter, JobArea, Person}.

By using WordNet, Thesaurus, and Swoogle, we identify common classes in CO, namely *People*, *Birth Control*, *Education*, *Food*, *Health*, *Property*, *Work*, *Hospital*, and *House Condition*. On the next stage, by overlapping the common parts, we determine the correspondence between classes in Ontology UV$_1$ (User view 1) and classes in ontology UV$_2$ (User view 2) with classes in CO. Figure 4, automatically generated in Protégé[8], show the relation between CO and UVs.

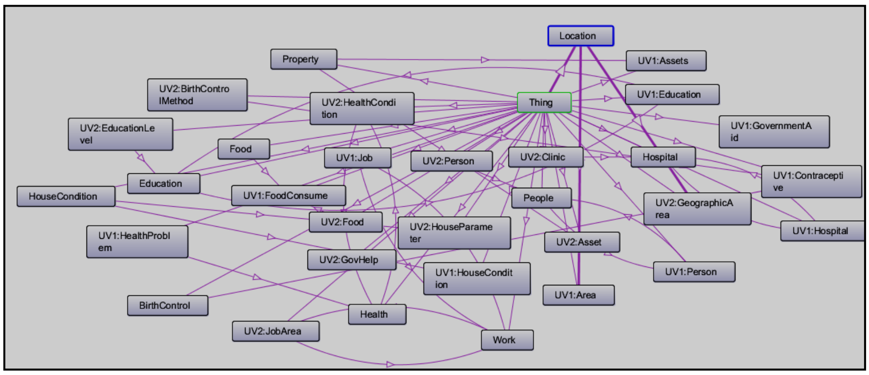


Fig. 4. The relation between UV's and CO

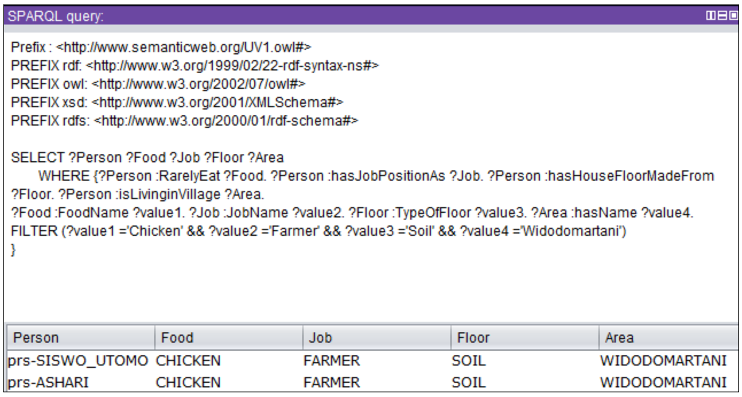
5 Used Technologies and Preliminary Results

Web Ontology Language (OWL) is a language for create ontologies to the web. OWL was designed for processing information and to provide a common way to process the content of web information. SPARQL[9] is a graph-matching query language. SPARQL can be used to express queries across diverse data sources. In Figures 5-7 we can see examples of the results of SPARQL queries. Based on Figure 5 we can see that Ontology UV$_1$ (data taken form Institution A) consists of classes Person, Food, Job, Floor and Area. UV$_1$ also includes the object properties "RarelyEat" (Chicken instance), "JobName" (Farmer instance) and TypeOfFloor (Soil instance). With SPARQL we get as result from UV$_1$ two people included in these criteria.

---

[8] http://protegewiki.stanford.edu/wiki/Protege4GettingStarted

[9] http://www.w3.org/TR/rdf-sparql-query/

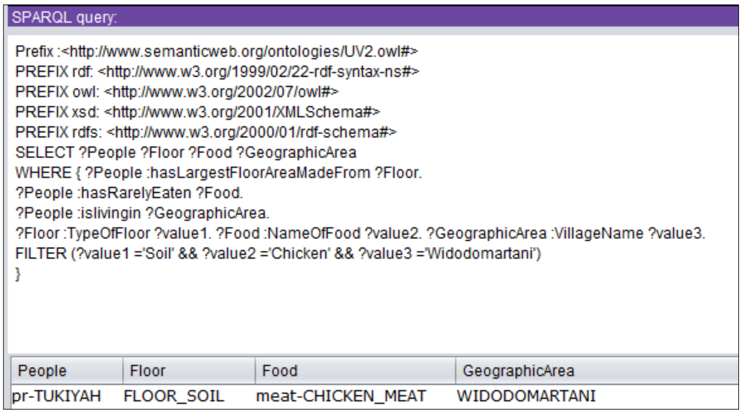


```
Prefix : <http://www.semanticweb.org/UV1.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?Person ?Food ?Job ?Floor ?Area
     WHERE {?Person :RarelyEat ?Food. ?Person :hasJobPositionAs ?Job. ?Person :hasHouseFloorMadeFrom
?Floor. ?Person :isLivinginVillage ?Area.
?Food :FoodName ?value1. ?Job :JobName ?value2. ?Floor :TypeOfFloor ?value3. ?Area :hasName ?value4.
FILTER (?value1 ='Chicken' && ?value2 ='Farmer' && ?value3 ='Soil' && ?value4 ='Widodomartani')
}
```

Person	Food	Job	Floor	Area
prs-SISWO_UTOMO	CHICKEN	FARMER	SOIL	WIDODOMARTANI
prs-ASHARI	CHICKEN	FARMER	SOIL	WIDODOMARTANI

Fig. 5. SPARQL result using UV_1



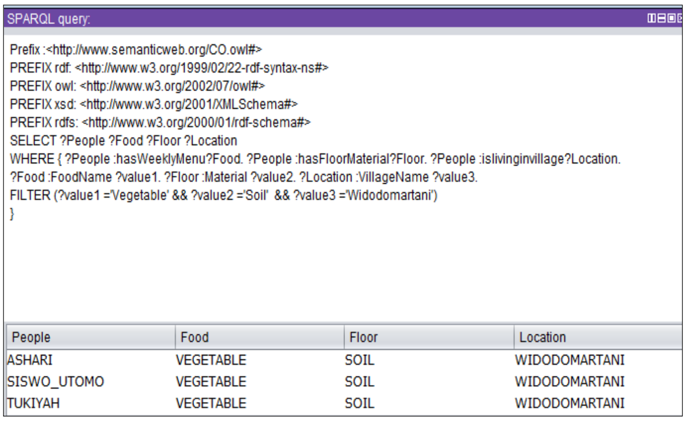
```
Prefix :<http://www.semanticweb.org/ontologies/UV2.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?People ?Floor ?Food ?GeographicArea
WHERE { ?People :hasLargestFloorAreaMadeFrom ?Floor.
?People :hasRarelyEaten ?Food.
?People :islivingin ?GeographicArea.
?Floor :TypeOfFloor ?value1. ?Food :NameOfFood ?value2. ?GeographicArea :VillageName ?value3.
FILTER (?value1 ='Soil' && ?value2 ='Chicken' && ?value3 ='Widodomartani')
}
```

People	Floor	Food	GeographicArea
pr-TUKIYAH	FLOOR_SOIL	meat-CHICKEN_MEAT	WIDODOMARTANI

Fig. 6. SPARQL result using UV_2

As we can see in Figure 6 ontology UV_2 (data taken from Institution B) consists of classes Person, Food, GeographicArea, and Floor (subclass of class House Condition) and also consists of object properties hasRarelyEaten (Chicken instance), isLivingIn (Widodomartani instance) and hasLArgestFloorAreaMadeFrom (Soil instance). Using SPARQL we get as result from UV_2 one person included in these criteria. It should be highlighted that poverty data in UV_1 and UV_2 was taken from the same village, Widodomartani. Based on the criteria used by Institution A and Institution B, implemented in the ontologies UV1 and UV2, the results returned by SPARQL queries are: Siswo Utomo and Ashari are poor people considering the ontology UV1, and Tukiyah is a poor person when considered the ontology UV2.

With common term in CO (see Figure 7), we can see that Siswo Utomo, Ashari and Tukiyah are poor people. With ontology alignment we determine the correspondence among concepts and implement the process of creating a new ontology based on two ontologies (UV1 and UV2) by overlapping the common parts.



```
Prefix :<http://www.semanticweb.org/CO.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?People ?Food ?Floor ?Location
WHERE { ?People :hasWeeklyMenu?Food. ?People :hasFloorMaterial?Floor. ?People :islivinginvillage?Location.
?Food :FoodName ?value1. ?Floor :Material ?value2. ?Location :VillageName ?value3.
FILTER (?value1 ='Vegetable' && ?value2 ='Soil'  && ?value3 ='Widodomartani')
}
```

People	Food	Floor	Location
ASHARI	VEGETABLE	SOIL	WIDODOMARTANI
SISWO_UTOMO	VEGETABLE	SOIL	WIDODOMARTANI
TUKIYAH	VEGETABLE	SOIL	WIDODOMARTANI

Fig. 7. SPARQL result in CO

Our future work will include functionalities that will allow users ask queries using JSP[10] (JavaServer Pages) and Jena[11] ontology API against OWL/RDF files. Through the ontology API, Jena provides a consistent programming interface for ontology applications.

6 Conclusion

Different communities have different perceptions and use different sets of terms (terminologies) to represent the same reality. The problem of it is how to share a different perception between communities and how to make a correspondence between different terms. In this research we used ontology alignment as a process to create a new ontology (common ontology) using a common set of terms by overlapping the common parts of the source ontologies. Using this approach it is possible to share different conceptualizations, different terminologies, and different meanings between different systems. We believe that ontology alignment is one of the best approaches to solve the problem of data and semantic heterogeneity.

Acknowledgment

We would like to acknowledge the support of the Erasmus Mundus EuroAsia program (2010-2012) for the research foundation of this research, and also to

---

[10] Javaserver Pages is a technology provides a simplified, fast way to create dynamic web content. http://www.oracle.com/technetwork/java/javaee/jsp/index.html.

[11] Jena provides a collection of tools and Java libraries to help user to develop semantic web. http://jena.apache.org/.

acknowledge Universidade do Minho and Universitas Gadjah Mada for the collaboration.

REFERENCES

[1] A. Malucelli and others, “Ontology-based services for agents interoperability,” *Faculdade de Engenharia, Universidade do Porto, Porto*, 2006.

[2] Y. Xue, “Ontological View-driven Semantic Integration in Open Environments,” The University of Western Ontario, 2010.

[3] J. Euzenat and P. Shvaiko, *Ontology matching*. Springer-Verlag New York Inc, 2007.

[4] A. Gangemi, D. Pisanelli, and G. Steve, “Ontology integration: Experiences with medical terminologies,” in *Formal ontology in information systems*, 1998, vol. 46, pp. 98–004.

[5] N. F. Noy, “Semantic integration: a survey of ontology-based approaches,” *ACM Sigmod Record*, vol. 33, no. 4, pp. 65–70, 2004.

[6] L. Stojanovic, N. Stojanovic, and J. Ma, “An approach for combining ontology learning and semantic tagging in the ontology development process: eGovernment use case,” *Web Information Systems Engineering–WISE 2007*, pp. 249–260, 2007.

[7] H. S. Pinto and D. N. Peralta, “Combining ontology engineering subprocesses to build a time ontology,” in *Proceedings of the 2nd international conference on Knowledge capture*, 2003, pp. 88–95.

[8] A. Pease, I. Niles, and J. Li, “The suggested upper merged ontology: A large ontology for the semantic web and its applications,” in *Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web*, 2002, vol. 28.

[9] D. Calvanese, G. De Giacomo, and M. Lenzerini, “A framework for ontology integration,” in *The Emerging Semantic Web—Selected Papers from the First Semantic Web Working Symposium*, 2002, pp. 201–214.

[10] H. S. Pinto and J. P. Martins, “A methodology for ontology integration,” in *Proceedings of the 1st international conference on Knowledge capture*, 2001, pp. 131–138.

[11] J. P. M. A. Silva and others, “Automatic and intelligent integration of manufacture standardized specifications to support product life cycle-an ontology based methodology,” 2009.

[12] H. S. Pinto, A. Gómez-Pérez, and J. P. Martins, “Some issues on ontology integration,” 1999.

[13] J. Euzenat, “Semantic precision and recall for ontology alignment evaluation,” in *Proc. 20th International Joint Conference on Artificial Intelligence (IJCAI)*, 2007, pp. 348–353.

[14] B. Chen, H. Tan, and P. Lambrix, “Structure-based filtering for ontology alignment,” in *Enabling Technologies: Infrastructure for Collaborative Enterprises, 2006. WETICE’06. 15th IEEE International Workshops on*, 2006, pp. 364–369.

[15] N. Choi, I. Y. Song, and H. Han, “A survey on ontology mapping,” *ACM Sigmod Record*, vol. 35, no. 3, pp. 34–41, 2006.

[16] N. F. Noy, “Ontology mapping,” *Handbook on ontologies*, pp. 573–590, 2009.

[17] Y. Kalfoglou and M. Schorlemmer, “Information-flow-based ontology mapping,” *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, pp. 1132–1151, 2002.

[18] T. Gruber, “What is an Ontology,” *Encyclopedia of Database Systems*, vol. 1, 2008.

A Proposal for a European Large Knowledge Repository in Advanced Food Composition Tables for Assessing Dietary Intake

Oscar Coltell[1,2], Francisco Madueño[1], Zoe Falomir[1], and Dolores Corella[2,3]

[1] Department of Computing Languages and Systems, Universitat Jaume I, Castellón, Spain
`{oscar.coltell, francisco.madueno, zfalomir}@uji.es`
[2] CIBER Physiopathology of Obesity and Nutrition (CIBEROBN),
Institute of Health Carlos III, Madrid, Spain
[3] Department of Preventive Medicine and Public Health, University of Valencia,
Valencia, Spain
`dolores.corella@uv.es`

Abstract. A proposal for designing and developing a European Repository of Knowledge on Advanced Food Composition Tables (FCTs), based on the existing national FCTs, is proposed in this paper. The requirements of the system, the interoperability strategies, and the cooperation of each national FCT for maintaining and updating the repository are discussed.

Keywords: Knowledge repositories, Food Composition Tables (FCTs), Joint Programming Initiative in A Healthy Diet.

1 Introduction

The study of the interaction between diet and the genome is crucial to prevent and treat cardiovascular diseases, some cancers, type 2 diabetes, etc. The assessment of a person's diet is a tedious task, and in practice, a portion of the intake information is evaluated and then the habitual participants' intake is extrapolated. In order to obtain enough statistical power to avoid measurement errors and changes in diet, it is necessary to obtain repeated measures of dietary information from a large number of participants over time. For extracting information regarding participants' diet, nutritionist use Food Frequency Questionnaires (FFQ), 24 hour dietary recalls (24HDRs), dietary records or dietary histories [1]. These surveys collect consumed foods or dishes, which can be transformed into energy and nutrient intake using Food Composition Tables (FCTs).

When conducting large multicenter studies in which individuals from several countries are involved, one limitation is the difficulty of data acquisition, harmonization and standardization in the different populations. In 2008, one pioneer initiative on this regard was carried out by the “European Food Information Resource AISBL” (EuroFIR

adfa, p. 1, 2011.

AISBL)[1], an International non-profit association (AISBL), whose aim was: "*the development, management, publication and exploitation of food composition data, and the promotion of international cooperation and harmonization through improved data quality, database searchability, standards development, dissemination and training for all users and stakeholders*". The research objective approached here is a proposal of a knowledge network repository, with four basic types of knowledge (food composition, dish composition, dietary patterns and diet-disease effects) which can enhance the EuroFIR project with new methods and techniques in the fields of large knowledge repositories, data mining, and ontology engineering.

Last June 14 in The Hague, the Joint Programming Initiative[2] (JPI) in "A Healthy Diet for a Healthy Life" conference was held and the 2010-2020 roadmap for harmonizing and structuring research efforts in the area of food, nutrition and health was presented. The goal of the JPI conference was to define the Strategic Research Agenda for the period 2011-2020 and beyond[3], which main aims are to provide a holistic approach to: (i) identify the key factors that affect diet-related diseases, (ii) discover new relevant parameters and mechanisms and (iii) define strategies that contribute to the development of actions, policies and innovative products suitable to reduce the burden of diet-related diseases. The JPI Agenda developed the corresponding subroadmap for each one of the three key interacting research areas that were identified and described in the previous Vision Document[4] of the JPI. The Research Areas (RA) are the following: RA1-Determinants of diet and physical activity; RA2-Diet and food production; and RA3-Diet-related chronic diseases.

Each research area roadmap in the Agenda presents two prime initiatives: for 2012-2014 and 2015-2019. The prime initiative for RA1 (2012-2014) is "*Establish a European transdisciplinary research network on determinants of dietary and physical activity behaviors and the relation with health and best practice implementation strategies for sustainable changes*". This initiative is a research challenge where the preparatory work is the collection, integration and assessment of monitoring systems, databases, determinants and outcome assessments. And one of the research needs to face the challenge is to establish and maintain an integrated trans-disciplinary database, with potential for secondary analysis by interested researchers with specific research hypothesis, assuming the initial data are collected according to best practice in biological, behavioral, socio-economic and environmental science traditions.

---

[1] EuroFIR. http://www.eurofir.net/. (Last access in August 6, 2012).

[2] JPI Conference: https://www.healthydietforhealthylife.eu/hdhlconference/ (Last access in August 6, 2012).

[3] The JPI Strategic Research Agenda for the period 2011-2020 and beyond. https://www.healthydietforhealthylife.eu/index.php?index=25. (Last access in August 6, 2012).

[4] The JPI Vision Paper (September 2010) https://www.healthydietforhealthylife.eu/ index.php?index=24. (Last access in August 6, 2012).

Technically speaking, the research challenge of creating a European FCT (EFCT) involves a technological challenge in the field of large databases and large repositories. The Scientific Advisory Board of the JPI, called DEDIPAC, claimed that the EFCT should not be a "data" or "information" database, but a knowledge network repository with contributions of at least 27 European countries. The specific challenge to face is to organize the existing knowledge, their supporting infrastructures and their associated management requirements of the databases containing national Food Composition Tables (FCT) and their integration in a large knowledge repository. Traditionally, FCTs were tables where a portion of each single food was decomposed in energy, macronutrients and other components that are not nutrients. The standard size of the portion is 100 g, but some FTCs take the edible part of the food (i.e., discarding the peel in oranges; in this case, 100 g of edible orange), and other FTCs take the whole food (i.e., the whole 100 g of orange, including the peel). Moreover, macronutrients are grouped in families, as lipids, proteins, carbohydrates; and no nutrients are minerals, vitamins and aminoacids. Usually, each FTC register contains around 50 components. However, the number of components may vary in each FTC. Regarding national and private (academic or enterprise) FCT creation, although they can be standardized and biochemically proved, they are usually different from country to country (or depending on the academic organization or enterprise aims and resources).

With the evolution of the information and communication technologies, FCTs were converted in databases and, later, Web services were added to allow on-line access to them. But the drawbacks of the traditional FTC were inherited by the FCT databases and emerged some specific problems as, for example, the lack of service due to site saturation or network breakdowns, the restricted access only to active members (who have paid the corresponding fee), the lack of programmed access (a set of procedures to manage queries coming from applications), the native language, and so on. That is the situation of the European FCT provided by the FAO[5] or EuroFIR[6].

The aim of this paper is to discuss a proposal for designing and developing a European Repository of Knowledge on Advanced FCTs and related knowledge (food composition, dish composition, dietary patterns and diet-disease effects, and semantic connections between them) based on the existing national FCTs, their system interoperability strategies, and the cooperation of each national FCT for maintaining and updating the repository.

For achieving this aim, the following strategies are discussed in this paper: (i) a process for retrieving data from the different national resources and populate the Repository (Section 2); (ii) the viability of the current software resources and protocols that can be used to integrate the different FTC databases (Section 3); and (iii) new methods and

---

[5] FAO. Food Composition Tables–Europe. http://www.fao.org/infoods/tables_europe_en.stm. (Last access in August 6, 2012).

[6] EuroFIR How to access FCDBs. http://www.eurofir.net/food_information/food_composition_databases /how_access_fcdbs. (Last access in August 6, 2012).

techniques for generating and extracting knowledge form the Repository (Section 4). Finally, some conclusions are provided.

2 Designing a Process for Retrieving Data and Populate the Repository

The process for retrieving data from the different national resources and populate the Repository can be very complex because the national FTC databases has been developed according to each country objectives, culture, funding and interests. Thus, data structures, nomenclatures, number of food components included or, even, formats and units (English or International Metric systems: e.g. quantities in grams vs. quantities in ounces) are not shared. Moreover, each database has different access protocols and restrictions (i.e., public vs. private access, human interface vs. programed interface or both, etc.) Therefore, before starting to discuss how we could apply the technical approach, previous political work should be done searching agreements for data sharing, open access protocols and medical and nutritional interests. Despite the above mentioned complexity, the process outlines can be described in a workflow composed by four steps:

STEP1: defining a Minimal Set of FCT data (MS-FCT). The MS-FCT is the common data that holds every FCT database in the same or approached format (no need of transformation or conversion). On the other hand, the Standard Set of FCT data (SS-FCT) must be defined. The SS-FCT is the standardized data that every FCT database should contain according strategic objectives of the knowledge repository (homogeneity, integration, interoperability).

STEP2: defining the knowledge levels in the repository. Initially, we have defined the following levels (see Fig. 1):

1. **Level 1: Food Composition**. Basic knowledge about the composition of each food but with the following variations: national FCT source, determination methods for each component, local and regional variations of the food, and original language.
2. **Level 2: Dish Composition**. Knowledge about the composition of dishes in single food, the standard portions (in Metrical and English measures) and their corresponding images, the corresponding recipes (the same food mixture is different ac-cording the cooking process), and the local and regional variations in recipes and portions.
3. **Level 3: Dietary patterns**. Knowledge about discovered dietary patterns in nutritional studies using data mining strategies. From dietary patterns, it would be possible to generate dietary models to apply in the kind of studies described in the JPI research areas prime initiatives.
4. **Level 4: Diet-disease effects**. Knowledge about associations and interactions between diet and disease (via genetic and phenotypic factors), recommendations for specific populations (i.e., celiac), high risk food for specific diseases, lowering risk food for specific diseases, etc.

All together should run in cooperation with every national FCT database trust, providing full access to authorized sources, level of service and frequent updates to guarantee the quality and accuracy of the provided knowledge in the repository.

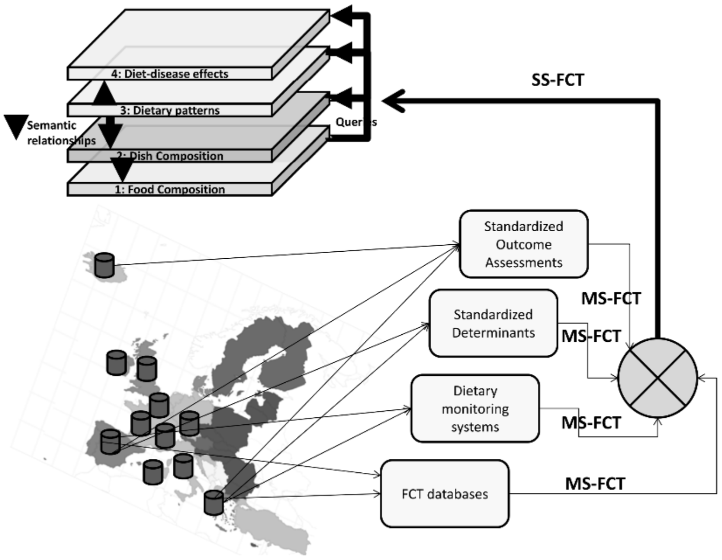


Fig. 1. Repository environment and functional structure. From each EU partner database, or set of databases (FCT, Dietary monitoring systems, Standardized Determinants and Standardized Outcome Assessments), a MS-FCT is provided. After some homogenization and integration processes, a SS-FCT is generated to update the Repository. The Queries path shows how queries between levels flow. Semantic relationships are defined only to immediate levels and show how to extract knowledge from the repository.

STEP3: studying, designing, developing and applying current software resources and protocols to integrate the different EU partners' FTC databases and other data (Fig.1), generating the corresponding sets of MS-FCT, for retrieving data from the different national resources.

STEP4: populating and maintaining the Repository, mainly injecting standardized data from the different national resources under the SS-FCT approach, but also using direct built-in methods and interfaces. It should be noted that the information is generated on national resources and not in the Repository.

3 The Viability of the Current Software Resources and Protocols

FCTs allow mapping foods or dishes with their corresponding energy and nutrients. In Nutritional Epidemiology, this is crucial due to the proved relation that exists between diet and some diseases [2], as for example, cardiovascular diseases [3-5], diabetes [6-7], and obesity [8-10], whose study requires large amounts of data for a statistical analysis. Then, the development of the proposed Large Knowledge Repository is certainly a colossal and challenging task evolving current technology and new technologies that undoubtedly have an initial cost but may pay off in the long term.

Previous works by our group [11], developed some medium scale projects in the area of medical informatics for automatizing nutritional questionnaires and calculating the nutritional composition of meals using several FCTs which used an ontology for translating the components in different FCTs to a common name. That ontology, named Nutriontology (NO), is running on an independent platform, which also contains all FTCs physical databases, applying interoperability strategies to manage the database access. Moreover, NO is part of a set of ontologies managed by an upper level ontology named NutriGenOntology (NGO). Other independent generic Web platform, named "Project", manage the set of automatized nutritional questionnaires and the participant's (and other data) database corresponding to one nutritional study. Thus, the communication between NO and a project are performed by Web services. Really, Project is a template which is instantiated in a particular platform as new nutritional studies are started and, then, the platform adopts the study name or acronym (i.e., Fituveroles, Obenutic, Obenomics, etc.) Therefore, we consider that this pilot system carried out by our group, which combines ontologies and web services in the appropriate manner, can be a start-up for achieving an integrated European FCT.

Besides, currently information repositories technology is rendered as insufficient for accomplish the integration and interoperability levels expected in such repositories, and the heterogeneity in the data is not efficiently managed. For example, the Semantic MediaWiki[7] do already consider the unit conversion problem at a very basic level. Another option, taking in account the very large scale of our proposal, is to define two wide strategies in both levels (Fig.1): level 1 with integration and interoperability; level 2 with homogenization. To integrate the different FTC databases, one suitable solution is combining semantic mappings for modelling FTC structures and semantic operations for retrieving data from the different national resources, and then, generating the corresponding MS-FTCs. Homogenization in the second level, under the SS-FCT approach, could foster the enhancement and specialization of existing data mining methods and techniques. Other solutions may be considered since some intelligent systems can cope with heterogeneity and interoperability in all levels. Then, it is too early for

---

[7] Semantic MediaWiki repository. http://semantic-mediawiki.org/wiki/Help:Custom_units#Converting_between_proportional_units. (Last access in August 5, 2012).

comparing the cost of addressing heterogeneity and interoperability versus the cost of homogenization in the proposed repository.

4 Developing new methods and techniques for generating and extracting knowledge form the Repository

It is necessary to define a standard language (i.e., XML-based language) for representing the Minimal Set of FCTs data and Standard Set of FCTs data, both including the basic four types of knowledge the Repository has to manage: food composition, dish composition, dietary patterns and diet-disease effects. But, the characteristics of these types of knowledge and the challenges derived from them must be identified.

The food composition knowledge tell us what elements are in one standard portion (100 g. of edible portion or net intake) of each food: macronutrients (proteins, fat and carbohydrates), micronutrients (aminoacids, minerals and vitamins), other components (water, alcohol, caffeine, etc.), and the corresponding total energy of the whole portion. In the biochemical analysis made for composing the FCT, each sample is taken from raw food, wherever possible with minor exceptions, to avoid nutrient alterations in cooking processes. Therefore, the primary source of the information is the food composition biochemical analysis performed by each national food authority. This kind of analysis is make once unless a new and better biochemical technique appears in market. The secondary source of information is the own FCT. It could be subject to change due to adding new food entries (the most usual) or reviewing the existing ones (very rarely). Moreover, there are some standards about FCT structure and organization. The derived challenge is, firstly, to homogenize FCT entries in a common set of components, nomenclatures and formats/units under the MS-FCT approach but keeping national differences; and secondly, to integrate and combine all national FCT entries in a maximal concept as it is the SS-FCT. The last one would cover lacks of data for each individual food in a FCT combining data from the rest of FCTs.

The dish composition knowledge describes the three main aspects of each dish: what food contains and in which quantity/proportion contributes each individual food, what cooking process has been applied, and what is the size of the portion. The proportion of each individual food determines the calculations of edible portions for obtaining the food composition from the FCT. The list of each individual food is not static due to national, regional, local and, of course, home variations, but keeping the main components (i.e., apple pie will not be more apple pie when apple is replaced by peach). Each kind of cooking process alters the properties of the food (i.e., vitamin or fiber degradation, fat substitution, etc.). Then, FCTs cannot be applied directly, but with cooking revisions. The size of the portion is the description of how big is and what quantity of food contains a dish. Here, a specific problem arises from the term “dish”, because we can have solid, liquid and semi-liquid food. Then, when we are describing a portion of solid food, we are using the traditional meaning of physical dish (or similar) and measures in grams or ounces/pounds. However, when we are describing a portion of liquid and semi-liquid food, we have to use different container as glass or cup, and measures in milliliters or liquid ounces/pints. Usually, portions are categorized as small, medium and big, where each category has assigned one quantity in weight or volume, but the quantity depends of the nature of food itself. Moreover, there are not any standard (or the facto standard) about dish structure and portions, but the cooking alterations are well studied and weighted. Therefore, the primary source of the information is composed by, in one hand, published tables of cooked food proprieties; and, on the other hand, published collections of recipes in books, journals, Web, etc. The derived challenge in this case is to define a Minimal Common Recipe Catalog (MCRC) which can be used in the scientific environment for assessing dish composition in the Repository. The MCRC should include the "official" composition of each dish plus cooking variants, standardized portions and units according the food state (solid, liquid, semi-liquid).

The dietary patterns knowledge show us common profiles of food intake in persons to whom dietary assessment questionnaires were administered. Dietary patterns usually are inferred from the participants in nutritional studies and, later, can be reviewed and organized to have well-established patterns. Therefore, the primary source of the information is the set of discovered dietary patterns, and the second source is the collection of scientific publications describing other patterns. The derived challenge in this case is to achieve a standard catalog of well-established patterns for making comparisons in each nutritional study.

The diet-disease effects knowledge show us the associations and interactions between diet and diseases, when diet may act as risk or protector factor over individuals with (genetic) susceptibility to particular disease. Really, associations and interactions are not analyzed taking in account a particular meal or food, but specific dietary patterns. So, dietary patterns and disease are strongly related. Therefore, the main source of the information is the set of statistically significant diet-disease associations and interactions discovered in the nutritional studies and published in journals. The derived challenge in this case is having the maximum and accurate knowledge as possible about diet-disease associations and interactions.

5 Conclusions

A framework for designing and developing a European repository of Knowledge for Food Composition Tables is proposed with in this paper and the scenarios and the steps for constructing this repository are also described. The main outline is to construct the knowledge base in a scalable way, moving from standardized knowledge towards population-dependent knowledge. The main challenge is to integrate repositories belonging to different national states (many issues due to the use of different data structures, different nomenclatures, and different formats and units). Moreover, FCTs are extended with three additional types of knowledge, dish composition, diet patterns and diet-disease effects, coming from other biomedical/biological data sources, for mining associations and interactions between diseases and food by means of dietary patterns.

A pilot approach was carried out by our group, which developed some medium scale projects in the area of medical informatics for automatizing nutritional questionnaires and calculating the nutritional composition of meals using several FCTs which used an ontology for translating the components in the different FCTs to a common name. Based on the success of this approach, we propose a solution to the integration of all European FCTs based on ontologies and web services, and asynchronous web technologies for assuring the minimal response time in knowledge queries, and for providing modular services, and the maximal underlying data organization.

Acknowledgements. This work has been partially funded by grants GEWIMICS (SAF2009-12304, MICINN), AGL2010-22319-C03 (MICINN), BEST/2011/261 (GVA), ACOMP/2011/145 (GVA), and CIBER “Physiopathology of Obesity and Nutrition” (ISCIII-FIS). CIBERobn is an initiative of the ISCIII.

References

1. Falomir Z., Arregui M., Madueño F., Coltell C., Corella D.: Automation of Food Questionnaires in Medical Studies: a state-of-the-art review and future prospects. Comp. Biol. Med. (in press, accepted on 25/07/2012 with DOI 10.1016/j.compbiomed.2012.07.008) (2012)
2. Feart C., Alles B., Merle B., Samieri C., Barberger-Gateau P.: Adherence to a Mediterranean diet and energy, macro-, and micronutrient intakes in older persons. J. Physiol. Biochem. (Epub ahead of print. PubMed PMID: 22760695) (2012)
3. Ganguly R., Pierce G.N.: Trans fat involvement in cardiovascular disease. Mol. Nutr. Food. Res. 56(7), 1090-1096 (2012)
4. de Oliveira Otto M.C., Mozaffarian D., Kromhout D., Bertoni A.G., Sibley C.T., Jacobs D.R. Jr, Nettleton J.A.: Dietary intake of saturated fat by food source and incident cardiovascular disease: the Multi-Ethnic Study of Atherosclerosis. Am. J. Clin. Nutr. 96(2), 397-404 (2012)
5. Hansen-Krone I.J., Enga K.F., Njølstad I., Hansen J.B., Braekkan S.K.: Heart healthy diet and risk of myocardial infarction and venous thromboembolism. The Tromsø Study. Thromb Haemost. 108(3). (Epub ahead of print. PubMed PMID: 22739999) (2012)
6. Rivellese A.A., Giacco R., Costabile G.: Dietary Carbohydrates for Diabetics. Curr. Atheroscler. Rep. (Epub ahead of print. PubMed PMID: 22847773) (2012)
7. Guldbrand H., Dizdar B., Bunjaku B., Lindström T., Bachrach-Lindström M., Fredrikson M., Ostgren C.J., Nystrom F.H.: In type 2 diabetes, randomisation to advice to follow a low-carbohydrate diet transiently improves glycaemic control compared with advice to follow a low-fat diet producing a similar weight loss. Diabetologia. 55(8), 2118-2127 (2012)
8. Corella D., Arnett D.K., Tucker K.L., Kabagambe E.K., Tsai M., Parnell L.D., Lai C.Q., Lee Y.C., Warodomwichit D., Hopkins P.N., Ordovas J.M.: A high intake of saturated fatty acids strengthens the association between the fat mass and obesity-associated gene and BMI. J. Nutr. 141(12), 2219-2225 (2011)
9. Bulló M., Garcia-Aloy M., Martínez-González M.A., Corella D., Fernández-Ballart J.D., Fiol M., Gómez-Gracia E., Estruch R., Ortega-Calvo M., Francisco S., Flores-Mateo G.,

Serra-Majem L., Pintó X., Covas M.I., Ros E., Lamuela-Raventós R., Salas-Salvadó J.: Association between a healthy lifestyle and general obesity and abdominal obesity in an elderly population at high cardiovascular risk. Prev. Med. 53(3), 155-161 (2011)

10. Foster G.D., Shantz K.L., Vander Veur S.S., Oliver T.L., Lent M.R., Virus A., Szapary P.O., Rader D.J., Zemel B.S., Gilden-Tsai A.: A randomized trial of the effects of an almond-enriched, hypocaloric diet in the treatment of obesity. Am. J. Clin. Nutr. 96(2), 249-54 (2012)
11. Fabregat A., Arregui M., Barrera E., Portolés O., Corella D., Coltell O.: NutriGeneOntology: A Biomedical Ontology for Nutrigenomics. In: Proceedings of the 2008 International Conference on Biomedical Engineering and Informatics; 2008, vol. 1, pp. 915-919. IEEE Computer Society, New York (2008)

Disambiguating automatically-generated semantic annotations for Life Science open registries

Antonio Jimeno-Yepes[1], Mara Pérez-Catalán[2], and Rafael Berlanga-Llavori[2]

[1] National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, USA
antonio.jimeno@gmail.com
[2] Universitat Jaume I,Castellón, Spain
mcatalan@icc.uji.es,berlanga@lsi.uji.es

Abstract. This paper presents our preliminary evaluation of the automatic semantic annotation of open registries. Conversely to traditional application of semantic annotation to scientific abstracts (e.g., PubMed), open registries contain descriptions that mix terminologies of Computer Science, Biomedicine and Bioinformatics, which makes their automatic annotation more prone to errors. Moreover, the extensive use of acronyms and abbreviations in these registries may also produce wrong annotations. To evaluate the impact of these errors in the quality of the automatically generated annotations we have built a Gold Standard (GS) with single-word annotations. Additionally, we have adapted a knowledge-based disambiguation method to measure the hardness in distinguishing right from wrong annotations. Results show that for some semantic groups the disambiguation can be performed with good precision, but for others the effectiveness is far from being acceptable. Future work will be focused on developing techniques for improving the semantic annotation of these poorly represented semantic groups.

1 Introduction

In recent years, open metadata registries have become a popular tool for researchers trying to locate resources in different domains, mainly in Life Sciences and Open Linked Data. These registries allow users to provide metadata about the resources in order to facilitate their discovery, which can be structured metadata, such as tags or categories, or free text descriptions. Although sophisticated standards have been proposed for annotating the resources, most of the metadata available in the registries are expressed in natural language, which makes more difficult the discovery of these resources in traditional search engines. Descriptions contain useful information about the resources and, moreover, they implicitly describe the features of the resources. Therefore, to facilitate the discovery of the most appropriate web resources, all these metadata has to be normalized in order to be automatically processed.

Semantic annotation techniques are frequently used to normalize the metadata. Semantic annotation (SA) is the process of linking the *entities* mentioned in a text to their *semantic descriptions*, which are stored in knowledge resources (KRs) such as thesauri and domain ontologies, like UMLS® Metathesaurus® and EDAM ontology [20] in Life Sciences. During the last years, we have witnessed a great interest in massively annotating biomedical information. Most of them are based on dictionary look-up techniques. These approaches try to find in the documents each text span that exactly matches some lexical forms of the terminological resource. Other approaches, like MetaMap [2] and EAGL [22], allow partial matching between text spans and lexical forms. Their main drawback is that precision is usually very low and they suffer from scalability issues. These annotators only base the matching on isolated text spans without taking into account the context of the matching, which is the main source of errors when annotating open collections.

Another issue that has to be taken into account in metadata normalization is that metadata in web resources registries usually contains vocabulary taken from different domains. For instance, in Life Sciences registries, the metadata contains words about medicine, bioinformatics and computers, with a high degree of overlapping between them. However, if the domains are not equally covered by the knowledge resources, some senses of some words can be disregarded and, therefore, the precision of the semantic annotations and, as consequence, also the quality of the retrieved resources may be affected. Thus, the quality of the semantic annotations becomes crucial in the discovery process.

There are two main problems that need to be addressed. One of them is ambiguity, since a term can be mapped to more than one concept or sense. The second one is the lack of coverage of the terminological resources. A term can be ambiguous but this might not be reflected in the terminological resource. As a consequence, there is no guarantee in many cases that even though the mapping is not ambiguous that is correct.

In this paper we study these issues in the context of the semantic annotation of open registries of Life Science resources, using the currently largest biomedical knowledge resource, that is, the NLM's UMLS [5].

2 Methods

We propose to study the effectiveness of unsupervised Word Sense Disambiguation (WSD) approaches. The definition of the concept is turned into a bag-of-words representation in which the words are weighted according to their relevance to the concept and related concepts. This concept profile is compared to the context of the ambiguous word and if it is over a trained threshold according to a similarity measure, then it is assigned the given concept. In this work, the window for the context of the ambiguous word is all the terms in the description of the registry.

The concept profiles are prepared based on the NLM's UMLS [5], which provides a large resource of knowledge and tools to create, process, retrieve, integrate and/or aggregate biomedical and health data. The UMLS has three main components:

- Metathesaurus, a compendium of biomedical and health content terminological resources under a common representation which contains lexical items for each one of the concepts, relations among them and possibly one or more definitions depending on the concept. In the 2009AB version, it contains over a million concepts.
- Semantic network, which provides a categorization of Metathesaurus concepts into semantic types. In addition, it includes relations among semantic types.
- SPECIALIST lexicon, containing lexical information required for natural language processing which covers commonly occurring English words and biomedical vocabulary.

Concepts are assigned a unique identifier (CUI) which has linked to it a set of synonyms which denote alternative ways to represent the concept, for instance, in text. Concepts are assigned one or more semantic types.

In the following section, we present the generation of the WSD profiles and present the similarity measures that will be used to compare the concept profiles and the context of the ambiguous words.

2.1 WSD profiles

Word sense disambiguation (WSD), given an ambiguous word in context, attempts to select the proper sense given a set of candidate senses. An example of ambiguity is the word *domain* which could either refer to *works or knowledge without proprietart interest* or, in biology, the *taxonomic subdivision even larger than a kingdom* or *a part of a protein*. The context in which *domain* appears is used to disambiguate it. WSD is an intermediary task which might support other tasks such as: information extraction (IE) [2], information retrieval (IR) and summarization [21].

WSD methods are based either on supervised learning or knowledge-based approaches [23]. Supervised methods are trained on examples for each one of the senses of an ambiguous word. A trained model is used to disambiguate previously unseen examples. Knowledge-based (KB) methods rely on models built based on the information available from available knowledge sources. In the biomedical domain, this would include the Unified Medical Language System (UMLS). In this scenario, the candidate senses of the ambiguous word are UMLS concepts. KB methods either build a concept profile [18], develop a graph-based model [1] or rely on the semantic types assigned to each concept for disambiguation [11]. These models are compared to the context of the ambiguous word being disambiguated. The candidate sense with highest similarity or probability is selected as the disambiguated sense.

Due to the scarcity of training data, KB methods are preferred as disambiguation methods. KB methods rely on information available in a terminological resource. Performance of knowledge-based methods depends partly on the knowledge resource, which usually is not built to perform WSD or IR tasks [14].

In our first WSD approach, the context words surrounding the ambiguous word are compared to a profile built from each of the UMLS concepts linked to the ambiguous term being disambiguated. This approach has been previously used by McInnes [18] in the biomedical domain with the NLM WSD corpus.

This algorithm can be seen as a relaxation of Lesk's algorithm [16], which is very expensive since the sense combination might be exponentially large even for a single sentence. Vasilescu et al. [24] have shown that similar or even better performance might be obtained disambiguating each ambiguous word separately.

A concept profile vector has as dimensions the tokens obtained from the concept definition or definitions if available, synonyms, and related concepts excluding siblings.

Stop words are discarded, and Porter stemming is used to normalize the tokens. In addition, the token frequency is normalized based on the inverted *concept* frequency so that terms which are repeated many times within the UMLS will have less relevance.

A context vector for an ambiguous term includes the term frequency; stop words are removed and the Porter stemmer is applied. The word order is lost in the conversion.

2.2 Similarity measures

We have compared the context vector of the term under evaluation (A) and the concept profile vector (B) based on the several similarity measures presented below. The length of the vectors is usually large due to the vocabulary size. But the context and profile vectors only have values for a limited number of entries and the others will have a value of zero.

One of these measures is the cosine similarity, shown in equation 1. The candidate concept with the highest cosine similarity is selected as candidate concept. This approach is used with UMLS based concept profiles [13, 18].

$$Cosine = \frac{A \cdot B}{\|A\| \|B\|} \tag{1}$$

Entailment, presented below, looks at the overlap between the two vectors and normalizes based on the number of tokens in the context vector. Compared to the cosine similarity, the overlap is based on counting the matches between both vectors instead of estimating the dot product. The matches are done considering the non-zero entries. This overlap is normalized by the length of context vector only to avoid a negative impact of a long concept profile.

$$Entailment(A, B) = \frac{|A \cap B|}{|A|} \tag{2}$$

The Jaccard coefficient measures similarity between sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets. Compared to entailment, the length of the concept profile is considered.

$$Jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{3}$$

Chi-square allows comparing two distributions. In our work, we compare the concept profile to the context vector. Chi-square has been used as a similarity measure in text categorization by Chen et al. [6] and we follow their formulation in this work.

$$\chi_v^2 = h\left[\sum_{i=1}^{n} \frac{A_i^2}{sum(A)(A_i + B_i)} + \sum_{i=1}^{n} \frac{B_i^2}{sum(B)(A_i + B_i)}\right] - h \tag{4}$$

$$sum(A) = \sum_{i=1}^{n} A_i \tag{5}$$

$$sum(B) = \sum_{i=1}^{n} B_i \tag{6}$$

$$h = sum(A) + sum(B) \tag{7}$$

2.3 Data set

In this paper, our aim is to analyze the impact of the automatic semantic annotations in the quality of the results of a retrieval system. To do that, we use a dictionary look-up semantic annotator [3] to automatically annotate the metadata of the resources registered in three Life Sciences registries: BioCatalogue [4], myExperiment [10] and SSWAP [9].

The semantic annotator is able to deal with several ontologies in order to cover as much as possible the different vocabularies that appear in the resources descriptions. In this work, the semantic annotator uses as knowledge resources (KRs): UMLS, EDAM (an ontology designed for Life Science open registries), myGrid (reference ontologies of BioCatalogue) and the entries of the Wikipedia that have as category some sub-category of the Bioinformatics category. A detailed description of the semantic annotator can be found in [19].

A preliminary analysis of the automatically generated semantic annotations suggests that concepts matching several words are usually unambiguous and are associated to a right sense. However, single word concepts are much prone to ambiguity and errors.

For this reason, we have manually created a Gold Standard (GS) with those annotations matching a single word. The GS has been curated by two people who have analyzed each combination of concept-word in each semantic annotation in the resources description, selecting the most appropriate concept in each case. The GS contains for each semantic annotation, represented as a triple $(concept, word, contextvector)$, a bit indicating if the sense is correct (1) or not (0). This GS contains 8863 single-word semantic annotations.

The whole catalogue contains 72958 semantic annotations, from which 42686 where annotated only with concepts from UMLS, 12269 were annotated with concepts from UMLS and the other KRs and 18003 were annotated with concepts from the other KRs but not from UMLS.

3 Results

We intend to evaluate the concept profiles and the similarity measures for filtering annotations in our data set. From our data set, we have selected the semantic groups of interest and split the set for each one of the semantic groups sets into 2/3 for training and 1/3 for testing. The semantics groups are the following: CONC (Concepts & Ideas), DISO (Disorders), LIVB (Living Beings) and PHYS (Physiology) as defined in the UMLS Semantic Network [17][3], while the groups CHED (Chemicals & Drugs) and PRGE (Proteins & Genes) follow the definition under the CALBC challenge[4]. CALBC groups definition is closer to our interests compared to the ones defined by the UMLS Semantic Network in these two cases.

Table 1 shows the distribution of semantic annotations of the GS per semantic group. Positive instances are the ones that are labeled with the specified semantic group and the negative ones are instances that should not be labeled with the semantic group. The distribution is usually skewed towards the negative class, i.e. the concept does not represent the correct sense of the word, except for the PRGE group in which the positive examples are more frequent. For example, in the service SMART registered in BioCatalogue, the word *domain* refers to protein domain and it has been annotated with the concepts *C1514562:PRGE*, that refers to the protein domain, and *C1883221:CONC*, that refers to the general concept of domain. Therefore, *C1514562* is the correct concept in this case and it is represented as a positive instance in the GS.

Semantic Group	Training	Positive	Negative	Testing	Positive	Negative
CHED	527	148	379	263	70	193
CONC	2139	598	1541	1068	283	785
DISO	180	6	174	90	4	86
LIVB	408	166	242	203	83	120
PHYS	169	44	125	84	22	62
PRGE	654	460	194	326	232	94

Table 1. Semantic group data set distribution

We would like to be able to decide if an annotation is correct given the measures presented above. We have trained a threshold for each of the measures based on the training set. This threshold is used to decide if the instance should

---

[3] http://semanticnetwork.nlm.nih.gov/SemGroups

[4] http://www.ebi.ac.uk/Rebholz-srv/CALBC/challenge_guideline.pdf

be labeled with the semantic group or not. The optimization measure has been the F-measure, while other measures could be considered. On the other hand, due to the skewness of the data, other measures as accuracy would not be as effective.

Table 2 shows the filtering performance of the different measures. Overall the similarity measures seem to perform similarly except for chi-square that performs better on average over the other measures. Chi-square shows a larger difference compared to other measures for the LIVB and PHYS semantic groups.

SG	Measure	Threshold	Precision	Recall	F-measure
CHED	chisquare	-3642.0724	0.4898	0.6857	0.5714
	cosine	0.9698	0.5169	0.6571	0.5786
	entailment	0.9269	0.4783	0.6286	0.5432
	jaccard	0.9961	0.4538	0.7714	0.5714
CONC	chisquare	-121.9878	0.2939	0.8693	0.4393
	cosine	1.0000	0.2647	1.0000	0.4186
	entailment	1.0000	0.2647	1.0000	0.4186
	jaccard	1.0000	0.2647	1.0000	0.4186
DISO	chisquare	-41397.6546	0.2500	0.2500	0.2500
	cosine	0.9956	0.1212	1.0000	0.2162
	entailment	0.9407	0.2000	0.5000	0.2857
	jaccard	0.9768	0.0000	0.0000	0.0000
LIVB	chisquare	-3416.2337	0.7349	0.8133	0.7722
	cosine	0.9995	0.4774	0.8916	0.6218
	entailment	0.9302	0.6173	0.6024	0.6098
	jaccard	0.9996	0.4774	0.8916	0.6218
PHYS	chisquare	-884.3186	0.4884	0.9545	0.6462
	cosine	1.0000	0.2619	1.0000	0.4151
	entailment	0.9662	0.3415	0.6364	0.4444
	jaccard	0.9855	0.4063	0.5909	0.4815
PRGE	chisquare	-173.5428	0.7099	0.9914	0.8273
	cosine	1.0000	0.7117	1.0000	0.8315
	entailment	1.0000	0.7117	1.0000	0.8315
	jaccard	1.0000	0.7117	1.0000	0.8315

Table 2. Semantic group results on the test set

4 Discussion

The results are interesting but there is still room for improvement. Among the evaluated measures, chi-square seems to perform better on average compare to the other measures. Cosine has been the preferred similarity measure in many biomedical disambiguation work [13] and would be interesting to evaluate chi-square in similar studies.

The best performing groups are LIVB and PRGE. In the case of LIVB, there are not only the species which have shown already easy to annotate [8], even though this semantic group includes in addition several population groups which seem more difficult to annotate. On the other hand, the best F-measure is obtained when all the cases are annotated as PRGE. This means that in addition to being difficult to annotate, the skewness is in favour of this semantic group.

DISO has a small set of positive cases related to the term *diabetes*. Most of the wrongly assigned terms are abbreviations like CA (California) or SIB (Swiss Bioinformatics Institute). Other mentions like *brain*, have been already identified in previous work [12] and different proposals for lexicon cleansing could be used. This semantic group has a reduced set of annotations which are relevant in our data set, which might indicate that the open registries include almost no mention of diseases.

CONC has the largest number of candidate instances from which only a small part is relevant to this semantic group and appears in large part of the example cases. In this first work, the context vector might be too broad to help decision making over annotations.

PHYS shows a large difference in performance with the chi-square measure. Looking at the examples, there is a limited number of terms used which seem to be always linked to PHYS. Examples of these terms are *pathway*, *transcription* and *transport*. Other terms annotated as PHYS rarely are labeled as PHYS in the gold standard. Among these terms, we find *interactions*, *size* or *status*.

Annotation of chemical entities has already proved to result in low performance [7]. CHED annotations seem to be complicated to filter properly. Again, there are sets of common terms that can be pre-filtered for this domain that in many cases are not related to the topic of interest. Examples of these terms are *products*, *CA* or *date*.

5 Conclusions and Future Work

We have introduced the problem of determining the correct sense of ambiguous terms depending on their context in the semantic annotations in open registries and evaluated the use of knowledge based methods used in disambiguation in the automatic annotation of these registries.

Better performance is required to use the filtered annotations in a retrieval system. We have worked with a large window, all the words in the definition of the registries, in the development of the context vector. A more restrictive window might provide a more focused context. In addition, we have seen that there are terms which seem to have a preferred sense in this data set. Chi-square performs better than other evaluated measures but has not been evaluated in biomedical WSD and could provide better performance than existing work.

We have evaluated knowledge-based WSD methods since, when we started this work, no training data was available. Given the current data set, trained conditional random fields approaches [15] could be evaluated on the annotated set.

Some direct follow-ups of this work are the refinement of particular details of the semantic annotator, such as the detection of locutions as entities that do not have to be annotated, the disambiguation of acronyms, the use of lexical patterns to recognise fragments that are entities as a whole, e.g. the citations, or the disambiguation of single words that are simplifications of multi-words. In addition, we are also considering the use of lexicon cleansing techniques to improve the lexicon.

6 Acknowledgments

This work was supported in part by the Intramural Research Program of the NIH, National Library of Medicine and by an appointment of A. Jimeno-Yepes to the NLM Research Participation Program sponsored by the National Library of Medicine and administered by the Oak Ridge Institute for Science and Education. This work has been also funded by the "Ministerio de Economía y Competitividad", project contract TIN2011-24147.

References

1. Eneko Agirre, Aitor Soroa, and Mark Stevenson. Graph-based word sense disambiguation of biomedical documents. *Bioinformatics*, 26(22):2889–2896, 2010.
2. A.R. Aronson and F.M. Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.
3. Rafael Berlanga, Victoria Nebot, and Ernesto Jimenez. Semantic annotation of biomedical texts through concept retrieval. In *BioSEPLN 2010*, 2010.
4. Jiten Bhagat, Franck Tanoh, Eric Nzuobontane, Thomas Laurent, Jerzy Orlowski, Marco Roos, Katy Wolstencroft, Sergejs Aleksejevs, Robert Stevens, Steve Pettifer, Rodrigo Lopez, and Carole A Goble. BioCatalogue: a universal catalogue of web services for the life sciences. *Nucleic acids research*, 38(Suppl 2):W689–94, jul 2010.
5. O. Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(Database Issue):D267, 2004.
6. Y.T. Chen and M.C. Chen. Using chi-square statistics to measure similarities for text categorization. *Expert Systems with Applications*, 38(4):3085–3090, 2011.
7. P. Corbett and P. Murray-Rust. High-throughput identification of chemistry in life science texts. *Computational Life Sciences II*, pages 107–118, 2006.
8. M. Gerner, G. Nenadic, and C.M. Bergman. Linnaeus: a species name identification system for biomedical literature. *BMC bioinformatics*, 11(1):85, 2010.
9. Damian DG Gessler, Gary S Schiltz, Greg D May, Shulamit Avraham, Christopher D Town, David Grant, and Rex T Nelson. SSWAP: A Simple Semantic Web Architecture and Protocol for semantic web services. *BMC Bioinformatics*, 10:309, 2009.
10. Carole A. Goble, Jiten Bhagat, Sergejs Aleksejevs, Don Cruickshank, Danius Michaelides, David Newman, Mark Borkum, Sean Bechhofer, Marco Roos, Peter Li, and David De Roure. myexperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Research*, 38(suppl 2):W677–W682, 2010.

11. S.M. Humphrey, W.J. Rogers, H. Kilicoglu, D. Demner-Fushman, and T.C. Rindflesch. Word sense disambiguation by selecting the best semantic type based on Journal Descriptor Indexing: Preliminary experiment. *Journal of the American Society for Information Science and Technology (Print)*, 57(1):96, 2006.
12. A. Jimeno, E. Jimenez-Ruiz, V. Lee, S. Gaudan, R. Berlanga, and D. Rebholz-Schuhmann. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC bioinformatics*, 9(Suppl 3):S3, 2008.
13. A. Jimeno-Yepes and A.R. Aronson. Knowledge-based biomedical word sense disambiguation: comparison of approaches. *BMC bioinformatics*, 11:565, 2010.
14. A. Jimeno-Yepes, R. Berlanga-Llavori, and D. Rebholz-Schuhmann. Ontology refinement for improved information retrieval. *Information Processing & Management*, 2009.
15. John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.
16. M. Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26. ACM, 1986.
17. A.T. McCray, A. Burgun, O. Bodenreider, et al. Aggregating umls semantic types for reducing conceptual complexity. *Studies in health technology and informatics*, (1):216–220, 2001.
18. Bridget McInnes. An unsupervised vector approach to biomedical term disambiguation: Integrating UMLS and Medline. In *Proceedings of the ACL-08: HLT Student Research Workshop*, pages 49–54, Columbus, Ohio, June 2008. Association for Computational Linguistics.
19. M. Pérez-Catalán, R. Berlanga, I. Sanz, and M.J. Aramburu. A semantic approach for the requirement-driven discovery of web resources in the Life Sciences. *Knowledge and Information Systems*, pages 1–20, 2012.
20. Steve Pettifer, Jon Ison, Matus Kalas, Dave Thorne, Philip McDermott, Inge Jonassen, Ali Liaquat, José M. Fernández, Jose M. Rodriguez, INB Partners, David G. Pisano, Christophe Blanchet, Mahmut Uludag, Peter Rice, Edita Bartaseviciute, Kristoffer Rapacki, Maarten Hekkelman, Olivier Sand, Heinz Stockinger, Andrew B. Clegg, Erik Bongcam-Rudloff, Jean Salzemann, Vincent Breton, Teresa K. Attwood, Graham Cameron, and Gert Vriend. The EMBRACE web service collection. *Nucleic Acids Research*, 38(suppl 2):W683–W688, 2010.
21. L. Plaza, A.J. Jimeno-Yepes, A. Díaz, and A. Aronson. Studying the correlation between different word sense disambiguation methods and summarization effectiveness in biomedical texts. *BMC bioinformatics*, 12(1):355, 2011.
22. P. Ruch. Automatic assignment of biomedical categories: toward a generic approach. *Bioinformatics*, 22(3):658–664, 2006.
23. M.J. Schuemie, J.A. Kors, and B. Mons. Word sense disambiguation in the biomedical domain: an overview. *Journal of Computational Biology*, 12(5):554–565, 2005.
24. F. Vasilescu, P. Langlais, and G. Lapalme. Evaluating variants of the Lesk approach for disambiguating words. In *Proceedings of the Conference of Language Resources and Evaluations (LREC 2004)*, pages 633–636, 2004.

Redundancy reduction for multi-document summaries using A* search and discriminative training

Ahmet Aker, Trevor Cohn and Robert Gaizauskas

University of Sheffield, UK

Abstract. In this paper we address the problem of optimizing global multi-document summary quality using A* search and discriminative training. Different search strategies have been investigated to find the globally best summary. In them the search is usually guided by an existing prediction model which can distinguish between good and bad summaries. However, this is problematic because the model is not trained to optimize the summary quality but some other peripheral objective. In this work we tackle the global optimization problem using A* search with the training of prediction model intact and demonstrate our method to reduce redundancy within a summary. We use the framework proposed by Aker et al. [1] as a baseline and adapt it to globally improve the summary quality. Our results show significant improvements over the baseline.

1 Introduction

Extractive multi-document summarization (MDS) aims to present the most important parts of multiple documents to the user in a condensed form [9, 13]. This is achieved by identifying a subset of sentences from the document collection which are concatenated to form the summary. Two common challenges in extractive MDS are: *search* – finding the best scoring summary from the documents – and *training* – learning the system parameters to best describe a training set consisting of pairs of documents and reference summaries.

In previous work the search problem is typically decoupled from the training problem. McDonald [14], for example, addresses the search problem by using Integer Linear Programming (ILP). In his ILP problem formulation he adopts the idea of Maximal Marginal Relevance (MMR) [5] to maximize the amount of relevant information in the summary and at the same time to reduce the redundancy within it. Others have also addressed the search problem using a variation of ILP [7, 8] but as well as using different approaches such as stack decoding algorithms [20], genetic algorithms [16] and submodular set function optimisation [12].

By separating search from training these approaches assume the existence of a predictive model which can distinguish between good and bad summaries. This is problematic because the model is not trained to optimize the summary quality but some other peripheral objective. The disconnect between the training and prediction settings compromises the predictive performance of the approach.

An exception is the work of Aker et al. [1], which proposes an integrated framework that trains the full prediction model directly with the search algorithm intact.

Their training algorithm learns parameters such that the best scoring *whole summary* under the model has a high score under an evaluation metric. However they only optimize the summary quality locally and do not take into account global features such as redundancy within the summary.

This paper addresses the redundancy problem within the integrated framework proposed by Aker et al. [1] and thus presents a novel approach to global optimization of summary quality. We present and evaluate our approach for incorporating a redundancy criterion into the framework. Our approach adapts the A* search to global optimization. The core idea of this approach is that redundant sentences are excluded from the summary if their redundancy with respect to the summary created so far exceeds a threshold. In our experiments this threshold is learned automatically from the data instead of being set manually as proposed in previous work.

The paper is structured as follows. Section 2 presents the work of Aker et al., [1], in detail. In Section 3 we describe our modifications to the framework proposed by Aker et al. and our proposed approach to address redundancy in extractive summarization. Section 4 describes our experimental setup to evaluate the proposed approach, and Section 5 the results. Finally, we conclude in Section 6.

2 Background

In this section we first review the work of Aker et al. [1] in detail, which is essential for the understanding of our modifications to their framework.

2.1 Summarization Model

A summarization model is used to score summaries. Summaries are ranked according to these scores, so that in search, the summary with the highest score can be selected. Aker et al. use the summarization model s to score a summary:

$$s(\mathbf{y}|\mathbf{x}) = \sum_{i \in \mathbf{y}} \phi(x_i)\lambda \tag{1}$$

where $\mathbf{x}$ is the document set, composed of k sentences, $\mathbf{y} \subseteq \{1 \ldots k\}$ is the set of indexes selected for the summary, $\phi(\cdot)$ is a feature function that returns a set of features values for each candidate summary and λ is the weight vector associated with the set of features. In search we use the summarization model to find the maximum summary $\hat{\mathbf{y}}$:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} s(\mathbf{y}|\mathbf{x}) \tag{2}$$

2.2 Search

In Aker et al. the creation of a multi-document summary is formulated as a search problem in which the aim is to find a subset of sentences from the entire set to form a summary. The search is also constrained so that the subset of sentences does not exceed the summary length threshold. In search, a search graph is constructed with edges representing the connections between the sentences and states with summaries.

Each node is associated with the information about the summary length and summary score. The authors start with an empty summary (start state) with length 0 and score 0 and follow an outgoing edge to expand it. A new state is created when a new sentence is added to the summary. The new state's length is updated with the number of words of the new sentence. The score of the state is computed under the summarization model described in the previous section. A goal state is any state or summary where it is not possible to add another sentence without exceeding the summary length threshold. The summarization problem is then finding the best scoring path (sum over the sentence scores on this path) between the start state and a goal state.

Aker et al. use the A* search algorithm [17] to efficiently traverse the search graph and accurately find the best scoring path. In A* search a best-first strategy is applied to traverse the graph from a starting state to a goal state. The search requires a scoring function for each state, here $s(\mathbf{y}|\mathbf{x})$ from Equation 1, and a heuristic function that estimates the additional score to get from a given state to a goal state. The search algorithm is guaranteed to converge to the optimal solution if the heuristic function is *admissible*, that is, if the function used to estimate the cost from the current node to the goal never overestimates the actual cost. The authors propose different heuristics with different run-time performances. The reported best performing heuristic is the "final aggregated heuristic". We use this heuristic as baseline and for our modification purposes.

2.3 Training

In Aker et al. training problem is formulated as one of finding model parameters, λ, such that the predicted output, $\hat{\mathbf{y}}$ closely matches the gold standard, $\mathbf{r}$. The quality of the match is measured using ROUGE [10]. In the training the standard machine learning terminology of loss functions, which measure the degree of error in the prediction, $\Delta(\hat{\mathbf{y}}, \mathbf{r})$ is adopted. The loss is formulated as $1 - R$ with R as being the ROUGE score. The training problem is to solve

$$\lambda = \arg\min_{\lambda} \Delta(\hat{\mathbf{y}}, \mathbf{r}) \tag{3}$$

where $\hat{\mathbf{y}}$ and $\mathbf{r}$ are taken to range over the corpus of many document-sets and summaries. The prediction model is trained using the minimum error rate training (MERT) technique [15]. MERT is a first order optimization method using Powell search to find the parameters which minimize the loss on the training data [15]. MERT requires n-best lists which it uses to approximate the full space of possible outcomes. A* search is used to construct these n-best lists and MERT to optimize the objective metric such as ROUGE that is used to measure the summary quality.

3 Addressing redundancy

To address redundancy within a summary we adopt the framework of Aker et al. [1] described in the previous section in that we re-use their summarization and training of the prediction model.

3.1 A* search with redundancy reduction

In this section we present our approach to dealing with redundancy within multi-document summaries, which implement the idea of omitting or *jumping over* redundant sentences when selecting summary-worthy sentences from the input documents. When sentences from the input documents are merged and sorted in a list according to their summary-worthiness, the generation of a summary starts by first including a top summary-worthy sentence into the summary, then the next one until a desired summary length is reached. If a sentence from the list is found to be similar to the ones already included in the summary (i.e. to be redundant), then this sentence should not be included into the summary, but rather *jumped over*. We integrate the idea of *jumping over* redundant sentences into the A* search algorithm described by Aker et al. The difference between our implementation and the one of Aker et al. is the integration of a function jump$(\mathbf{y}, y)$ into the search process. We use this function to jump over a sentence with the index y when it is redundant with respect to the summary $\mathbf{y}$. Thus compared to Aker et al. we do not only skip a sentence if it is too long as it is the case in Aker et al., but also when it is redundant compared to the summary created so far. In our work we replace the jump conditions of Aker et al. with:

$$\text{lengthConstraintsOK} \wedge \text{jump}(\mathbf{y}, y) == F \tag{4}$$

where lengthConstraintsOK represents the situation when the next sentence does not violate the summary length in Aker et al. and jump$(\mathbf{y}, y) == F$ the case where the next sentence is not redundant and therefore not to be jumped over.

Jump based on redundancy threshold (JRT): We use the similarity score of a sentence x_i with respect to the summary $\mathbf{y}$ and a similarity or redundancy threshold R to decide whether to jump over the sentence or not. In general we jump over a sentence x_i if its similarity score is above R (see Algorithm in 1). The similarity scores are computed using the sim(., .) function shown in Equation 5.

Algorithm 1 Jump when similarity score is above a threshold R, $jump(\mathbf{y}, x_i)$

```
Require: require a similarity or redundancy threshold $R$
if sim$(\mathbf{y}, x_i) \leq R$ then
    return $F$
end if
return $T$
```

$$sim(\mathbf{y}, x_j) = \frac{1}{n} \sum_{l=1}^{n} \frac{|ngrams(\mathbf{y}, l) \bigcap ngrams(x_j, l)|}{|ngrams(x_j, l)|} \tag{5}$$

where ngrams$(\mathbf{y}, n)$ is the set of n-grams in summary $\mathbf{y}$ and ngrams(x_j, n) in sentence x_j respectively. This method returns 0 if $\mathbf{y}$ and x_j do not share any n-grams. When all n-grams of x_j are found in the list of n-grams of $\mathbf{y}$ the method returns 1. Note that we use this function to only see how many n-grams of x_j are found in $\mathbf{y}$. The other direction is less important for our purpose. The idea of omitting redundant sentences if their redundancy score exceeds a threshold has already been introduced in previous work [4, 11, 18, 19]. However, in contrast to these studies, in which the redundancy threshold is set manually, we learn it automatically.

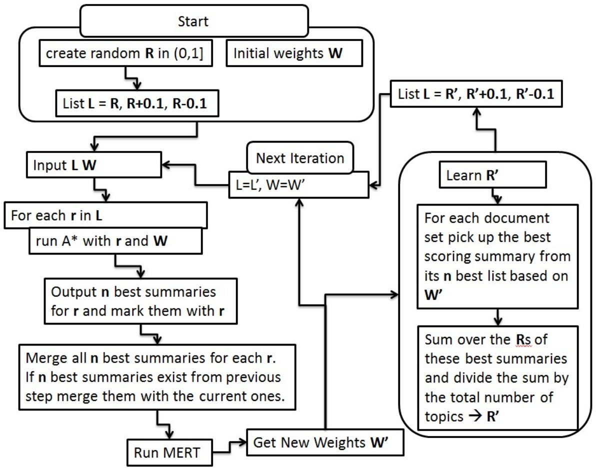

Start
create random **R** in (0,1]
Initial weights **W**
List **L** = **R**, **R+0.1**, **R-0.1**
Input **L W**
For each **r** in **L**
run A* with **r** and **W**
Output **n** best summaries for **r** and mark them with **r**
Merge all **n** best summaries for each **r**. If **n** best summaries exist from previous step merge them with the current ones.
Run MERT
Get New Weights **W'**
Next Iteration
L=L', W=W'
List **L** = **R'**, **R'+0.1**, **R'-0.1**
Learn **R'**
For each document set pick up the best scoring summary from its **n** best list based on **W'**
Sum over the **Rs** of these best summaries and divide the sum by the total number of topics → **R'**

Fig. 1. Learning the redundancy threshold R. The learning procedure starts in the box denoted with *Start*.

To learn the redundancy threshold R we make use of the entire framework (search and training) and proceed as shown in Figure 1. In the beginning (the top left of the figure) we create a random $R \in (0, 1]$. In addition to this R we generate two further values: $R + 0.1 \leq 1$ and $R - 0.1 > 0$. These two additional numbers are used to move R towards its optimum value. All three Rs are used to generate n best summaries using A* search. In the A* search we also require a prediction model to score the sentences. For this we start with an initial prediction model (initial feature weights W). For each of the R values (denoted with r in the figure) we then create an n best list using A* search leading to $3 \times n$ summaries. If there are summaries from a previous step we extend the new n best list with them, so that in training the entire history of n best lists is provided. For each summary its corresponding R value is known. Next, these n best summaries are input to MERT to train new weights W', i.e. a new prediction model. After obtaining W' we can pick up the summary from the n best summaries created for each document set MERT has used to come up with W'. We sum the R values of those summaries (in total m for m document sets) and divide the sum by m to obtain the new R'. We replace R with R' and W with W' and repeat the entire process until no new summaries are added to the n best list, when the process stops. Depending on which R was used to generate the best summaries (R, $R + 0.1$ or $R - 0.1$), the optimal value for R ((R that leads to best summaries under the ROUGE metric)) will choose its direction either towards > 0 or ≤ 1.

4 Experimental settings

In this section we describe the data used in the experiments, our summarization system and the training and testing procedure.

4.1 Data

For training and testing we use the freely available image corpus described in [3]. The corpus contains 296 images of static located objects (e.g *Eiffel Tower*, *Mont Blanc*) each with a manually assigned place name and object type category (e.g. *church*, *mountain*). For each place name there are up to four model summaries that were extracted manually from existing image descriptions taken from the *VirtualTourist* travel community website. Each summary contains a minimum of 190 and a maximum of 210 words.

4.2 Summarization system

To generate summaries for each of the 296 document sets we use an extractive, query-based multi-document summarization system. It is given three inputs: a query (place name, e.g. *Westminster Abbey*), the object type associated with an image (e.g. *church*) and a set of web-documents retrieved using the place name as query. The summarizer uses the following features described in [2, 1]:

- ***sentencePosition***: Position of the sentence within its document. The first sentence in the document gets the score 1 and the last one gets $\frac{1}{n}$ where n is the number of sentences in the document.
- ***inFirst5***: Binary feature indicating whether the sentence is one of the first 5 sentences of the document.
- ***isStarter***: A sentence gets a binary score if it starts with the query term (e.g. *Westminster Abbey*) or with the object type, e.g. *The church*.
- ***LMProb***: The probability of the sentence under a bi-gram language model. We trained a separate language model on Wikipedia articles about locations for each object type, e.g., *church*, *bridge*, etc. When we generate a summary about a location of type church, for instance, then we apply the church language model on the related input documents.[1]
- ***DepSim***: Similar to *LMProb* we trained a separate dependency pattern model using Wikipedia articles about locations for each object type. As in *LMProb* we use these models to score the input sentences. A sentence is scored based on the number of patterns it contains from the model.
- ***sentenceCount***: Each sentence gets assigned a value of *1*. This feature is used to learn whether summaries with many sentences are better than summaries with few sentences or vice versa.
- ***wordCount***: Number of words in the summary, to decide whether the model should favor long summaries or short ones.

---

[1] For our training and testing sets we manually assigned each location to its corresponding object type.

Table 1. ROUGE scores. In each row the results were obtained with the prediction model trained on the metric of that row.

Recall	Aker et al. [1]	JRT
R2	0.094	**0.109***
RSU4	0.146	**0.167***

Table 2. Example summary about the query *Akershus Castle*.

Norwegian Royalty have been buried in the Royal Mausoleum in the castle. During the 17th and 18th century the castle fell into decay, and restoration work only started in 1899. The Akershus castle and fortress are located on the eastern side of the Oslo harbor. The fortress was first used in battle in 1306. The original Akershus Castle is located inside the fortress. Akershus Fortress (Norwegian: Akershus Festning) is the old castle built to protect Oslo, the capital of Norway. The fortress was built in 1299, and the meaning of the name is 'the (fortified) house of (the district) Aker'. In the 1600s a castle (or in norsk, "slott") was built. In the reign of Christian IV the medieval stronghold was converted into a Renaissance castle and the fortifications were extended. Guided tours of the fortress in the summer, all year on request. The services are announced in the newspapers and are open to all. During World War II, several people were executed here by the German occupiers. The fortress was reconstructed several times to withstand increasing fighting power. The castle is well positioned overlooking Oslo's harbour. The fortress was strategically important for Oslo and therefore for Norway as well.

5 Results

We use 191 document sets for training and 105 for testing. When training the prediction model we use ROUGE as a metric to maximize because it is also used for automatic summary evaluation in DUC[2] and TAC.[3] In particular, following DUC and TAC we use ROUGE 2 (R-2) and ROUGE SU4 (R-SU4) for both in training and testing. R-2 computes the number of bi-gram overlaps between the automatic and model summaries. R-SU4 measures uni-gram overlaps between two text units but also bi-grams composed of non-contiguous words, with a maximum of four words between the words. The results of our experiments are shown in Table 1.

As shown in Table 1 the results achieved with the JRT method where we learn a redundancy threshold R automatically are better than the ones obtained using the setting without the idea of jump. The JRT method significantly[4] ($p < 0.001$) outperforms the method of Aker et al..[5]

The values of the learnt redundancy threshold R differ for different ROUGE metrics: for R2 this is 0.5338 and for RSU4 0.4675. The different R values are expected given the different properties of R2 and RSU4. Compared to R2 the redundancy threshold for RSU4 is more strict which reflects the way RSU4 works. As mentioned in Section 4, RUS4 measure the uni-gram overlap between two text units but also bi-grams where gaps of up to four words are allowed between the words. This means that RSU4 is able to capture more similarities between sentences than R2, where single word overlaps are not captured. In R2 gaps within a bi-gram are allowed. For example bi-grams

---

[2] http://duc.nist.gov/

[3] http://www.nist.gov/tac/

[4] We use a two-tail paired T-test to compute significance test.

[5] We have also studied different alternative methods to the JRT one to be used in the jump$(.,.)$ function such as favoring the following sentence to the current one if it is less redundant than the current one or combining the redundancy scores with the actual raw scores of the sentences and jumping only over the current sentence if the combined score is less than the combined score of the following sentence. However, the results by these alternative methods led only to moderate improvement over the baseline. For this reason we do not report those results.

Table 3. Readability evaluation results: Each cell shows the percentage of summaries scoring the ranking score heading the column for each criterion in the row as produced by the summary method indicated by the subcolumn heading – Aker et al. (RW) and JRT. The numbers indicate the percentage values averaged over the three people.

	5		4		3		2		1	
Criterion	RW	JRT	RW	JRT	RW	JRT	RW	JRT	RW	JRT
clarity	6.2	22.4	41.7	73.5	29.2	2.0	20.8	0	2.1	2.0
coherence	6.2	28.6	18.8	42.9	33.3	24.5	37.5	4.1	4.2	0
focus	6.2	26.5	33.3	61.2	29.2	12.2	29.2	0	2.1	0
grammar	4.2	12.2	58.3	67.3	12.5	4.1	20.8	14.3	4.2	2.0
redundancy	4.2	8.2	8.3	61.2	2.1	12.2	41.7	18.4	43.8	0

Table 4. Readability evaluation results: Each cell shows the percentage of summaries scoring the ranking score $>= 4$ for each criterion in the row as produced by the summary method indicated by column heading – Aker et al. (RW) and JRT. The numbers indicate the percentage values averaged over the three people.

Criterion	RW	JRT
clarity	47.9	95.9
coherence	25	71.5
focus	39.5	87.7
grammar	30.2	79.5
redundancy	12.5	69.4

AB and $A??B$ are identical in RSU4, but not in R2. Consequently, a stricter redundancy threshold is required in RSU4 than in R2. This fact illustrates also that there cannot be a single R for every ROUGE metric and highlights the importance of learning it for each of the ROUGE metrics separately.

From the example summary about the query *Akershus Castle* shown in Table 2 we can see that the summary does capture a variety of facts about the castle such as when the castle was built, where it is located, etc. This type of essential information about the castle occurs only once in the summary. What is repeated in most of the sentences are referring expressions such as the name of the place (*Akershus Castle*) or the object type (*the castle* or *the fortress*). Sentences containing referring expressions are more likely to contain relevant information about the castle in the model summaries than sentences which do not contain such expressions. The redundancy thresholds are set to allow some repetition in the summary, which means that MERT learned to allow referring expressions to be repeated in the summary, so it can maximize the ROUGE metrics.

We also evaluated our summaries using a readability assessment as in DUC and TAC. DUC and TAC manually assess the quality of automatically generated summaries by asking human subjects to score each summary using five criteria – *grammaticality, redundancy, clarity, focus* and *structure*. Each criterion is scored on a five point scale with high scores indicating a better result [6]. In the evaluation we asked three people to assess the summaries. Each person was shown 100 summaries (50 from each summary type selected randomly from the entire test set of 105 places). The summaries were shown in a random way. The results of the manual evaluation are shown in Table 3. Table 4 shows percentage values of summaries which achieved scores at levels four or above.

We see from Table 3 that JRT type summaries perform much better than in the Aker et al. setting where summaries are generated without redundancy detection. The percentage values at levels 5 and 4 (see Table 4) show that the JRT summaries have more clarity (95.9% of the summaries), are more coherent (71.5% of the summaries), have better focus (87.7% of the summaries) and grammar (79.5% of the summaries) and contain less redundant information (69.4% of the summaries) than the ones generated in the $wordLimit$ setting (47.9%, 25%, 39.5%, 30.2% and 12.5%). The substantial improvement in redundancy from the Aker et al. setting to JRT demonstrated that incorporating a jump into a summarization system adds to redundancy reduction but also improves other quality aspects of the summary.

6 Conclusion

In this paper we proposed and evaluated an automatic method for improving the global quality of extractive multi-document summaries by means of reducing the redundancy within summaries. We used the framework proposed by Aker et al. [1] as a baseline because it uses a combined search and training approach to maximize the summary quality locally and adapted it for global optimization. We demonstrated that our proposed method, JRT, for redundancy reduction improves the quality of the summary over the baseline as indicated by the ROUGE metric and manual evaluation. In JRT we jump over sentences which are more similar than a similarity threshold R learnt automatically. We have seen that the properties of different ROUGE metrics require different redundancy thresholds, so that R must be learned for each ROUGE metric separately. The automatically determined R values appeared to be neither too strict nor too generous as they allow referring expressions to be redundant in the output summary but not whole factual assertions. This reflects the fact that in the model summaries the sentences containing referring expressions are also those which contain the most relevant information about a query.

In future work we intend to address several issues arising from this work. First, we intend to incorporate semantic knowledge into computation of the redundancy scores. Currently, when learning the R value we purely use surface level comparison and compute the redundancy score between a sentence and a summary using uni and bi-gram lexical overlaps. By doing this we can only capture the repetition of information units if they are expressed in the same way. We believe that the results can be further improved if techniques to detect semantic overlaps are also used. Second, we aim to address the issue of information flow, which is currently missing in the output summaries. From the example summary we can see that the summary reads like the bag of sentences. By integrating flow into the A* search algorithm we hope to improve the readability of the summaries.

References

1. Aker, A., Cohn, T., Gaizauskas, R.: Multi-document summarization using A* search and discriminative training. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. pp. 482–491. Association for Computational Linguistics (2010)

2. Aker, A., Gaizauskas, R.: Generating image descriptions using dependency relational patterns. Proc. of the ACL 2010, Upsala, Sweden (2010)
3. Aker, A., Gaizauskas, R.: Model Summaries for Location-related Images. In: Proc. of the LREC-2010 Conference (2010)
4. Barzilay, R., McKeown, K., Elhadad, M.: Information fusion in the context of multi-document summarization. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. pp. 550–557. Association for Computational Linguistics (1999)
5. Carbonell, J., Goldstein, J.: The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 335–336. ACM (1998)
6. Dang, H.: Overview of DUC 2005. DUC 05 Workshop at HLT/EMNLP (2005)
7. Gillick, D., Favre, B.: A scalable global model for summarization. In: Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing. pp. 10–18. Association for Computational Linguistics (2009)
8. Gillick, D., Riedhammer, K., Favre, B., Hakkani-Tür, D.: A global optimization framework for meeting summarization. In: Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on. pp. 4769–4772. IEEE (2009)
9. Jones, K.: Automatic summarizing: factors and directions. Advances in Automatic Text Summarization pp. 1–12 (1999)
10. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out: Proc. of the ACL-04 Workshop pp. 74–81 (2004)
11. Lin, C., Hovy, E.: From single to multi-document summarization: A prototype system and its evaluation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 457–464. Association for Computational Linguistics (2002)
12. Lin, H., Bilmes, J.: Multi-document summarization via budgeted maximization of submodular functions. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 912–920. Association for Computational Linguistics (2010)
13. Mani, I., Maybury, M.: Advances in automatic text summarization. the MIT Press (1999)
14. McDonald, R.: A study of global inference algorithms in multi-document summarization. Advances in Information Retrieval pp. 557–564 (2007)
15. Och, F.: Minimum error rate training in statistical machine translation. Proc. of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1 p. 167 (2003)
16. Riedhammer, K., Gillick, D., Favre, B., Hakkani-Tür, D.: Packing the meeting summarization knapsack. Proc. Interspeech, Brisbane, Australia (2008)
17. Russell, S., Norvig, P., Canny, J., Malik, J., Edwards, D.: Artificial intelligence: a modern approach. Prentice hall Englewood Cliffs, NJ (1995)
18. Saggion, H.: A robust and adaptable summarization tool. Traitement Automatique des Langues 49(2) (2008)
19. Sauper, C., Barzilay, R.: Automatically generating wikipedia articles: A structure-aware approach. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1. pp. 208–216. Association for Computational Linguistics (2009)
20. Yih, W., Goodman, J., Vanderwende, L., Suzuki, H.: Multi-document summarization by maximizing informative content-words. In: Proceedings of IJCAI. vol. 7 (2007)

A Dependency Relation-based Method to Identify Attributive Relations and Its Application in Text Summarization

Shamima Mithun and Leila Kosseim

Concordia University
Department of Computer Science and Software Engineering
Montreal, Quebec, Canada
{s_mithun, kosseim}@encs.concordia.ca

Abstract. In this paper, we propose a domain and genre-independent approach to identify the discourse relation called *attributive*, included in Grimes' relation list [7]. An attributive relation provides details about an entity or an event or can be used to illustrate a particular feature about a concept or an entity. Since attributive relations describe attributes or features of an object or an event, they are often used in text summarization (e.g. [2]) and question answering systems (e.g. [12]). However, to our knowledge, no previous work has focused on tagging *attributive* relations automatically. We propose an automatic domain and genre-independent approach to tag attributive relations by utilizing dependency relations of words based on dependency grammars [3]. In this paper, we also show how attributive relations can be utilized in text summarization. By using a subset of the BLOG06[1] corpus, we have evaluated the accuracy of our attributive classifier and compared it to a baseline and human performance using precision, recall, and F-Measure. The evaluation results show that our approach compares favorably with human performance.

1 Introduction

According to [15], "Discourse relations - relations that hold together different parts (i.e. proposition, sentence, or paragraph) of the discourse - are partly responsible for the perceived coherence of a text". In a discourse, different kinds of relations such as *contrast*, *causality* or *elaboration* may be expressed. For example, in the sentence *"If you want the full Vista experience, you'll want a heavy system and graphics hardware, and lots of memory"*, the first and second clauses are related through the discourse relation *condition.* The use of discourse relations have been found useful in many applications such as document summarization (e.g. [1, 2, 13]) and question answering (e.g. [10, 12]). However, these relations are often not considered in computational language applications because domain and genre-independent robust discourse parsers are very few.

---

[1] http://ir.dcs.gla.ac.uk/test_collections/blog06info.html

In this paper, we propose a domain and genre-independent approach to identify the discourse relation called *attributive*, included in Grimes' relation list [7]. An attributive relation provides details about an entity or an event. For example, in *Mary has a pink coat.*, the sentence exhibits an attributive relation because it provides details about the entity *coat*. Attributive relations can also be used to illustrate a particular feature about a concept or an entity - e.g. *Picasa makes sure your pictures are always organized.* The sentence of this example also contains an attributive relation since it is describing a particular feature of the entity *Picasa*. Even though attributive relations are often used in summarization (e.g. [13]) and question answering systems (e.g. [12]), to our knowledge, no previous work has focused on tagging *attributive* relations automatically. We propose an automatic domain and genre-independent approach to identify whether a sentence contains an attributive relation by utilizing dependency relations of words based on dependency grammars [3]. In this paper, we also show how attributive relations can be utilized in text summarization and how our tagger has been evaluated in that context.

2 Related Work

Currently, to identify discourse relations automatically from multi-documents, only a few approaches are available. The most notable ones are the SPADE parser [14], Jindal et al.'s approach [8], and HILDA [6].

The SPADE parser [14] was developed within the framework of RST (Rhetorical Structure Theory). The SPADE parser identifies discourse relations within a sentence by first identifying elementary discourse units (EDU)s, then identifying discourse relations between two EDUs (clauses) by following the RST theory. However, the attributive relation is not included within these relations.

Another discourse parser is presented in [8]. This parser focuses on tagging the comparison relation. In order to label a clause as containing a *comparison* relation, [8] used a set of keywords and annotated texts, and generate patterns for comparison sentence mining. A Naïve Bayes classifier is then used using the patterns as features to learn a 2-class classifier (comparison and non-comparison). This approach is used in our summarization system (Section 4.2) to tag intra-clausal comparison relations; but again, it does not deal with attributive relations.

Another notable work is that of [6] who designed the discourse parser called HILDA[2] (HIgh-Level Discourse Analyzer) which can tag discourse relations at the text level. First, this parser extracts different lexical and syntactical features from the input texts. Then the parser is trained using the RST Discourse Treebank[3] (RST-DT) corpus. This parser consists of two SVM classifiers. The first classifier finds the most appropriate relation between two textual units and the second classifier verifies whether two adjacent text units should be merged to

---

[2] HILDA: http://nlp.prendingerlab.net/hilda

[3] http://www.isi.edu/ marcu/discourse/Corpora.html

form a new subtree. However, the source of the parser is not publicly available and again does not tag attributive relations.

Other notable works on discourse parsing and discourse segmentation are proposed by (e.g. [11, 16]). However, the attributive relation is not tagged by any of these approaches. Discourse parsing systems are being developed in other languages than English such as [4] for Spanish.

3 A Method based on Dependency Relations

According to [12], an attributive relation provides details about an entity or event. It can be used to illustrate a particular attribute or feature about a concept or an entity. For example, *Subway sells custom sandwiches and salads.* - contains an attributive relation since it provides an attribute about *Subway.* This relation has been used successfully by [12] in question answering and natural language generation. However, currently, no automatic approach is available to identify attributive relations.

To develop our method to identify attributive relations, we have performed a corpus analysis of 200 attributive sentences from the BLOG06 corpus[4].

A first analysis of our development set showed that 83% of the time, attributive relations occur within a clause; as opposed to many other discourse relations that span across clauses. Due to this, our approach is based on the analysis of single clauses. To identify attributive relations automatically, similarly to Fei et al.'s work [5], we have used dependency relations of words based on dependency grammars [3].

Table 1. Sample Dependency Relations between Words (taken from [5])

Relation Name	Description	Examples	Parent	Child
subj	subject	I will go	go	I
obj	object	tell her	tell	her
mod	modifier	a nice story	story	nice

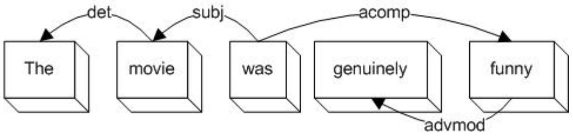
Dependency relations of words are defined based on dependency grammars [3]. They refer to the binary relations between two words where one word is the parent (or head) and the other word is the child (or modifier). In this representation, one word can be associated with only one parent but with many children (one word can modify only one other word, but a word can have several modifiers). Therefore, when the dependency relations of a sentence is created it will be in the form of a tree (called a dependency tree). Typical dependency relations are shown in Table 1.

---

[4] BLOG06 is a TREC test collection, created and distributed by the University of Glasgow to support research on information retrieval and related technologies. BLOG06 consists of 100,649 blogs which were collected over an 11 week period (a total of 77 days) from late 2005 and early 2006. The total size of collection is 25 gigabytes. In this corpus, blogs vary significantly in size, ranging from 44 words to 3000 words.

Different words of a sentence can be related using dependency relations directly or based on the transitivity of these relations. For example, the dependency relations of the sentence *"The movie was genuinely funny."* as produced by the Stanford parser[5] is shown in Figure 1.

Fig. 1. Dependency Relations for the Sentence: *The movie was genuinely funny.*

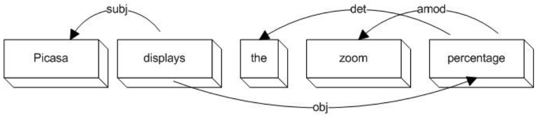


The head of the arrow points to the child, the tail comes from the parent, and the tag on the arrow indicates the dependency relation type. For example, in Figure 1, both words *movie* and *funny* are modifiers of the word *was*. While, the word *movie* is the subject of the word *was*, the word *funny* is a direct adjectival complement (`acomp`) to the word *was*. With the help of dependency relations, it is possible to find how different words of a sentence are related.

In order to develop our classifier, we have first parsed the sentences of our development set using the Stanford parser. A manual analysis of these parses showed that to be classified as an *attributive* sentence, the topic of the sentence needs to be the descendant of a verb and be in a subject or object relation with it. However, the topic and the verb can be related in several ways; which we describe by 3 heuristic rules:

Heuristic 1: The Topic is a Direct Nominal Subject: The *topic* is a direct nominal subject, a noun phrase that is the syntactic subject of the *verb* (e.g., `subj` in the Stanford parser).

Fig. 2. Example of Heuristic 1 to Tag the Attributive Relation



For example, the sentence *"Picasa displays the zoom percentage"* contains an attributive relation where the topic "*Picasa*" is directly related to the verb "*displays*" using the dependency relation `subj` (shown in Figure 2). This is the most frequently encountered dependency relation which occurs within a clause in our attributive development set and accounts for 42% of the development set.

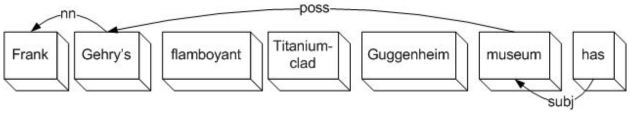
Heuristic 2: A Noun is the Syntactic Subject and the Topic is a Modifier of the Noun: A noun is the syntactic subject of the sentence and the *topic* is a modifier of the noun. This heuristic rule accounts for modifiers that can be a noun compound modifier (e.g., `nn` in the Stanford parser),

---

[5] http://nlp.stanford.edu/software/lex-parser.shtml

a propositional modifier (e.g., `prep` in the Stanford parser) or a possession modifier (e.g., `poss` in the Stanford parser).

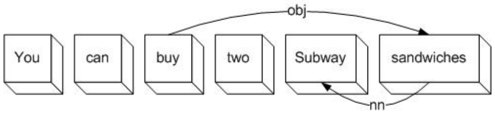
Fig. 3. Example of Heuristic 2 to Tag the Attributive Relation



For example, the sentence *"Frank Gehry's flamboyant, titanium-clad Guggenheim Museum has a similar relationship to the old, masonry city around it."* contains an attributive relation where the noun "*Museum*" is the subject of the sentence and the topic "*Frank Gehry*" is a possession modifier of the noun "*Museum*" (a partial dependency tree is shown in Figure 3). These dependency relations account for 38% of the development set.

Heuristic 3: A Noun is the Syntactic Direct Object and the Topic is a Modifier of the Noun: A noun is the syntactic direct object of the *verb* (e.g., `obj` in the Stanford parser) and the *topic* is a modifier of the noun. Under this heuristic rule, a modifier can be a noun compound modifier (e.g., `nn` in the Stanford parser).

Fig. 4. Example of Heuristic 3 to Tag the Attributive Relation



For example, the sentence *"You can buy two Subway sandwiches for $7.99 on sunday."* contains an attributive relation where the noun "*sandwiches*" is the object of the verb "*buy*" and the *topic* "*Subway*" is a modifier of the noun '*sandwiches*" (a partial dependency tree is shown in Figure 4). These relations account for 16% of the development set.

Given a sentence and a topic, our rule-based classifier tries to determine if any of the 3 heuristics shown above are applicable. If this is the case, it tags the sentence as attributive.

The next section will discuss how attributive relations can be used in blog summarization and how our approach has been evaluated in that context.

4 Evaluation

To evaluate our attributive tagger, we have performed both an intrinsic and an extrinsic evaluation.

4.1 Intrinsic Evaluation

For the intrinsic evaluation, we have evaluated the performance of our attributive classifier against a manually created gold standard using precision (P), recall (R), and F-Measure (F). For this evaluation, since no standard dataset was available, we have developed our own test set containing 400 sentences from the BLOG06 corpus; where two annotators manually tagged 200 sentences as attributive and 200 as non-attributive. Discrepancy between annotators was settled through discussion to arrive at a consensus. It must be noted that both the development and the test sets contain no common sentences.

In this evaluation, we have also calculated and compared the baseline and human performance with our classifier's performance. These were computed as follows: the baseline method tags a sentence as attributive if the topic of the sentence is the direct nominal subject (i.e. heuristic rule 1 in Section 3). This method was chosen because it was the most frequently encountered dependency relation in our attributive development set (42% of the times). On the other hand, to evaluate the human performance to tag attributive relations, we asked two human participants to annotate 100 sentences from the test corpus. These 100 sentences were randomly selected from the corpus where 50 sentences are positive examples (e.g. attributive) and 50 sentences are negative examples (e.g. non-attributive). At the end, human performance was compared with the gold standard using precision, recall and F-measure.

Table 2. Intrinsic Evaluation of the Attributive Tagger

	Precision	**Recall**	**F-Measure**
Attributive Classifier	77%	76%	77%
Baseline	39%	67%	49%
Human Performance	79%	88%	83%

Table 2 shows the evaluation results of our attributive classifier. The table also shows the baseline and human performance for identifying attributive relations. We can see that the performance of the human participants (F-Measure = 83%) is much higher than the baseline (F-Measure = 49%). Our attributive classifier (F-Measure = 77%) performs better than the baseline and is a little weaker than human participants.

From the evaluation results, we can see that the precision and the overall F-Measure score of human participants are not very high (around 80%). We suspect that the reason behind this is that even though attributive relations are useful in natural language research, this relation is not well recognized and humans may not be very familiar with it. To verify this, we have calculated the inter-annotator agreement in tagging attributive sentences using Cohen's kappa. The results show that inter-annotator agreement is moderate according to [9] with a kappa value of 0.51, which seems to support our hypothesis.

4.2 Extrinsic Evaluation

To do the extrinsic evaluation, we have tested our attributive relation identification approach with our BlogSum summarizer [13] and have evaluated its effect on the summaries generated. Let us first describe the summarizer we used and how the tagger was used.

BlogSum BlogSum is a domain-independent query-based blog summarization system that uses intra-sentential discourse relations within the framework of schemata. The heart of BlogSum is based on discourse relations and text schemata.

Text schemata are patterns of discourse organization used to achieve different communicative goals. Text schemata were first introduced by McKeown [12] based on the observation that specific types of schemata are more effective to achieve a particular communicative goal. Schema-based approaches were also used by other researchers in the context of question answering and text generation to generate relevant and coherent text. However, schema-based approaches are usually domain-dependent where the domain knowledge is pre-compiled and explicitly represented in knowledge bases or is used for structured documents (e.g. Wikipedia articles).

BlogSum works in the following way: First candidate sentences are ranked using the topic and question similarity to give priority to topic and question relevant sentences. Since BlogSum works on blogs, which are opinionated in nature, to rank a sentence, the sentence polarity (e.g. positive, negative or neutral) is calculated using a subjectivity score. The subjectivity score of a sentence is also used to calculate its relevance to the question. To extract and rank sentences, our approach calculates a score for each sentence using the features shown below:

$$Sentence\ Score = Question\ Similarity + Topic\ Similarity + |SubjectivityScore|$$

where, question similarity and topic similarity are calculated using cosine similarity based on words *tf.idf* and subjectivity score is calculated using a dictionary-based approach using the MPQA lexicon[6], which contains more than 8000 entries of polarity words.

Then sentences are categorized based on the discourse relations that they convey. This step is critical because the automatic identification of discourse relations renders BlogSum independent of the domain. This step also plays a key role in content selection and summary coherence as schemata are designed using these relations. For predicate identification, BlogSum considers 28 discourse relations including the attributive relation. Then four different approaches are used to identify these predicates: a) the SPADE parser [14] (see Section 2); b) a comparison relations classifier adapted from [8] (see Section 2); c) a topic-opinion discourse relation tagger, and d) our own attributive tagger described in Section 3. It is to be noted that an analysis of 221 random summary sentences from the

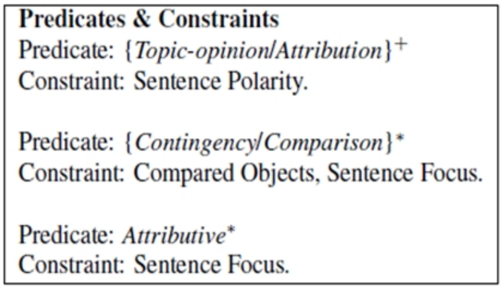
---

[6] MPQA: http://www.cs.pitt.edu/mpqa

BLOG06 corpus shows that 32% of the sentences were tagged by our attributive tagger.

In order not to answer all questions the same way, BlogSum uses different schemata to generate a summary that answers specific types of questions. Each schema is designed based on giving priority to its associated question type and subjective sentences as summaries for opinionated texts are generated. Each schema specifies the types of predicates and the order in which they should appear in the output summary for a particular question type.

Fig. 5. A Sample Discourse Schema used in BlogSum



Predicates & Constraints
Predicate: $\{$*Topic-opinion/Attribution*$\}^+$
Constraint: Sentence Polarity.

Predicate: $\{$*Contingency/Comparison*$\}^*$
Constraint: Compared Objects, Sentence Focus.

Predicate: *Attributive**
Constraint: Sentence Focus.

Figure 5 shows a sample schema that is used to answer *reason* questions (e.g. “Why do people like Picasa?”). According to this schema, one or more topic-opinion or attribution predicates followed by zero or many contingency or comparison predicates followed by zero or many attributive predicates can be used[7].

Finally the most appropriate schema is selected based on a given question type; and candidate sentences fill particular slots in the selected schema based on which discourse relations they contain.

Extrinsic Evaluation within BlogSum To evaluate the performance of our tagger in an extrinsic evaluation, we used it within BlogSum. In these experiments, we used the original ranked list of candidate sentences before applying the discourse schema, called OList, as a baseline, and compared them to the BlogSum-generated summaries with and without the tagger. We used the Text Analysis Conference (TAC) 2008 opinion summarization dataset[8] which is a subset of BLOG06. The TAC 2008 opinion summarization dataset consists of 50 questions on 28 topics; on each topic one or two questions were asked and 9 to 39 relevant documents were given. For each question, one summary was generated by OList and two by BlogSum and the maximum summary length was restricted to 250 words.

---

[7] Following [12]’s notations, the symbol / indicates an alternative, * indicates that the item may appear 0 to n times, + indicates that the item may appear 1 to n times.

[8] http://www.nist.gov/tac/

With this dataset, we have automatically evaluated how BlogSum performs using the standard ROUGE-2 and ROUGE-SU4 measures. For this experiment, on each question, two summaries were generated by BlogSum; one using the attributive tagger and the other without using the attributive tagger. In this experiment, ROUGE scores are also calculated for all 36 submissions in the TAC 2008 opinion summarization track. Table 3 shows the evaluation results.

Table 3. Extrinsic Evaluation of the Attributive Tagger

System Name	**ROUGE-2 (F)**	**ROUGE-SU4 (F)**
TAC Average	0.069	0.086
OList - Baseline	0.102	0.107
BlogSum without Attributive Tagger	0.113	0.115
BlogSum with Attributive Tagger	0.125	0.128
TAC Best	0.130	0.139

The table shows that BlogSum performs better than OList, and performs better with the use of the attributive tagger using both ROUGE-2 and ROUGE-SU4 metrics. Without using the attributive tagger, BlogSum misses many question relevant sentences whereas the inclusion of the attributive tagger helps to incorporate those relevant sentences into the final summary. This result indicates that our attributive tagger helps to include question relevant sentences without including noisy sentences thus improving the summary content. These results also confirms the correctness and usefulness of our tagger.

Compared to the other systems that participated to the TAC 2008 opinion summarization track, BlogSum performed very competitively; its F-Measure score difference from the TAC best system is very small. Both BlogSum and OList performed better than the TAC average systems.

5 Conclusion and Future Work

In this paper, we have presented a domain and genre-independent approach to identify attributive discourse relations which provides attributes or features of an object or an event. We have utilized dependency relations of words to identify these relations automatically. Evaluation results show that our approach achieves an F-Measure of 77% on our test-set of blogs, which compares favorably with humans and is much higher than the baseline. We have also showed that attributive relations can be used successfully in an application such as blog summarization to generate informative and question-relevant summaries.

As future work, we would like to evaluate the accuracy of each heuristic and analyze further the performance of our classifier with the goal of improving its performance and deal with attributive relations than span across clauses.

Acknowledgement

The authors would like to thank the anonymous referees for their valuable comments on a previous version of the paper. This work was financially supported by NSERC.

References

[1] Bosma, W.: Query-Based Summarization using Rhetorical Structure Theory. *In Proceedings of the 15th Meeting of Computational Linguistics in the Netherlands CLIN*, (2004), Leiden, Netherlands.

[2] Blair-Goldensohn, S.J., McKeown, K.: Integrating Rhetorical-Semantic Relation Models for Query-Focused Summarization. *In Proceedings of the Document Understanding Conference (DUC) Workshop at NAACL-HLT 2006*, (2006), New York, USA.

[3] de Marneffe, M.C., Manning, C.D.: The Stanford Typed Dependencies Representation. *In Coling 2008: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, 1–8. (2008), Manchester. U.K.

[4] da Cunha, I., SanJuan, E., Torres-Moreno, J-M., Lloberes, M., Castellón, I.: DiSeg 1.0: The First System for Spanish Discourse Segmentation. *J. Expert Systems with Applications*, 39(2):1671–1678 , 2012.

[5] Fei, Z., Huang, X., Wu, L.: Mining the Relation between Sentiment Expression and Target Using Dependency of Words. *PACLIC20: Coling 2008: Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, 257–264 (2008), Wuhan, China.

[6] Feng, V. W., Hirst, G.: Text-level Discourse Parsing with Rich Linguistic Features. *In Proceedings of the The 50th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-2012)*, (2012), Jeju, Korea.

[7] Grimes, J.E.: The Thread of Discourse. Cornell University, NSF-TR-1, NSF-GS-3180, 1972, Ithaca, New York.

[8] Jindal, N., Liu, B.: Identifying Comparative Sentences in Text Documents. *SIGIR'06: In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 244-251 (2006), Washington, USA.

[9] Landis, R.J., Koch. G.G.: A one-way components of variance model for categorical data. *J. Biometrics*, 33(1):671–679, 1977.

[10] Marcu, D.: From Discourse Structures to Text Summaries. *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization.* 1997, 82–88, Madrid, Spain.

[11] Marcu, D.: The Rhetorical Parsing of Unrestricted Texts: A Surface-based Approach. *J. Computational Linguistics*, 26(3):395–448, 2000.

[12] McKeown, K.R.: Discourse Strategies for Generating Natural-Language Text. *J. Artificial Intelligence*, 27(1):1–41, 1985.

[13] Mithun, S.: Exploiting Rhetorical Relations in Blog Summarization. *PhD thesis*, Department of Computer Science and Software Engineering, Concordia University, Montreal, Canada, 2012.

[14] Soricut, R., Marcu, D.: Sentence Level Discourse Parsing using Syntactic and Lexical Information. *NAACL'03: In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 149–156 (2003), Edmonton, Canada.

[15] Taboada, M.: Discourse Markers as Signals (or not) of Rhetorical Relations. *J.Pragmatics*, 38(4):567–592, 2006.

[16] Tofiloski, M., Brooke, J., Taboada, M.: A Syntactic and Lexical-Based Discourse Segmenter. *In Proceedings of Proceedings of the 47th Annual Meeting of ACL* 2009, PA, USA.

Using biomedical databases as knowledge sources for large-scale text mining

Fabio Rinaldi, Institute of Computational Linguistics,
University of Zurich, Switzerland

Abstract

In this paper we discuss how terminological knowledge extracted from biomedical databases can be used effectively in large-scale processing of the biomedical literature. We briefly present an integrated information extraction and text mining environment which is capable of reliably identifying and disambiguating several categories of relevant domain entities, which can then constitute relevant indexing entries in order to allow efficient retrieval of relevant documents and passages. Additionally the system generates ranked lists of candidate interactions among the detected entities, which can be useful for several purposes, from assisted literature curation to question answering systems.

1 Introduction

The rapid increase of novel scientific results in the domain of molecular biology renders it necessary to collect this information in structured repositories, so that it becomes easily accessible to the end users. Well-known databases like UniProt, Mint, IntAct, BioGrid, collect information about proteins and their interactions. PharmGKB [4, 12] curates knowledge about the impact of genetic variation on drug response for clinicians and researchers. The Comparative Toxicogenomics Database (CTD) collects interactions between chemicals and genes in order to support the study on the effects of environmental chemicals on health [5]. A significant amount of manual effort is needed in order to extract from the literature the information required to accurately fill those databases (a process referred to as "curation"). Text mining solutions are increasingly requested to support the process of curation of biomedical databases.

The OntoGene project[1] focuses on the improvement of biomedical text mining through the usage of advanced natural language processing techniques. Our approach relies upon information delivered by a pipeline of NLP tools, including sentence splitting, tokenization, part of speech tagging, term recognition, noun and verb phrase chunking, and a dependency-based syntactic analysis of input sentences [11, 8]. The results of the entity detection feed directly into the process of identification of interactions.

---

[1] `http://www.ontogene.org/`

Different implementations of the OntoGene system have been used for participation in several well-known text mining shared tasks, such as BioCreative, CALBC and BioNLP, obtaining always competitive results. For example, in the BioCreative 2009 challenge the OntoGene system obtained the best results for protein-protein interactions [10]. More recently, within the scope of the SASEBio project (Semi-Automated Semantic Enrichment of the Biomedical Literature), we have developed a user-friendly interface (ODIN: OntoGene Document INspector) which can be used by database curator to inspect the results of the text mining system. The interface is designed to simplify the interaction of the user with the text mining system, allowing for example modification of incorrect results. The system can then learn based upon this interaction.

In the rest of this short paper we briefly describe the OntoGene pipeline architecture and the ODIN interface for assisted curation.[2]

2 Information Extraction

Biomedical terminological resources can be leveraged for construction of large-scale knowledge bases. One example is KaBOB (Knowledge Base of Biology), a large RDF store based upon 17 prominent biomedical daabases. KaBOB contains 5.6-billion RDF-triples [1]. Similar kinds of integrated data networks can be used for knowledge discovery purposes through usage of semantic web technologies (see for example [2]).

In our own work we have used such databases as knowledge sources for the process of semi-automated information extraction. In the rest of this section we describe the OntoGene Text Mining pipeline which is used to (a) provide all basic preprocessing (e.g. tokenization) of the target documents, (b) identify all mentions of domain entities and normalize them to database identifiers, and (c) extract candidate interactions.

2.1 Preprocessing and Detection of Domain Entities

Several large-scale terminological resources are used by the OntoGene system in order to detect names of relevant domain entities in biomedical literature (proteins, genes, chemicals, diseases, etc.) and ground them to widely accepted identifiers assigned by the original database, such as UniProt Knowledgebase, National Center for Biotechnology Information (NCBI) Taxonomy, Proteomics Standards Initiative Molecular Interactions Ontology (PSI-MI), Cell Line Knowledge Base (CLKB), etc.

From the original databases we extract preferred names and synonyms for each term, together with its unique identifier. This information is used to annotate the input documents using an efficient lookup procedure. A term normalization step is used to take into account a number of possible surface variations of the terms. The same normalization is applied to the list of known terms at the beginning of the annotation process, when it is read into memory, and to the candidate terms in the input text, so that a matching between variants of the same term becomes possible despite the differences in the surface strings [8]. For more technical details of the OntoGene terminology recognition process, see [7].

---

[2] Readers interested in more details are invited to consult the journal publications available from the OntoGene web site.

The terminological resource obtained as described above is used to annotate biomedical text in a relatively straightforward way. First, in a preprocessing stage, the input text is transformed into a custom XML format, and sentences and tokens boundaries are identified. For this task, we use the LingPipe tokenizer and sentence splitter which have been trained on biomedical corpora. The tokenizer produces a granular set of tokens, e.g. words that contain a hyphen (such as ‘Pop2p-Cdc18p’) are split into several tokens, revealing the inner structure of such constructs which would allow to discover the interaction mention in “Pop2p-Cdc18p interaction”. Tagging of terms is performed by sequentially processing each token in a sentence and, if it can start a term, annotate the longest possible match (partial overlaps are excluded). In the case of success, all the possible IDs (as found in the term list) are assigned to the candidate term.

Ambiguity is a serious problem for several types of entities. For example names of some proteins and genes can refer to several different database identifiers. For example, *hemoglobin* can refer to human hemoglobin or to mouse hemoglobin (or to any other species). Besides, even in humans there are several different types of hemoglobin. Using knowledge about the organisms which are the focus of the experiments described in each paper we can disambiguate to a large extent entities such as proteins and genes. In the OntoGene pipeline we apply an approach which we first described in [3]. We first create a ranked list of ’focus’ organisms based on all mentions of proteins, genes, cell lines and organisms in the paper. In the disambiguation process we remove all the IDs that do not correspond to an organism present in the list. Additionally, the scores provided for each organism can be used in ranking the candidate IDs for each entity. Such ranking is useful in a semi-automated curation environment where the curator is expected to take the final decision. However, it can also be used in a fully automated environment as a factor in ranking any other derived information, such as interactions where the given entity participates.

2.2 Detection of Interactions

Mentions of relevant domain entities in a given text span are used by the OntoGene system to create candidate interactions. The selected text span can vary from a sentence to a larger observation window. Simple co-occurrence in the selected text span is a low-precision, but high-recall indication of a potential relationship among those entities. In order to obtain better precision the OntoGene system uses the syntactic structure of the sentence, and the global distribution of interactions in the original database. In this section we describe in detail how candidate interactions are ranked by our system, according to their relevance for the original database.

The OntoGene system creates an initial ranking of the candidate relations from the selected text span using only the frequency of the respective entities with the following formula:

$$relscore(e_1, e_2) = (f(e_1) + f(e_2))/f(E)$$

where $f(e_1)$ and $f(e_2)$ are the number of times the entities e_1 and e_2 are observed in the abstract, while $f(E)$ is the total count of all identifiers in the abstract. An additional zone-based boost might be used in some cases (e.g. for entities mentioned in the title).

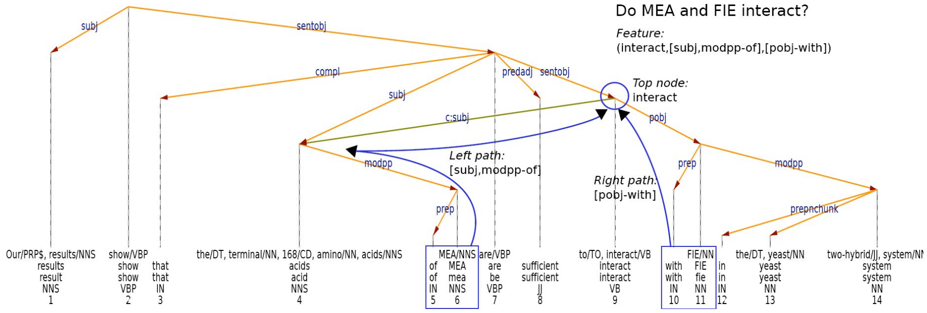
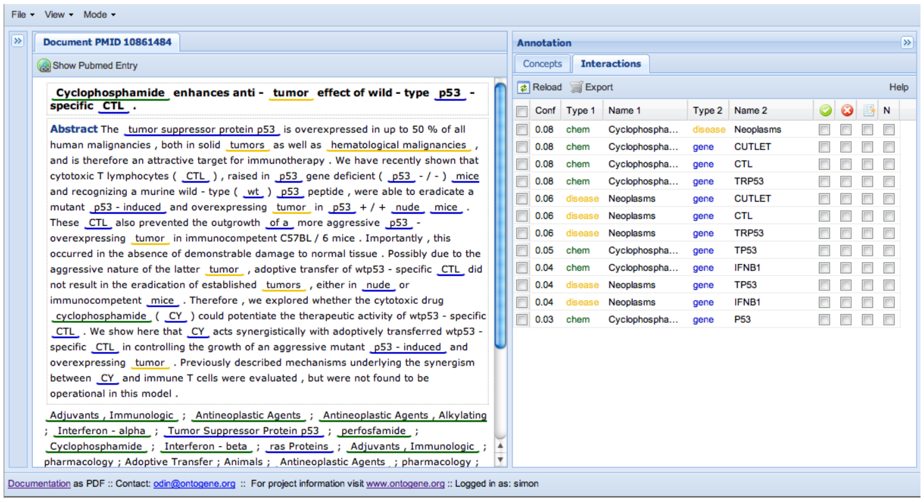


Figure 1: Example of sentence analysis and detection of an interaction.

The OntoGene pipeline makes use of an internally developed dependency parser [13] in order to parse all sentences in the input documents. The information derived from the dependency analysis is used to improve on the baseline ranking for candidate interaction. Besides, the syntactic analysis provides useful information for the extraction of the interaction type. Given two terms identified in the same sentence, a collector traverses the tree from each of the two terms upwards to the lowest common parent node, recording all intermediate nodes and dependency paths along the route. An example of such a traversal can be seen in Figure 1. Such traversals have been used in many PPI applications, they are commonly called tree walks or paths.

Each candidate interaction is assigned a score, obtained by combining several features, including: (1) *Syntactic path*, which encodes the information provided by the dependency structure between the two entities in the candidate interaction; (2) *Known interaction*: in order to better distinguish between 'novel' interactions (more important for the curation process) and 'older' interactions (already known, thus less important for the curation process), we penalize interactions that are already reported in the reference databases, in proportion to their 'age' (date at which the interaction was first reported); (3) *Novelty score*: we also use linguistic clues in order to to distinguish between sentences that report the results detected by the authors (e.g. *"Here we report that..."*) from sentences that report background results. Interactions in 'novelty' sentences are scored higher than interactions in 'background' sentences; (4) *Zoning*: different structural zones of the paper have often different levels of relevance. We observed that novel interactions are often mentioned in the abstract and the conclusions, while the introduction and methods section are less likely and therefore get lower scores; (5) *Pair salience*: the frequency of mentions in the paper of each of the entities in the candidate pair is an important indicator of the relevance of that interaction in the paper. Scores from each feature are then combined and normalized to the `[0,1]` range, in order to produce a ranking for the candidate interactions.

The results of the OntoGene text mining system are made accessible through a curation system called **ODIN** ("OntoGene Document INspector") which allows a user to dynamically inspect the results of the text mining pipeline. An experiment in interactive curation has been performed recently in collaboration with the PharmGKB database [4, 12]. The results of this experiment are described in [6]. [9] provides fur-


File
View
Mode
Document PMID 10861484
Show Pubmed Entry
Annotation
Concepts
Interactions
Reload
Export
Help
Documentation as PDF :: Contact: odin@ontogene.org :: For project information visit www.ontogene.org :: Logged in as: simon

Cyclophosphamide enhances anti - tumor effect of wild - type p53 - specific CTL .

Abstract The tumor suppressor protein p53 is overexpressed in up to 50 % of all human malignancies , both in solid tumors as well as hematological malignancies , and is therefore an attractive target for immunotherapy . We have recently shown that cytotoxic T lymphocytes (CTL) , raised in p53 gene deficient (p53 - / -) mice and recognizing a murine wild - type (wt) p53 peptide , were able to eradicate a mutant p53 - induced and overexpressing tumor in p53 + / + nude mice . These CTL also prevented the outgrowth of a more aggressive p53 - overexpressing tumor in immunocompetent C57BL / 6 mice . Importantly , this occurred in the absence of demonstrable damage to normal tissue . Possibly due to the aggressive nature of the latter tumor , adoptive transfer of wtp53 - specific CTL did not result in the eradication of established tumors , either in nude or immunocompetent mice . Therefore , we explored whether the cytotoxic drug cyclophosphamide (CY) could potentiate the therapeutic activity of wtp53 - specific CTL . We show here that CY acts synergistically with adoptively transferred wtp53 - specific CTL in controlling the growth of an aggressive mutant p53 - induced and overexpressing tumor . Previously described mechanisms underlying the synergism between CY and immune T cells were evaluated , but were not found to be operational in this model .

Adjuvants , Immunologic ; Antineoplastic Agents ; Antineoplastic Agents , Alkylating ; Interferon - alpha ; Tumor Suppressor Protein p53 ; perfosfamide ; Cyclophosphamide ; Interferon - beta ; ras Proteins ; Adjuvants , Immunologic ; pharmacology ; Adoptive Transfer ; Animals ; Antineoplastic Agents ; pharmacology ;

☐	Conf	Type 1	Name 1	Type 2	Name 2	✓	✗	✎	N
☐	0.08	chem	Cyclophospha...	disease	Neoplasms	☐	☐	☐	☐
☐	0.08	chem	Cyclophospha...	gene	CUTLET	☐	☐	☐	☐
☐	0.08	chem	Cyclophospha...	gene	CTL	☐	☐	☐	☐
☐	0.08	chem	Cyclophospha...	gene	TRP53	☐	☐	☐	☐
☐	0.06	disease	Neoplasms	gene	CUTLET	☐	☐	☐	☐
☐	0.06	disease	Neoplasms	gene	CTL	☐	☐	☐	☐
☐	0.06	disease	Neoplasms	gene	TRP53	☐	☐	☐	☐
☐	0.05	chem	Cyclophospha...	gene	TP53	☐	☐	☐	☐
☐	0.04	chem	Cyclophospha...	gene	IFNB1	☐	☐	☐	☐
☐	0.04	disease	Neoplasms	gene	TP53	☐	☐	☐	☐
☐	0.04	disease	Neoplasms	gene	IFNB1	☐	☐	☐	☐
☐	0.03	chem	Cyclophospha...	gene	P53	☐	☐	☐	☐

Figure 2: Entity annotations and candidate interactions on a sample PubMed abstract

ther details on the architecture of the system. Figure 2 shows a screenshot of ODIN.

3 Conclusion

In this paper we briefly described the OntoGene text mining system, targeted at the extraction of entities and relationships from the biomedical literature. The OntoGene pipeline leverages upon manually curated resources and is capable of reliably identifying entity and relationships which can optionally be delivered using standard semantic-web formats such as RDF or OWL. The long-term vision of the project is a deeper integration of databases and literature.

Acknowledgments

This research is partially funded by the Swiss National Science Foundation (grant 100014-118396/1) and Novartis Pharma AG, NIBR-IT, Text Mining Services, CH-4002, Basel, Switzerland.

References

[1] Michael Bada, Kevin Livingston, and Lawrence Hunter. An ontological representation of biomedical data sources and records. *Bio-Ontologies*, 2011.

[2] Huajun Chen, Li Ding, Zhaohui Wu, Tong Yu, Lavanya Dhanapalan, and Jake Y. Chen. Semantic web for integrated network analysis in biomedicine. *Briefings in Bioinformatics*, 10(2):177–192, 2009.

[3] Thomas Kappeler, Kaarel Kaljurand, and Fabio Rinaldi. TX Task: Automatic Detection of Focus Organisms in Biomedical Publications. In *Proceedings of the BioNLP workshop, Boulder, Colorado*, pages 80–88, 2009.

[4] T.E. Klein, J.T. Chang, M.K. Cho, K.L. Easton, R. Fergerson, M. Hewett, Z. Lin, Y. Liu, S. Liu, D.E. Oliver, D.L. Rubin, F. Shafa, J.M. Stuart, and R.B. Altman. Integrating genotype and phenotype information: An overview of the PharmGKB project. *The Pharmacogenomics Journal*, 1:167–170, 2001.

[5] C.J. Mattingly, M.C. Rosenstein, G.T. Colby, J.N. Forrest Jr, and J.L. Boyer. The Comparative Toxicogenomics Database (CTD): a resource for comparative toxicological studies. *Journal of Experimental Zoology Part A: Comparative Experimental Biology*, 305A(9):689–692, 2006.

[6] Fabio Rinaldi, Simon Clematide, Yael Garten, Michelle Whirl-Carrillo, Li Gong, Joan M. Hebert, Katrin Sangkuhl, Caroline F. Thorn, Teri E. Klein, and Russ B. Altman. Using ODIN for a PharmGKB re-validation experiment. *Database: The Journal of Biological Databases and Curation*, 2012.

[7] Fabio Rinaldi, Kaarel Kaljurand, and Rune Saetre. Terminological resources for text mining over biomedical scientific literature. *Journal of Artificial Intelligence in Medicine*, 52(2):107–114, June 2011.

[8] Fabio Rinaldi, Thomas Kappeler, Kaarel Kaljurand, Gerold Schneider, Manfred Klenner, Simon Clematide, Michael Hess, Jean-Marc von Allmen, Pierre Parisot, Martin Romacker, and Therese Vachon. OntoGene in BioCreative II. *Genome Biology*, 9(Suppl 2):S13, 2008.

[9] Fabio Rinaldi, Gerold Schneider, and Simon Clematide. Relation mining experiments in the pharmacogenomics domain. *Journal of Biomedical Informatics*, 2012. *doi:10.1016/j.jbi.2012.04.014*.

[10] Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Simon Clematide, Therese Vachon, and Martin Romacker. OntoGene in BioCreative II.5. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(3):472–480, 2010.

[11] Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Michael Hess, and Martin Romacker. An Environment for Relation Mining over Richly Annotated Corpora: the case of GENIA. *BMC Bioinformatics*, 7(Suppl 3):S3, 2006.

[12] Katrin Sangkuhl, Dorit S. Berlin, Russ B. Altman, and Teri E. Klein. PharmGKB: Understanding the effects of individual genetic variants. *Drug Metabolism Reviews*, 40(4):539–551, 2008. PMID: 18949600.

[13] Gerold Schneider. *Hybrid Long-Distance Functional Dependency Parsing*. Doctoral Thesis, Institute of Computational Linguistics, University of Zurich, 2008.

Exploiting the UMLS Metathesaurus in the Ontology Alignment Evaluation Initiative

Ernesto Jiménez-Ruiz, Bernardo Cuenca Grau, and Ian Horrocks

Department of Computer Science, University of Oxford
`{ernesto,berg,ian.horrocks}@cs.ox.ac.uk`

Abstract. In this paper we describe how the UMLS Metathesaurus—the most comprehensive effort for integrating medical thesauri and ontologies—is being used within the context of the Ontology Alignment Evaluation Initiative (OAEI). We also present the obtained results in the Large BioMed track of the OAEI 2011.5 campaign where the reference alignments are based on UMLS. Finally, we propose a new reference alignment based on the harmonisation of the outputs of the systems participating in the OAEI Large BioMed track.

1 Introduction

The Ontology Alignment Evaluation Initiative[1] (OAEI) is an international campaign for the systematic evaluation of ontology matching systems —software programs capable of finding correspondences (called *alignments*) between the vocabularies of a given set of input ontologies [22, 7, 9, 23]. The matching problems in the OAEI are organised in several tracks, with each track involving different kinds of test ontologies [7]. The ontologies in the largest test case in the OAEI 2011 contain only 2,000–3,000 classes; however, ontology matching tools have significantly improved in the last few years and there is a need for more challenging and realistic matching problems for which suitable reference alignments exist [22, 7].

UMLS-Metathesaurus (UMLS) [1] is currently the most comprehensive effort for integrating medical thesauri and ontologies, including the National Cancer Institute Thesaurus (NCI) [12, 11], the Foundational Model of Anatomy (FMA) [19] and the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [24], which are large-scale and semantically rich ontologies. NCI, FMA and SNOMED CT are gradually superseding the existing medical classifications and are becoming core platforms for accessing, gathering, and sharing biomedical knowledge and data. Hence, matching such large ontologies represents a very interesting challenge for the OAEI initiative.

In this paper we describe how the UMLS correspondences between NCI, FMA and SNOMED CT have been used as reference alignments for the new *Large BioMed track*[2] in the OAEI initiative. Furthermore we present the results obtained in the OAEI 2011.5 campaign for this track and we propose a new reference alignment based on the harmonisation of the outputs of the participating ontology matching systems.

---

[1] `http://oaei.ontologymatching.org/`

[2] `http://www.cs.ox.ac.uk/isg/projects/SEALS/oaei/`

Table 1. The notion of "Joint" in the MRCONSO file from the UMLS distribution.

CUI	Language	Source	Entity
C0022417	ENG	FMA	Joint
			Set_of_joints
		SNOMED CT	Joint_structure
		NCI	Joint
			Articulation

Table 2. UMLS-based alignment between FMA, NCI and SNOMED CT for the notion of "Joint".

Ontology pair	Generated Alignments
FMA ∼ NCI	$\langle 1, \mathit{FMA}{:}\mathit{Joint}, \mathit{NCI}{:}\mathit{Joint}, 1.0, \mathit{equiv}\rangle$ $\langle 2, \mathit{FMA}{:}\mathit{Joint}, \mathit{NCI}{:}\mathit{Articulation}, 1.0, \mathit{equiv}\rangle$ $\langle 3, \mathit{FMA}{:}\mathit{Set\_of\_joints}, \mathit{NCI}{:}\mathit{Joint}, 1.0, \mathit{equiv}\rangle$ $\langle 4, \mathit{FMA}{:}\mathit{Set\_of\_joints}, \mathit{NCI}{:}\mathit{Articulation}, 1.0, \mathit{equiv}\rangle$
FMA ∼ SNOMED CT	$\langle 5, \mathit{FMA}{:}\mathit{Joint}, \mathit{SNOMED}{:}\mathit{Joint\_structure}, 1.0, \mathit{equiv}\rangle$ $\langle 6, \mathit{FMA}{:}\mathit{Set\_of\_joints}, \mathit{SNOMED}{:}\mathit{Joint\_structure}, 1.0, \mathit{equiv}\rangle$
SNOMED CT ∼ NCI	$\langle 7, \mathit{SNOMED}{:}\mathit{Joint\_structure}, \mathit{NCI}{:}\mathit{Joint}, 1.0, \mathit{equiv}\rangle$ $\langle 8, \mathit{SNOMED}{:}\mathit{Joint\_structure}, \mathit{NCI}{:}\mathit{Articulation}, 1.0, \mathit{equiv}\rangle$

2 The UMLS-based reference alignments

Ontology alignments are often conceptualised as tuples with the form $\langle id, e_1, e_2, n, \rho\rangle$, where id is a unique identifier for the mapping, e_1, e_2 are entities in the vocabulary of the integrated ontologies, n is a numeric confidence measure between 0 and 1, and ρ is a relation between e_1 and e_2, typically subsumption (i.e., e_1 is more specific than e_2) and equivalence (i.e., e_1 and e_2 are synonyms) [8]. The OAEI initiative uses an RDF format to represent the alignments[3] [6] containing the aforementioned elements. Alternatively, OAEI alignments are also represented as OWL 2 subclass and equivalence axioms with the mapping identifier (id) and confidence (n) added as OWL 2 annotation axioms [4].

Although the standard UMLS distribution does not directly provide sets of alignments (in the OAEI sense) between the integrated ontologies, it is relatively straightforward to extract alignment sets from the information provided in the distribution files [15]. Concretely, we have processed the *MRCONSO*[4] file, which contains every entity in UMLS together with its *concept unique identifier* (CUI), its source vocabulary (e.g. FMA), its language (e.g. English), and other attributes not relevant for the OAEI. Table 1 shows an excerpt from the *MRCONSO* file associated to the notion of "Joint".

It follows from Table 1 that the notion of "Joint" is shared by FMA, SNOMED CT and NCI. In particular, FMA contains the entities $Joint$ and Set_of_joints, NCI the entities $Articulation$ and $Joint$, and SNOMED CT only the entity $Joint\_structure$. All these entities have been annotated with the same CUI C0022417 and therefore, according to UMLS's intended meaning, they are synonyms. Then, for each pair of entities $e1$ and $e2$ from *different* sources and annotated with the same CUI, we have

---

[3] `http://alignapi.gforge.inria.fr/format.html`

[4] `http://www.ncbi.nlm.nih.gov/books/n/nlmumls/ch03/`

Table 3. UMLS-based alignments

Ontology pair	**Original alignments**	**Unsatisfiabilities**	**Refined alignments**
FMA $\sim$ NCI	3,024	655	2,898
FMA $\sim$ SNOMED CT	9,072	6,179	8,111
SNOMED CT $\sim$ NCI	19,622	20,944	18,322

Table 4. Results for the Large BioMed track in the OAEI 2011.5 campaign.

System	**Size**	**Unsat.**	**Refined UMLS**			**Original UMLS**			**Time (s)**
			P	**R**	**F**	**P**	**R**	**F**	
LogMap	2,658	**9**	**0.868**	0.796	**0.830**	**0.875**	0.769	0.819	126
GOMMA$_{\text{bk}}$	2,983	17,005	0.806	**0.830**	0.818	0.826	**0.815**	**0.820**	1,093
GOMMA$_{\text{nobk}}$	2,665	5,238	0.845	0.777	0.810	0.862	0.759	0.807	960
LogMapLt	3,466	26,429	0.675	0.807	0.735	0.695	0.796	0.742	57
CSA	3,607	$>10^5$	0.514	0.640	0.570	0.528	0.629	0.574	14,068
Aroma	4,080	$>10^5$	0.467	0.657	0.546	0.480	0.647	0.551	9,503
MapSSS	2,440	33,186	0.426	0.359	0.390	0.438	0.353	0.391	$>10^5$

generated the corresponding (equivalence) UMLS-based alignments with a confidence value of 1.0 (see Table 2).

The integration of new resources in UMLS combines expert assessment and sophisticated auditing protocols [1, 3, 10]. However, it has been noticed that UMLS-based alignments lead to a large number of unsatisfiable classes if they are represented as OWL 2 axioms and integrated with the input ontologies [15, 14]. For example the integration of SNOMED CT and NCI via UMLS-based alignments leads to more than 20,000 unsatisfiable classes. To address this problem, we have presented in [14] a refinement of the (original) UMLS-based alignments that do not lead to (many) unsatisfiable classes (see Table 3). This refinement is based on the alignment repair module of the ontology matching system LogMap [14, 16].

3 Results of the Large BioMed track in the OAEI 2011.5

In this section we briefly present the obtained results in the Large BioMed track of the OAEI 2011.5 campaign.[5] We have only evaluated the FMA-NCI matching problem, where the used versions of FMA and NCI contains 78,989 and 66,724 classes, respectively. The original and refined UMLS-based alignments (see Table 3) has been used as reference to evaluate the efficiency of participating ontology matching systems.

Table 4 summarizes the obtained results where systems has been ordered according to the F-measure against the refined UMLS-based reference alignment. LogMapLt —a simple ontology matcher—has been used as a base-line. Besides precision (P), recall (R), F-measure (F) and runtimes we have also evaluated the coherence of the alignments when reasoning together with the input ontologies.[6] Note that we have evaluated

---

[5] `http://www.cs.ox.ac.uk/isg/projects/SEALS/oaei/2011.5/`

[6] We have used the OWL 2 reasoner HermiT [20]

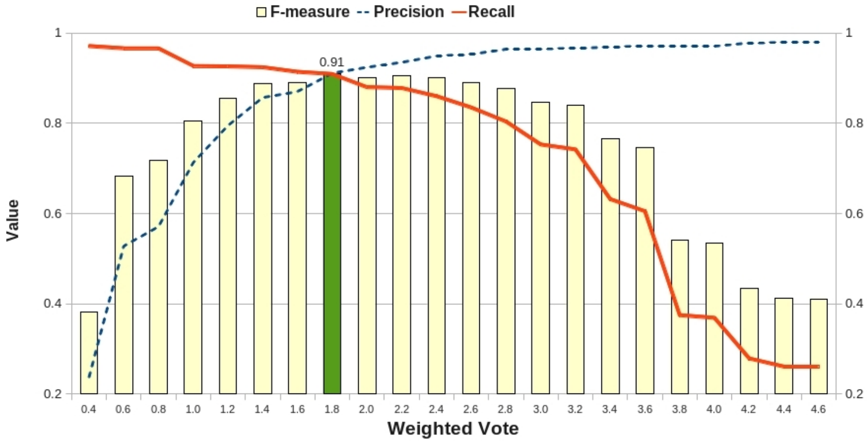


Fig. 1. Harmonised alignments for the FMA-NCI matching problem of the OAEI 2011.5.

GOMMA [17] with two different configurations. GOMMA$_{bk}$ uses UMLS-based background knowledge, while GOMMA$_{nobk}$ has this feature deactivated.

GOMMA (with its two configurations) and LogMap are a bit ahead in terms of F-measure with respect to Aroma [5], CSA [25] and MapSSS [2], which could not top the results of the base-line LogMapLt. GOMMA$_{bk}$ obtained the best results in terms of recall, while LogMap provided the best results in terms of precision and F-measure. The use of the original UMLS-based reference alignment did not imply important variations. Since the original set contains more mappings, precision and recall slightly increases and decreases, respectively. It is worth mentioning, however, that GOMMA$_{bk}$ improves its results when comparing with the original UMLS-based reference alignment and provides the best F-measure.

Regarding mapping coherence, only LogMap generated an ‘almost’ clean output in all three tasks. Although GOMMA$_{nobk}$ also provides highly precise output correspondences, they lead to a huge amount of unsatisfiable classes.

4 Towards a silver standard reference alignment

The original UMLS-based reference alignment, as shown in Section 2, contains errors (i.e. lead to large number of unsatisfiable classes when integrated with the input ontologies). On the other hand, the refined UMLS-based reference alignment is based on the (incomplete) alignment repair techniques of the ontology matching systems LogMap [14, 16], which may fail to detect and discard the appropriate alignments. Thus, in order to turn the extracted UMLS-based reference alignments into an agreed-upon gold standard expert assessment would be needed, which is almost unfeasible for large alignment sets. We have opted to move towards a *silver standard* by harmonising the outputs of different matching tools over the relevant ontologies. Similar silver standards have been developed for named entity recognition problems [21, 13].

We have harmonised the outputs of the systems participating in the OAEI 2011.5 FMA-NCI matching problem. Each system has been associated a weighted vote based on its precision w.r.t. the refined UMLS-based reference alignment (see Table 4). For example, LogMap and MapSSS have been associated the weights 0.868 and 0.426, respectively. Note that systems participating with two versions (e.g. GOMMA and LogMap) have been only considered once in the voting process.

Figure 1 summarises the evolution of the F-measure, Precision and Recall for the harmonised alignment depending on the minimum required votes. For example the harmonised alignment set requiring 4.0 points of weighted votes has a precision of 0.971 and a recall of 0.369 w.r.t. the refined UMLS-based reference alignment. As expected precision increases and recall decreases as the required votes increase.

We have selected the harmonised alignment set with the highest F-measure (0.91) as the “first” silver standard of the FMA-NCI matching problem. This set contains 2,890 alignments that have been ”at least” voted by two systems with weight 0.90. Note that this harmonised alignment has not been yet refined and it is known to lead to more than 14,000 unsatisfiable classes when integrated with FMA and NCI.

5 Future work

In the OAEI 2012 campaign[7] we also intend to evaluate the SNOMED-NCI and FMA-SNOMED matching problems using the correspondent UMLS-based reference alignments (see Table 3). We will also create harmonised silver standards alignments and we will evaluate the participating systems against them. This comparison will be very useful to analyse how different a system is with respect to the others.

Finally, we also intend to combine different reasoning and diagnosis tools such as ALCOMO[8] [18] to generate error-free refinements of both the UMLS-based reference alignments and the harmonised silver standards.

Acknowledgements

This work was supported by the Royal Society, the EU FP7 project SEALS and by the EPSRC projects ConDOR, ExODA, LogMap and Score!. We also thank the organisers of the OAEI evaluation campaigns for providing test data and infrastructure.

References

1. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic acids research 32 (2004)
2. Cheatham, M.: MapSSS results for OAEI 2011. In: Proceedings of the 6th Ontology Matching Workshop. pp. 184–189 (2011)
3. Cimino, J.J., Min, H., Perl, Y.: Consistency across the hierarchies of the UMLS semantic network and metathesaurus. J of Biomedical Informatics 36(6) (2003)

---

[7] `http://www.cs.ox.ac.uk/isg/projects/SEALS/oaei/2012/`

[8] `http://web.informatik.uni-mannheim.de/alcomo/`

4. Cuenca Grau, B., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P., Sattler, U.: OWL 2: The next step for OWL. Journal of Web Semantics 6(4), 309–322 (2008)
5. David, J., Guillet, F., Briand, H.: Association Rule Ontology Matching Approach. Journal of Semantic Web Information Systems 3(2), 27–49 (2007)
6. David, J., Euzenat, J., Scharffe, F., dos Santos, C.T.: The Alignment API 4.0. Semantic Web 2(1), 3–10 (2011)
7. Euzenat, J., Meilicke, C., Stuckenschmidt, H., Shvaiko, P., Trojahn, C.: Ontology Alignment Evaluation Initiative: six years of experience. J Data Semantics (2011)
8. Euzenat, J.: Semantic precision and recall for ontology alignment evaluation. In: Proc. of the 20th International Joint Conference on Artificial Intelligence, IJCAI. pp. 348–353 (2007)
9. Euzenat, J., Ferrara, A., van Hage, W.R., Hollink, L., Meilicke, C., Nikolov, A., Ritze, D., Scharffe, F., Shvaiko, P., Stuckenschmidt, H., Sváb-Zamazal, O., Trojahn dos Santos, C.: Results of the Ontology Alignment Evaluation Initiative 2011. 6th OM workshop (2011)
10. Geller, J., Perl, Y., Halper, M., Cornet, R.: Special issue on auditing of terminologies. Journal of Biomedical Informatics 42(3), 407–411 (2009)
11. Golbeck, J., Fragoso, G., Hartel, F.W., Hendler, J.A., Oberthaler, J., Parsia, B.: The National Cancer Institute's Thésaurus and Ontology. J. Web Sem. 1(1), 75–80 (2003)
12. Hartel, F.W., de Coronado, S., Dionne, R., Fragoso, G., Golbeck, J.: Modeling a description logic vocabulary for cancer research. Journal of Biomedical Informatics 38(2) (2005)
13. Jiménez-Ruiz, E., Rebholz-Schuhmann, D., Lewin, I.: Exploitation of cross-references between terminological resources within the CALBC context. In: 1st Intl. Workshop on Exploiting Large Knowledge Repositories, DEXA Workshops (2011)
14. Jiménez-Ruiz, E., Cuenca Grau, B.: LogMap: Logic-based and Scalable Ontology Matching. In: Proc. of the 10th International Semantic Web Conference (ISWC). pp. 273–288 (2011)
15. Jiménez-Ruiz, E., Cuenca Grau, B., Horrocks, I., Berlanga, R.: Logic-based assessment of the compatibility of UMLS ontology sources. J Biomed. Sem. 2 (2011)
16. Jiménez-Ruiz, E., Cuenca Grau, B., Zhou, Y., Horrocks, I.: Large-scale interactive ontology matching: Algorithms and implementation. In: Proc. of ECAI (2012)
17. Kirsten, T., Gross, A., Hartung, M., Rahm, E.: GOMMA: a component-based infrastructure for managing and analyzing life science ontologies and their evolution. Journal of Biomedical Semantics 2, 6 (2011)
18. Meilicke, C.: Alignment Incoherence in Ontology Matching. Ph.D. thesis, University of Mannheim, Chair of Artificial Intelligence (2011)
19. Mejino Jr., J.L.V., Rosse, C.: Symbolic modeling of structural relationships in the foundational model of anatomy. In: Proc. of First International Workshop on Formal Biomedical Knowledge Representation (KR-MED 2004). pp. 48–62 (2004)
20. Motik, B., Shearer, R., Horrocks, I.: Hypertableau Reasoning for Description Logics. Journal of Artificial Intelligence Research 36, 165–228 (2009)
21. Rebholz-Schuhmann, D., Jimeno Yepes, A., Van Mulligen, E.M., Kang, N., Kors, J., Milward, D., Corbett, P., Buyko, E., Beisswanger, E., Hahn, U.: CALBC Silver Standard Corpus. J Bioinform Comput Biol. pp. 163–179 (2010)
22. Shvaiko, P., Euzenat, J.: Ten challenges for ontology matching. In: On the Move to Meaningful Internet Systems (OTM Conferences) (2008)
23. Shvaiko, P., Euzenat, J.: Ontology matching: State of the art and future challenges. IEEE Trans. Knowl. Data Eng. 99 (2011)
24. Spackman, K.: SNOMED RT and SNOMED CT. Promise of an international clinical ontology. M.D. Computing 17 (2000)
25. Tran, Q.V., Ichise, R., Ho, B.Q.: Cluster-based similarity aggregation for ontology matching. In: Proc. of 6th Ontology Matching Workshop. pp. 142–147 (2011)

`KB_Bio_101`: A Repository of Graph-Structured Knowledge

Vinay K. Chaudhri, Michael Wessel, and Stijn Heymans

Artificial Intelligence Center, SRI International, Menlo Park, CA, 94025

1 Introduction

The goal of Project Halo is to develop a "Digital Aristotle" — a reasoning system capable of answering novel questions and solving advanced problems in a broad range of scientific disciplines and related human affairs [3]. As part of this effort, SRI has created a system called Automated User-Centered Reasoning and Acquisition System (AURA) [12], which enables educators to encode knowledge from science textbooks in a way that it can be used for answering questions by reasoning.

A team of biologists is currently using AURA to encode a popular biology textbook that is used in advanced high school and introductory college courses in the United States [15]. The knowledge base called `KB_Bio_101` is an outcome of this effort and contains concept taxonomy for the whole textbook and detailed rules for 20 chapters of the textbook. The current focus in the project is to expand the `KB_Bio_101` to cover all the 56 chapters of the book by December 2013. In the longer-term, `KB_Bio_101` will be expanded both in expressiveness and coverage. In terms of expressiveness, the Project Halo team is investigating the use of defaults, exceptions, negations, disjunctions and a process language. In terms of scope, the KB will likely be expanded to cover multiple textbooks potentially spanning a full undergraduate curriculum.

AURA uses a knowledge representation and reasoning system called Knowledge Machine (KM) [8]. KM supports a variety of representation features that include a facility to define classes and organize them into a hierarchy and define concept partitions (disjointness and covering axioms), ability to define relations (also known as slots) and organize them into a relation hierarchy, support for nominals, a facility to define horn rules, a procedure language, a situation mechanism, and a STRIPS representation for actions. KM performs reasoning by using inheritance, description-logic style classification of individuals, backward chaining over rules, and a heuristic unification. In addition, KM can use its situation mechanism and STRIPS representation of actions to simulate their execution. While the AURA team has experimented with the use of all of these features, the current core of AURA leverages only a small subset. The Project Halo team has invested significant effort to identify these core features and to specify them in a declarative manner. One example of such an effort is the work to specify the heuristic unification in KM using an answer set programming framework [7]. The net result of these efforts is that the team is now able to export the `KB_Bio_101` in a variety of standard declarative languages, for example, first order logic with equality [9], SILK [11], description logics (DLs) [5] and answer set programming [10].

The `KB_Bio_101` is a central component of an electronic textbook application called Inquire Biology [2] aimed at students studying from it. SRI has worked with teachers and students to collect a large number of questions that are of practical interest for this application. Working from those questions, the team has formulated logical reasoning tasks that must be performed by a reasoner.

The `KB_Bio_101` presents a unique opportunity for us to test our reasoners and to motivate further development. Recognizing that logical reasoning is only one component of the overall task of answering questions, the team at SRI is in the process of formulating similar challenges for knowledge representation [1] and natural language generation [6] which are also centered on `KB_Bio_101`. Taken collectively, these multiple challenges position us to make major leaps in AI in general, and knowledge-based question answering in particular.

2 Representation of Graphs in a Standard DL Syntax

There are two problems that need to be addressed to provide a representation of graphs in the DLs: defining a syntax for describing graphs and defining a family of graph expressiveness layers. We explain this in more detail next.

In principle, role value maps would be needed in order to truthfully represent the content of the `KB_Bio_101`. Role-value maps are a standard-way of expressing graph-structured descriptions in DL syntax. Unfortunately, unrestricted role value maps quickly lead to undecidability. There are decidable variants of role value maps, e.g. the restricted role-value-maps in a description logic with existential restrictions and terminological cycles ($\mathcal{EL}$ with cyclical TBoxes) of Baader [4], and we will check the applicability of this work to `KB_Bio_101`.

In recent work on description graphs [14] and description graph logic programs [13], a DL knowledge base is extended using a graph structure. While this proposal allows representation of graphs, it does not extend the conventional DL syntax in a graceful manner in that the conventional syntax can be completely abandoned in favor of this new syntax. The OWL export of `KB_Bio_101` extends the conventional syntax of OWL to encode graph structures.

Restrictions in description graphs prohibit the use of certain forms of cycles are too severe for `KB_Bio_101` which needs cyclicity in addition to the ability to express graphs. While the work on description graphs acknowledges the need for more expressive formalisms that go beyond tree structures, the nature of `KB_Bio_101` is sufficiently different from the setting in description graphs that it requires further research and could prove to be a data set that drives research beyond the current state.

3 Reasoning with Graph-Structured Descriptions

Similarity reasoning and relationship reasoning are two tasks that are of great practical interest to our application. In a similarity reasoning task, we are given two graph structured descriptions A and B, and the task is to compute new descriptions that correspond to their intersection and difference.

In the relationship reasoning task, we first create an ABOX by instantiating each concept in the TBOX, and then given two individuals A and B, we wish to compute all possible paths of a certain length between those individuals.

4 Summary

An initial version of the `KB_Bio_101` in OWL is now available. We are interested in identifying collaborators interested in exploiting this KB in the context of their tool set. We will work with them to first define an acceptable translation, and then participate in an experimental evaluation of the results of the reasoning tasks suggested above.

5 Acknowledgments

This work has been funded by Vulcan Inc.

References

1. Deep knowledge representation and reasoning challenge. `https://sites.google.com/site/dkrckcap2011/` and `https://sites.google.com/site/2nddeepkrchallenge/`.
2. Inquire: An Intelligent Textbook. `http://aivideo.org/2012/`.
3. Project Halo. `http://www.projecthalo.com`.
4. Franz Baader. Restricted role-value-maps in a description logic with existential restrictions and terminological cycles. In *Proceedings of the International Workshop on Description Logics (DL 2003)*, 2003.
5. Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, 2nd edition, 2007.
6. Eva Banik, Claire Gardent, Donia Scott, Nikhil Dinesh, and Fennie Liang. KBGen Text Generation from Knowledge Bases as a New Shared Task. In *International conference on Natural Language Generation*, 2012.
7. Vinay K. Chaudhri and Tran C. Son. Specifying and Reasoning with Underspecified Knowledge Bases Using Answer Set Programming. In *Proc. of International Conference on Knowledge Representation and Reasoning (KR)*, 2012.
8. Peter E. Clark and Bruce Porter. Knowledge machine userss guide. Technical report, University of Texas at Austin.
9. Melvin Fitting. *First-Order Logic and Automated Theorem Proving*. Springer, 1996.

10. M. Gelfond and V. Lifschitz. Logic programs with classical negation. In D.Warren and Peter Szeredi, editors, *Logic Programming: Proceedings of the Seventh International Conference*, pages 579–597, 1990.
11. Benjamin N. Grosof. SILK: Higher Level Rules with Defaults and Semantic Scalability. In Axel Polleres and Terrance Swift, editors, *Web Reasoning and Rule Systems, Third International Conference (RR 2009)*, volume 5837 of *Lecture Notes in Computer Science*, pages 24–25. Springer, 2009.
12. David Gunning, Vinay K. Chaudhri, Peter Clark, Ken Barker, Shaw-Yi Chaw, Mark Greaves, Benjamin Grosof, Alice Leung, David McDonald, Sunil Mishra, John Pacheco, Bruce Porter, Aaron Spaulding, Dan Tecuci, and Jing Tien. Project Halo Update – Progress Toward Digital Aristotle. *AI Magazine*, Fall 2010.
13. Despoina Magka, Boris Motik, and Ian Horrocks. Modeling Structured Domains using Description Graphs and Logic Programming. In *Proceedings of the International Workshop on Description Logics (DL 2012)*, 2012.
14. Boris Motik, Bernardo Cuenca Grau, Ian Horrocks, and Ulrike Sattler. Representing Ontologies using Description Logics, Description Graphs and Rules. *Artificial Intelligence*, 173:1275–1309, 2009.
15. Jane B. Reece, Lisa A. Urry, Michael L. Cain, Steven A. Wasserman, Peter V. Minorsky, and Robert B. Jackson. *Campbell Biology.* Benjamin Cummings, 9th edition, 2011.

If it's on web it's yours!

Abdul Mateen Rajput

Life Science Informatics, Bonn University, Bonn, Germany

1 Introduction:

Text mining is an emerging field and there are many applications of this field since the rate of information production has increased many folds in recent past. Despite exponentially rate of data production we are still struggling for the answer of the question which can satisfy our needs as it has been said that we are drowning in sea of data while dying of thirst for knowledge. One important area which seeks answer from massive datasets is biomedical sciences, where text mining facilitates to add value and provides different procedures to analyze bulk data being produced either after each new experiment of microarray, fMRI etc or by scientific publications.

To explore the knowledge from data one needs to have access to it to get valuable information [datasets may vary in size and it depends upon the questions you are going to ask from it]. The availability of some datasets is usually restricted to the provider and user may sometime doesn't find the correct dataset he/she is interested in, though it may be browsable on the web but not available as repository to apply natural language processing and text mining tools and user finds difficulties to achieve what is required. There are many web crawlers (HTTrack[1], GRUB[2] etc) but the problem with these programs is they bring too much noise and uncleaned data. The cleaning of this data is also an issue and usually takes more time than downloading. In the current paper we discuss a smart approach to make clean dataset from any online website. The resultant dataset could be any file format you are interested in and the method will provide you different possibilities to extract from many layers of web pages. The methodology we are going to discuss is freely available and following programs are required for it:

- Mozilla Firefox[1]
- DownThemALL, Firefox Plugin[2]
- Notepad++[3]
- Linkgopher, Firefox plugin[4] /GREP (shareware) [5]

2 Methods:

The initial steps of the corpora creation requires to look for the pattern of the hyperlinks of the data you are interested in and if the links of data is available on one page

---

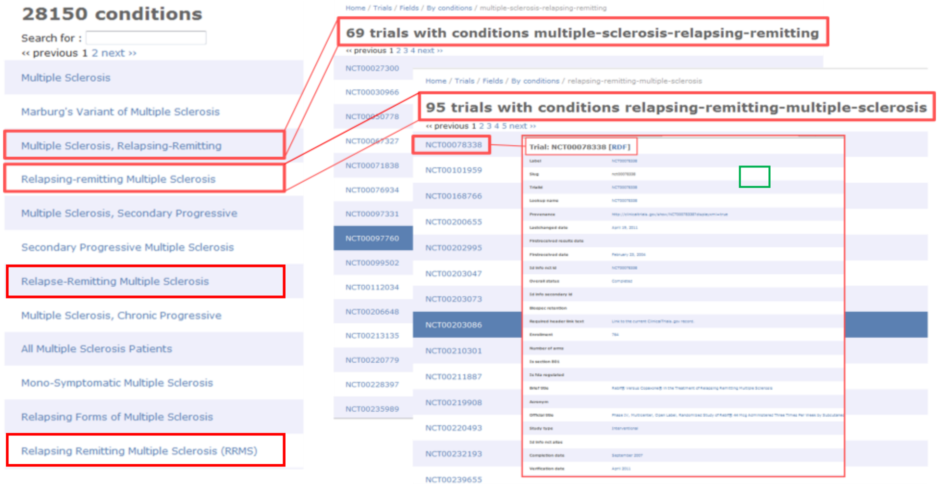
[1] http://www.httrack.com/

[2] http://www.gnu.org/software/grub/

then DownThemALL can automatically detects the links and you can start downloading instantly. If the actual data is under few layers of web pages then you can download the source pages and then actual data by combining all the source html pages and extracting links via LinkGopher or by using Grep program. The good feature of Grep is that it will also bring the data within the proximity of upto 5 lines from the actual search term.

3 Use Case:

The use case discusses the task we did with linkedCT.org [6], which is a RDF processed repository of clinicaltrials.gov [7]. We needed to download all the clinical trials associated with a particular disease and those clinical trials were stored under 4 different names (Multiple sclerosis Relapsing-Remitting, Relapsing-remitting Multiple Sclerosis, Relapse-Remitting Multiple Sclerosis, Relapsing Remitting Multiple Sclerosis). The actual data we were looking for was stored under 2 html pages where all the label of clinical trials associated with the disease state was mentioned (see figure 1). We stored the source html pages of actual clinical trials (4 pages associated with the disease titles) and then merge them together so we can have all the names of files on one html page. We found that the pattern of RDF storage and the page where it contains the link of it doesn't differ much and there is a similar pattern for each RDF file associated with the webpage link. Further we extracted all the links by using LinkGopher from the merged page and then looked at the patterns of RDF and html page. After finding out the pattern we simply replace the keywords with the one which was associated with RDF and then downloaded all the RDF files by simply using DownThemALL.



28150 conditions
Search for :
« previous 1 2 next »
Multiple Sclerosis
Marburg's Variant of Multiple Sclerosis
Multiple Sclerosis, Relapsing-Remitting
Relapsing-remitting Multiple Sclerosis
Multiple Sclerosis, Secondary Progressive
Secondary Progressive Multiple Sclerosis
Relapse-Remitting Multiple Sclerosis
Multiple Sclerosis, Chronic Progressive
All Multiple Sclerosis Patients
Mono-Symptomatic Multiple Sclerosis
Relapsing Forms of Multiple Sclerosis
Relapsing Remitting Multiple Sclerosis (RRMS)

Home / Trials / Fields / By conditions / multiple-sclerosis-relapsing-remitting
69 trials with conditions multiple-sclerosis-relapsing-remitting
« previous 1 2 3 4 next »
NCT00027300
NCT00030966
NCT00050778
NCT00067327
NCT00071838
NCT00076934
NCT00097331
NCT00097760
NCT00099502
NCT00112034
NCT00206648
NCT00213135
NCT00220779
NCT00228397
NCT00235989

Home / Trials / Fields / By conditions / relapsing-remitting-multiple-sclerosis
95 trials with conditions relapsing-remitting-multiple-sclerosis
« previous 1 2 3 4 5 next »
NCT00078338
NCT00101959
NCT00168766
NCT00200655
NCT00202995
NCT00203047
NCT00203073
NCT00203086
NCT00210301
NCT00211887
NCT00219908
NCT00220493
NCT00232193
NCT00239655

Trial: NCT00078338 [RDF]

Fig. 1. The overall view of the dataset. We needed many different RDFs (in green box) stored under different pages, description page of clinical trial, label page of different clinical trial and on top the disease page.

4 Conclusion:

We have used this method with several different websites and collect a large repository for using different text analytics tools. However, the procedure also has some limitation (doesn't work with Java links) and you have to carefully find out the patterns of dataset etc. On the contrary the good thing is that it is freely available and very quick rather than clicking the links and saving it manually.

5 Reference:

1. *Firefox*. Available from: http://www.mozilla.org/en-US/firefox/new/.
2. *DownThemAll*. Available from: http://www.downthemall.net/.
3. *Notepad++*. Available from: http://notepad-plus-plus.org/.
4. *LinkGopher*. Available from: https://addons.mozilla.org/en-us/firefox/addon/link-gopher/.
5. *Windows Grep*. Available from: http://www.wingrep.com/.
6. *LinkedCT*. Available from: http://linkedct.org/.
7. *ClinicalTrials*. Available from: http://www.clinicaltrials.gov/.