A Multi-perspective Termino-ontological Resource for Information Immersion System Design

Vanessa Andréani[1] and Thomas Lebarbé[2]

[1,2] Laboratoire LIDILEM (EA 609), Université de Grenoble, Domaine universitaire, 1180 avenue centrale, 38400 Saint Martin d'Hères, France
[1] TecKnowMetrix SAS, 4 rue Léon Béridot, ZA Champfeuillet, 38500 Voiron, France
[1] va@tkm.fr, [2] thomas.lebarbe@u-grenoble3.fr

Abstract. More and more structures have to deal with huge amounts of data, and it may be difficult to find the relevant information. To address this issue, documents must be indexed efficiently. Resources such as termino-ontological resources can be used, since they can provide a model of these documents, and index them by means of this model. We present a model and a building method for a multi-perspective termino-ontological resource that provides a structure to be used in an information immersion system. This system will allow to access a set of documents by different entry points according to users' needs.

Keywords: natural language processing, information retrieval, termino-ontological resource, modeling, indexing, named entities

1 Introduction: Context

In order to face huge amounts of information, "information retrieval" and "information extraction" systems have been largely studied and developed. We suggest a third, complementary system: "information immersion" to give users both automatisms and a certain freedom of interpretation.

1.1 Application Context

TecKnowMetrix (TKM) is a startup that offers consulting services in innovation strategy that range from state-of-the-art studies to competition analyses. This requires the analysis of large corpora containing mainly patents and scientific articles. For such consultancies, TKM experts need to detect relevant information from a given corpus, often containing several thousands of documents. The type of information needed varies a lot according to the kind of study they perform.

Yet, although the current tool used by TKM experts offers possibilities of data search, analysis and visualization, it may not be adapted to each type of study because of a limited flexibility. Finding the relevant information can be long and tedious, because the user cannot access the corpus from his own point of view, which is the one he takes within the perspective of a given study.

Hence, these moving professional constraints require a system that allows experts to immerse themselves inside a corpus according to different access points. These points must be able to respond to the specific needs of a given

study. Since we are not able to anticipate the whole set of the users' needs, we have designed a new system, flexible enough to adapt to a wide range of situations.

1.2 Corpus

The numerous corpora built until now form a set of several millions of documents. This set is our study corpus, and named entities (NE's) are clearly delimited within the document header. We consider NE's as "*the set of person names, organization names and place names in a given text*" [1]. Although NE's should only comprise "*entities for which one or many rigid designators […] stands for the referent*" [2], they also include "*temporal expressions and some numerical expressions such as amounts of money and other types of units*" [2] for practical purposes. Since the documents we work on are published by actors involved in scientific and technical fields, we focus on organization and person names, places and dates.

Each document in the set is associated with:

- its author, as an individual person;
- the organization that published it (and its hierarchy);
- the organization's geographical information;
- the date of publication.

All these NE's are already identified, normalized [3] and classified according to their nature by a semi-automated process.

We consider that exploiting them to design an immersion system is relevant, since they characterize all documents. Each kind of NE contains a given type of information, and all these pieces of information complete one another. If we add to this the data provided by representative terms extracted from the document's text, these NE's can be considered as a set of features qualifying a document. We consider a document's representative terms as the ones that convey the set of topics addressed in this document, and form its theme [4]. A given document is written by an author, published by an organization at a given date and in a given place, and deals with a topic represented by its terms.

2 Multi-perspective Termino-ontological Resource: a Model for Representing Information

Named entities (NE) are salient in our corpus, their use is systematic, and they convey relevant information. As such, and coupled with important terms of a document, they can be used to model and structure the set of documents.

Preserving the diversity of information conveyed by each kind of NE is crucial, since this diversity will ensure that the new system is flexible. Hence, the model must be consistent and at the same time reproduce this diversity, since this is a way to give the user several access points to a corpus.

To achieve this goal, we have designed a multi-perspective termino-ontological resource (TOR) that allows to highlight the different corpora's richness, and to index all documents in the set. This model of TOR is adapted to any corpus containing relevant named entities, provided that they can be identified

and classified according to their type. Such a model can be used in information retrieval systems.

2.1 Ontologies, Terminologies and Termino-ontological Resources

TOR can take various forms and differ from one another depending on the application they are designed for. These resources are used differently by several research fields; hence, defining the notion of TOR formally and in a generic way is irrelevant [5].

Nonetheless, we can consider that a TOR brings together characteristics of ontologies, and others of terminologies. According to [6], an ontology is "a standardized specification representing classes of the objects acknowledged as existing in a field". Moreover, an ontology is designed to meet the needs of a specific application, and to be used in a computational context. On the other hand, a terminology lists a field's terms and relations between them [6]. Therefore, we could roughly differentiate ontologies from terminologies by considering that an ontology involves a field's concepts, while a terminology is about a field's terms. As such, we regard them as complementary in their use.

Finally, a TOR can be defined in extension by the different forms it can take. For instance, a TOR can be a thesaurus for automatic indexing systems, an ontology for an organizational memory, a terminology repository for technical data management systems, etc. [5]

Our TOR describes qualified, characterized and categorized entities linked by qualified, characterized and categorized relations. Limiting the kinds of entities and relations reduces software complexity and increases system efficiency.

2.2 The Multi-perspective Termino-ontological Resource Model

In order to preserve the diversity of information provided by named entities in our set of documents, our TOR is made of five different perspectives, or facets, each one of them corresponding to a given type of information. The document is placed at the center of the TOR, and refers to each of the five facets, thus linking the facets together *via* the document. Four of the five facets list the NE's according to their type: 1) authors, 2) organizations, 3) places, 4) dates. The fifth contains relevant terms of each document text.

Modeling the set of documents by means of several but few autonomous, yet dependent facets allows the user to cross different types of information according to the way he needs to access the documents. For instance, if he wants to know which organizations collaborate on a given theme, he can access the data by crossing the organizations perspective with the thematic and authors facets. This way, he can see which organizations work together via individuals who work for two or more organizations.

2.3 Construction Method

Our TOR is built and stored in a relational database, each facet corresponding to a distinct sub-database. They are linked together by the documents' id's, the

documents being at the center of our resource. The data is stored by means of two methods. On the one hand, NE's are integrated in the first four facets of our TOR *via* a normalization process. This normalization allows to "clean up" and standardize the NE's. Moreover, particularly for organization NE's, normalization allows to detect organizations' hierarchy and make it explicit, which is the first step to modeling the data conveyed by the documents [3]. Furthermore, this normalization step corrects typing and spelling mistakes before importing the data in the tables, which allows to gather variants of a same NE under a same normalized name. For now, our normalization system is in the final stage of development, and was evaluated in its automatic aspect. It correctly normalizes 84% of named entities, according to users' expectations [3]. The link between the raw and the normalized data is maintained. These raw versions are regarded as potential graphical variants of a given entity.

On the other hand, the thematic facet is built from extraction methods, such as n-grams ("segments répétés" in [7]). In this statistical method, we consider that n-grams appearing more than once in a text may represent its theme. They are therefore extracted and stored as such, leaving interpretation to the user.

Globally, the construction of our TOR is composed of different steps. For each new study, the different documents are imported into the database, and represent our raw data. Named entities are first automatically normalized. Secondly, the aided normalization phase produces fully reliable data. Terms are then extracted from documents. Finally, this "clean" and structured information is integrated in our TOR. From here, the user will be able to search the database through the TOR *via* an interface to get relevant documents according to his needs for the current study.

Hence, this multi-perspective TOR is a way to index a set of documents so that users can exploit it in an optimal way in an immersion system.

3 Exploiting the Termino-ontological Resource: Perspectives and Conclusion

Thanks to our multi-perspective termino-ontological resource, and to the indexing model it allows, analysts will be able to access the set of documents by different entry points, each entry point potentially combining from 1 to 5 different facets of the TOR.

This number of combinations allows the user to access the data according to his specific needs at a given time and for a given type of study. Unlike enumerative systems such as Google, we intend to place the user in a human-machine interface where he navigates within a graphical network of structured information, and can access the contents during his surfing. Moreover, the simplicity of the system guarantees that the user knows and understands the information network he is provided with.

Our immersion system will present an interface to the user, who will enter a query related to one or a combination of facets of our TOR. The tool will return cognitive or geographical representations mapping the corresponding documents. Each item on a map will give access to the corresponding textual information. Moreover, selecting a document or a subset of documents, the user can access other documents linked to the first ones according to different criteria.

In order to meet the professional constraints, the next aspect is therefore to

provide analysts with a proper way to visualize the selected information [8]. For instance, a geographical map will be efficient to show most advanced countries in biofuel research, but will be irrelevant if the user wants to know the names of universities and companies collaborating in pharmacology: a relational network will be preferred here.

Hence, our forthcoming work will deal with the relevance of visualization tools to be integrated to the immersion system according to the type of selected information. This way, analysts can make good use of documents representation by our multi-perspective TOR. In the same manner, this type of system could be used to search into any corpus containing named entities if they are important items, no matter the domains concerned. For instance, we plan to adapt our system to the Manuscrits de Stendhal en Ligne (Online Stendhal Manuscripts) [9], which explore literature texts.

The immersion system is under development. The TOR will then be evaluated by its usage and not merely as a model.

References

1. Poibeau, T.: Deconstructing Harry, une évaluation des systèmes de repérage d'entités nommées. Revue de la SEE (2001)
2. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. In: Sekine S. and Ranchhod E. (eds.) Named Entities: Recognition, classification and use, pp. 3–28, John Benjamins publishing company (2009)
3. Andréani, V., Lebarbé, T.: Named entity normalization for termino-ontological resource design: mixing approaches for optimality. In: Proceedings of JADT 2010, *to appear*
4. Pichon, R., Sébillot, P.: Différencier les sens des mots à l'aide du thème et du contexte de leurs occurrences: une experience. In: Proceedings of TALN 1999, pp. 279–288 (1999)
5. Bourigault, D., Aussenac-Gilles, N.: Construction d'ontologies à partir de textes. In: Proceedings of TALN 2003, T2, pp. 25-50 (2003)
6. Reymonet, A., Thomas, J., Aussenac-Gilles, N.: Modélisation de ressources termino-ontologiques en OWL. In: Journées Francophones d'Ingénierie des Connaissances (IC 2007), Francky Trichet (Eds.), Cépaduès Editions, pp. 169-180 (2007)
7. Lafon, P., Salem, A.: L'Inventaire des segments répétés d'un texte. In: Mots, Nr. 6, pp. 161-177 (1983)
8. Roy, T.: Visualisations interactives pour l'aide personnalisée à l'interprétation d'ensembles documentaires. Thèse de doctorat de l'Université de Caen Basse-Normandie (2007)
9. Lebarbé, T., Meynard, C.: Nouvelles pratiques éditoriales, nouvelles lectures: les enjeux de l'édition électronique de manuscrits littéraires. In: Mémoire du Livre, Nr. 1 (2009)