Named entity normalization for termino-ontological resource design: mixing approaches for optimality

Vanessa Andréani [1] [2], Thomas Lebarbé [2]

[1] TecKnowMetrix SAS – 4, rue Léon Béridot – 38500 Voiron – France

[2] Laboratoire LIDILEM (EA 609) – Université Stendhal – Grenoble – France

Abstract

This paper introduces a method and a process to normalize named entities. Normalization is a useful prior step in many fields of information processing, such as information retrieval or language resources constitution. It enables to obtain reliable data, on which further processes can rely. Our method mixes three kinds of approaches, which we regard as complementary. The combination of endogenous, exogenous and anthropogenous methods allows to achieve satisfactory results and correctly normalized entities. As such, these entities are integrated in a multi-perspective termino-ontological resources.

Keywords: named entities, normalization, endogenous processes, exogenous processes, anthropogenous processes, pattern matching, similarity metrics, mutual information, termino-ontological resource

1. Introduction

We introduce here a method and a system to normalize named entities. It is based on several approaches, which we consider as complementary: endogenous methods, exogenous ones and anthropogenous ones. This system aims at normalizing data that will be integrated in a multi-perspective termino-ontological resource. Only the exogenous methods have been evaluated, the global approach shall be evaluated by the end of the year.

TecKnowMetrix (TKM) is a startup that offers consulting services in innovation strategy that range from state-of-the-art studies to competition analyses. They all require the analysis of large corpora containing mainly patents and scientific articles. These corpora are built by our analysts and gather scientific production related to the customers' fields of interest.

For such consultancies, TKM experts need to identify the actors of a given field of activity. Scientific and technical documents refer to named entities that are isolated in dedicated fields or tables in a database: article author(s), patent owner(s), inventor(s) or applicant(s).

Just as (Poibeau, 2001: 65) does, we consider named entities as «the set of person names, organization names and place names in a given text» [1]. Although named entities should only comprise «entities for which one or many rigid designators […] stands for the referent» (Nadeau and Sekine, 2009: 5), they also include «temporal expressions and some numerical expressions such as amounts of money and other types of units» (Nadeau and Sekine, 2009: 5) for practical purposes. Since we mainly work on actors involved in a given field, we focus on organization and person names, places and dates.

[1] Translated from French by the author.

The problem lies in the way named entities are recorded: the user has to "clean them up" since they contain noise or can be written in different ways. This generates errors in counts or graphics: if the same organization name is spelled in three different ways, these three occurrences will be counted as three different organizations instead of one. This "cleaning-up" allows to obtain reliably usable data.

Hence, the aim of normalization is to gather different names that refer to the same entity, and to identify and correct variants and spelling mistakes. More formally, we define named entity normalization as a process that turns several occurrences of a same type into a standard form. Up to now, TKM experts had to carry out this normalization process manually, which is time-consuming, especially when dealing with studies containing thousands of documents. After importing documents in TKM's database, the expert had to check each organization name in a spreadsheet, that is in a static list, to make sure the spelling was right, and had to bring together duplicate records. Detecting duplicate records by hand only with the help of alphabetical sorting takes a long time, is not user-friendly, and is particularly inaccurate.

Our work allows to reduce this task and makes it easier and faster. Now the experts can carry it out within the framework of a normalization process, where they are in charge of validating normalizations suggested by the system. This way, the expert user is totally integrated in the process, which is based on a strong human-machine interaction, since he has to validate the inferences on the data at several stages of the process. Therefore, the system we have designed is not entirely automated, but semi-automated. The user controls the data, and makes it truly reliable and suited to his needs.

This human-machine interaction implies to take into account the way the user interacts with the system. Hence, ergonomics notions such as usability and utilisability must be considered. We have led a study amongst the expert and novice users so as to design the most relevant processes and interfaces, thus reducing the cognitive load for the users.

2. Different processes for normalization

2.1. Global process

Since we work on structured data, localizing named entities comes down to fetching them in the dedicated fields or tables in the database. The major difficulty lies in cleaning them up from mistakes and noise that can cause interferences during further processes taking these named entities as input. Organization names are most impacted by these problems; that is why we will focus on them while describing our system. A raw organization name such as: *Division of Gynecologic Oncology*, *Department of Obstetrics and Gynecology*, *University of Kenttucky Medical Center*, *800 Rose Street*, *Whitney-Hendrickson Building*, *Lexington 40536*, *USA* should be normalized by the end of the process as: *Univ Kentucky*.

Experts have defined this form as the one corresponding to their needs as main organization name. In most cases, users only need this main name. Despite this fact, the remainder of the raw name is also parsed according to a hierarchical structure and normalized; here, *Medical Center* will be considered as secondary organization name, *Department of Obstetrics and Gynecology* as third name, *Division of Gynecologic Oncology* as fourth and last name, and *Lexington USA* as geographical localization.

To achieve this result, and to address the different types of problems, our system mixes several techniques that can be grouped into three kinds of approaches: techniques requiring external data (exogenous processes), endogenous processes, and anthropogenous processes.

The most general name is extracted and rewritten from the raw name with rules and patterns (Fig. 1 step 1). Then similarity metrics are applied to correct duplicate records. At this point, the user has to validate the system's suggestions (steps 2, 2'). Finally, mutual information and surface calculation are used to parse problematic names. The user must confirm the relevance of propositions (steps 3, 3') before names are integrated in the database, along with the corresponding raw name, and linked to the documents they were extracted from. We have chosen a methodological presentation rather than a chronological one to describe this system.

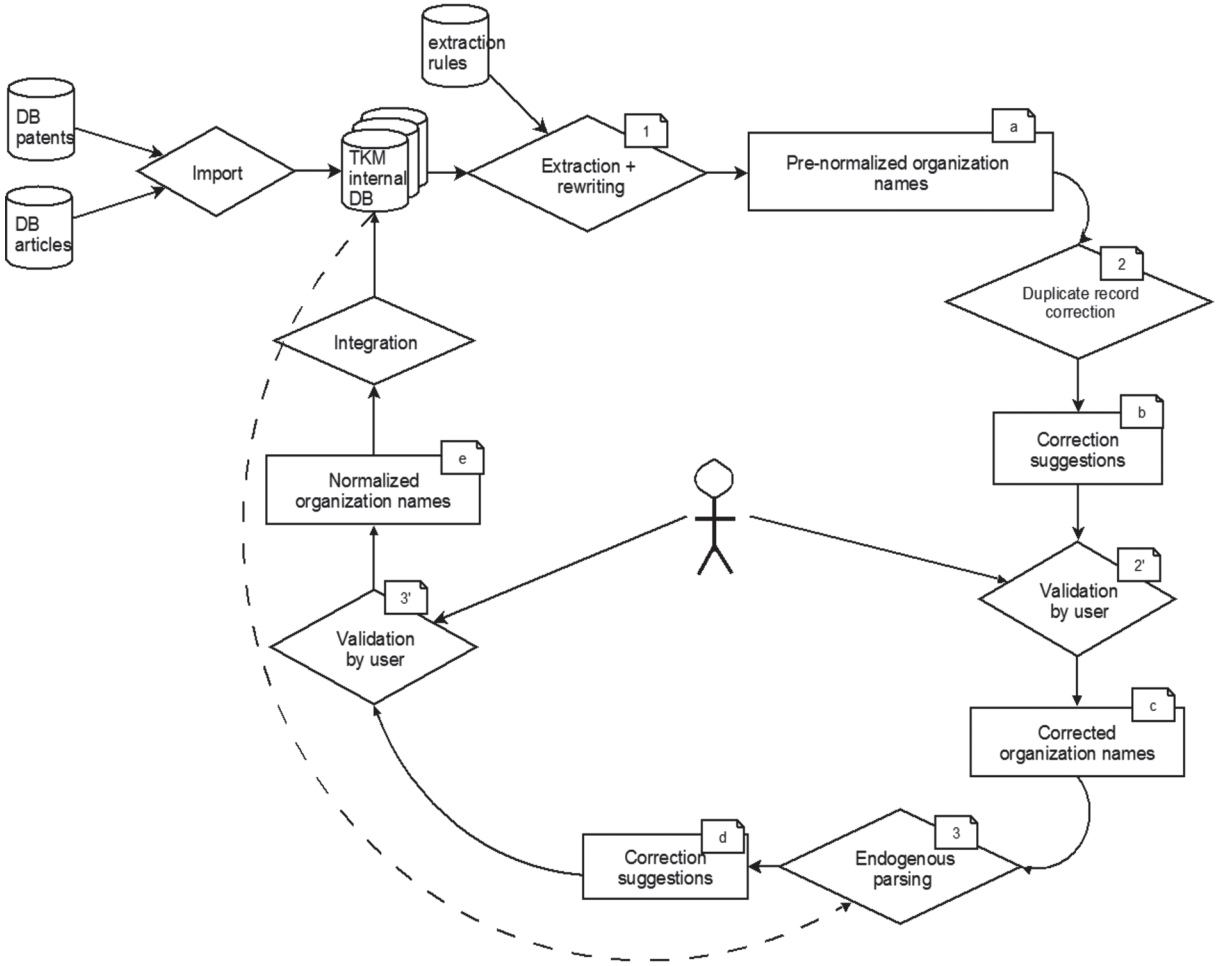

Figure 1: Global normalization process

2.2. Endogenous processes: from Levenshtein to structured mutual information

Endogenous qualifies «anything coming from the inside, originating from inside the object, organism, system, or set studied» [2]. Hence, in this framework, an endogenous process uses data from the corpus itself to process the corpus. Works such as (Bourigault, 1993; Bourigault and Frérot, 2006) and (Vergne, 2004) show that endogenous techniques are a way of solving problems for which external resources or rule-based methods show their weaknesses. The Levenshtein edit distance (LED) detects duplicate records due to spelling mistakes and allows

[2] Le petit lexique de la complexité : http://www.mcxapc.org/static.php?file=lexique.htm&menuID=lexique. Translated from French by the author.

to correct them. Surface calculation demarcates relevant sequences inside a raw organization name. Finally, structured mutual information detects repeated structures to parse an organization name. While the LED is a well-known method to detect similarities between two strings, surface calculation and structured mutual information are original approaches.

2.2.1. Levenshtein Edit Distance

Because of spelling mistakes or differences in writing conventions, variants of the currently processed named entity can be found in the database. Looking for similarity between names enables to detect and correct duplicate records. The use of similarity metrics is a way to exploit information found in data to process to detect this similarity: it consists in comparing two or more strings to determine whether they are similar. This comparison can have several base units according to the method used, *e.g.* characters as in the LED (Levenshtein, 1966), words as in the WHIRL method (Elmagarmid et al., 2007), windows of *n* characters, phonemes, etc. Using LED in our normalization process (Fig. 1, step 2) enables to match names containing mistakes with "correct" names. This way, these pairs can be highlighted so that users see them and can confirm the fact that the two names refer to the same entity.

The LED is an absolute value, which can cause problems when the strings in a pair are short: whereas a distance of 1 is not much for 20-character-long strings, it may be significant when the strings are 3 or 4 character-long. Hence, we compute a relative LED that considers the mean length of the two strings in a given pair.

We also observed that the strong recurrence of words designating institutions distorted the LED results. Two organization names can contain *Univ* or *Institute*, despite the fact they refer to two distinct entities. This is the case for the pair *Univ Stanford – Univ Salford,* in spite of a small relative distance. To reduce the number of false positives, and therefore noise, we lowered the weight of these words, which are not significant in this context and as such, can be considered as stop words. After this weighting, the relative distance for *Univ Stanford – Univ Salford* is beyond the limit for the names to be considered as duplicate records. This computing is efficient for detecting this kind of mistakes, but does not work in all cases.

2.2.2. Surfaces

Another approach of the data consists in considering them as recurrent term sequences. The longer the sequence, the less probable it is. Hence, if a given sequence appears, it may allow to recognize recurring blocs and by this, to parse relevant organization names. By correlating term sequence length and frequency of this sequence in the database, surface calculation enables this parsing when a name contains a named entity and one of its sub-entities.

Weighting sequence frequency by its length prevents from giving too much importance to a sequence's frequency, which can be a problem when dealing with compound names such as *New York*. For an organization name such as *New York Univ Medical Center*, the sequences will be computed according to Fig. 2:

Center is excluded from the computation since it is a lower-level institution name and does not belong to the main organization name. Without this weighting by the sequences' length, *York Univ* would have been detected as main name, because this sequence is the most frequent since the University of York exists independently from the University of New York.

This process allows to parse correctly a certain number of names, but is not always accurate. The global system needs another process to rely on when this one is inadequate.

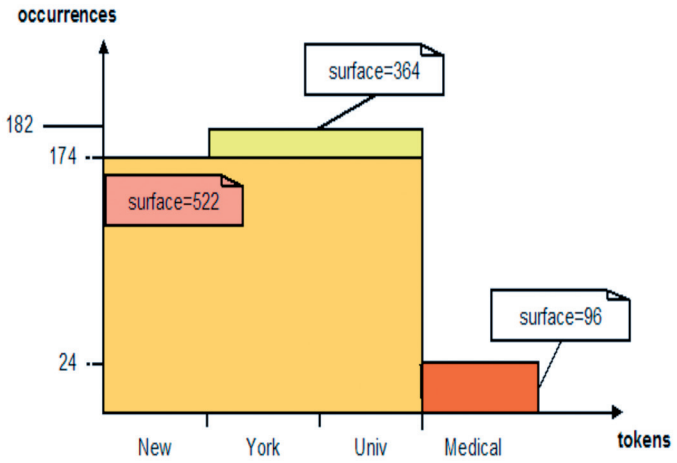

Fig. 2: Surface calculation for "New York Univ Medical Center"

2.2.3. Mutual information on linguistic structures

According to (Church and Hanks, 1990: 23), and considering *x* and *y* as two words, mutual information «compares the probability of observing x and y together (the joint probability) with the probability of observing x and y independently (chance)» [3]. This metric allows to measure the level of association between two words, for instance to know if they form a fixed expression, a compound, etc. In our framework, measuring the level of association between some words can indicate the limit between the major entity and one of its sub-entities.

We interfaced the mutual information (MI) calculation with the chunks theory. A chunk is a minimal, non-recursive constituent (Abney, 1991). A constituent includes a head and its satellites, which are "attracted" to it. From this, we can assume that a head's satellite should co-occur more often with its head than with other words, and particularly with other heads.

A sequence such as ***Univ*** *Colorado Health Sciences* ***Center*** can be parsed into two chunks, whose heads are *Univ* and *Center*. To decide if the satellites, i.e. *Colorado*, *Health* and *Sciences*, depend on the first or the second head, which would enable us to parse and determine the most general organization name, we can adapt the MI computation.

In practical terms, we compute, for each satellite, its mutual information with each one of the two heads. Concerning our example, we start with computing MI of *Colorado* with *Univ*, and MI for *Colorado* and *Center*. We assume that the highest MI indicates which head the satellite must be associated with. This method enables us to associate *Colorado* with *Univ*, and *Health* and *Sciences* with *Center*, and therefore to select *Univ Colorado* as main organization name, and *Health Sciences Center* as sub-entity. Although this method cannot guarantee a perfect normalization for each name, it is quite efficient and gives satisfactory results when associated with surface calculation.

Whereas computing the Levenshtein edit distance is relatively simple, algorithmic complexity increases with surface calculation and even more with MI on linguistic structures. Hence, we do not go further in endogenous processes, estimating we have reached a balance between cost due to calculation complexity and gain.

[3] More formally : $I(x, y) = \log_2 (P(x, y)/(P(x)P(y))$.

2.3. Exogenous processes: relying on external data

The adjective exogenous qualifies «what comes from the outside, originating from outside the object, organism, system, or set studied. Opposed to: endogenous» [4]. In the context of information processing or natural language processing, an exogenous process finds information necessary for processing data in external resources, which are "plugged" into the process. Such a process can be a pattern-based system, and information sources can be dictionaries, ontologies, etc. Whereas notions of exogenous and endogenous are opposed in theory, endogenous processes and exogenous ones can be complementary in their use.

The normalization process, in phase 1 (Fig. 1), uses a pattern-based system to solve a certain type of problems found in the data. In that, we follow principles used in the Caderige Project (Alphonse et al., 2004), which is dedicated to information extraction in medical texts. They first apply a normalization step, founded on the use of patterns including synonymy triggers to detect equivalent gene names. Although fields, nature of data, and global goals are different, we can take up again the general principle of patterns.

This stage in the process allows to remove and place in distinct fields what users consider as noise in most cases, and to rewrite the remainder so that it fits the standards. For instance, concerning the organization name *Department of Medical Mycology*, *Vallabhbhai Patel Chest Institute*, ***University of Delhi***, *Delhi*. *vpci@delnet.ren.nic.in*, everything except *University of Delhi* will be considered as noise, and we will only keep in the field "organization name" the main name, i.e. the one standing for the entity that includes all the others.

This step is based on two lexica; a multilingual lexicon containing grammatical words, such as *of* or *the*, and a multilingual list of institution names, such as *University*, *Institute* or *Corporation*, considered as trigger words since they indicate the presence of a named entity. Patterns are designed using these lexica, and recognize specific lexical and syntactic structures that allow to clean up a raw organization name.

Exogenous processes show two main advantages on endogenous ones. First, using symbolic methods relying on external resources enables to design very precise patterns after observing data in the corpus. Moreover, although designing these rules may be time-consuming, their computing is much less costly. Applying first an exogenous process "cleans up" the data, so that endogenous processes are applied afterwards only to data that could not be normalized by our exogenous method. Limits of this approach lie in the fact that some sequences cannot be rewritten and parsed properly, because of their syntactic structure. Furthermore, spelling mistakes are not corrected. Let us take our earlier example:

Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, ***University of Kenttucky*** *Medical Center, 800 Rose Street, Whitney-Hendrickson Building, Lexington 40536, USA.*

This stage will normalize it as *University of Kenttucky Medical Center*, which shows two problems:

- The mistake in *Kenttucky*, typed with two *t*'s instead of one, was not corrected;
- The persistence of noise, since we could not eliminate *Medical Center*, despite the trigger word *Center*.

[4] Le petit lexique de la complexité : http://www.mcxapc.org/static.php?file=lexique.htm&menuID=lexique. Translated from French by the author.

Our patterns, efficient on most of our data, cannot solve these difficulties, since we use very few external resources patterns can rely on. Unlike Khalid et al. (2008) or Jijkoun et al. (2008), who worked on named entity normalization in "ordinary", unspecialized texts using Wikipedia, we decided to use the minimum amount of external resources such as lexica or dictionaries. A large coverage resource of this kind would be too expensive to build in specialized fields.

2.4. Anthropogenous processes: integrating the user in the process

So far, we have named *endogenous* the processes using information coming from data themselves, and *exogenous* the processes using information coming from the outside. Here we introduce the question of information brought by the human in the normalization process. Hence, we call *anthropogenous* the processes using information coming from the human. Anthropogenous is not a term used in our field of study, and we borrow it from biology [5].

More formally, we define *anthropogenous,* as opposed to *endogenous* and *exogenous*, and as qualifying what impacts the object, organism, system or set studied, and comes from a human factor. In our framework, an anthropogenous process relies on humans' skills, and particularly on their interpreting skills. It places the human at the center of a system as an interpreting element, who possesses knowledge that cannot be formalized.

In order to make the processes we described more efficient and reliable, we gave the user a predominant place in the normalization process. He interacts at two points in the normalization, when totally automating the decisions becomes too risky. His role is crucial for each of these steps, since each part of the process relies on the results obtained during the previous one. Hence, however fine-grained the computing, we always need an interpreting and validating user to ensure data reliability.

The global idea in the user's participation is to consider him/her as an active contributor in the process to make sure normalized data will fit his needs. However, his contribution must remain limited to prevent him from cognitive overload and from spending too much time on a task we try to lighten. Hence, whereas the system automatically processes named entities, the user's main task comes down to validation of these processed data. The fact that the users' decisions are saved to be reused for new normalization is a crucial element of our process. It is worth pointing out that these anthropogenous processes are applied to a given study at a given time. This fact prevents the user from doing too many validations at a time, since the whole system is used "on demand". This way, he will process a maximum of about five thousand names at a time.

The first point of interaction for the user is step 2' (Fig. 1). He validates or invalidates suggestions of corrections by the Levenshtein edit distance (LED). Each pair of names whose relative LED is below an empirically-set threshold is proposed to the user as potential duplicate records. The user has to decide whether the system is right, and if so, has to choose the correct record. If none of them is acceptable, he can type in a new name.

Moreover, the user can decide whether a decision must be saved in a dedicated table or not. A decision can be appropriate only in a given context, but might not be consistent with the whole database and with all the new organization names that will be added in the future. For instance, a name like *Academy of Sciences* can be changed into *Chinese Academy of Sciences* for a particular study, but it would be inappropriate to turn each occurrence of *Academy of Sciences*

[5] In biology, *anthropogenous* usually qualifies phenomena caused by the intervention of human beings.

in the database into *Chinese Academy of Sciences*: there may also be publications from the Hungarian Academy of Sciences, etc. with the raw name *Academy of Sciences*.

The user also has a part is step 3' (Fig. 1). His actions are similar to the ones in step 2', since he has to confirm or invalidate data processed in step 3. The major difference is that he must confirm a name's parsing instead of validating an association between two names.

The user is a reliable information source for data processing. This allows him to act on the way data will be processed and used in future normalization processing. Each decision taken is stored and can be used again in subsequent normalizations. The fact that the user does not take a decision is also taken into account. All this enables to create fully reliable data. Since our goal was to make the normalization task easier and faster, we deliberately limited users' interaction to a few actions: validating, invalidating, and, if needed, correcting some names.

3. Evaluation

3.1. Different kinds of evaluation

As said earlier, our system mixes several approaches and involves an active participation from the user. As such, it implies different evaluation methods, some of them especially dedicated to the efficiency of human-machine interactions.

For steps 1, 2 and 3 (Fig. 1), we can use measures traditionally used for natural language processing systems in conferences like Text REtrieval Conference (TREC) or Message Understanding Conference (MUC). For instance, we chose to evaluate steps 1, 2 and 3 with measures derived from precision and recall.

As for steps 2' and 3', that is the ones involving users, evaluating them demands other measures, since precision and recall would be irrelevant and would not take into account the interactive nature of processes. Other parameters must be taken into account, like rate of competence and rate of effectiveness: they mainly consist in measuring the number of corrections that are made by the user (Luzzati, 1996). We can also use cognitive ergonomics, which describe the way information is perceived by users (Bellies, 2002).

Finally, we will evaluate the system as a whole, using all criteria we cited.

3.2. Evaluation of the pattern-based system

Up to now, we carried out a full evaluation for the first part of our system, i.e. our system of rules founded on patterns. We established a protocol to reliably measure our first results. It is based on the standards we defined with expert users, and we assume that if a name fits these standards, then it is "correct" as normalized name. If it does not, the error type is identified and counted. We tested our system on 1000 organization names. These publications come from several specialized fields, such as biochemistry, medicine, nanotechnologies, etc.

We applied our pattern-based system on these names, and designed an evaluation interface to systematize the evaluation process and avoid human mistakes as much as possible. It is worth pointing out that that out of these 1000 raw organization names, 86,4% do not fit the standards. This confirms the fact that partly automating normalization is useful. After applying the pattern-based system on these names, the rate of correct names reaches 84%. Errors are mainly due to noise found in processed names. We assume that integrating endogenous and anthropogenous processes in our next evaluation will strongly improve this rate, since they

reduce noise. However, this normalization results in an important gain of time for users. This gain will be precisely evaluated on short term.

4. Conclusion

Through this paper, we addressed the complex problem of named entity normalization. Normalization can strongly improve performances for different tasks in information processing, such as information extraction or information retrieval (Alphonse et al., 2004; Poibeau, 2003). Consequences implied by a precise normalization are very important when information retrieval tasks must be as accurate as possible, as it is the case for TKM.

To obtain reliable results and quality data, a normalization process needs three sources of information: 1) endogenous approaches that take into account data specificity, particularly in specialized fields, but are very costly in terms of computation; 2) exogenous methods for very precise patterns, their computing being cheaper, but lacking objectivity and reliability; 3) anthropogenous processes that are strongly reliable, but can be time-consuming for users.

Hence, these three methods are complementary, and their combination is the best way to obtain a quality named entity normalization. By proposing a normalization process based on these approaches, we take into account simultaneously data specificity and users' needs. The need for this normalization task is due to the importance of named entities in texts: without accurate named entity normalization, it is difficult to identify them and group them together.

Named entities are entry points to texts. These texts often link them together or with other textual sequences (contexts) that qualify, describe or evaluate them. Representing these relations in a termino-ontological resource will allow to detect and represent an important amount of information about named entities and texts that contain them.

Through this normalization, we are also building a multi-perspective termino-ontological resource (TOR) that will be used to access a given corpus in several ways, according to users' immediate needs. This TOR will be made up of five different perspectives, according to the nature of elements they contain: organizations, authors, dates, places, and terms extracted from text bodies, which are representative of their subject. According to the type of study they work on, users do not always need to explore the corpus the same way. For instance, if a user needs to rank countries according to their productivity on a given topic, he accesses the TOR by terms, places and possibly organization names. On the other hand, if he wants to know collaborations between organizations through authors they have in common, he enters the TOR by organizations names and author names.

Building this TOR is unthinkable without normalizing data first: it guarantees that data integrated is reliable, so that any process using this TOR is accurate. Hence, this normalization task should be applied before any textual information processing, in order to ensure its relevance and to provide the user with the best access possible to text contents.

References

Abney S. (1991). Parsing by chunks. In Robert, C., Berwick, S., Abney, P. and Tenny, C., editors, *Principle-Based Parsing: Computation and Psycholinguistics*, volume 44 of *Studies in Linguistics and Philosophy*. Dordrecht: Kluwer, pp. 72-79.

Alphonse E., Aubin S., Bessières P., Bisson G., Hamon T., Lagarrigue S., Nazarenko A., Nédellec C., Ould Abdel Vetah M., Poibeau T. and Weissenbacher D. (2004). Extraction d'information

appliquée au domaine biomédical – apprentissage et traitement automatique de la langue. In *Proceedings of CIFT*.

Bellies L. (2002). *La conception: processus d'élaboration et d'évaluation de représentations pour l'action*. Thèse de Doctorat en Ergonomie, Ecole Pratique des Hautes Etudes, Paris.

Bourigault D. (1993). An endogenous Corpus Based Method for Structural Noun Phrase Disambiguation. In *Proceedings of Conference of the European Chapter of ACL (EACL)*, pp. 81-86.

Bourigault D. and Frérot C. (2006). Acquisition et évaluation sur corpus de propriétés de sous-catégorisation syntaxique. *TAL*, 47: 141-154.

Church K. and Hanks P. (1990). Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics*, vol. 16 (1): 23-29.

Elmagarmid A.K., Ipeirotis P.G. and Verykios V.S. (2007). Duplicate Record Detection : A Survey. In *IEEE Transactions on Knowledge and Data Engineering*, 19, pp. 1-16.

Jijkoun V., Khalid M.A., Marx M. and de Rijke M. (2008). Named Entity Normalization in User Generated Content. In *Proceedings of SIGIR 2008 – Workshop on Analytics for Noisy Unstructured Text Data*.

Khalid M.A., Jijkoun V. and de Rijke M. (2008). The Impact of Named Entity Normalization on Information Retrieval for Question Answering. In *LNCS* 4956.

Levenshtein V. (1966). Binary codes capable of correcting deletions, insertins and reversals. *Soviet Physics Doklady*, 10: 707-710.

Luzzati D. (1996). *Le dialogue verbal homme-machine, études de cas.* Paris : Masson.

Nadeau D. and Sekine S. (2009). A survey of named entity recognition and classification. In Sekine, S. and Ranchhod, E., editors, *Named Entities*, Special Issue of Lingvisticae Investigationes, 30, 1: 3-28.

Poibeau T. (2001). Deconstructing Harry, une évaluation des systèmes de repérage d'entités nommées. *Revue de la société d'électronique, d'électricité et de traitement de l'information.*

Poibeau T. (2003). *Extraction automatique d'information: du texte brut au web sémantique*. Paris: Hermès.

Vergne J. (2004). Découverte locale des mots vides dans des corpus bruts de langues inconnues, sans aucune ressource. In *JADT2004* 2, pp. 1158-1164.