Automatic Construction of a TMF Terminological Database Using a Transducer Cascade

Chihebeddine Ammar
MIRACL & Sfax University
Sfax, Tunisia
chihebeddine.ammar@fss.rnu.tn

Kais Haddar
MIRACL & Sfax University
Sfax, Tunisia
kais.haddar@fss.rnu.tn

Laurent Romary
INRIA & Humboldt University
Berlin, Germany
laurent.romary@inria.fr

Abstract

The automatic development of terminological databases, especially in a standardized format, has a crucial aspect for multiple applications related to technical and scientific knowledge that requires semantic and terminological descriptions covering multiple domains. In this context, we have, in this paper, two challenges: the first is the automatic extraction of terms in order to build a terminological database, and the second challenge is their normalization into a standardized format. To deal with these challenges, we propose an approach based on a cascade of transducers performed using CasSys tool of the Unitex linguistic platform that benefits from both: the success of the rule-based approach for the extraction of terms, and the performance of the TMF standard for the representation of terms. We have tested and evaluated our approach on an Arabic scientific and technical corpus for the Elevator domain and the results are very encouraging.

1 Introduction

The automation of terminology will reduce the time and cost that usually takes terminological database construction. It will also help us to construct terminological databases with broad coverage, especially for recent concepts and poor language coverage (Arabic for example). On the other side, the representation of terminological data in a standard format allows the integration and merging of terminological data from multiple source systems, while improving terminological data quality and maintaining maximum interoperability between different applications.

One of the very rich in terminology working area are the scientific and technical documents. They cover several scientific and technical fields, so, we will need several terminological databases, one for each field. For this reason, we decided to work on a specific domain: the elevators.

To automate any process, we need a framework. The choice of this framework is not an easy task. In fact, many frameworks exist, based on: formal grammars, logical formalism, discrete mathematics, etc. The rule-based approach requires: a thorough study of the characteristics of terms and construction of necessary resources such as dictionaries, trigger words and extraction rules.

Finite automata and in particularly transducers are often used in the Natural Language Processing (NLP). The general idea is to replace the rules of formal grammars with representation forms. Transducers offer a particularly nice and simple formulation, and prove their capability of representing complex grammars due to their graphic representation. They have a success for the extraction of named entities (NE) and terms. In fact, precision is more important for rule-based systems.

Another issue is to decide which standard will we choose to model our terminological databases, which standard will best represent scientific and technical terms and which model to use, onomasiological or semasiological?

Our main objective is to create a standardized terminological resource from a corpus of Arabic scientific and technical documents (patents, manuals, scientific papers) able to support automatic text processing applications. Our approach is based on a cascade of transducers performed using CasSys tool of Unitex. It aims to extract and annotate under standardized TMF (Terminological Markup Framework) form technical terms of elevator field. The first step is a pre-treatment to resolve some problems of the Arabic language (e.g. agglutination). The second step is to extract and annotate terms. And the final one is a post-treatment consisting of cleaning documents.

Proceedings of Recent Advances in Natural Language Processing, pages 17–23,
Hissar, Bulgaria, Sep 7–9 2015.

This paper is organized as follows. Section 2 is devoted to the presentation of the previous work. We present, in section 3, the characteristics of Arabic scientific and technical terms. In section 4, we argue the choice of terminology model. In section 5, we present our approach. Section 6 is devoted to experimentation and evaluation and we conclude and enunciate some perspectives in section 7.

2 Previous Work

Three methods for building a terminological knowledge base exist: manual, semi-automatic and automatic. In the literature, there are some terminological databases for scientific and technical fields, most of them were constructed manually or semi-automatically.

For instance, the multilingual terminology of the European Union, IATE[1], contains 8,4 million terms in 23 languages covering EU specific terminology as well as multiple fields such as agriculture or information technology. The multilingual terminology portal of the World Intellectual Property Office, WIPO Pearl[2], gives access to scientific and technical terms in ten languages, including Arabic, derived from patent documents. It contains 15,000 concepts and 90,000 terms. Since WIPO has not a collection of Arabic patents, Arabic terms are often translations from the WIPO translation service. In (Lopez and Romary, 2010b), the authors developed a multilingual terminological database called GRISP covering multiple technical and scientific fields from various open resources.

Three main approaches are generally followed for extraction: rule-based (or linguistic) approach, training based (or statistic) approach and hybrid approach. What distinguishes the approaches mentioned, is not the type of information considered, but their acquisition and handling. The linguistic approach is based on human intuition, with the manual construction of analysis models, usually in the form of contextual rules. It requires a thorough study of the types of terms, but it has a success for the extraction of NE and terms. In fact, precision is more important for symbolic systems.

In previous work on non scientific and technical documents, there are those who used linguistic methods based on syntactic analysis (see for instance (Bourigault, 1992) and (Bourigault, 1994)). But the most used approach is the hybrid approach combining statistical and linguistic techniques (Dagan and Church, 1994).

The most recent work on scientific and technical documents were mainly based on purely statistical approaches. They used standard techniques of information retrieval and data extraction. Some of them use machine learning tools to extract header metadata using support vector machines (SVM) (Do et al., 2013), hidden markov models (HMM) (Binge, 2009), or conditional random fields (CRF) (Lopez, 2009). Others use machine learning tools to extract metadata of citations (Hetzner, 2008), tables (Liu et al., 2007), figures (Choudhury et al., 2013) or to identify concepts (Rao et al., 2013). All these approaches rely on previous training and natural language processing.

The need to allow exchanges between reference formats (Geneter, DXLT, etc.) has brought to the birth of the standard ISO 16642, TMF, specifying the minimum structural requirements to be met by every TML (terminological Markup Language).

3 Characteristics of Arabic Scientific and Technical Terms

Our study corpus contains 60 Arabic documents: 50 patents, 5 scientific papers and 5 manuals and installation documents of elevators collected from multiple resources: manuals from the websites of elevator manufacturers, patents from multiple Arabic intellectual property offices and scientific papers from some Arabic journals. All of these documents are text files and contain a total number of 619k tokens.

This corpus will allow us to construct the necessary resources such as dictionaries, trigger words and extraction rules and to study the characteristics of Arabic terms. Indeed, we noted the existence of some semantic relationships among terms of our collection, such as synonymy.

In fact, some terms have the same signified and different signifiers. For example, مصعد بدون وزن معادل signifies مصعد بدون وزن عكسي "elevator without counterweight". Here, the two terms have the same part (مصعد بدون وزن "elevator without weight") and two synonymous words (معادل "equivalent" and عكسي "reverse"). Another type of semantic relationships is the hierarchical relationship in two ways. Firstly, from the generic term to the specific term(s) (from hy-

[1] http://iate.europa.eu

[2] http://www.wipo.int/wipopearl/search/home.html

peronym to hyponym). For example, hyperonym: مركبة "vehicle", hyponyms: مصعد "elevator", عربة "car". Secondly, from the all to the different parts (from holonym to meronyms). For example, holonym: مصعد "elevator", meronyms: عربة "car", باب "door", زر "button", etc.

Some factors make the automatic analysis of Arabic texts a painful task, such as: the agglutination of Arabic terms. In fact, the Arabic language is a highly agglutinative language from the fact that clitics stick to nouns, verbs, adjectives which they relate. Therefore, we find particles that stick to the radicals, preventing their detection. Indeed, textual forms are made up of the agglutination of prefixes (articles: definite articleال "the", prepositions:ل "for", conjunctions: و "and"), and suffixes (linked pronouns) to the stems (inflected forms: أبوابه "its doors", أبواب "doors" + ه "its").

Another problem is the ambiguity which may be caused by several factors. For example, Arabic language is one of the Semitic languages that is defined as a diacritized language. Unfortunately, diacritics are rarely used in current Arabic writing conventions. So two or more words in Arabic can be homographic. Such as the word يعد (without diacritics) that could be (if we add diacritics): يعُد "return", يُعِدّ "prepare" or يَعُدّ "count".

Despite documents of our corpus are in Arabic language, some of them have a literal translation of key terms and technical words. These translations can be in English or French and are usually of a very high quality because they are made by professional human translators. They facilitate the task of our terminological database implementation (Language Section and Term Section of the TMF model) and make it multilingual.

4 TMF Terminological Model

The terminology is interested in what the terms mean: notions, concepts, and words or phrases that they nominate. This is the notional or conceptual approach. Motivated from the terminology industrial practice, the Terminological Markup Framework (TMF[3]) (Romary, 2001) was developed as a standard for onomasiological (sense to term) resources. In this paper, we need a generic model able to cover a variety of terminological resources. That is why we consider that the standard TMF is the most appropriate for our terminological database. The meta-model of the standard TMF is defined by logical hierarchical levels. It thus represents a structural hierarchy of the relevant nodes in a linguistic description. The meta-model describes the main structural elements and their internal connections.

It is combined with data categories (ISO 12620[4]) from a data category selection (DCS). Using the data model based on ISO 16642 allows us to fulfill the requirements of standardization and to exploit Data Category Registry (DCR) following the ISO 12620 standard for facilitating the implementation of filters and converters between different terminology instances and to produce a Generic Mapping Tool (GMT) representation, i.e. a canonical XML representation. The main role of our terminological extractor is to automatically generate terms in GMT format and create a normalized terminological database of scientific and technical terms.

Figure 1 shows an example of scientific terminological entry (Multi-car elevator) in the form of an XML document conforming GMT in three languages (Arabic, French and English).

5 Proposed Approach

The extraction method of Arabic terms that we advocate is rule-based. In fact, the rules that are manually built, express the structure of the information to extract and take the form of transducers. These transducers generally operate morphosyntactic information, as well as those contained in the resources (lexicons or dictionaries). Moreover, they allow the description of possible sequences of constituents of Arabic terms belonging to the field of elevators. The approach that we propose to extract terms for the field of elevators is composed of two steps (Figure. 2): (i) identifying the neces-

[3]ISO 16642:2003. Computer Applications in Terminology: Terminological Markup Framework

[4]ISO 12620:2009. Terminology and Other Language and Content Resources – Specification of Data Categories and Management of a Data Category Registry for Language Resources

```
<?xml version="1.0" encoding="utf-8"?>
<tmf>
    <struct type="TE">
        <feat type="EntryIdentifier">32</feat>
        <feat type="SubjectField">Elevator</feat>
        <feat type="Definition">مصعد بسيارات تعلق معا</feat>
        <struct type="LS">
            <feat type="Lang">Anglais</feat>
            <struct type="TS">
                <feat type="Term">multi-car elevator</feat>
                <feat type="Synonym">multi-deck elevator</feat>
            </struct>
        </struct>
        <struct type="LS">
            <feat type="Lang">Arabe</feat>
            <struct type="TS">
                <feat type="Term">مصعد متعدد المقصورات</feat>
                <feat type="Synonym">مصعد متعدد العربات</feat>
            </struct>
        </struct>
        <struct type="LS">
            <feat type="Lang">Francais</feat>
            <struct type="TS">
                <feat type="Term">ascenseur multi-voiture</feat>
            </struct>
        </struct>
    </struct>
</tmf>
```

Figure 1: Terminological entry conforming GMT

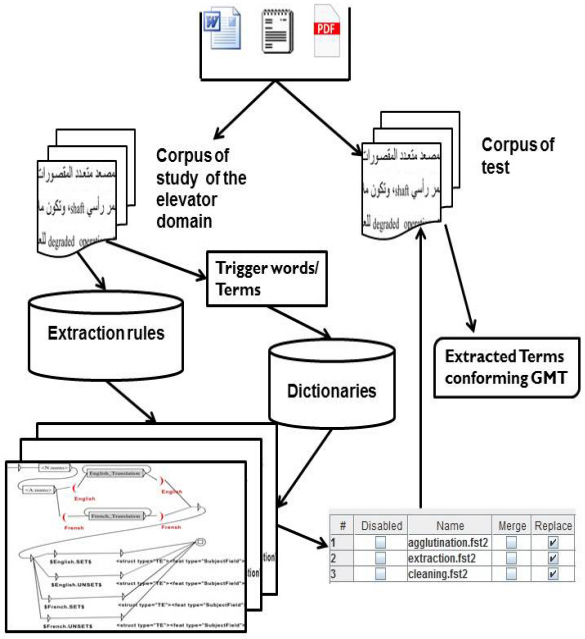

Figure 2: Proposed approach

sary resources to identify terms to extract, (ii) the creation of a cascade of transducer each of which has its own role.

In the following, we detail the different resources and steps of our approach.

5.1 Necessary Linguistic Resources

For our approach, we construct linguistic resouces from our study corpus, such as dictionaries, trigger words and extraction rules (syntactic patterns). In the following, we present these resources.

5.1.1 Dictionaries

For the domain of elevator, subject of our study, we identified the following dictionaries: a dictionary of inflected nouns and their canonical forms, a dictionary of inflected verbs, a dictionary for adjectives, a dictionary for trigger words of the domain and dictionaries of particles, possessive pronouns, demonstrative pronouns and relative pronouns. The structure of the various dictionary entries is not the same. It can vary from one dictionary to another. It must contain the grammatical category of the entry (noun, adjective), but, according to the dictionary, it may contain also: gender (masculine, feminine or neutral) and number (singular, dual, plural or broken plural), definition (defined or undefined), case (accusative, nominative or genitive) or mode (indicative, subjunctive or jussive), person (1st person, 2nd person or 3rd person) and voice (active or passive).

5.1.2 Trigger Words

The extraction rules generally use morphosyntactic information such as trigger words for the detection of the beginning of a term. We opted for increasing the number of rules and triggers in order to have as efficient as possible extraction system. We identified 162 trigger words, some of them can trigger the recognition of up to 5 terms. For this reason we classified them in classes.

5.1.3 Extraction Rules

To facilitate the identification of the necessary transducers for the extraction of terms, we have built a set of extraction rules. Indeed, they give the arrangement of the various constituents of terms in a linear manner easily transferable as graphs. We identified 12 extraction rules. Table 1 shows some of them. Four grammatical features are attribuated here: gender (masculine (m) or feminine (f)), number (singular (s), dual (d) or plural (p)), definition (defined (r) or undefined (n)) and case (accusative (a), nominative (u) or genitive (i)).

Examples of trigger words are: تحريك "mobilization" for the rules **R1** and **R5**, صيانة "mainte-

Rule number	Extraction rules
R1	<Pattern 1>:=<Trigger word> <N:nums><PREP>(<N:nums>)$^{+}$
R2	<Pattern 2>:=<N:nums> <PREP><N:nufs>[<Adj:nufs>]
R3	<Pattern 3>:=<Trigger word> <N:nums><Adj:nums>
R4	<Pattern 4>:= <Trigger word> <N:nufs><N:rums>
R5	<Pattern 5>:= <Trigger word> <N:nufp><N:rums>

Table 1: Some extraction rules of Arabic patent terms

nance" for the rule **R3** and رفع "lifting" for the rule **R4**. Table 2 shows some extracted terms due to the precedent extraction rules (here identified by their number in Table 1).

Rule number	Extracted terms
R1	مصعد بدون وزن عكسي "elevator without counterweight"
R2	مصعد ببكرة محزوزة "elevator with splined roller"
R3	مصعد آلي "automatic elevator"
R4	عربة المصعد "elevator car" لوحة التحكم "contor panel"
R5	أحبال الرفع "hoisting ropes"

Table 2: Terms extracted due to extraction rules

5.2 Implementation of Extraction Rules

We created three types of transducers. The first one is the transducer of pre-treatment solving Arabic prefixes and suffixes agglutination. To recognize the agglutinative character, we should enter inside the token. As Unitex works on a tokenized version of the text, it is not possible to make queries entering within the tokens, except with morphological filters or the morphological mode which is more appropriate in our case. To do this, we must define the whole portion of grammar using the symbols < and > as presented in Figure. 3). The transducer annotate every part of the agglutinated token with appropriate grammatical category.

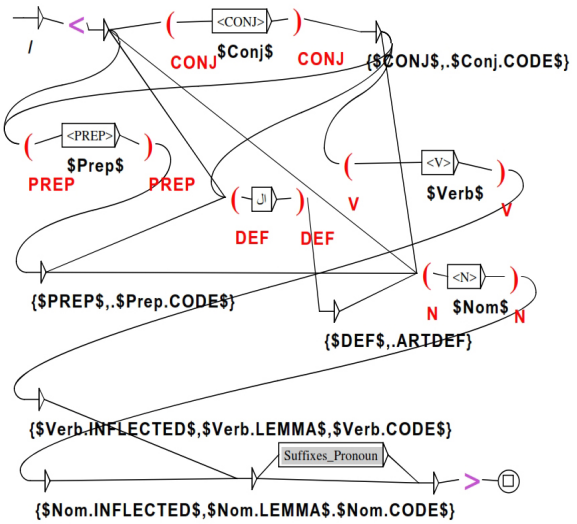

Figure 3: Transducer of resolution of agglutination

The second transducer, as shown in Figure. 4, includes all subgraphs of term extraction and annotation under the GMT format ("extraction_trasducers" box). In order to improve terms extraction, trigger words are regrouped into the "trigger_words" box.

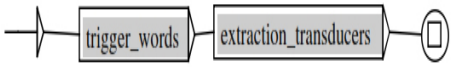

Figure 4: The main extraction transducer

Figure. 5 shows one of the transducers that extract and annotate terms. It also recognizes the French or English translation of terms (if available) thanks to the "French_Translation" and "English_Translation" subgraphs and annotate them in a new Language Section (LS) in the GMT format as shown in Figure. 1.

The final transducer is a post-treatment transducer consisting on document cleaning: its role is to delete all text remains (which is not XML). Figure. 6 is an overview of this transducer. The subgraph "XML" recognize all the XML element that could be contained by the <struct type="TE"> GMT element.

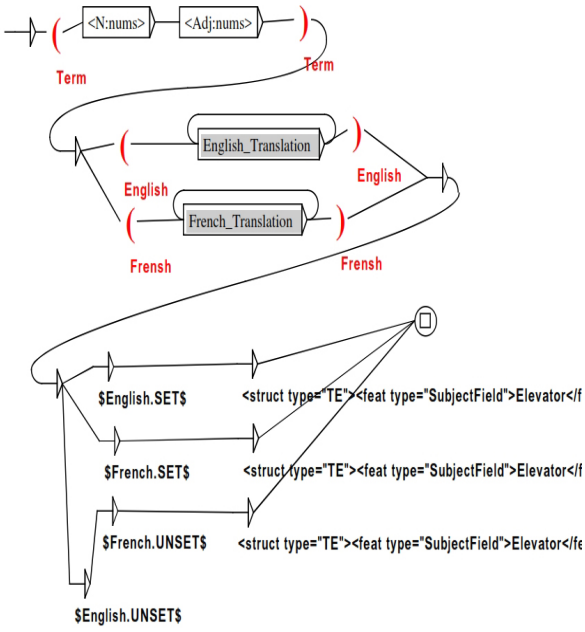

Figure 5: Example of extraction subgraph

6 Experimentation and Evaluation

Our test corpus contains 160 Arabic documents from multiple resources: 100 patents, 50 scientific papers and 10 manuals and installation documents of elevators, with a total number of 1.6m tokens. Our transducers are called in a specific order in a transducer cascade which is directly implemented in the linguistic platform Unitex[5] using the CasSys tool (Friburger and Maurel, 2004). Each graph adds its own annotations due to the mode "Replace". This mode provides, as output, a recognized term surrounded by a GMT annotation defined in the transducers.

In order to conduct an evaluation, we applied the cascade implemented on the test corpus. We manually evaluated the quality of our work on the test corpus. The total number of terms is 852. Table 3 gives an overview of the obtained results.

Terms	Extracted terms	Erroneous terms
852	827	59

Table 3: Overview of the obtained results

The obtained results are satisfactory, the transducers were able to cover the majority of terms with a precision of 0.95 and a recall of 0.97 with a F-score of 0.95. We therefore find that the proposed method is effective.

[5]Unitex: http://www-igm.univ-mlv.fr/ unitex/

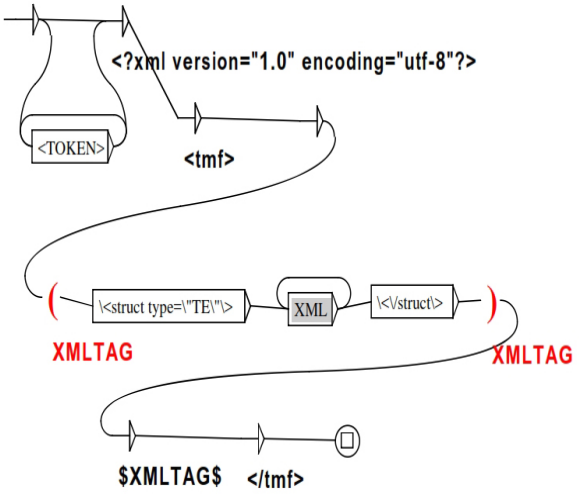

Figure 6: Post-treatment transducer

The noise can be caused by the absence of diacritics in our corpus and dictionaries, which could create ambiguity problem. It may also be caused by the absance of high granularity features of our dictionary entries. For this reason, we will try to add other semantic and grammatical features to our dictionary entries to improve our results. Despite the good results, we were forced to spend our terminological database to a terminologist to correct erroneous terms and their definitions. We believe that the automatic integration and merging of our database with other existing databases can help us to automatically correct errors.

7 Conclusion

In this paper, we built a set of transducers. Then we generated a cascade allowing extraction of scientific and technical terms. Extracted terms were represented in a standardized format (GMT). The generation of this cascade is performed using the CasSys tool, built-in Unitex linguistic platform. The operation of the transducer cascade required the construction of resources such as dictionaries.

In the immediate future, we will create a transducer cascade to extract bibliographic data and metadata of citations, tables, formulas and figures from scientific and technical documents and patents. We will also extract terms using a satatistic approach. Finally, we will try to combine the two approaches in a hybrid one.

Acknowledgments

We would like to thank Alexis Neme and Denis Maurel for helpful discussion and advice.

References

Cui Binge. 2009. Scientific literature metadata extraction based on HMM. *CDVE 2009*, pages 64-68. Luxembourg, Luxembourg.

Didier Bourigault. 1992. Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases. In *Proceedings of the 14th International Conference on Computational Linguistics COLING'92*, volume 3, pages 977-981. Nantes, France.

Didier Bourigault. 1994. LEXTER, Un Logiciel d'Extraction de TERminologie. Application à l'Acquisition de Connaissances à Partir de Textes. *Doctoral thesis*. Ecole des hautes Etudes en Sciences Sociales.

Erik Hetzner. 2008. A simple method for citation metadata extraction using hidden markov models. *JCDL 2008*, pages 280-284. New York, USA.

Huy H.N. Do, Muthu K. Chandrasekaran, Philip S. Cho and Min Y. Kan. 2013. Extracting and Matching Authors and Affiliations in Scholarly Documents. *JCDL 2013*. Indianapolis, Indiana, USA.

Ido Dagan and Ken Church. 1994. Termight: Identifying and Translating Technical Terminology. In *Proceedings of the 4th Applied Natural Language Processing Conference ANLP'94*, pages 34-40. Stuttgart, Germany.

Laurent Romary. 2001. An Abstract Model for the Representation of Multilingual Terminological Data: TMF Terminological Markup Framework. *TAMA 2001*. Antwerp, Belgium.

Nathalie Friburger, Denis Maurel. 2004. Finite-state transducer cascades to extract named entities in texts. In *Theoretical Computer Science*,volume 313, pages 93-104.

Patrice Lopez. 2009. GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. *ECDL 2009*. Corfu, Greece.

Patrice Lopez and Laurent Romary. 2010b. GRISP: A Massive Multilingual Terminological Database for Scientific and Technical Domains. *LREC 2010*. La Valette, Malta.

Pattabhi R.K. Rao, Sobha L. Devi, Paolo Rosso. 2013. Automatic Identification of Concepts and Conceptual relations from Patents Using Machine Learning Methods. *ICON 2013*, pages 18-20. Noida, India.

Sagnik R. Choudhury, Prasenjit Mitra, Andi Kirk, Silvia Szep, Donald Pellegrino, Sue Jones and C Lee Giles. 2013. Figure Metadata Extraction from Digital Documents. *ICDAR 2013*, pages 135 - 139. Washington, USA.

Ying Liu, Kun Bai, Prasenjit Mitra and C. Lee Giles. 2007. Tableseer: automatic table metadata extraction and searching in digital libraries. *JCDL 2007*, pages 91-100. Vancouver, Canada.