Conceptual Structure of Automatically Extracted Multi-Word Terms from Domain Specific Corpora: a Case Study for Italian

Elisa Lavagnino
LCI - Télécom Bretagne &
CeRTeM - Università degli Studi di Genova
elisa.lavagnino
@telecom-bretagne.eu

Jungyeul Park
LINA
Université de Nantes
jungyeul.park
@univ-nantes.fr

Abstract

This paper is based on our efforts on automatic multi-word terms extraction and its conceptual structure for multiple languages. At present, we mainly focus on English and the major Romance languages such as French, Spanish, Portuguese, and Italian. This paper is a case study for Italian language. We present how to build automatically conceptual structure of automatically extracted multi-word terms from domain specific corpora for Italian. We show the experimental results for extracting multi-word terms from two domain corpora ("natural area" and "organic agriculture"). Since this work is still ongoing, we discuss our future direction at the end of the paper.

1 Introduction

Great progress has been recently obtained on using text analysis to extract terms in a specific field. The study of texts helps in finding and organizing textual segments representing conceptual units. A corpus is a collection of texts stored in an electronic database. Texts have been selected to be representative of a particular goal. A corpus must be balanced in quality and quantity contents: in order to be representative of a domain, texts have to cover all the possible communicative situations. Generally, in a specialised domain, users share contents and they normally can understand and communicate with each others without ambiguities. However, when different communities get in touch the possibility of misunderstanding arises because of terminological variation. This variation can be detected at a conceptual level or at the formal one. Our approach tries to overcome this problem by collecting different text typologies. Texts may be extracted from different sources which can be classified as their specialisation level, their contents, their pragmatic application, etc. In our case, we are interested in using different texts, in order to analysis the result of automatic extraction in different communicative situations to improve its functioning.

A term can be simple if composed by one word, or complex if composed by several words. This paper focuses on extracting and conceptually structuring multi-word terms for Italian. Collet (2000) affirmed that a complex term (multi-word term in our terminology) is a complex unit whose components are separated by a space and are syntactically connected. The resulting unit denominates a concept which belongs to the language for special purposes (LSP). Texts on any domain are easily available on the Web these days. To create a corpus representing a field, materials should be, however analysed and re-elaborated in order to resolve eventual problems arising the transfer of data. In particular, a corpus have to be processed in order to classify the composing units. This classification represents the first step towards terminological extraction. Terminologists must often look through many texts before finding appropriate ones (Agbago and Barrire, 2005). L'Homme (2004) presents guidelines for choosing terminology such as domain specificity, language originality, specialization level, type, date, data

Proceedings of the 2nd Workshop on Cognitive Aspects of the Lexicon (CogALex 2010), pages 48–55,
Beijing, August 2010

evaluation.[1]

Since interaction between domains increases consistently, domain specificity is a crucial point to considerer during the creation of a corpus. Text typologies and communicative situations reflect their peculiarity to terms. A concept can be represented differently if the level of specialisation of a text or the context changes. Here, we consider the context as the frame in which the communication takes place. For example, the domain of "natural area", Italian language is really interesting because terms register a high level of variations due to the different contexts.

The LSP changes as the society evolves. Terms can register the diachronic variation due to the development of a certain domain. The evolution of a domain influences also the terminologies which form LSP. Terminological evolution generates variations in the conceptual representation which should be observed in order to detect terms and their variants and to establish relations between them. For example, the domain of "organic agriculture" is now evolving and changing because of political choices. This affects the terminology and the eventual creation of new forms. The affix *bio-* which can be use as a variant of almost all multi-word terms concerning the biological production such as *metodo di produzione biologica* ('method of organic production') becomes *metodo bio* and *prodotto biologico* ('organic product') becomes *prodotto bio* or just *bio*.

In this paper, we present an approach for extracting automatically multi-word terms (MWT) from domain specific corpora for Italian. We also try to conceptually structure them, that is we build the 'conceptual' structure of variations of multi-word terms where we can learn dynamics of terms (Daille, 2002). Conceptual structure in this paper limits to the semantic relationships between terms such as **Hyperonomy**, **Antony**, **Set of**, and **Result** between multi-word terms and we currently implement only hyperonomy relations.

Actually, this paper is based on our efforts on automatic multi-word terms extraction and its conceptual structure for multiple languages. At present, we mainly focus on English and the major Romance languages such as French, Spanish, Portuguese, and Italian. This paper is a case study for Italian language. The remaining of this paper is organized as follows: We explain how to automatically extract and conceptually structure multi-word terms from domain specific corpora in the next section. We also describe some implementation issues and current advancement. Since this work is still on-going, we discuss our future direction in the last section.

2 Automatically Extracting and Conceptually Structuring Multi-Word Terms

2.1 ACABIT

To extract automatically multi-word terms from domain specific corpora and conceptually structure them for Italian, we adapt existing ACABIT which is a general purpose term extractor. It takes as input a linguistically annotated corpus and proposes as output a list of multi-word term candidates ranked from the most representative of the corpus to the least using the log-likelihood estimation.[2] ACABIT is currently available for English and French as different programs for each language. Fundamentally, ACABIT works as two stages: `stat` and `tri`. At the `stat`, it allows us to identify multi-word terms in corpora to calculate the statistic. At the `tri`, it allows us to sort and conceptually structure them based on base terms. For the moment, we reimplement universal `stat` for major Romance languages. We explain the more detailed issues of our reimplementation of ACABIT for Italian in Section 2.3.

2.2 Base Term and its Variations

For automatic multi-word term identification, it is necessary to define first the syntactic structures which are potentially lexicalisable (Daille, 2003). We refer to these complex sequences as **base terms**. For Italian, the syntactic structure of base terms is as follows (where Noun_1 is a head):

[1] The translated text is adapted from Agbago and Barrire (2005)

[2] `http://www.bdaille.fr`

Noun$_1$ Adj *area protetta* ('protected area'), *azienda agricola* ('agricultural company')

Noun$_1$ Noun$_2$ *zona tampone* ('buffer area')

Noun$_1$ *di* (Det) Noun$_2$ *sistema di controllo* ('control system'), *conservazione dei biotopi* ('biotope conservation')

Besides these base term structures, there is also [Noun$_1$ *à* V$_{inf}$] for example for French. For Italian, there might be [Noun$_1$ *da* V$_{inf}$] such as *prodotto biologico da esportare* ('organic product to export') which is rather phraseology and not a term. Consequently, we define only three base term structures for Italian for now.

ACABIT for Italian should spot variations of base terms and puts them together. For example, there are **graphical variations** such as case differences and the presence of an optional hyphen inside of base term structures, **inflexional variations** where *aree protette* ('protected areas') should be considered as the variation of *area protetta* ('protected area'), or **shallow syntactic variations** which only modifies function words of the base terms, such as optional character of the preposition and article such as *sistema di informazione* and *sistema informativo* ('information system').

To conceptually structure identified multi-word terms, ACABIT for Italian should put together syntactic variations which modify the internal structure of the base term: internal modification and coordination. Internal modification variations introduce the modifier such as the adjective in [Noun$_1$ *di* Noun$_2$] structure or a nominal specifier inside of [Noun$_1$ Adj] structure. For example, *qualità ambientale* ('environmental quality') and *elevata qualità ambientale* ('high environmental quality') for [Noun$_1$ *di* Noun$_2$] structure and *ingrediente biologico* ('organic ingredient') and *ingrediente d'origine biologico* ('organic origin ingredient') for [Noun$_1$ Adj] structure . Coordination variations coordinate or enumerate the base term structure, for example *habitat ntaurali* ('natural habitat') and *habitat naturali e quasi naturali* ('natural and almost natural habitat')

2.3 Implementation

To keep consistent with the original ACABIT and to take an advantage of by directly using a certain part of existing modules, we use the input and the output formats of ACABIT. The input format of ACABIT requires the lemmatized forms of words for detecting inflexional variations of multi-word terms. For example, putting together inflexional variations such as *area protetta* and *aree protette* ('protected area(s)') is easily predictable by using their lemmatized forms. The original version of ACABIT for French uses BRILL's POS tagger[3] for POS tagging and FLEMM[4] for restoring morpho-syntactic information and lemmatized forms. And for English, it uses BRILL's POS tagger and CELEX lexical database[5] as a lemmatiser.

Since we are reimplementing ACABIT for multiple languages and we want to use the homogeneous preprocessing for ACABIT, we use TREETAGGER[6] which annotates both of part-of-speech tags and lemma information as preprocessor for . Moreover, TREETAGGER is available for several languages. We, then adapt the result of TREETAGGER for the input format for ACABIT. We use French POS tagger's tagset (Étiquettes de Brill94 Français INALF/CNRS) for every language, we convert TREETAGGER tagset into BRILL's tagset.[7]

Figure 1 shows the example of the input format of ACABIT in XML makes use of which conforms to Document Type Definition (DTD) in Figure 2. In Figure 1, POS tags are followed by morpho-syntactic information and the lemmatized form of a token in each <PH>.[8] TREETAGGER provide only lemmatized forms with POS information, instead of providing its main

[3] http://www.atilf.fr

[4] http://www.univ-nancy2.fr/pers/namer/Telecharger_Flemm.htm

[5] http://www.ldc.upenn.edu/

[6] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[7] http://www.lirmm.fr/~mroche/Enseignements/FdD_M2P_old/Etiqueteur/tags.html#francais_inalf

[8] For the convenience of the notations, accented characters are sometimes presented as 'e and 'a for è and à, respectively in the Figure.

morphological features such as gender, number, person and case as FLEMM in the previous version of ACABIT. We simply introduce dummy morphological information because it is not actually used in ACABIT. Note that è/SYM/è in Figure 1 is not correctly POS-tagged by TREE-TAGGER. It is one of flexional forms of *essere* ('be') instead of the symbol (SYM). However, we do not perform any post-processing to correct errors and we leave it as it is analyzed for the moment.

In Figure 2, <CORPUS> is for the name of the corpus, <RECORD> is for different texts which are usually from separate files, <INFO> has a format like <INFO>00/CAR/00 -/-0001800/SBC/0001800</INFO> with the year of text creation 00 and the file identification 0001800. <TITLE> is for the title, <AB> is for the text body, and <PH NB="*num*"> is for sentence identification.

ACABIT proposes as output a list of multi-word terms ranked from the most representative of the corpus using log-likelihood estimation (Dunning, 1993) and their variations in the corpus. It also shows the semantic relation between multi-word terms. The example of the output is given in Figure 3. A base term, for example *area protetto* ('protected area') is put together with its syntactic variations "*area naturale protetto* ('natural protected area') and *area marino protetto* ('marine protected area'). We can rewrite them as **Hyperonomy** (*area naturale protetto*) = *area protetto*" or **Hyperonomy** (*area marino protetto*) = *area protetto* because ACABIT identifies that *area protetto* is a hypernym of *area naturale protetto* and *area marino protetto* as <MODIF>ied terms of <BASE> terms . Likewise, a base term *prodotto biologico* ('organic product') has its syntactic variation: internal modification such as *prodotto non biologico* ('non-organic product'), *prodotto alimentare non biologico* ('non-organic alimentary product'), and *prodotto ittico biologico* ('organic fishing product'), and coordination like *prodotto biologico e non biologico* ('organic and non-organic product'). Moreover, there are **Antonym** relation described as LINK type="Neg" between the base terms and some of its syntactic variations such as *prodotto non biologico* and *prodotto alimentare non biologico*. Note that output of ACABIT in Figure 3 only contains canonical forms of multi-word terms.

2.4 Experiments

Creation of domain specific corpora: For our experiments we crawl two domain corpora of "natural area" domain which consists of 17,291 sentences and 543,790 tokens from *Gli E-Quaderni*[9] and 47,887 sentences and 1,857,914 tokens from *Parchi*[10]. We also crawl in the Internet to create the corpus of "organic agriculture" which consists of 5,553 sentences and 150,246 tokens from *National legislations* and *European legislations* for organic agriculture[11].

Automatic evaluation: Table 1 shows the statistics of experimental results from each domain. Since our domain corpora are mutually related, we count the common multi-word terms and there are 600 unique terms (base terms + variations) shared in both corpora. This is 18.74% of the number of terms in "organic agriculture". Figure 4 shows example of these common terms.

2.5 Current advancement

Till now, we reimplement only stat for multiple languages. To conceptually struture them, we still borrow tri of the previous ACABIT. We have not implemented yet full features of stat for Italian neither because of the lack of morpho-syntactic rules.

For example, the preposition inside of the term of [Noun_1 *di* Noun_2] structure might be equivalent to a prefix-added Noun_2 such as *deterioramento dopo la raccolta* ('rot after harvest') vs. *deterioramento post-raccolta* ('post-harvesting rot'). Likewise, the morphological derivation

[9] http://www.parks.it/ilgiornaledei parchi/e-quaderni-federparchi.html

[10] http://www.parks.it/federparchi /rivista/

[11] http://www.sinab.it/index.php?mod =normative_politiche&smod=comunitarie &m2id=189&navId=196 and http://www.sinab.it/index.php?mod=normative_politiche &smod=nazionali&m2id=189&navId=197, respectively.

```
<?xml version="1.0" encoding="UTF-8"?>
<CORPUS>
<RECORD>
<INFO>00/CAR/00 -/- 0001800/SBC/0001800</INFO>
<TITLE> </TITLE>
<AB>
<PH NB="0"> La/DTN:_:s/la presente/ADJ:_:p/presente Ricerca/SBP:_:s/Ricerca
'e/SYM/'e frutto/SBC:_:s/frutto di/PREP/di un/DTN:_:s/un lavoro/SBC:_:s/lavoro
realizzato/ADJ2PAR:_:s/realizzare da/PREP/da una/DTN:_:s/una
pluralit'a/ADJ:_:s/pluralit'a di/PREP/di soggetti/SBC:_:p/soggetto -/SYM/-
pubblici/ADJ:_:p/pubblico ,/, privati/ADJ:_:p/privato ,/, del/DTN:_:s/del
mondo/SBC:_:s/mondo della/DTN:_:s/della ricerca/SBC:_:s/ricerca e/COO/e
dell'/DTN:_:s/dell' associazionismo/SBC:_:s/associazionismo -/SYM/-
sul/DTN:_:s/sul tema/SBC:_:s/tema agricoltura/SBC:_:s/agricoltura ,/,
ambiente/SBC:_:p/ambiente ,/, aree/SBC:_:p/area protette/ADJ:_:p/protetto ,/,
occupazione/SBC:_:p/occupazione ./.
</PH>
...
</AB>
</RECORD>

<RECORD>
...
</RECORD>
</CORPUS>
```

Figure 1: Example of the input of ACABIT

```
<!ELEMENT CORPUS (RECORD)*>
<!ELEMENT RECORD (DATE?, TITLE?, INFO?, AB)>
<!ELEMENT DATE (#PCDATA)>
<!ELEMENT INFO (#PCDATA)>
<!ELEMENT TITLE (#PCDATA)>
<!ELEMENT AB (PH)*>
<!ELEMENT PH (#PCDATA)>
<!ATTLIST PH NB CDATA #IMPLIED>
```

Figure 2: DTD for the input format of ACABIT

Domain	Total # of extracted multi-word terms	Unique # of terms (base terms + variations)	Unique # of terms (base terms + variations) without hapax
"Natural Area"	34,665	21,119 (16,182+4,937)	4,131 (3,724+407)
	120,633	63,244 (46,421+16,823)	12,674 (11,481+1,193)
"Organic Agriculture"	10,071	3,201 (2,509+692)	1,737 (1,431+306)

Table 1: Experimental results

```
<?xml version="1.0" encoding="UTF-8"?>
<LISTCAND>
...
<SETCAND new_ident="3" loglike="4839.794" freq="183">
<LINK type="Neg" old_ident1="3" old_ident2="3_0"></LINK>
<LINK type="Neg" old_ident1="3" old_ident2="3_1"></LINK>
        <CAND old_ident="3_0">
        <NA freq="38">
                <MODIF> <TERM> prodotto non biologico </TERM>
                </MODIF>
        </NA>
        </CAND>
        <CAND old_ident="3_1">
        <NA freq="4">
                <MODIF> <TERM> prodotto alimentare non biologico </TERM>
                </MODIF>
        </NA>
        </CAND>
        <CAND old_ident="3">
        <NA freq="2">
                <COORD> <TERM> prodotto biologico e non biologico </TERM>
                </COORD>
        </NA>
        <NA freq="1">
                <MODIF> <TERM> prodotto ittico biologico </TERM>
                </MODIF>
        </NA>
        <NA freq="138">
                <BASE> <TERM> prodotto biologico </TERM>
                </BASE>
        </NA>
        </CAND>
</SETCAND>
...
<SETCAND new_ident="6" loglike="6757.769" freq="260">
    <CAND old_ident="6">
    <NA freq="234">
        <BASE> <TERM> area protetto </TERM>
        </BASE>
    </NA>
    <NA freq="23">
        <MODIF> <TERM> area naturale protetto </TERM>
        </MODIF>
    </NA>
    <NA freq="3">
        <MODIF> <TERM> area marino protetto </TERM>
        </MODIF>
    </NA>
    </CAND>
</SETCAND>
<SETCAND new_ident="881" loglike="1855.26" freq="39">
    <CAND old_ident="881">
    <NA freq="39">
        <BASE> <TERM> pratica agricolo </TERM>
        </BASE>
    </NA>
    </CAND>
</SETCAND>
...
</LISTCAND>
```

Figure 3: Example of the output

attivitá economiche sostenibili (‘economical sustainable activity’)
conservazione del paesaggio (‘landscape preservation’)
danno ambientale (‘environmental damage’)
elemento naturalistico (‘naturalistic element’)
equilibrio naturale (‘natural equilibrium’)
denominazione d’origine protetta (‘protected origin denomination’)
denominazione d’origine controllata (‘controlled origin denomination’)

Figure 4: Example of common terms shared in both “natural area” and “organic agriculture”

of Noun_2 in [Noun_1 *di* Noun_2] structure might imply a relational adjective such as *acidità del sangue* (‘acidity of the blood’) vs. *acidità sanguigna* (‘blood acidity’). Figure 5 shows examples of rules of morpho-syntatic variations between noun and adjectival endings for Italian, which they are independently provided as external properties file for Italian. In Figure 5, endings `-zione` (nominal) and `-tivo` (adjectival) mean that if there are adjective ended with `-tivo` like *affermativo*, the system searches for the morphological derivation of a noun ended with `-zione` like *affermazione* and put them together. Only partial rules of morpho-syntatic variations for Italian are presently integrated. We try to find the exhaustive list in near future.

3 Discussion, Conclusion and Future Work

In general, manual retrieval and validation of terms is labor intensive and time consuming. The automatic or semi-automatic methods which works on text in order to detect single or multi-word terms relevant to a subject field is referred to as term extraction. Term extraction produces the raw material for terminology databases. It is a process which is likely to produce significant benefits in terms individuation. The reasons which justify term extractions are:

1. building glossaries, thesauri, terminological dictionaries, and knowledge bases; automatic indexing; machine translation; and corpus analysis rapidly.

2. Indexing to automatize information retrieval or document retrieval.

3. Finding neologism and new concepts.

Term extraction systems are usually categorized into two groups. The first group is represented by the linguistically-based or rule-based approaches use linguistic information such as POS and chunk information to detect stop words and to select candidate terms to predefined syntactic patterns. The second group is represented by the statistical corpus-based approaches select n-gram sequences as candidate terms. The terms are selected by applying statistical measures. Recently, these two approach are combined.

We implement ACABIT for Italian, which uses the combined method to extract multi-word terms and structure them automatically. We introduce base term structures and their linguistic variation such as graphical, inflexional, and shallow syntactic variations. We also consider the modification of the structure of base terms such as internal modification using adjective and coordinate variations. We evaluate on two domain specific corpora mutually related “natural area” and “organic agriculture” to extract multi-words terms and we find 600 unique terms shared in both copora. This paper is based on our efforts on automatic multi-word terms extraction and its conceptual structure for multiple languages and this is a case study for Italian language. For the moment, we reimplement universal `stat` for major Romance languages. Most of previous work on extracting terms, especially for multiple languages are focusing on single-word terms and they are also often based on statistical approach with simple morphological patterns, for example Bernhard (2006), and Velupillai and Dalianis (2008).

Nominal ending	Adjectival ending	Examples
-zione	-tivo	*affermazione* ('affirmation') / *affermativo* ('affirmative')
-zione	-ante	*comunicazione* ('communication') / *comunicante* ('communicable')
-logia	-metrico	*ecologia* ('ecology') / *econometrico* ('econometric')
-gia	-gico	*enologia* ('enology') / *enologico* ('enologic')
-a	-ante	*cura* ('treat') / *curante* ('treating')
-	-bile	*cura* ('treat') / *curabile* ('treatable')
-ia	-peutico	*terapia* ('therapy') / *terapeutico* ('therapeutic')
-	-le	*vita* ('life') / *vitale* ('vital')
-	-tico	*acqua* ('water') / *acquatico* ('aquatic')

Figure 5: Example of rules of morpho-syntatic variations (noun-adjective)

Since this work is still on-going, we consider only **Hyperonomy** relations as the conceptual relation where a relative adjective modifies inside of the base term with [Noun_1 Adj] or [Noun_1 *di* Noun_2] structures. We also consider **Antonym** only with negative adverbs like *non*. There are still **Antonym** (e.g. *solubilità micellare* ('micellar solubilization') vs. *insolubilità micellare* ('micellar insolubilisation')), **Set of** (e.g. *piuma d'anatra* ('duck feather') vs. *piumaggio dell'anatra* ('duck feathers')), **Result** (e.g. *filettaggio del salmone* ('salmon filleting') vs. *filetto di salmone* ('salmon fillet')) relationships. ACABIT for French detects conceptual relations by using morphological conflating which implements stripping-recording morphological rules. We are planning to add these conceptual relationships in ACABIT for Italian in near future.

Acknowledgment

The authors would like to thank Béatrice Daille who kindly provide to us with ACABIT, for her valuable remarks on an earlier version of this paper. We also thank the four anonymous reviewer for their constructive comments.

References

Agbago, Akakpo and Caroline Barrière. 2005. Corpus Construction for Terminology. *Corpus Linguistics 2005*. Birmingham, United Kingdom, July 14-17, 2005.

Bernhard, Delphine. 2006. Multilingual Term Extraction from Domain-specific Corpora Using Morphological Structure. In *11th Conference of the European Chapter of the Association for Computational Linguistics*. Trento, Italy, April 3-7, 2006.

Collet, Tanja. 2000. *La réduction des unités terminologiques complexes de type syntagmatique*. Ph.D. Dissertation. Université de Montréal.

Daille, Béatrice. 2002. *Découvertes linguistiques en corpus*. Habilitation à diriger des recherches. Université de Nantes.

Daille, Béatrice. 2003. Conceptual structuring through term variations. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*. Sapporo, Japan. July 7-12, 2003.

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61-74.

L'Homme, Marie-Claude. 2004. *La terminologie : principes et techniques*, Les Presses de l'Université de Montréal.

Velupillai and Dalianis (2008).

Velupillai, Sumithra and Hercules Dalianis. 2008. Automatic Construction of Domain-specific Dictionaries on Sparse Parallel Corpora in the Nordic Languages. In *Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization*. Manchester, United Kingdom. August 23, 2008.

Williams, Geoffrey Clive. 2003. From meaning to words and back: Corpus linguistics and specialised lexicography. *ASp* 39-40. http://asp.revues.org/1320.