Applying Lexical Constraints on Morpho-Syntactic Patterns for the Identification of Conceptual-Relational Content in Specialized Texts.

Couturier, Jean-François, Neuvel, Sylvain, Onscope inc.
651 Notre-Dame Ouest, Montréal, Qc, Canada, H3C 1J1. jfcouturier@onscope.com, sneuvel@onscope.com
Drouin, Patrick, Observatoire de linguistique Sens-Texte,Université de Montréal;
C.P. 6128, succursale Centre-ville, Montréal, Qc, Canada, H3C 3J7. patrick.drouin@umontreal.ca

0. Abstract

In this paper, we describe a formal constraint mechanism, which we label Conceptual Constraint Variables (CCVs), introduced to restrict surface patterns during automated text analysis with the objective of increasing precision in the representation of informational contents. We briefly present, and exemplify, the various types of CCVs applicable to the English texts of our corpora, and show how these constraints allow us to resolve some of the problems inherent to surface pattern recognition, more specifically, those related to the resolution of conceptual or syntactic ambiguities introduced by the most frequent English prepositions.

1. Introduction

In recent years, a number of teams have worked on defining models for the automatic extraction of terminology (Daille, 1994; Bourigault, 1994; Jacquemin, 2001) and conceptual relations (Meyer & al., 1999; Morin, 1999; Condamines & Rebeyrolle, 2001; Marshman et L'Homme, 2005) in order to build knowledge bases that can be used for various applications: information retrieval, ontology building, terminology work, etc. The domain of intellectual property, particularly the description of wares and services, tends to show characteristics that would make it a good candidate for natural language processing (NLP). In this project, NLP is used for (1) automatic classification of wares and services within the Nice classification (www.wipo.int), (2) computer-assisted translation of the descriptions of wares and services as well as (3) filtering techniques applied to information retrieval in the domain of trademark searching.

The aim of this study is to provide an effective method for the automatic extraction of terms and conceptual relations between terms in this domain in order to gain access to the informative content of texts. The corpus used for the experiments is built using a sample of documents taken from the Canadian Trademark Register (roughly 12 million words) describing wares and services covered by trademarks. Our hypothesis is the following: descriptions can be analysed and processed as a sublanguage. Based on that assumption, we can propose a methodology for extracting uniterms, multiterms and conceptual relations specific to the area of wares and services of trademarks. In order to achieve this, we use a hybrid technique of linguistic features combined with statistical and corpus-based tools. Candidate terms are extracted using standard statistical methods like bigrams, log-likelihood and mutual information. The results of the statistical processing are filtered using a combined term formation pattern technique (Frantzi & Ananiadou, 1997; Drouin, 2003) and term frontier approach (Bourigault, 1994). Conceptual relations are identified using lexical-syntactic patterns that were collected and formalized following a thorough manual analysis of the sublanguage being processed. The tool identifies, amongst others, the following major relations: HYPONYMY (… *footwear* ***namely*** *running shoes, hiking boots…*), FUNCTION (…*electrical apparatus* ***for*** *recording, reproducing, amplifying and processing sound…*), COMPOSITION (…a *decorative device consisting of a base* ***which contains*** *electronic apparatus…*), DOMAIN (…*software and users manuals sold therewith ,* ***for use in*** *the fields of call management , contact management…*), USER (…*message pads* ***for use by*** *pharmacists…*) and NEGATION (… *knit goods ,* ***except*** *hats and caps…*). Patterns are represented using a tripartite label containing the following elements: <HEAD_SUB_REL, FORM, RELATION> where HEAD_SUB_REL is syntactic relation between head and subordinated terms, RELATION is the conceptual relation and FORM is a lexical-syntactic pattern. The output of the analysis is a tagged version of the input text where all relations are encoded with such triple tags and followed by the scope of the relation, as in the example below[1]:

(1)

source: *…containers made of plastic materials for use by pharmacists in the dispensing of pills , tablets , capsules , liquids and other forms of medication…*

output:

```
containers
[COMPLEMENT/MADEOF/COMPOSITION] made of [/COMPLEMENT]
        [C0--] plastic materials [--C0]
[COMPLEMENT/USEBY/USERS] for use by [/COMPLEMENT]
        [C2--] pharmacists [--C2]
[COMPLEMENT/IN/FUNCTION] in [/COMPLEMENT]
[C3--] the dispensing
        [PREPOSITION/OF/SPEC] of [/PREPOSITION]
        [P4--] pills [COMMA],[/COMMA] tablets
        [COMMA],[/COMMA] capsules [COMMA],[/COMMA]
        liquids
        [CHARNIERE/TYPEOF/HYPONYM] and other forms of
        [/CHARNIERE] [H5--] medication [--H5]
        [--P4]
[--C3]
```

2. Conceptual Constraint Variables

The constraints under discussion relate to the third and final part of the tags described above, the RELATION element. They are designed to increase the granularity of the analysis, to enable a finer definition of the conceptual

[1] For the purpose at hand, it is sufficient to define the scope of a conceptual marker as the segment of text located to its right that is semantically or conceptually dependent on it.

content in the analyzed texts. CCVs fall into different categories and can be defined along two major axis, the first one concerning the formal aspect of the constraint, the second having to do with its function. CCVs can thus be *lexical* (composed of one or many lexical elements), *syntactic* (relating to word order and/or parts-of -peech), or *morpho-syntactic* in nature (going beyond part-of-speech to include phonological or orthographical aspects of words). Functionally speaking, CCVs are defined as *endomorphic* if they participate in defining the formal content of a conceptual marker, and *contextual* if they restrict the application of a marker by specifying elements outside the marker's boundaries. In the examples to follow, we use what we consider to be variants of the marker FOR to exemplify all 6 flavors of CCVs.

2.1 Lexical Constraints

Lexical constraints are simply composed of a list of words, or even a single word. The only obligatory element of our marker FOR can actually be defined as an endomorphic lexical CCV, as shown in the example below, where the boldface element is simply a label for the CCV and the element "Lex:" introduces its lexical content.

(2) ccv1: **FOR**
{ Lex: *for* }

The same marker can also involve lexical CCV containing lexical items that can only optionally be found within its boundaries, as with the expressions <*designed for*>, <*especially for*>, etc., which are, within the limits of our sublanguage, equivalent to the bare preposition. Two of these endomorphic lexical CCV are given below. Their actual use and their respective positions within the marker will be discussed in section 3.

(3) ccv2: **ESPECIALLY**
{ Lex: *especially* | *specifically* | *primarily* | *essentially* }

ccv3: **DESIGNED**
{ Lex: *designed* | *adapted* | *constructed* | *made* | *fitted* }

2.2 Syntactic constraints

Syntactic CCVs only refer to the linear position of an expression and/or to the part-of-speech of adjacent lexical items. The default, and least precise interpretation of our marker FOR, which we call "specifier", or SPEC, is a simple restriction on the reference of the preceding element. It is found in phrases like : *wires for computers, food for cats*, etc. and can in part be defined using a syntactic CCV that restricts the following word to a plural noun. Here the element "Syn:" denotes the syntactic nature of the restriction, NNS is the plural noun POS-tag used in the Penn Tree Bank.

(4) ccv4: **FOLLOWED_BY_NNS**
{ Syn: __ NNS}

2.3 Morpho-Syntactic Constraints

Morpho-syntactic CCVs specify some of the (morpho)phonological features of words. For example, the next two CCVs require a singular noun whose ending is either *–ion* or *–ment*, or a noun or gerund (VBG) ending in *–ing*. Both CCVs are useful in determining whether the scope of a given FOR marker denotes the function of the preceding element. We discuss these two CCVs further in the next sections.

(5) ccv5: **ION/MENT**
{ M-Syn: X(*ion* | *ment*)/NN }

ccv6: **ING**
{ M-Syn: X*ing*/(NN | VBG) }

2.4 Complex Conceptual Constraints

CCVs, as we discussed, can contain lexical, syntactic, or morpho-syntactic information. CCVs can also contain information found on more than one level of representation, much as what we find in auto-modular theories of grammar (Sadock, 1991, for example). Furthermore, CCVs can call upon one another in their definition. A syntactic CCV, for example, can specify that the lexical element in a given position must match the specifications of another lexical or morpho-syntactic CCV, as in the example below.

(6) ccv7: **FOLLOWED_BY_ACTION**
{ M-Syn: X = **ION/MENT** | **ING**
Syn: __ DET? X **OF**? NP }

CCV7 indicates that the following element corresponds to the morpho-syntactic patterns defined in the two CCVs in (5), followed by an optional *of* and a noun phrase. Similarly, two of the following three CCVs are defined using the lexical pattern of the first one. (Question marks indicate optional elements, AP and NP respectively stand for adjectival phrase and nominal phrase.)

(7) ccv8: **LEX_DOMAIN**
{ Lex: *field* | *domain* | *sector* | *industry* }

ccv9: **DOMAIN_OF**
{ Lex: X = **LEX_DOMAIN** Y = **OF**
Syn: __ DET? X Y NP }

ccv10: **X_DOMAIN**
{ M-Syn: X = **LEX_DOMAIN**
Syn: __ DET? (AP | NP) X }

CCV8 is a list of lexical items having to do with domains of application; CCV9 indicates that the following phrase must have the form: *the* (optional) (*field* | *domain* | *sector* | *industry*) *of*, followed by a noun phrase; and CCV10 basically requires a noun phrase that ends with one of the lexical items of LEX_DOMAIN.

2.5 Endomorphic and Contextual Constraints

Going back to our marker FOR, its various conceptual interpretations can be defined using lexical, syntactic or morphosyntactic CCVs, some of which described in the previous sub-sections. In other words, a Conceptual Constraint Pattern (CCP), that is to say a surface pattern associated with a precise conceptual interpretation, can be exhaustively defined as a set of Conceptual Constraint Variables[2]. These CCVs can be used to specify elements that are included in the marker itself, we refer to such

[2] A Conceptual Constraint pattern defined by multiple CCVs is thus a Complex Conceptual Constraint Pattern, or CCCP, which represents Unambiguous Surface Semantic Restrictions, or USSR.

constraints as endomorphic (or more precisely endomorphic to a given conceptual constraint pattern). Conversely, a CCV can be used to specify the context, be it syntactic, lexical, or morpho-syntactic, in which a marker receives a given interpretation. The marker FOR can be interpreted as introducing either a FUNCTION of the preceding item, its DOMAIN of application, its USERS, etc. Each of these interpretations corresponds to a conceptual pattern exhaustively composed of CCVs. For example, the triple tag <COMPLEMENT/FOR/FUNCTION> (FOR denoting the function of the item that precedes it) is triggered by the conceptual pattern in (8) and covers expressions such as those found in (9)[3]: (the CCV labeled USEIN refers to the word *use* followed by the word *in*.)

(8) **<COMPLEMENT/FOR/FUNCTION>**
Endo : **ESPECIALLY**? **DESIGNED**? **ESPECIALLY**? **FOR USEIN**?
Context : **FOLLOWED_BY_ACTION**

(9) *apparatus <for> cleaning carpets*
apparatus <designed for> the treatment of cancer
apparatus <especially designed for use in> polishing floors

The endomorphic portion of the pattern indicates what must or may appear inside the marker itself, and the contextual part specifies that this pattern must immediately precede the pattern described by the CCV **FOLLOWED_BY_ACTION**, i.e. a word ending in *–ing*, *-ion* or *–ment* and belonging to specific morpho-syntactic categories. Similarly, FOR will be interpreted as introducing the DOMAIN of application of the preceding item when it matches one of the conceptual patterns in (10) and covers expressions such as those found in (11).

(10) **<COMPLEMENT/FOR/DOMAIN> #1**
Endo : **ESPECIALLY**? **DESIGNED**? **ESPECIALLY**? **FOR USEIN**?
Context : **X_DOMAIN**

<COMPLEMENT/FOR/DOMAIN> #2
Endo : **ESPECIALLY**? **DESIGNED**? **ESPECIALLY**? **FOR USEIN**? **DOMAIN_OF**
Context : NP

(11) a) *apparatus <for> the dairy industry*
apparatus <adapted for use in> the scientific fields

b) *apparatus <for the field of> science*
apparatus <adapted for use in the domain of> mining

Both conceptual patterns are composed of CCVs described earlier. It is interesting to note that linearity imposes a different structure to conceptual patterns covering very similar expressions. Pattern #2, covers expressions in which words like *domain* or *field*, covered by CCV **LEX_DOMAIN**, are followed by a prepositional phrase. The word included in **LEX_DOMAIN** and the following preposition can therefore be included inside the boundaries of the marker since they do not add any information. Pattern #1, on the other hand, matches expressions in which the word included in **LEX_DOMAIN** is the head of its noun phrase and, thus, to the right of the relevant words. It can therefore not be included inside the marker and must be part of the contextual portion of the pattern rather than in its endomorphic section. The marker FOR can also be interpreted as introducing a USER, the type of individuals for whom the element to the left is intended. Three additional CCVs are necessary to define the relevant Conceptual Constraint Patterns.

(12) ccv11: **LEX_USER**
{ Lex: *men* | *women* | *children* | *pharmacists* | etc.}

ccv12: **FOLLOWED_BY_USER**
Lex: X = **LEX_USER**
Syn: __ X

ccv13: **USEBY**
Lex: X = *use* ; Y = *by*
Syn: __ X Y

(13) **<COMPLEMENT/FOR/USERS > #1**
Endo : **ESPECIALLY**? **DESIGNED**? **ESPECIALLY**? **FOR USEBY**
Context : **FOLLOWED_BY_NNS**

<COMPLEMENT/FOR/USERS > #2
Endo :**ESPECIALLY**? **DESIGNED**? **ESPECIALLY**? **FOR**
Context : **FOLLOWED_BY_USER**

A given instance of the marker FOR is thus interpreted as introducing a USER if it includes the segment "for use by" followed by a plural noun; or if FOR is immediately followed by a lexical element included in **LEX_USER**. The CCPs in (13) covers examples such as (14):

(14) *footwear <for>men, women and children*
apparatus< especially designed for> children
apparatus < intended for use by> pharmacists

A summary of the various CCPs included under the label FOR is provided in Table 1.

3. Results and Evaluation

The manual evaluation of the results obtained using this technique indicates a high degree precision of the analysis without significant impact on recall and performance of the overall process. Of the various interpretations of the marker FOR, FUNCTION is by far the most frequent, followed by SPEC, DOMAIN, and finally USER. 97% of tokens received the correct interpretation and 7 tokens were interpreted incorrectly. Of those 7, 1 is a false positive for the FUNCTION interpretation, the rest are all cases in which the marker received the SPEC interpretation instead of FUNCTION or USER. Since SPEC (see definition in Table 1) is the least precise and the default interpretation of the marker FOR, it is easy to understand why almost all interpretation errors are included under this label. If, for example, the lexical item *accountant* is not included in the CCV labeled **LEX_USER**, the software simply cannot distinguish between the phrases *software for accountants* and *software for nuclear reactors* and both expressions will receive the SPEC interpretation. Furthermore, the extremely low frequency of the USER interpretation has

[3] Obviously, the patterns defined in these examples, as well as the CCVs we describe are somewhat simplified versions of their actual self. A complete definition using the actual CCVs used in our application would introduce a level of complexity that is unnecessary given our current purposes.

made the creation of a somewhat thorough lexical set for **LEX_USER** rather difficult, hence the very low recall of the USER interpretation. The distribution, precision and recall for the various CCPs of FOR are given in Table 2.

FUNCTION
Endo : **ESPECIALLY? DESIGNED? ESPECIALLY? FOR USEIN?**
Context : **FOLLOWED_BY_ACTION**
Ex:

- *apparatus <for> cleaning carpets*
- *apparatus <designed for> the treatment of cancer*
- *apparatus <especially designed for use in> polishing floors*

DOMAIN
Endo : **ESPECIALLY? DESIGNED? ESPECIALLY? FOR USEIN?**
Context : **X_DOMAIN**
OR
Endo : **ESPECIALLY? DESIGNED? ESPECIALLY? FOR USEIN?**
DOMAIN_OF
Context : NP
Ex:

- *apparatus <for> the dairy industry*
- *apparatus <adapted for use in> the scientific fields*
- *apparatus <for the field of> science*

USER
Endo : **ESPECIALLY? DESIGNED? ESPECIALLY? FOR USEBY**
Context : **FOLLOWED_BY_NNS**
OR
Endo : **ESPECIALLY? DESIGNED? ESPECIALLY? FOR**
Context : **FOLLOWED_BY_USER**
Ex:

- *footwear <for>men, women and children*
- *apparatus< especially designed for> children*
- *apparatus < intended for use by> pharmacists*

SPEC
Endo : **ESPECIALLY? DESIGNED? ESPECIALLY? FOR**
Context : **FOLLOWED_BY_NNS**
Ex:

- *wires <for> computers*
- *food < for> cats*
- *rubber tires <especially designed for> mountain bikes*

Table 1: Conceptual Constraint Patterns
of the marker FOR

Interpretation	*Distribution*	*Precision*	*Recall*
FUNCTION	69.4%	99.4%	98.1%
DOMAIN	2.6%	100%	100%
USER	1.3%	100%	50%
SPEC	26.7%	90.3%	99.4%

Table 2: Precision and Recall for the CCPs
of the marker FOR

4. Conclusion

The use of CCVs in the analysis of text yields a granularity in the analysis that is sufficient for the extraction and interpretation of information contained in descriptions of wares and services. The increase in precision resulting from the technique allows us to reuse the output for further NLP processing such as the development of a specialized knowledge base, the building of an ontology, computer-assisted translation and the expansion of a translation memory with increased reliability. The advantages of the mechanisms we propose are twofold. First, the use of CCVs in the analysis of text describing wares and services obviously increases the precision of the analysis when compared to a scenario in which every instance of the marker is given the same vague interpretation. Second, the use of CCVs allows us to process relatively complex linguistic structures using extremely simple mechanisms. It allows us to use only information contained in the text without recourse to complex representational schemes, unification grammars, or dictionaries or thesaurus containing conceptual or semantic information. On the other hand, it seems clear that an approach such as ours can only be realistically implemented when dealing with a specific sub-language characterized by a limited number of structures and conceptual relations. The efficiency of the analysis is also proportional to one's knowledge of the sub-language in question, and presupposes a detailed examination of the data. Given the size of the corpora, it is often difficult to achieve an exhaustive analysis of a given pattern, especially when the CCVs that define that pattern are lexical in nature.

5. References

Bourigault, Didier. (1994). Extraction et structuration automatique de terminologie pour l'aide à l'Acquisition des connaisssances à partir de textes. *RFAI'94*, pp. 397-408.

Condamines, A and Rebeyrolle, J. (2001). Searching for and identifying conceptual relationships via a corpus-based approach to a Terminological Knowledge Base (CTKB): Method and Results. In Didier Bourigault, C. Jacquemin and M-C. L'Homme (Eds.), *Recent Advances in Computational Terminology.* Amsterdam : John Benjamins, pp. 127-148.

Daille, Béatrice. (1994). *Approche mixte pour l'extraction de terminologie: statistique lexicale et filtres linguistiques*. Thèse en informatique fondamentale. Université de Paris 7. Paris.

Drouin, Patrick (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, vol. 9, no 1, pp. 99-117.

Frantzi, K.T. and S. Ananiadou. (1997). Automatic Term Recognition Using Contextual Cues. *Proceedings of the 3rd DELOS Workshop*. Zurich, offprint.

Jacquemin, Christian. (2001). *Spotting and Discovering Terms Trough Natural Language Processing.* Cambridge: MIT Press.

Marshman, E. et M.-C. L'Homme. (2005). Disambiguation of Lexical Markers of Cause and Effect. *Proceedings of LSP 2005*. Bergamo, Italy.

Meyer, I. K. Mackintosh, C. Barrière, and T. Morgan. (1999). Conceptual sampling for terminographical corpus analysis. *Proceedings of 5th International Congress on Terminology and Knowledge Engineering (TKE99)*. Innsbruck, Austria, pp 256-267.

Morin, E. (1999). *Extraction automatique de liens sémantiques entre termes à partir de corpus de textes techniques.* PhD Thesis in Computer science, Université de Nantes.

Sadock, Jerrold M. (1991). *Autolexical Syntax: A Theory of Parallel Grammatical Representations*. Studies in Contemporary Linguistics, Chicago: University of Chicago Press.