Lexical sets and Text-Processing

Christian CHAMPENDAL
Faculté des Lettres et Sciences Humaines - University of Nice (UNSA)
Nice, France
champend@unice.fr

Thierry PITARQUE
Equipe ASTRE, UNSA-CNRS
06903 Sophia Antipolis-Cedex France
pitarque@unice.fr

Abstract

The extraction of lexical sets from a corpus in Digital Signal Processing (DSP) has been detailed before on general sets, with direct ELT applications. In this contribution, a more specialized set is investigated to illustrate the possibility of actually using the results in more "intelligent" Text-Processing.

1 Introduction

The **Lexical Set Extraction** technique was presented before[1]. From a sample corpus in Digital Signal Processing, lexical sets are extracted, the most important ones are investigated to establish recurring predicative relations within the domain. In this contribution, another set from the same corpus is studied, to briefly illustrate the technique. Possible applications in Text Processing are then assessed.

1.1 Corpus

Results are extracted from a sample corpus of 135 articles in DSP, which represents about 500 pages, 3900 paragraphs, 16 500 lines and 350 000 word occurrences (2 million characters with a size of 3,27 Moctets). Articles were selected from a CD-Rom, ICASSP'98 International Conference on Acoustics, Speech and Signal Processing according to quality and diversity. Lexical sets are then easily obtained from the lists of lemmas and their occurrences. From all the corpus sets, about 100 are quantiatively prominent and their study will ensure a reasonable coverage of the domain.

2 Lexical set [*FILTER*]

This more specialized set is frequently found in the corpus. It is constituted of only four **lexical units** : *filter/filters* are verb-noun homographs, *filtering* is a verb-noun-gerund one, *filtered* is a verb-adjective homograph. Most of the 1441 **occurrences** are nominal with few verbal combinations. Hence, the different units of the set can be predicted in the nominal **positions:** ***Co***, C_1, C_2 (complements), $Circ_n$ (circumstancials) or ***NC*** (noun complement).

```
[Filter](1441/4)
(809)filter (62)filtered
(146)filtering    (424)filters
```

As for **collocations**, elements of the set are found as modifiers or heads of nominal, adjectival, prepositional phrases and compounds, some of which are listed below:

```
adaptive filter(60)
adaptive filtering(14)
filter bank(151)
fir filter(55)
modulated filter bank(23)
weighted median filter (13)
polyphase synthesis filter(8)
polyphase synthesis filter bank(6)
```

Some interesting side phenomena are observed for this set: **acronyms** *(*EKF= *extended Kalman filter,* QMFs=quadrature mirror filters*, WM=Wiener median filters)*; **affixates** with or without dash (*subfilter, anti-filter, post-filter)*; **agglutinates** *(filterbanks),* and **coinage** *("liftered*" meaning *low-pass filtered)*. Two parallel collocative **paradigms** built on *filter* 'Object' and *filtering* 'Activity' coexist *(digital filtering/filter(s); kalman filtering/filter).* The quantitative importance of these phenomena underlines the necessity to process them thoroughly

3 Uses in Text-Processing

Concordances of verbal **and** nominal units and their correlations are studied with a view to extracting an **Extended Predicative Formula** that lists complements and circumstancials according to frequency for all predicative members of the set[2]

---

[1] fully in [Champendal03] and in a previous contribution [Champendal04 Oct] , both drawing on general lexical set [*USE*]. ELT and especially ESP classroom applications are direct: lexical acquisition, syntactic and semantic structuration.

[2] Order of complements and circumstancials is obviously flexible, numbers correspond to statistically

Co <filter>C_1 C_{irc1} $C_{irc2}$$C_{irc3}$ C_{irc4}

Co = *{we, algorithm, filter, weight,...}*
C_1 = *{data, image, signal, noise, procedure, sample, waveforms,...}*
C_{irc1} = **from** *{error, source, signal}*
C_{irc2} = **in/over** *{domain, region, frequency band...}*
C_{irc3} = **to do** *{imaging, noise removal...}*
C_{irc4} = **by use of /using** *{statistics, optimization techniques...}*

Verbal units represent only 2% of all occurrences. They are mostly passive; a small number of infinitives and gerunds are also found in the corpus.

*In WI, pitch-cycle waveforms **are filtered** **in** the evolution domain **to** decompose the signal **into** two waveform surfaces, one characterising voiced speech and a second representing unvoiced speech.*

*In an array processing application in [10] an additional noise determining transducer was used to cancel noise and interference, but in other circumstances it becomes appropriate **to filter** noise **from** signals and it has been shown in this paper that it is possible to consider **filtering** **in** the two dimensional third order cumulant domain*

The sets' **connectivity** (ie. affinity with other prominent sets of the corpus), is automatically detected as combined occurrences of two sets are counted in the *Associated Units* files facility of the **Z-text®** software. **Clusters** are then studied as a priority within the corpus: they will enable prediction of recurrent relations in the domain.Highest clustering obtains with the following sets: *[OPTIM], [DESIGN], [FILTER] (reflexive), [RESULT], [SIGN], [SHOW], [REQUIR], [USE], [WEIGH]*. Normally, varied nominal, verbal and participial solutions are present.

[FILTER+OPTIM]

filter~optimal:14	filters~optimal:5
filter~optimization:7	filters~optimization:3
filter~optimize:2	filtering~optimal:4
filter~optimum:2	

---

established positions. Co is 'human', 'object' or 'abstract', C_1 'object' or 'abstract'; $Circ_n$ can be expressed or implied and recoverable in context

The studied set [*FILTER*] is mostly nominal. Nominal clusters (ie. connection of mostly nominal sets) produce mainly Noun-Noun or Adjective–Noun associations, but this set also clusters with verbal sets *([SHOW], [USE], [REQUIRE]...*), generating predicative relations where it occupies subject or direct object positions. Various connections with [*USE*] are presented below:

[FILTER+USE]

filter~use:8	filter~useful:3
filter~uses:2	filter~using:27
filters~use:5	filters~using:14

Extended Predicative Formula: <Use>

C_0=Subject **<Use>C_1**=Complement**(Od)**
$Circ_n$=Circumstantial**(Circ/Adv)**

*The algorithm **uses** two neighbouring pixels, one left and the other top, **as** a pioneering block **to** search for the best matched blocks **in**side a pre-defined window.*

a) Verbal use

C_0=filter <use>C_1 ($Circ_n$)

C_0 <use>C_1=filter ($Circ_n$)

*The echo-path model **use**s a single pole single zero digital **filter***

C_{01}=filter <be-**used**>(C_{10}) ($Circ_n$)

*In prediction, a **filter** is **use**d to estimate future values of a signal from prior observations.*

b) Nominal use

<**use**> of **C_1=filter** ($Circ_n$)

*This standard allows the **use** of any wavelet **filter** up to a length of 32 taps.*

c) Adjectival-participle use

C_{01}= filter <**used**> ($Circ_n$)

<modif+**used**>**C_{01}=filter** ($Circ_n$)

*Block diagram of the constrained adaptive **filter use**d to determine the Wiener solution.*
*Our iterative algorithm for state estimation is based on the Expectation- Maximization (EM) algorithm and outperforms the widely **use**d Extended Kalman **filter** (EKF)*

d) Circumstancial link <Use>

C_0 <**use**>C_1 **Circ$_3$=to filter**

These weights, in turn, were ***used*** *t<u>o</u>* ***<u>filter</u>*** *<u>the multi-tone signal resulting in the estimate shown in Fig1</u>*

Circumstancial link <Filter>

C_0 **<filter>**C_1
Circ$_4$=(by) using/by use of Z

Digital image enhancement and noise ***filtering*** ***<u>by use of</u>*** *<u>local statistics</u>.*

-Ing participle forms generate typical ambiguities where a similar surface portion of the sentence, can refer to different underlying relations[3]

C_0=filter <using>C_1 (Circ$_n$)

A ***filter*** ***<u>using</u>*** *<u>an 11 8 grid filter for F1 and a 5x5 average for F2</u> was trained on images of faces degraded by AWGN with □ 2 = 200.*

4 Conclusions

Combined with reduction of predicative relations to their core relations, the technique helps optimize various NLP applications:data and text-mining, computer-aided generation of abstracts, multilingual corpus alignment (an equivalent sample corpus was easily obtained in French, using the same technique and the sets connected together with their results: **English**: [*FILTER*]<> **French**:[*FILTRE*]). In the next stage, top-down expert knowledge will help to shape legal ontologies of the domain.

References

CHAMPENDAL C. , 2004 October; Language teaching and corpus linguistics, IATEFL – CALL Review, ISSN 1028-428

CHAMPENDAL C., 2004 November ; Outils linguistiques pour Ingénieurs, Actes du Colloque TICE Méditerrannée – ISDM (www.isdm.org)

CHAMPENDAL C. , 2003 ; Analyse d'un corpus d'articles en Traitement Numerique du Signal en vue de modelisation linguistique et d'application TIC ; Thèse de Sciences du langage, directeur: Henri Zingle.

CHAMPENDAL C. 1999 Morphosyntaxe du verbe anglais; contribution au JILA'99 ; Nice 24-25 juin 1999.

MEL'□UK Igor A., CLAS A, POLGUÈRE A, 1995 *Introduction à la lexicologie explicative et combinatoire*; Ed Duculot.

TRIBBLE C. 2000, Practical uses for language corpora in ELT in A Special Interest in Computers, P. Brett and G. Motteram Eds, IATEFL.

TdL99a Travaux du LILLA , 1999 ; Henri Zingle Numero special– La modelisation des langues naturelles – Aspects theoriques et pratiques.

ZINGLE H, 1998 ZTEXT : un outil pour l'analyse de corpus in *Travaux du LILLA n°3*, pp 69-78.

[3]Occurrences being limited, some sentences are manipulated to obtain the required string ***filter+using.*** Even marginal topicalized possibilities are accounted for*:* C_0 <filter>C_1 Circ$_4$=<using>C_1.
This entire data file we filter (by) using a Hamming weighted (mis)matched filter