

To increase the coding rate through an integrated approach of techniques of textual analysis

Come elevare il tasso di codifica attraverso un approccio integrato di più tecniche di analisi testuale

Manuela Murgia,

Valentina Talucci

ISTAT, Via C. Balbo, 16 - 00184 ROMA, e-mail: murgia@istat.it , talucci@istat.it

Riassunto: Il presente lavoro nasce da un'esigenza empirica molto precisa: migliorare le procedure di codifica automatica utilizzando in modo integrato più approcci di analisi testuale. L'obiettivo è di riuscire ad attribuire un codice alfanumerico ad una lista di informazioni testuali molto articolata, attraverso l'integrazione di procedure di normalizzazione e pulizia dei dati di input con procedure di codifica automatica che, sulla base di determinate regole, consentano il riconoscimento dei testi di input e la successiva attribuzione di un codice univoco. I software utilizzati sono: Actr, Taltac, Sas, Visual Basic.

Keywords: Automated Coding, Textual analysis techniques.

1. Methods and data

The automation of the coding process of textual responses has already been adopted in Istat for a lot of surveys to code data according to different official classifications (Industry, Occupation, Education, etc.). A generalised software system, ACTR (Automatic Coding by Text Recognition) has been used for all the coding activities providing results coherent with those of other NSI (Macchia et al., 2006). This job reports the methodology adopted for the treatment of a particular set of textual data represented by almost 3 million records of descriptions of Industry sectors provided by enterprises to the Chamber of Commerce (CCIA). The aim is to assign to each description its Industry code according to the Ateco classification. These data present some difficulties to be automatically processed: *i)* the Industry descriptions are very long: they are much longer than those usually collected in statistical surveys and 50% of them exceed the maximum length managed by ACTR (200 bytes); *ii)* they contain a lot of misspelled words; *iii)* they contain a lot of redundant and meaningless information that are absolutely not useful for the attribution of the Ateco code. For all these reasons, it has been designed a procedure that integrates different software packages for textual analysis. It is constituted of two conceptual steps:

1. the deletion of redundant texts from CCIA descriptions, using Taltac (Trattamento Automatico Lessico-Testuale per l'Analisi del Contenuto) (Taltact2, 2006) SAS and Visual Basic (VB);
2. the automatic coding of the pre-processed descriptions, using ACTR (ACTRv3, 1998)

2. Results and perspectives

The identification of redundant texts of the descriptions was the most complex job. It consisted of the joint analysis of the CCIA archive and of the so called '*Ateco dictionary*' which is the informative base used by ACTR that contains all the Ateco classification codes (about 900 categories and 17 sections) associated to almost 25,000 descriptions. Some descriptions are taken from official classification manuals, while others are collected (and then coded) in statistical surveys. The textual matching between these two lists was carried out using Taltac which performed a sort of "union" of the two sets. The texts not included in this union were considered "typical" of each list and those "typical" of the CCIA file were deleted using a VB software procedure. The identification and deletion of redundant texts were made in two phases: first, all the redundant 'segments' (sequences of words of a certain length and frequency) were deleted and the resulting file was coded; secondly, single redundant 'words' were deleted from those texts left un-coded in the first step and the coding procedure performed afterwards. The deletion of words was performed as the last step since it affected the readability of the descriptions, making more difficult the subsequent quality analysis of the codes assigned by ACTR. The textual matching provided 2,108 redundant segments and 19,142 redundant words. The procedure was tested on a random sample of 30,000 descriptions of the CCIA file. Before coding, texts longer than 200 bytes were cut in three parts according to full stop or semicolon that enterprises used to describe more Industry sectors. For these cases the procedure assigned more codes for a single description. A subsequent analysis will establish the prevalent one. Results are shown in the following table.

Table 1: *Unique codes assigned after the deletion of redundant texts*

<i>Sample of 30,000 records</i>	Coded texts		Not coded texts	
	N.	%	N	%
Original sample	6,243	20.81	23,757	79.19
Phase 1: Sample after deletion of segments	12,489	41.63	17,511	58.37
Phase 2: Sample after deletion of words	14,389	47.96	15,611	52.04

Results can be considered very good and, if the quality analysis confirm them, the procedure will be applied to the universe of descriptions. These positive results depends also on the fact that the integrated approach allowed to update and improve the '*Ateco Dictionary*' as well as the entire coding environment thus enhancing the coding rate.

References

- ACTRv3 User Guide - (1998), Statistics Canada.
 S. Macchia, S. Mastroluca (2006) The automatic coding process in the 2001 Italian General Population Census: efficacy and quality. *International Conference on Quality in Official Statistics (Q2204)*, Mainz.
 Taltac2, (2006) Guida per gli utenti.