Coding the spoken language through the integration of different approaches of textual analysis

Stefania Macchia[1], Manuela Murgia[1], Valentina Talucci[1]

[1] ISTAT- Via Cesare Balbo, 16 - 00184 Rome - Italy

macchia@istat.it - murgia@istat.it - talucci@istat.it

Abstract

The aim of this study is to assign a code to very long and complex textual descriptions provided by the Italian Chambers of Commerce according to the official classification of the Industry sector (ATECO). This is done through the integration of different software packages, which are based on different methodologies for text normalisation and analysis. Mainly, two softwares are used: the first is Taltac, for the texts' context and structure analysis; the other one is ACTR, for the text processing aimed at assigning codes. Other software (SAS and MySQL), were used to make the integration performing at best.

Keywords: automated coding, textual analysis techniques, industry sector.

1. Introduction

The aim of this study is to code the Chambers of Commerce Industry (CCIA) descriptions in order to update the ISTAT Enterprise Register that represents the universe of reference for business surveys. These descriptions can be considered a peculiar type of spoken language being not very structured, full of technical words and of expressions typical of this Register.

Usually, to automatically code survey textual variables according to official classifications a generalised software system ACTR v3 (*Automated Coding by Text Recognition*) is used with very good results in terms of coding rates. For survey variables like Occupation, Municipality, Education and Industry sector an ACTR application is created: any file containing free-text answers to these variables is submitted to the proper coding environment and coded texts are produced as output.

This time it was not possible to submit the CCIA file to the relative coding environment (called ATECO 2002) because texts were physically longer than those accepted by the system (200 bytes). Besides, this extremely high length was due, in many cases, to meaningless texts which contain information not useful for assigning a code. Finally, CCIA texts contained so many types of misspelled words to make impossible their recognition by ACTR and, therefore, their coding.

To overcome all these difficulties it was decided to make an integrated use of different software packages that perform the treatment of texts according to different approaches and aims. The main software used were:

- Taltac (*Trattamento Automatico Lessico-Testuale per l'Analisi del Contenuto*) (Taltac2, 2006), that performs a textual analysis with the aim of determining the characteristics of the content and of the structure of texts;

- ACTR which is designed to make strings recognition inside a text in order to assign a unique code.

Other software, like SAS and MySQL, have been used to make this integration performing at best.

This paper is articulated as follows:

- Chapter 2 talks about the ISTAT experiences with ACTR and the functioning of this software.
- Chapter 3 describes in details all the problems met to code the CCIA file and how they have been solved. It also provides a description of Taltac as well as the final results.

2. The processing of open-ended questions' texts in statistical surveys

Survey questionnaires designed in ISTAT mainly contain closed questions. Sometimes open-ended questions, that let respondents answer in their own words, are also used to measure economic, demographic or social phenomena. The resulting answers have to be treated with the purpose of assigning them a code according to official classifications.

The coding activity of responses to open-ended questions was made manually until some years ago, but it was a time-consuming job and did not guarantee the process standardisation. That is why in 1998 ISTAT decided to test an automated coding system. The software selected was ACTR, developed by Statistics Canada. The choice fell on ACTR because it is a generalised system, independent from the language, and already successfully used by other National Statistics Institutes (Tourigny and Moloney, 1995).

2.1. ACTR software system

ACTR's philosophy lies on methods originally developed at US Census Bureau (Hellerman, 1982), but uses matching algorithms developed at Statistics Canada (Wenzowski, 1988).

The coding activity follows a quite sophisticated phase of text standardisation, called *parsing*, that provides 14 different functions such as characters mapping, deletion of trivial words, definition of synonymous, suffixes removal, etc.. The *parsing* aims at removing grammatical or syntactical differences so to make equal two different descriptions with the same semantic content.

The parsed response to be coded is then compared with the parsed descriptions of the dictionary, the so called *reference file*. If this search returns a perfect match, called *direct match*, a *unique* code is assigned, otherwise the software uses an algorithm to find the best suitable partial (or fuzzy) matches, providing an *indirect match*.

In practice, in the latter case the software takes out of the *reference file* all the descriptions that have at least one parsed word in common with the parsed response and assigns them a standardised score that ranges between 0 and 10 (10 corresponds to a perfect match). This score is calculated as a function of the weight of each single common word; the weight is inversely correlated to the frequency of occurrence of the word in the dictionary (*reference file*). Then, the system arranges by decreasing scores the descriptions extracted from the *reference file* and compares them with some user-defined threshold parameters. The result could be:

- a *unique* match, if a unique code is assigned to a response phrase;
- *multiple* matches, if several possible codes are proposed;

- a *failed* match, if no matches are found.

The first case does not require a human intervention, while the other ones have to be evaluated by expert coders.

As already mentioned ACTR is a generalised system. It is generalised with respect to the language and the classification used. This means that the construction of the coding environment must be carried out by the user who has to adapt the system to the Italian language and to each classification and, at the end, to test it.

The construction of the coding dictionary (*reference file*) is the heaviest activity, since its quality and its size deeply affect the performance of the automated coding. This activity is aimed at processing the texts of the official classification manual so as to reproduce the respondents' natural language as close as possible. This is done in different steps, the most important of which is the integration of manual's descriptions with empirical response patterns taken from previous surveys. Not less important is the automation of rules for the assignment of codes to make the system working as a human coder would do.

2.2. Automated coding applications developed in ISTAT

Numerous coding applications for different classifications have been developed in ISTAT. The most important are referred to the following variables:

- Occupation
- Industry
- Education level
- Country/Nationality
- Municipalities.

They have been used in different surveys and for the 2001 Population and Industry Censuses with very good results.

The performance of the automated coding is measured according to two indicators:

- *Recall rate* (coding rate): percentage of codes automatically assigned.
- *Precision rate*: percentage of correct codes automatically assigned.

According to these two indicators, the performance of ISTAT coding applications were satisfactory and the results coherent with those obtained by other National Statistical Offices (De Angelis et al., 2000).

Focusing on the Industry application, the *reference file* contains almost 30,000 descriptions corresponding to approximately 900 categories and 17 sections. The automated coding has always performed better for business surveys than for households or individuals surveys. This is because the concept of Industry sector is closer to the first target than to the latter.

To exemplify, the recall rate obtained in the Population Census was 53.6%, while in the Intermediate Industry Census it was 58.8% (Macchia and Mastroluca, 2004).

3. Processing the Chambers of Commerce Industry descriptions

As already mentioned CCIA descriptions constitute one of the main informative source to update the ISTAT Enterprises Register. It is therefore very important to univocally code them.

The aim of this job is to assign the Industry code to each description, according the new Industry classification (ATECO 2007). The descriptions to be coded are stored in an archive of almost 6 million records and shows some difficulties to be automatically processed:

i) the CCIA 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.

Excessive length as well as redundancy mainly depend on the way and on the reasons why the information about the Industry sector is collected. In survey questionnaires a structured question is used together with specifications for respondents on how to describe their company's Industry sector in a synthetic way and with examples of correct and proper descriptions. Vice versa, when an entrepreneur describes to the Chambers of Commerce what his company does, there are no specifications on how to do it and he tends to go deeply in details and also to specify other concepts like the company mission, its juridical status, etc.. This is because the collection of information by the Chambers of Commerce has not the statistical purpose of surveys to measure an economic phenomena.

For all these reasons a strategy was designed to automatically treat these descriptions. It includes two logical steps:

1) the *identification* and *deletion* of redundant texts from CCIA descriptions;
2) the *automatic coding* of descriptions pre-processed in the previous step.

A software procedure was then developed which integrates different software packages for textual analysis, respectively:

1) Taltac, SAS and MySQL for the first step;

2) ACTR for the second one.

To set up this procedure a simple random sample of 60,000 descriptions was extracted from the CCIA archive, half of it constituted of texts originally longer then 200 bytes and another half shorter.

3.1. Taltac software system

As already mentioned the identification of redundant parts of descriptions was carried out using Taltac.

Taltac is a software for statistical-textual analysis that allows to know the content of a text without reading it and, after recent developments, also to know its main structures notwithstanding its content. It analyses words and relations among each others inside a text, according to a lexical-metrical approach. This approach provides information about the content of the text, its dimensions in terms of words (called *graphical forms*) and about its main grammatical or semantic categories.

It also calculates some index of statistical relevance which indicates how "strong or peculiar" a word or sequences of words are inside a text.

For the present job, the use of Taltac was limited to some of its numerous functions: the functions of *search* and *extraction* of information from the text and the *union of lists* function, which provides the union of two sets of texts.

In more details, Taltac was used to extract *graphical forms* and *repeated segments* from both files, CCIA archive and ATECO dictionary. According to Taltac terminology a *graphical form* is a sequence of characters included in the alphabet which lays among two weak-separators (character not included in the alphabet). To state it differently, it could be identified as a word. *Repeated* s*egments* are sequences of adjacent *graphical forms* of an established length that lies among two strong separators.

To extract *repeated segments* it is necessary to set the value of some Taltac parameters. Those used for this application were:

- the length of the segment, or, equivalently, the number of *graphical forms* it contains,
- the frequency of each *graphical form* inside the text.

After few empirical trials, the length parameter was set at 6 while the frequency at 2. The choice of this last low value was due to the need of finding any kind of "redundant" segment to be deleted from the CCIA texts to make them as short as possible (and therefore usable by the coding system).

Graphical forms and *repeated segments* have been extracted from the CCIA file and from the ATECO dictionary. The resulting lists were compared using the *union of lists* function. The words or the segments not included in this union were considered "typical" of each list and those "typical" of the CCIA file were deleted from the descriptions (see par. 3.2).

3.2. The identification and deletion of redundant texts from CCIA descriptions

Following a visual analysis of some of the sample descriptions, it came out that redundant parts of the texts were both single words (i.e. some adjectives, adverbs, or substantives etc. which had no relations with Industry sector) and sequences of words expressing redundant concepts. For instance, typical descriptions were "*the company was founded with the mission of producing shoes since 1995*" or "*the company is a non-profit-making association having the purpose of producing and selling hand made toys*". In these long texts, the only information relevant for the definition of an Industry sector were respectively "*producing shoes*" and "*producing and selling hand made toys*".

To test the sequence of steps to be performed, an empirical study was conducted using Taltac. From this study, it was clear that after deleting single words from descriptions they were often not so easily readable, because they had no more an accomplished grammatical structure. The same thing did not happen after deleting sequences of words expressing redundant concepts. For this reason, the identification and deletion of redundant texts was performed in two sequential steps:

1. first of all, sequences of words expressing redundant concepts were identified and deleted and the resulting "cleaned" descriptions were submitted to the automatic coding;
2. then, only not coded descriptions were processed for the identification and deletion of single words and then submitted again to the automated coding.

The *reference file* built for the automatic coding application was considered the *corpus* containing the whole set of relevant texts with regard to the Industry sector. A comparative analysis (union of lists, see par. 3.1) between this corpus and the CCIA archive was carried out to identify the redundant texts to be deleted from the descriptions of this last file.

Taltac was used to identify *segments* in the two archives. To each segment the system assigns an index of relevance which is maximum when the segment contains only words having a meaningful content and that are present only inside the segment (and not in rest of the *corpus*).

Segments extracted from the two archives were sorted according to their frequency and then compared: all *segments* with a frequency higher than 50 present in CCIA archive and not present in our *corpus* were considered redundant and deleted.

This analysis succeeded in identifying 2,108 *redundant segments*, which were deleted from the CCIA archive with a MySQL procedure.

The same logic was used to identify redundant words. Taltac was used to extract from the reference *corpus* and from the CCIA archives the respective vocabularies (whole of different words with their corresponding frequency).

The words not in common in the two vocabularies were 22,206. They were analysed to identify which of them could be used to enrich the *reference file* (they could be considered synonymous of words already present in the coding dictionary) and which instead could be deleted. As a result of this step, 19,142 words were considered redundant and deleted from the CCIA archive.

3.3. The results of automated coding of CCIA descriptions

The procedure on the original sample was performed separately on two files: the first one containing descriptions shorter than 200 bytes and the other one with longer texts, as 200 bytes is the threshold length to be submitted to ACTR.

Both files were processed in order to delete redundant segments. After the deletion phase texts were submitted to the automated coding. It has to be said that before entering the coding step texts longer than 200 bytes were cut in four parts according to full stop or semicolon; this means that more than one code could be assigned to each company. This was considered a correct solution because it was noted that texts contained descriptions relative to more than one Industry sector and therefore it was important to assign at least one code. For those cases receiving multiple codes a subsequent analysis will establish which is the prevalent one.

The following table shows the results of the automated coding on the two sub-samples (longer and shorter than 200 bytes, each of them regarding 30,000 companies) obtained before and after the deletion of redundant segments. These results are at level of single company (to each company at least an ATECO code was assigned). Naturally, the sample with the descriptions longer than 200 bytes gave no unique codes before being processed with the described procedure.

Sample with descriptions originally shorter than 200 bytes	**Number of coded Companies**	**Recall rate %**
Unique codes before deletion of redundant segments	12,486	41.6
Unique codes after deletion of redundant segments	18,772	62.4
Sample with descriptions originally longer than 200 bytes		
Unique codes before deletion of redundant segments	0	0
Unique codes after deletion of redundant segments	11,844	39.5

Table 1. *Coding rates obtained before and after deletion of redundant segments*

As it can be seen, the deletion of redundant segments had a positive impact on the recall rate (20% higher), providing results that, on the first sub-sample, are even higher than those obtained in previous experiences. For this reason, it was decided not to delete redundant words from this sub-sample: the gain in recall rate obtained after the words deletion would have not balanced the lost in clarity of descriptions that is fundamental to make the quality analysis of the codes assigned by ACTR.

Sample with descriptions originally longer than 200 bytes	**Number of coded companies**	**Recall rate %**
Unique codes before deletion of redundant words	11,844	39.5
Unique codes after deletion of redundant words	13,917	46.4

Table 2. *Results obtained on the second sub-sample after the deletion of redundant words.*

The above *Table.2* shows that also in this case, the deletion step (redundant words) had a positive impact on the recall rate (7% higher), even if it did not allow to obtain such a high value as that of the first sub-sample because the great part of redundant words was already contained in segments previously deleted.

A last analysis was made to check the number of companies that received more than one ATECO code: only 3,780 coded companies out of 32,689 (11.6%) obtained multiple codes. This would speed up the following step aimed at identifying the prevalent code.

In conclusion, the average recall rate on the complete sample was 54.4% (average between 62.4% of shorter descriptions and 46.4% on longer ones), which was considered absolutely satisfactory by the responsible of the ISTAT Enterprises Register. This procedure will be soon optimised from a technical point of view, to be used for the entire universe of enterprises and for possible further samples of texts with a similar problematic nature.

References

Appel M. and Hellerman E. (1983). Census Bureau experience with Automated Industry and Occupation Coding. In American Statistical Association*, Proceedings of Section on Survey Research Methods*, pages 32-40.

Bolasco S. (1999). *L'analisi multidimensionale dei dati*. Roma, Carocci ed., pages 179-248.

De Angelis R., Macchia S., Mazza L. (2000). Applicazioni sperimentali della codifica automatica: analisi di qualità e confronto con la codifica manuale. *Rivista di statistica ufficiale - Quaderni di ricerca Istat n. 1/2000.*

Hellermann E. (1982). Overview of the Hellerman I&O Coding System. Internal document, US Bureau of the Census, Washington.

Kalpic D. (1994). Automated coding of census data. *Journal of Official Statistics,* Vol. 10: 449-463.

Knaus R. (1987). Methods and problems in coding natural language survey data. *Journal of Official Statistics*, Vol. 1: 45-67.

Lyberg L. and Dean P. (1992): Automated Coding of Survey Responses: an international review. In Conference of European Statisticians, Work session on Statistical Data Editing, Washington DC.

Macchia S. and Mastroluca S. (2004): The automatic coding process in the 2001 Italian General Population Census: efficacy and quality. In *Proceedings of Q2004 International Conference on Quality in Official Statistics*, Mainz.

Taltac2. (2006). *Guida per gli utenti*.

Tourigny J. Y. and Moloney J. (1995). The 1991 Canadian Census of Population experience with automated coding. In *United Nations Statistical Commission, Statistical Data Editing, 2*.

Wenzowski M. J. (1988). ACTR – A Generalised Automated Coding System. *Survey Methodology*, Vol. 14: 299-308.