

The automatic coding of Economic Activities descriptions for Web users

Cecilia Colasanti, Stefania Macchia, Paola Vicari
Istat, cecolasa@Istat.it, macchia@Istat.it, vicari@Istat.it

Abstract

ACTR (Automatic Coding by Text Recognition) system was introduced in Istat in '90s to code data collected in statistical surveys through a batch process, producing good results. ACTR was used in different contexts and with different classifications. Being a generalised system, to be used it is necessary to build up the informative base for each classification; as larger the dictionary is as better the results are.

In a lot of surveys the economic activity is one of the most important collected variables, its classification is very complex and its structure has several hierarchical levels. The coding activity performed manually was so heavy that it was decided to build an automatic coding application which was widely used with very good results both from the quantitative and the qualitative point of view.

With the last revision the new classification (Nace rev.2) is deeply different from the previous one, so updating the coding application was a very heavy task.

On the other hand, the classification is more suitable to the present economic context and is unique for all the users. For this reasons it was decided to implement a new tool to support the Istat Web site users in identifying the code corresponding to the activities performed by their enterprises. This tool integrates the potentialities of ACTR in textual matching with the web technologies, guaranteeing a more performing solution than simply navigating in the official classification manual.

Keywords: automatic coding, classifications, web site

1. The textual descriptions coding in Istat

The automation of the coding process of textual variables collected in several statistical surveys has been introduced in Istat since the early '90s and it produced good results both in terms of quality of data and of savings of time and human resources.

A generalised system was selected for this purpose: ACTR (Automatic Coding by Text Recognition), developed by Statistics Canada which processes in batch data sets of textual descriptions. It is independent from the language and the classification, so its customisation according to both these aspects must be performed by the users who are charged with the construction of the informative bases corresponding to each classification to be used.

This is the heaviest job and consists in restructuring the descriptions of the classification manuals so as to make them short, clear, unambiguous and as close as possible to the way respondents are used to speak and in including in the informative

base synonymous and pre-coded descriptions given by respondents to previous surveys which detected the phenomenon.

Concerning the text processing, 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*. According to a proper measure of similarity between the texts to be coded and descriptions of the reference file and depending on some user-defined threshold parameters defining the range of acceptance, the system produces the following possible results:

- 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.

In Istat several applications concerning different classifications (Economic Activities, Occupation, Education level, Country of birth/Nationality, Municipalities, Causes of deaths) have been implemented and used in different surveys and, some of them, for the 2001 Census. The results obtained were very satisfactory in terms of *recall rate* (percentage of coded texts on the total of texts to be coded) and of *precision rates* (percentage of correct codes on the total of automated assigned codes), which represents an enhancement of quality of data with respect to manual coding, which did not guarantee the standardisation of the process.

Until now, ACTR was used to code data collected in statistical surveys, which is the aim it was designed for, but in the recent period a new need came out concerning the Economic Activities classification which has been deeply revised in 2007: supporting Istat web site users in identifying the code corresponding to the activity performed by their enterprises.

2. The Ateco coding application

Ateco is the national version of the Nace, the European classification of the economic activities; Ateco is used in Istat in all the business surveys and in several household ones. For this reason Ateco was one of the first coding applications developed and tested.

Concerning the Ateco classification, the first application was built for the 1991 release, which had 5 hierarchical levels (Sections, Divisions, Groups, Classes and Categories), corresponding to a total of 1,668 definitions.

In that occasion, the first decision was to include in the dictionary the descriptions corresponding to all the hierarchical levels, so as to be able to assign codes both to generic responses (corresponding to the highest levels) and to specific ones (corresponding to the lowest levels).

As already mentioned, the construction of the informative base corresponding to the classification to be used is the heaviest job the statistician has to carry out when uses a generalised software coding system, like ACTR.

The first steps, aimed at re-elaborating the official manual in order to simplify descriptions, defining synonymous and eliminating exception clauses, led to a dictionary of 6,783 texts. Then a sample of empirical responses collected in the quality survey of the 1991 Population Census was used in order to test the system performances and to integrate the informative base with the subset of them not automatically coded due to a lack of the dictionary. After this activity, the dictionary grew up to 8,860 texts. Afterwards, a second different file was used to test and enrich the coding application, concerning the Intermediate Industry Census (Long Form survey), which allowed to add 9,288 empirical responses.

Another informative source used to implement the coding application was the classification of products (PRODCOM) which lists, for each economic activity recognised by the Ateco classification, all the correspondent products. As made for the Economic Activities classification, the enrichment of the dictionary followed a step aimed at re-elaborating the official PRODCOM manual. Following this job, more than 500 new texts were added in the Ateco dictionary.

Since then, the Ateco coding application was used for several surveys, obtaining the results shown in the following table. As it can be seen, the automated coding results have always been higher in business surveys than in households or individuals surveys. This is due to the fact that the concept of economic activity is closer to respondents of the first type of surveys than to the latter one.

Table 1 - Economic activities coding application results

	Recall rate	Precision rate
I Labour Force Pilot survey	43.5	85.0
I 2001 Population Census Pilot survey	51.2	93.7
II 2001 Population Census Pilot survey	51.9	90.0
2001 Population Census (for Institutional Households forms)	53.6	92.3
2001 Industry Census	80.7	-

Also the texts collected in these surveys were used to enrich the coding application, which led to a dictionary of 27,306 descriptions.

Then, in 2003, an updating of the coding application was made to incorporate the changes in Ateco 2002, but, in that case, Ateco 2002 was not very different from Ateco 1991.

3. The present economic activities classification and its ACTR coding application

The present economic activities classification – Ateco 2007 – is the national version of Nace Rev. 2, the European economic activities classification. Nace Rev. 2 is the result of a complex revision activity at international level. The revision involved two aspects: a) a process of convergence among the main economic activities

classifications: Isic, Nace and Naics (the classification adopted by the North-American countries); b) the need of a classification that reflects the changes in the present economy.

In the new classification new concepts have been introduced, and new activities have been created to reflect different forms of production and emerging new activities.

In order to satisfy all these needs, the detail of the classification has substantially increased (from 514 to 615 classes and from 883 to 918 categories for the national version).

In order to have an idea of the impact of changes on the official statistics due to the implementation of Nace Rev. 2, the four digit codes that split in two or more new codes are around the 45 per cent.

The new economic activities classification adopts the same classification principles of the previous one, therefore the way to classify the enterprises doesn't change.

The update of the ACTR application with the new economic activities classification was a long and complex process because the descriptions of the *reference* file had to be analysed and the *parsing* rules managing synonymous had to be confirmed, because not all of them would have been coherent with the new classification.

This process was made of different steps and problems:

1. only a part of the old classification at five digit level (around the 65 per cent) directly translated in the new one. The other part had to be checked description by description,
2. since the classification was very different, some descriptions have been completely re-examined; in some cases it was necessary to divide old descriptions (e.g.: "Repair and installation of pumps") because a part is now in a code (Repair, group 33.1) and the other part is in a different code (Installation, group 33.2),
3. completely new activities were introduced,
4. it was necessary to delete some old descriptions because obsolete (281 texts).

Moreover it was necessary to check the old descriptions because the new classification is more detailed than the previous one; as a consequence it can happen that an old description could be split in two or more different codes.

After the complete revision, the amount of the new texts introduced in the dictionary was around 3,000; the texts deleted were 281; the revisions in the dictionary were around 800. The changes in the classification were so deep that it was necessary to introduce more than 200 revisions in the *parsing*.

Since the new classification was more complex than the previous one, it was decided to carry on some special surveys on specific fields (ICT, R&D, Professional, scientific and technical activities; Office administrative and support activities). The descriptions collected in these surveys were used to enrich the dictionary.

In table 2, it is possible to see the dimensions of the economic activities dictionaries corresponding to the different classification releases.

Table 2 - Dimensions of the dictionaries of economic activities application

	Texts in the dictionary
Ateco '91	27,306
Ateco 2002	30,745
Ateco 2007	33,587

Different types of quality tests were conducted to verify the performances of the application just implemented; these tests confirmed very good results both for *recall rate* and *precision rate* (see Ferrillo, Macchia, Vicari, 2008).

In the occasion of this revision – Nace Rev.2 – it was clear that the users need as many tools as possible in order to apply the new classification. As a consequence, it was natural to think to adapt the already realised ACTR for the website.

The users of this technology are numerous: not only the Chambers of Commerce and the Statistics offices but also the private citizens who have to start a new activity or, more simply, have to declare their code in order to pay taxes.

It is important to remember that Ateco 2007 is used by all the administrative sources in Italy and that it is the first economic activities classification to be unique for Istat and for all the administrative sources.

For this purpose, a new tool has been developed which integrates the potentialities of ACTR in textual matching with the web technologies, guaranteeing a more performing solution than simply navigating in the official classification manual.

4. ACTR on Web

The application allows the user to describe his economic activity with a free text and to identify his Ateco code, through the textual matching based on the ACTR algorithm.

In addition, the management of the users' queries has been implemented with the aim of:

1. monitoring the performance of the application in terms of *recall* and *precision* rates;
2. using the not coded descriptions to enrich the coding dictionary.

4.1 The adaptation of ACTR batch coding application to Web users needs

From the content point of view, there would be no difference between the ACTR batch application and the Web one, because in both cases an economic description is given with the purpose of obtaining the corresponding classification code.

On the contrary, from a technical point of view, an ad hoc architecture was designed to manage two different IT environments: the Web Server of the Istat Web site and the Windows XP Server where ACTR runs (see par. 4.2).

On the other hand, some important differences have been introduced in the Web application as regards to the batch one. This was due to satisfy two particular Web users needs:

- while in a batch process the main aim is that of maximising the number of unique codes assigned to each processed description, for a Web user it could be more useful to have at disposal a list of definitions, with the corresponding codes, among which selecting the most pertaining one;
- in the activity of coding data collected in statistical surveys, it is useful to identify not only the maximum detail level codes but also codes at higher hierarchical levels conforming to the dissemination policies; on the contrary, a Web user necessarily needs to identify the complete code corresponding to his economic activity.

For all these reasons, the Web application has been implemented with the following characteristics, different from the batch procedure:

- the user-defined threshold parameters, defining the range of acceptance of the indicator which measures the similarity between the description to be coded and those of the dictionary, were modified so as to enhance the possibility that, in

case of *indirect match*, the system produces a *Multiple* results instead of a *Unique* one;

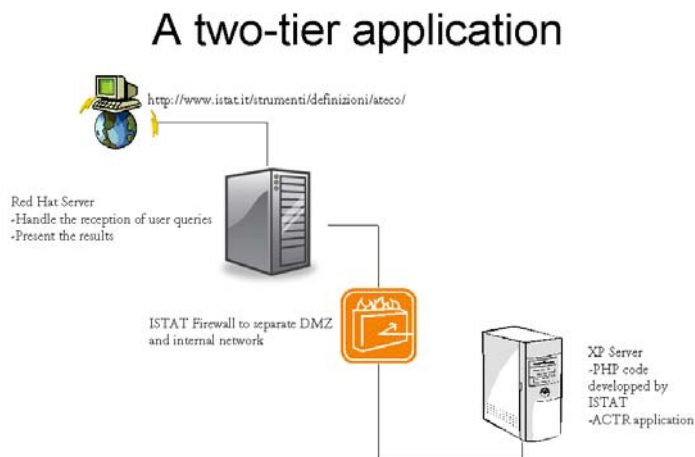
- in case of *Multiple* results, the maximum number of proposed descriptions has been raised at 7, instead of 5, like in the batch application;
- the used informative base is not the complete dictionary, like in the batch application, but a subset of it containing only descriptions corresponding to the maximum detail level codes; this implies that, if the user provides a generic description of his activity, proper error messages suggest how to describe it in a more correct way.

In addition, another function has been implemented: the list of users queries is stored and sent with a weekly periodicity to the classification experts, with the purpose of using them as a precious source of information to enrich and update the coding application.

4.2 The Web architecture

As already said, ACTR was designed to manage descriptions collected in statistical surveys in a batch process, so it was necessary to design an architecture which included it in a web environment.

The web interface to ACTR is a two-tier application.



The presentation layer is hosted on a Red Hat Linux server and is composed of two sections: one handling the reception of user queries and the other one presenting the results. In the first module the user is allowed to type a string which is limited to 200 characters in length, describing his economic activity. The second module then provides the output in the form of a selectable list of matching activity descriptions (up to 7 items). The user must select the item that best describes his economic activity and confirm the selection. If no matching activity is found, the ACTR system raises a warning message which is handled by the web application, suggesting the user to provide a better or more detailed description. Online help pages for the application are also stored on the presentation server.



The application layer is hosted on a Windows XP server where several software modules are deployed:

- The PHP code used to parse user queries from the HTML pages in the presentation layer and forwarding them to the ACTR software. Queries have been serialized due to the ACTR not accepting multiple requests in parallel.
- The ACTR application itself, which processes the query and returns an output which may fall in one of the above mentioned cases (Unique, Multiple or Failed match).
- The PHP processing engine used to format the output and interface the web server providing the end user with the result of the query.

Before the official launch day, several load tests have been conducted on the application using the JMeter performance testing tool. Tests have been made simulating a number from 10 to 1000 contemporaneous queries on the ACTR web application. The think time is from one and half second (10 contemporaneous accesses) to one and half minutes (1000 contemporaneous accesses).

The data shown in table 3 concern:

- minimum, medium, maximum think time of a statistical sample;
- the think time standard deviation;
- the throughput, that is the data quantity transmitted in one second;
- the percentage of error in collecting data.

The think time and standard deviation are expressed in milliseconds.

Table 3 – Load tests of the ACTR on Web application

N. of contemporaneous processes	10	50	100	200	300	500	600	700	1000
Maximum think time	2997	15337	32110	64821	90262	163054	184324	208558	210760
Minimum think time	289	322	407	428	304	377	349	395	0
Medium think time	1591	7537	17034	31184	45053	85746	93842	108970	125794
Standard deviation	868	4374	9298	18893	26095	45584	52258	59689	66104
Throughput	3.3/sec	3.2/sec	3.0/sec	3.0/sec	3.3/sec	3.1/sec	3.2/sec	3.3/sec	4.7/sec
Percentage of error	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	4.29%	31.10%

Subsequently, a user access monitoring system has been developed using Webalizer, as well as the function used to periodically (currently once a week) dump the set of processed queries and automatically forward this to the classification experts, who enrich and update the ACTR informative base. A scheduled procedure to automatically handle updates to the ACTR application has been implemented.

4.3. The WEB application results

The new tool was available on the Istat website from the 26th of May 2008. The ACTR web application had a big success; in fact, more than eight months passed and the queries continue to be more than 10,000 a week.

At the same time the queries analysis suggested a series of improvements in the dictionary and in the *parsing* rules.

Table 4 - Number of queries of the ACTR on web application

Date	N. queries	Date	N. queries
09/06/08	9,225	29/09/08	13,091
16/06/08	6,290	06/10/08	11,268
20/06/08	10,386	13/10/08	10,000
27/06/08	10,327	20/10/08	10,075
07/07/08	10,535	27/10/08	11,568
14/07/08	10,925	03/11/08	12,077
21/07/08	10,304	10/11/08	9,806
28/07/08	8,813	17/11/08	10,302
04/08/08	8,044	24/11/08	10,738
11/08/08	5,514	01/12/08	10,900
18/08/08	1,803	09/12/08	9,928
25/08/08	3,069	15/12/08	8,135
01/09/08	4,579	22/12/08	9,927
08/09/08	8,632	29/12/08	4,204
15/09/08	9,233	05/01/09	6,178
22/09/08	8,690	12/01/09	12,971

In the meantime the availability of this tool allowed to save a substantial amount of time because, before its implementation, the Ateco experts in Istat were called very often to solve problems connected to the classification, while, since the web application is available, the questions to the e-mail dedicated to the new Ateco diminished.

Unfortunately the users don't read carefully the suggestions in order to obtain the correct result, moreover they make a lot of spelling mistakes. A large part of failed results is due to: a) the spelling mistakes, b) users that write numbers of code instead of descriptions of their activities.

For this reason a first improvement of application was made, concerning the inhibition from writing numbers, while the possibility of reducing spelling mistakes through a spelling corrector, which should process the descriptions before submitting them to ACTR, is under study.

Conclusions and perspectives

The good results obtained with the ACTR on web application let think that this function can be considered a pivot experience which could be extended to other classifications, as, from a technical point of view, the implemented architecture can be easily adapted to other informative bases.

In addition, as the users queries are a rich source of information to enrich the coding dictionary, the hypothesis of using specific software for textual analysis for this purpose is under study. The aim is providing a support to the analysis of failed matches through a specific software. In particular, the selected system should make the analysis of the not coded descriptions easier by:

- identifying the descriptions which contain words or sequences of words which can be pertaining in order to define the economic activity;
- aggregating descriptions with similar contents, according to the previous phase;
- selecting the descriptions to be submitted to experts according to their frequency of occurrence which should be higher then a pre-defined value.

A possible software system which could be used for this purpose is Taltac (*Trattamento Automatico Lessicale e Testuale per l'Analisi del Contenuto*), because it provides a series of functions to carry out the above mentioned tasks.

References

- Bolasco S. (1999) *L'analisi multidimensionale dei dati*, Carocci ed., Roma
- Eurostat (2007) NACE Rev. 2. Introductory Guidelines, division Statistical governance, quality and evaluation
- Eurostat (2006) Regulation (EC) No 1893/2006 of the European Parliament and of the Council of 20 December 2006, *Official Journal of the European Union*, L 393/1
- Ferrillo A., Macchia S., Vicari P. (2008) *Different quality tests on the automatic coding procedure for the Economic Activities descriptions*, proceedings of Q2008 European Conference on Quality in Official Statistics
- Lyberg L., 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., Murgia M. (2002) *Coding of textual responses: various issues on automated coding and computer assisted coding*, proceedings of JADT 2002: 6es Journées Internationales d'Analyse Statistique des Données Textuelles
- Wenzowski M. J. (1988) *ACTR – A Generalised Automated Coding System*, Survey Methodology, vol. 14, pp 299-308