Integration between automatic coding and statistical analysis of textual data systems

Stefania Macchia, Manuela Murgia, Paola Vicari

Istat, via Cesare Balbo, 16 – 00184 Rome - Italy

Abstract

This paper describes a new experience, made in Istat, about the integration between the automatic coding system, used to code textual variables, and instruments of textual analysis. In particular it describes how ACTR [1] software, specific for coding applications, can be integrated with the software Taltac2, specific for statistical textual analysis, so to use them together to improve the coding results of texts not collected in surveys. In particular, this paper is about the treatment of the economic activity descriptions provided by the Web site users who search for an Ateco code for different reasons (payment of taxes, administrative matters and so on). These descriptions cannot be totally coded by the ACTR system because some of them may use expressions not covered by the application while some others may contain meaningless information. The integrated use of ACTR and Taltac2 will help in identifying these two typologies of descriptions in order to add the first one to the informative base and to discard the other one.

Keywords: Automated coding, textual analysis, Nace

1. Automated coding of economic activities descriptions in Istat

Since the beginning of the 1990s in Istat, economic activities descriptions provided by respondents of statistical surveys have been coded automatically, using a system called ACTR, produced by Statistics Canada. ACTR is a generalised system, which means that it provides the functions to normalize texts and the matching algorithms, but not the informative base regarding the classifications involved, which must be built by the users.

The construction of the application regarding the economic activities classification (realized the first time for Ateco 1991) [2] has been a heavy job because the classification structure is very complex as it consists of many hierarchical levels. In addition, it was necessary to integrate more sources of information to implement a rich informative base (Macchia, 2001).

In recent years, the application has been updated according to the new classification revisions. This was a very hard job, especially for the Ateco 2007 edition, because new concepts have been introduced at the highest level of the classification and new details have been created to take into account different forms of production and emerging new industries (Ferrillo et al., 2008).

It must also be pointed out that the informative base of an automated coding system is not a static database, because it must be continuously updated according to the language used by

[1] ACTR – Automatic Coding by Texts Recognition.

[2] Ateco is the Italian version of Nace; the present Ateco 2007 is the national version of the Nace Rev. 2.

respondents. In the ACTR environment, for instance, the informative base is constituted of a dictionary and a series of parsing files:

- the dictionary contains lists of descriptions associated to the classification codes. For each code it is possible to have *'n'* descriptions corresponding to different ways of describing the concept associated to the code;
- the parsing files contain information which is used by the system for the text normalisation, for instance synonymous of words, or of couple of words, contained in the dictionary.

So, each time the system is used to code data of a survey, the coding results must be analysed to verify whether some failures of the system (errors in assigning codes or not coded cases) can be solved through the enrichment of the informative base, adding some new descriptions in the dictionary or inserting further synonymous in the parsing files.

It can be seen in Fig. 1 how the Ateco dictionary grew in terms of number of descriptions following its updating according to the new classification revisions and its enrichment due to its usage to code data of different surveys.

	Texts in the dictionary
Ateco 1991	27,306
Ateco 2002	30,745
Ateco 2007	34,180

Figure 1: Dimensions of the dictionaries of economic activities application

The results of the Ateco coding application obtained on data of different surveys are shown in Fig. 2. They are measured according to two parameters:

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

	Economic activities coding application	
	Recall rate (%)	*Precision rate (%)*
Intermediate Industry Census	58.8	91.0
Population Census 1991/Quality survey	54.5	85.0
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	–
2008 Eusilc	52.0	–

Figure 2: Economic activities coding application results

These results can be considered satisfactory, even if they are always 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 second ones.

2. New informative sources to be coded through the Ateco coding application

As already mentioned, ACTR has always been used to code descriptions provided by respondents of statistical surveys (as it was planned for this purpose), that means to process in batch short and structured texts. As a matter of fact, responses given in survey questionnaires always contain descriptions very short and expressed according to ad hoc specifications given in the same questionnaire.

Recently, two new needs came out:

- coding descriptions collected by other administrative institutions (Chambers of Commerce - Cciaa);
- providing a new tool for Istat Web site users who needed to identify the Ateco code corresponding to their economic activity.

The first need was due to the necessity of updating the Istat enterprises Register. Knowing that Cciaa descriptions constitute one of the main sources of information able to satisfy this need, it was very important to univocally code them.

The second need depended on the fact that Ateco 2007 is used by all the administrative sources in Italy since it is the first economic activities classification to be unique for Istat and for all the public offices. Therefore there is a great number of users who can be interested in identifying their Ateco code, not only the Chambers of Commerce and the Statistics offices but also private citizens who have to start a new activity or, more simply, have to declare their code in order to pay taxes.

As both these sources (Cciaa and Web file) contain descriptions which are not collected for statistical purposes as surveys (that have to measure an economic phenomenum), they have different characteristics than those processed until now. For this reason it was necessary to define new strategies to integrate the ACTR functions and to guarantee the systematic enrichment of its database.

2.1. The strategy defined to code Chambers of Commerce descriptions

Just a few words will be spent to describe the strategy adopted to optimize the coding results of the Chambers of Commerce descriptions, because this matter has been widely treated in another job (Macchia et al., 2008).

The need of providing a support to ACTR to code the Chambers of Commerce archive depends on the fact that, as said before, this administration does not collect the economic activity descriptions for statistical purposes. Therefore these descriptions don't have the characteristics proper of those collected in surveys, but:

- they are often very long;
- they contain a lot of misspelled words;
- 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 fact that when an entrepreneur describes to the Chambers of Commerce what his company does, he can go deeply in details and also specify other concepts like the company mission, its juridical status, etc.

For all these reasons, it was immediately clear that submitting these descriptions to the automatic coding system without any previous treatment would have not guaranteed acceptable results.

As a matter of fact, the recall rate obtained running ACTR on a subset of these descriptions was 41.6%, which is very low if compared to the performance obtained on data collected in business surveys.

It was necessary to identify redundant information in the Chambers of Commerce texts and to delete them from descriptions in order to make their content suitable for the automated coding. For this purpose a procedure was designed which integrated the potentialities of a software for statistical analysis of textual data, Taltac2 [3] (Bolasco, 2000), with those of ACTR.

In more details, Taltac2 was used to extract graphical forms and repeated segments from both files, Cciaa archive and Ateco dictionary. According to Taltac2 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, a graphical form could be identified as a word. Repeated segments 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 Taltac2 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 Cciaa 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 Cciaa 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 the intersection were considered "typical" of each list and those "typical" of the Cciaa file were deleted from the descriptions.

After this treatment, the descriptions were submitted to ACTR obtaining a recall rate of 62.4% (versus the previous one of 41.6%).

2.2. The implemented strategy for coding Web descriptions

The automated coding of the economic activity through the Web tools follows different principles than the classical batch coding. Therefore, when designing the ACTR Web application, the peculiar needs of the Web users have to be taken into account. They can be synthetically described as follows:

- 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 suitable one;
- when coding data collected in statistical surveys, it is useful to identify not only the codes corresponding to the lowest hierarchical level (complete codes), but also codes at higher levels conforming to the dissemination policies. On the contrary, a Web user exclusively needs to identify the complete code corresponding to his economic activity.

[3] Taltac2 – Trattamento Automatico Lessicale e Testuale per l'Analisi del Contenuto di un Corpus.

For these reasons some important differences have been introduced in the Web application as regards to the batch one. To better understand these differences, some details must be reported concerning the batch application.

The coding activity performed by ACTR 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, with a batch process, 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.

For the Web application it was necessary to design an architecture which included ACTR in a Web environment. The Web interface to ACTR is a two-tier application. The application layer is hosted on a Windows XP server where several software modules are deployed, which parse user queries from the html pages in the presentation layer and forward them to the ACTR software.

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. The second module then provides the output in the form of a selectable list of matching activity descriptions. 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.

In addition, in order to satisfy the Web users' needs, the Web application has been implemented with the following characteristics:

- the user-defined threshold parameters, defining the range of acceptance of the similarity measure 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* result instead of a *unique* one;
- in case of *multiple* result, the maximum number of proposed descriptions has been raised at 7, instead of 5 like in the batch application;
- the informative base underneath the application 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.

The new tool was available on the Istat Website from the end of May 2008. The ACTR Web application had a big success; in fact, as time goes by, the number of queries per week continues to grow, being actually more than 15,000 a week (see Fig. 3).

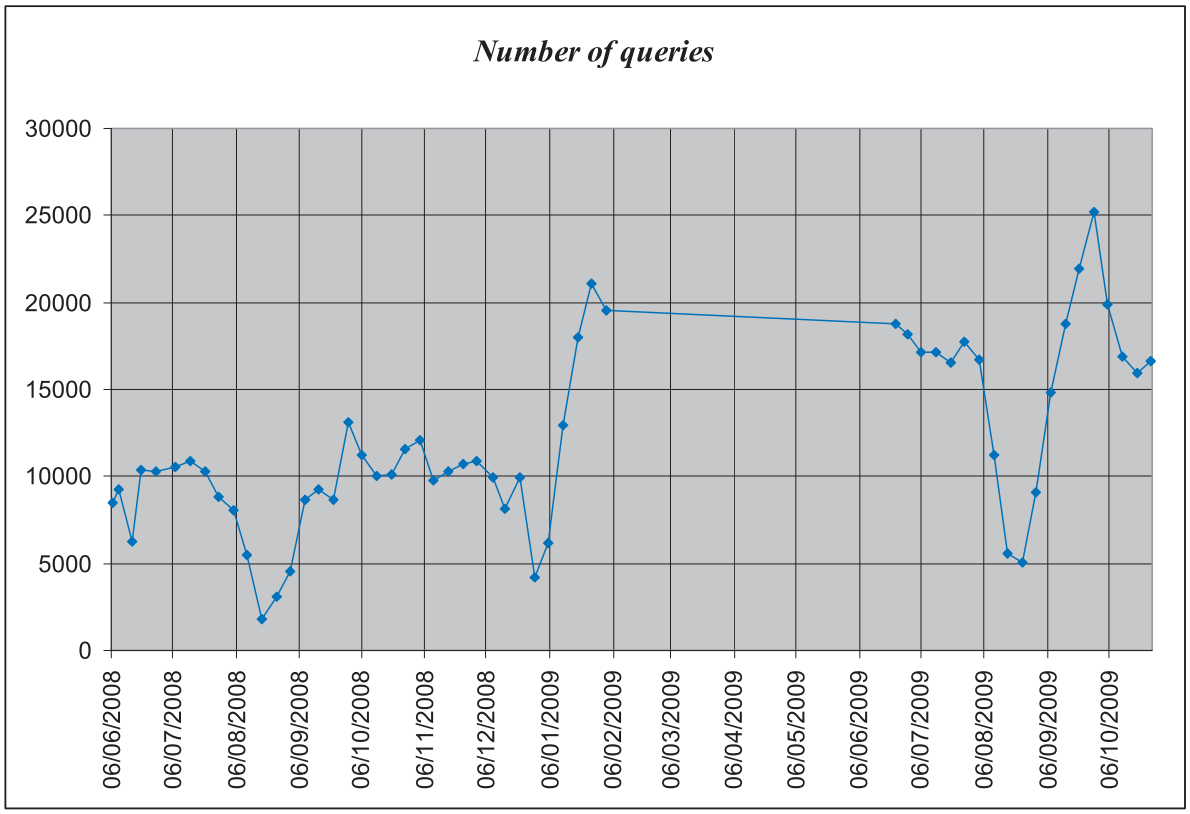

Figure 3: Number of queries of the ACTR on Web application

The great amount of descriptions provided by the users constitute a precious informative source to be used to enrich the ACTR dictionary and improve the application performance.

For this reason, another function has been implemented: the list of these queries is stored and weekly sent to the classification experts who are supported by an automated procedure aimed at identifying, inside the descriptions, the informative content useful to assign a code not intercepted by ACTR due to some dictionary lack (see par. 2.3).

2.3. The strategy to exploit Web users' descriptions to enrich the Ateco coding environment

In order to use at best the information collected by the Web for updating and enriching the Ateco dictionary, we can adopt a strategy quite similar, in its base principles, to that used for the application involving the Cciaa archive. In particular, we can use Taltac2 to identify, from both sources, lists of words and of sequence of words – segments – to be compared through the intersection function and then use the results of this function to improve the Ateco coding environment (see Fig. 4).

Different from the already used strategy is the way results of the intersection function can be used to reach this aim. In fact, if for the Cciaa application we used the intersection to discard redundant information, we can now use it to both discard and keep new information from the Web descriptions. This is justified by the assumption that Web descriptions cannot be totally coded by the Ateco coding application because of two main and opposite reasons: on one side they contain redundant or meaningless information in terms of the Ateco classification, but, on the other side, they contain useful information to be added to the Ateco dictionary to improve the coding results.

Graphically:

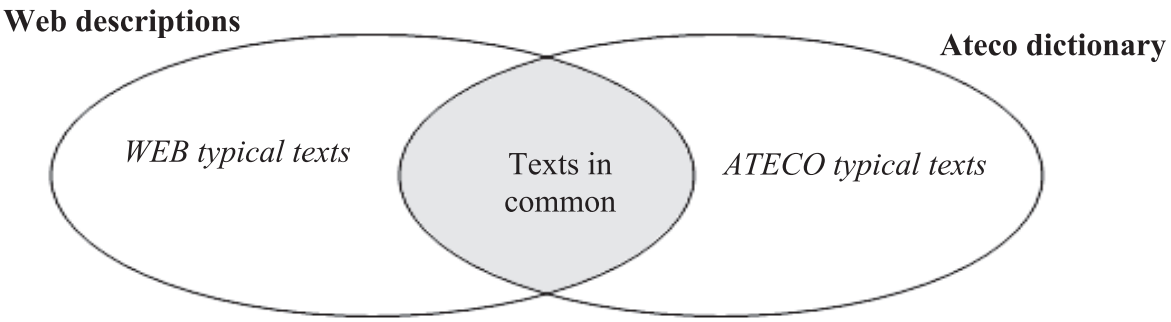

Figure 4: Intersection between Web descriptions list and Ateco dictionary

Web typical descriptions can be divided into two subsets, namely, the subset of "useful information" and that of "not useful information" in terms of the coding application. The following picture shows what just said (see Fig. 5).

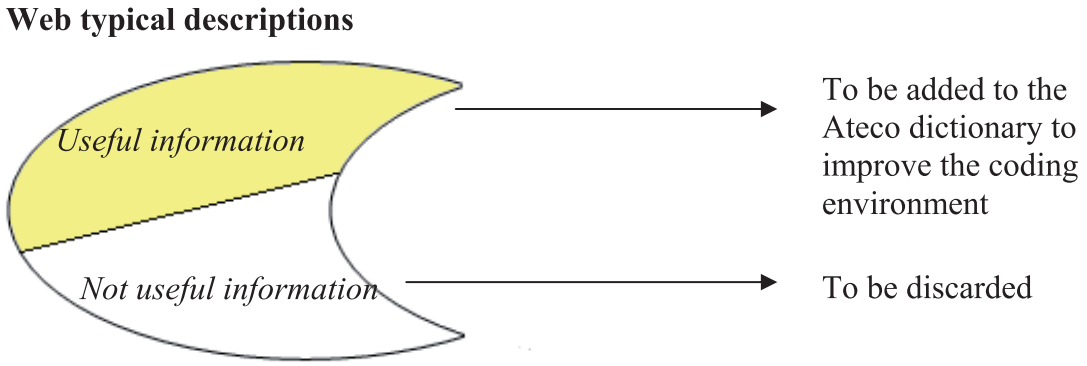

Figure5: Composition of Web descriptions after the intersection with the Ateco dictionary

If the assumption is correct, it is necessary to design a strategy able to define, in an automatic way, to which subset a Web description belongs to. In particular, the strategy should be based on a self-learning system that from an initial manual step, where all the descriptions are manually analysed and assigned to the right subset, will constantly reduce the manual intervention until it becomes just a residual activity and all the job is made automatically.

In order to reduce the manual intervention, we can use the "semantic tagging" function provided by Taltac2 that allows to search for concepts in a text (a textual analysis) after they have been defined in a vocabulary (lexical analysis). In this way it is possible to identify both useful concepts to be added to the Ateco dictionary (vocabulary in Taltac2 terminology) and not useful concepts that will constitute a different vocabulary. By using the semantic tagging, together with a system of intersection functions, we can reach the aim of automatically keep and discard the proper information. In fact, the intersection between new lists of Web queries, periodically downloaded from the Web, and the vocabulary of not useful concepts will enlarge this vocabulary thus reducing the list of not useful words to be manually analysed. Similarly, the

lists of useful concepts, which are determined by subtracting the not useful concepts dictionary from the set of total typical Web descriptions, will be more and more automatically generated, making marginal the manual analysis.

In more details the procedure will work according to the following steps.

- Step one: Intersection between the Web queries list and ACTR dictionary. Typical Web descriptions are taken for the next steps.
- Step two: Manual analysis (semantic tagging) of typical Web descriptions to identify the subset of not useful words → creation of the dictionary of not useful concepts.
- Step three: Identification of the subset of useful words by subtracting the dictionary of not useful words from the entire list of Web descriptions. This subset will be added to the Ateco dictionary.
- Step four: A new list of Web queries is downloaded and intersected with the dictionary of not useful concepts. In this way not useful concepts are automatically eliminated from the subsequent steps.
- Step five: Further intersection between the list coming from step four and the Ateco dictionary. In this way common concepts comprehensive also of the new ones added in the previous step are automatically individuated.

The procedure goes back to step two.

It is important to notice that, by repeating these steps, it seems that manual analysis tends to disappear. This is true if the language variability of the Web respondents is decreasing or at least constant. We know that is not true (Macchia et al., 2003) because the human language is under a continuous transformation and because new economic activities arise as time goes by. This means that a manual analysis step will always be present without lowering down the efficiency of the procedure.

3. Conclusions

This paper highlights the importance of the integration between different software systems of texts treatment. Particularly, it shows how systems that treat textual information according to totally different aims can be used jointly taking advantages just of their different approaches and methodologies on which they are based.

One of these advantages, for Istat, is the possibility of using different sources of information to update its informative bases. This is a very important aspect since till now only information coming from surveys has been used while, with this methodology, information that is not collected according to some standards are immediately suitable for the Istat coding systems.

References

Bolasco S. (1999). *L'analisi multidimensionale dei dati*. Roma: Carocci.

Bolasco S. (2000). TALTAC: un environnement pour l'exploitation de ressources statistiques et linguistiques dans l'analyse textuelle. Un exemple d'application au discours politique. In Rajman M. and Chappelier J.C., editors, *JADT 2000*, Lausanne, 9-11 Mars, tome 2, pp. 342-352.

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.

Eurostat (2007). Nace Rev. 2. Introductory Guidelines, division Statistical governance, quality and evaluation.

Ferrillo A., Macchia S. and Vicari P. (2008). Different quality tests on the automatic coding procedure for the Economic Activities descriptions Q2008 European Conference on Quality in Survey Statistics Roma 08-11/11/08.

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. (2001). 'Integration of sources to build a dictionary for Automated Coding of Industry'. Riunione scientifica del gruppo SIS Classificazione e Analisi dei dati (Palermo, 5-6 luglio 2001).

Macchia S., Murgia M. and Perrone D. (2003). 'The influence of language variability on the enrichment of the coding system dictionary of Occupation'. Convegno intermedio SIS 2003 - Analisi Statistica Multivariata per le Scienze Economico-Sociali, le Scienze Naturali e la Tecnologia (Napoli, 9-11 giugno 2003).

Macchia S., Murgia M. and Talucci V. (2008). Coding the spoken language through the integration of different approaches of textual analysis. In Heiden, S. and Pincemin, B., editors, *JADT 2008*, Lyon, 12-14 Mars, pp. 745-752.

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