How the text mining measures complex phenomena in official statistics

Come il text mining misura fenomeni complessi nelle statistiche ufficiali

Bolasco Sergio, Pavone Pasquale
Dipartimento Memotef Università di Roma "La Sapienza",
sergio.bolasco@uniroma1.it;
Scuola Superiore S.Anna di Pisa, pasquale.pavone@sssup.it

Riassunto:
Il presente lavoro si propone, attraverso strumenti di text mining, di misurare quantitativamente caratteristiche dell'attività quotidiana descritta nei diari individuali dell'indagine Istat sull'Uso del Tempo (TUS). In particolare, vengono studiati fenomeni riguardo la localizzazione delle attività relazionali riconducibili al "comunicare con". Le maggiori potenzialità di un'analisi condotta direttamente su informazioni in linguaggio naturale sono dovuti alla migliore "risoluzione" della misurazione, in quanto l'analisi dei concetti è più flessibile, precisa e accurata di quella basata su codifiche. Ciò migliora la produzione, anche in forma tradizionale, di statistiche ufficiali aprendo nuove prospettive nella valutazione di fenomeni complessi quali sono quelli da misurare nel calcolo dei bilanci del tempo.

Keywords: text mining, information extraction, ETL, linguistic resources

1 Introduction

The applications of textual statistics[1] handling information expressed in natural language (unstructured textual data) in the same way as classical structured (quantitative and / or categorical) data have increased in recent years. The greatest potential for the direct analysis of textual information depends on the better "resolution" of the measurement, because analyses based on concepts are more flexible, precise and accurate than those conducted through keywords or coding. This paper aims, through lexical and textual analysis, to measure quantitatively the characteristics of the everyday activities of individuals described in the diaries of the Istat Time Use Survey (TUS). The survey aims to establish a free text daily diary to describe the activities carried out during the day. For the first time in the survey of 2002-2003 (Romano, 2007), Istat has agreed to acquire the full text of individual diaries, thereby providing an archive of great importance, not only in size (over 50,000 diaries, equivalent to 16,000 pages of text) but especially in content, clearing the way for numerous developments. The limits resulting from the ambiguity of natural language are largely resolved at the start of the treatment, by appropriate tools for this type of data[2]. Each application of the models and

[1] Lebart *et al.* (1998); Aureli & Bolasco (2004), Dulli et al. (2004); www.jadt.org : online JADT Proceedings, 2000--2010.

[2] There are several software for the natural language processing and automatic analysis of texts, which differ according to the type of analysis to be conducted. In this study, considering the statistical purpose of the analysis, we used the TaLTaC2 software, which

techniques of text mining is characterized by strong multidisciplinary integration involving statistics, computer science and linguistics in equal measures.
We will illustrate the procedure adopted to automatically extract information from the non-structured text of the diaries, record them in a structured way (as a Boolean or frequency) in a matrix of individual data and then cross the variables generated by the textual analysis with the categorical characteristics of individuals in order to produce official statistics. In particular, phenomena concerning the intensity of social interaction – that can be related to the "who you are communicating with " – and the different locations of this type of activity are regarded here. The study is conducted by considering individuals as units of analysis, where the diary of a day is regarded as a single context (see Bolasco *et al.* 2007).

2 Definition of the resource "place" and relational activities

The places of individual daily activities described in the diaries are captured through a general model presented in our previous work (Bolasco and Pavone 2010). This model allows us to identify a wide variety of adverbial locutions indicating place, based on the linguistic structure of a prepositional syntagm, as follows:

PREPOSITION (ADJECTIVE) SUBSTANTIVE (ADJECTIVE)

where the adjectives are in brackets because their presence is optional and / or repeated. For example, starting from the primary term "home", the model recognizes sequences such as "at home", "my second home", "*nella mia casa futura* (in my future home)". The whole syntagm may be repeated several times, with the adjectival function of the first noun, for example: <on the seat | of the car>; <*alla festa* | *di compleanno* | *di un amico* (at the Birthday Party | of a friend)> (Table 1).

Table 1 - Examples of expressions of place from the model

PREP	**POSS**	**AGG**	**SOST**	*PREP*	*POSS*	*AGG*	*SOST*
a			casa				
davanti a			casa				
nella	mia	seconda	casa				
nella	mia		casa			futura	
a			casa		mia		
a			casa	di	mia		madre
a			casa	del			vicino
vicino (a)			casa				

The model, based on a hybrid system consisting of rules and dictionaries, is done in two stages. An initial exploratory phase of training, used to develop the basic components of the model and a second application stage to detect their actualization in the corpus of the TUS. The application of the model is divided into: i) the launch of a query, consisting in a single regular expression composed of 39 sequences in the OR for a total of over 150 relations (rules) between the concepts expressed by 16 semantic dictionaries able to extract locutions, ii) the evaluation of the entities found, iii) the calculation of the occurrences of each term, for a total number of occurrences (redundant) equal to

stands for Automatic Treatment for Lexical and Textual Analysis of the Contents of a Corpus, developed from research at the University of Rome "La Sapienza " (Bolasco 2010; www.taltac.it)

22% of the entire corpus. These sequences, as space-time modifiers, were divided ex-post into sub-thematic classes, distinguishing between activities "at home" (his own, with relatives, friends or others) from activities "away" related to movement (walking, cycling, on public transport, ...) or activities related to roles-places (at the hairdressers, newsagents, ...) or linked to different environments / sites (in the office, at the bank, in a shop, among the market stalls, ...).

Relational activities are identified by studying a sequence of two "components", interlaced by the keyword <*con*> (in some cases <*a*>). In particular, the first component of the verbal type, limited to verbs expressing communication "talk / communicate with", and "call / tell". These concepts have been captured even when expressed in similar terms (phone call, phone) in a compound verb phrase: "make (a) p." or "be (on) the p.". The second component is the "who", ie the actor who is addressed by the speaker. Several classes of actors already defined (parents, spouses, children, grandchildren, grandparents, friends etc.) are used to reconstruct the sequence, even with more complex expressions such as: <*parlo di politica con mia moglie* (I'm talking about politics with my wife)>. For a list of verbs and actors considered, see Bolasco *et al*. (2007).

3 Measuring the characteristic places of relational activities

After having defined the entities and their concepts, created thematic dictionaries, the search in the text was based on the construction of complex queries, using regular expressions, in order to identify the sequences in the diaries that realise these activities in relation to different categories of actors (relatives / friends) in conjunction with the different classes of place identified by the model as described above. In particular, in our case the set of queries takes the following form:

```
"CATSEM(Verb) LAG3 CATSEM(Prep) LAG4 CATSEM(Actor#) LAG8 WH LAG3 CATSEM(Place#) LAG2 |"
```

where CATSEM denotes the classes of: i) verbs of "communicating", ii) actors ("relatives / friends"), iii) prepositions "*con/a/tra/in* (with / to / between / in)", iv) places ("own home / home of other people / other places / means of transport"). The LAG # expresses the maximum number of words in the interval between two operands of the expression and the token <|> denotes the end of the sentence. Some examples of the sequences extracted are shown in Table 2.

Table 2 – Some examples of sequences

raccontato a mia moglie cosa ho fatto oggi WH a casa mia |
litigo con mia sorella WH a letto a casa |
parlo con mio marito WH a casa di amici |
giocato a calcio con mio fratello e i nostri amici WH parco |
parlavo con i miei familiari con l' autoradio accesa WH in macchina |
chiacchierato con gli amici § ho ascoltato la radio WH a casa mia |
gioco con un amichetto § WH a casa della nonna |
chiacchierato con amici e parenti aspettando gli sposi § WH al ristorante |
chiacchiero con degli amici e ascolto musica WH in corriera |

Each query captures an instance, whenever the sequence is present in the diary. The result of the query produces a new variable that measures the presence / absence (or frequency) of the entity for each individual. This new structured information can be placed in connection with the individual a priori information, such as structural variables (age, sex, marital status, education level) to produce traditional statistics.
By applying this model to a sub-sample (10,000 units) of the Istat survey, we obtain a statistic of the type shown in Table 3, corresponding to 18,628 sentences.

Table 3 – Sentences concerning relation activities with parents/friends of the sub-sample by gender, age groups and type of place (percentage values)

Relation activities	Men					Women					Total place
	Age groups					Age groups					
	14-24	25-44	45-64	65+	Total	14-24	25-44	45-64	65+	Total	
with relatives at own home	2.4	9.0	7.6	4.5	*23.5*	2.9	14.7	10.9	6.2	*34.7*	***58.2***
with relatives at home of other people	0.1	0.2	0.1	0.1	*0.4*	0.1	0.2	0.1	0.1	*0.4*	***0.8***
with relatives in other places	0.3	1.8	1.5	0.5	*4.2*	0.5	2.4	1.6	0.7	*5.2*	***9.4***
with relatives on a mean of trasport	0.3	1.6	1.4	0.5	*3.8*	0.5	2.4	1.6	0.7	*5.2*	***9.0***
with friends at own home	0.3	0.5	0.2	0.1	*1.1*	0.5	0.8	0.5	0.4	*2.1*	***3.3***
with friends at home of other people	0.0	0.3	0.1	0.0	*0.5*	0.1	0.3	0.1	0.1	*0.6*	***1.0***
with friends in other places	2.2	3.0	1.7	1.1	*8.0*	1.8	1.7	0.6	0.3	*4.5*	***12.5***
with friends on a mean of trasport	1.1	1.2	0.5	0.4	*3.2*	1.2	1.0	0.2	0.2	*2.7*	***5.9***
Total gender by age	**6.7**	**17.6**	**13.2**	**7.2**	***44.7***	**7.6**	**23.4**	**15.6**	**8.7**	***55.3***	***100.0***

References

Aureli E., Bolasco S. (a cura di) (2004) *Applicazioni di analisi statistica di dati testuali* Casa Editrice Università "La Sapienza", Roma.

Bolasco S. (2010). *Taltac2.10 Sviluppi, esperienze ed elementi essenziali di analisi automatica dei testi*, LED, Milano.

Bolasco S., Canzonetti A., Capo F. M. (2005) *Text mining: uno strumento strategico per imprese e istituzioni*, CISU, Roma.

Bolasco S., D'Avino E., Pavone P. (2007) Analisi dei diari giornalieri con strumenti di statistica testuale e text mining, in: *I tempi della vita quotidiana. Un approccio multidisciplinare all'analisi dell'uso del tempo,* Romano, M. C. (ed.), ISTAT, Roma, 309-340.

Bolasco S., Pavone P. (2010) Automatic Dictionary and Rule-Based Systems for Extracting Information from Text, in: *Data Analysis and Classification* Proceedings of the 6th Conference of the Classification and Data Analysis Group of the Società Italiana di Statistica, Palumbo, F. , Lauro, C. N. , Greenacre, M. (Eds.), Springer, Berlin-Heidelberg, 189-198.

Dulli S., Polpettini P., Trotta M. (2004) *Text mining: teoria e applicazioni,* Franco Angeli, Milano.

Lebart L., Salem A., Berry L. (1998) *Exploring textual data,* Kluwer Academic Publ., Dordrecht.

Romano M. C. (ed.) (2007) *L'uso del tempo* - Indagine multiscopo sulle famiglie "Uso del tempo" - Anni 2002-2003, Collana: Informazioni, n. 2, ISTAT, Roma.