Textual Data Analysis tools for Word Sense Disambiguation

Simona Balbi[1], Agnieszka Stawinoga[2]

Università "FedericoII" di Napoli
[1] simona.balbi@unina.it, [2] agnieszka.stawinoga@unina.it

Abstract

The ambiguity of words is a crucial question when dealing with an automatic analysis of documentary data bases. In Text Mining, Word Sense Disambiguation is the task of giving a particular sense to a term with different meanings both in the case of coincidental and polysemous homographs. In literature, the proposed solutions are mainly based on two elements: some knowledge related to the term and the context in which the term appears. Limiting the related knowledge to grammatical tagging, and to an analysis of collocations, here we focus our attention on identifying the context, in a data driven approach. Our framework is based on Textual Data Analysis and we assume that language and knowledge can be modeled as networks of words and the relations between them. The aim of this paper is to propose an extension of the strategy for building lexical sources in Balbi *et al*. (2012), in order to deal with ambiguous words. The methodological basis is given by the joint use of lexical Correspondence Analysis and Network Analysis. Our idea is investigating the neighborhood of ambiguous terms, with respect to the different latent semantic components, emerging thanks to Correspondence Analysis of a training set of documents, in order to build some rules, useful in solving WSD problems in the entire *corpus*.

Keywords: ambiguity, network analysis, correspondence analysis, homographs.

1. Introduction

In Natural Language Processing, ambiguity is a well-known problem since the first attempts to realize automatic translators. In Text Mining, the task of giving a particular sense to a term with different meanings is still an open research field, under the label of Word Sense Disambiguation (WSD). In literature, the existing solutions are mainly based on two elements: some knowledge related to the term and the context in which the term appears (Kasture and Agrawal, 2012). Supervised, or unsupervised, algorithms integrated with electronic dictionaries are common solutions for dealing with the problem of word-sense disambiguation. These algorithms usually have acceptable performance when meanings are totally different, but they frequently have poor results in the case of finer sense distinctions.

Here we propose a new strategy for performing the Word Sense Disambiguation task, in a semi-supervised approach. The methodological frame is given by a joint use of lexical Correspondence Analysis and Network Textual Analysis, as proposed in (Balbi *et al.*, 2012). Working on a training set of documents, our idea is to partition the neighborhood of ambiguous terms, represented by an ego-network, with respect to the different latent semantic components emerging thanks to Correspondence Analysis. Once identified the *alters* of the ambiguous word, related with a peculiar sense, it is possible to build rules for the sense tagging of the word.

The remainder of the paper is organized as follows. In Section 2, we present a quick overview on some central issues related to ambiguity. Some methodological recalls on the statistical tools we intend to use are in Section 3 (Correspondence Analysis in Sec. 3.1. and Network

Analysis in Sec. 3.2), and subsequently in Section 4 we propose the new disambiguation strategy. In Section 5, we show the effectiveness of our proposal on a real data application. In the concluding remarks (Section 6), we introduce some potential improvement of our proposal, based on the choice of a proper metric for measuring the similarity of the terms.

2. Word Sense Disambigation

Every time the same sequence of characters has more than one sense, we have a problem of ambiguity. Sometimes the meanings are similar (the so called polysemous homographs), sometimes the meanings are totally different (i.e. coincidental, or true, homographs), according to their historical origin in the most common cases.

In literature, one typical example of both coincidental and true homographs is given by the word "bass" (figure 1):

a) "bass" indicates various trimly shaped fishes of both fresh and salt water
b) "bass" indicates:
   i. a low-frequency sound,
   ii. instruments in the bass range,
   iii. a classical male singing voice.

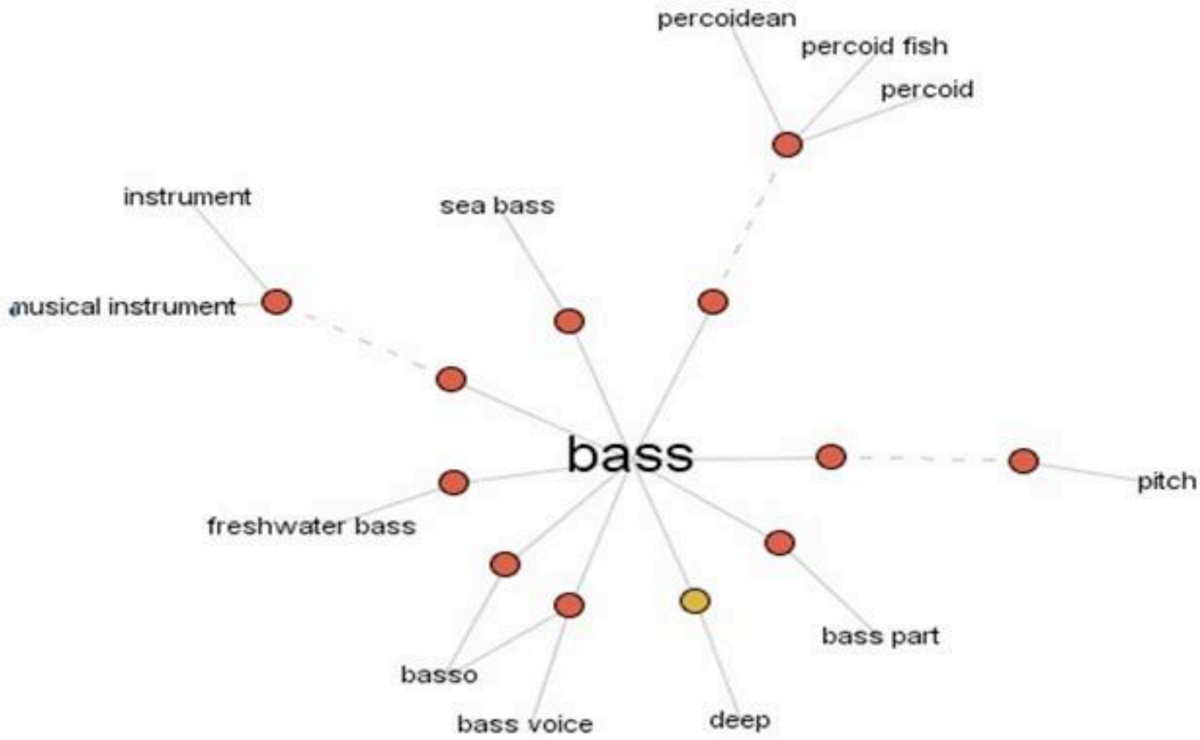

Figure 1. The word BASS and its meaning (by Visual Thesaurus.com)

If we consider a), and b) we have a true homograph, while b) is a case of polysemy.

An ambiguous word can be connected with words belonging to different topics. This circumstance is well-known in the natural language processing field and many algorithms have been developed for dealing with word-sense disambiguation, usually with the help of an internal dictionary, and/or performing supervised, or unsupervised, classification. Those algorithms try to solve the problem in the pre-processing step, while the corpus is transformed in a structured data base. They have acceptable performance when they properly deal with true homographs: the different sense can be identified by considering the context in which the word is used. If we find "bass" in an article of *Sport Fishing Magazine*, we have good reasons for thinking that we have the case a), while if we find *bass* in the program of the Lyric Theater of London, we have something related with the situation b). However, in the case of finer sense distinctions (as in b) i., ii., or iii. in the Lyric Theater program), disambiguation algorithms usually have poor results (Stevenson and Wilks, 2003).

3. The methodological frame

As the first aid in disambiguating a word is given by the context in which the word is used, our attention focuses on the relations among words. Following Popping (2000) in introducing Network Text Analysis, we assume that language and knowledge can be modeled as networks of words and the relations between them. In a previous work, we showed how relations can be manifest or latent, and how both kinds of relations influence the links of a word with other words (Stawinoga and Balbi, 2012). Latent relations are objective functions which are not directly visible, but they can be extracted by means of correspondence analysis. In contrast, manifest relations are the intersubjective relations studied by tools of social network analysis (deNooy, 2003). Therefore we put our contribution in the frame of the statistical analysis of relational data, taking into account the longstanding debate about the proper method for analyzing such a kind of data.

3.1. Correspondence Analysis

Correspondence analysis (CA) is a descriptive multivariate method (Lebart *et al*., 1984) aiming at representing the structural relationships between rows and columns in a two-way table. Although it can be applied on any table, provided that there are nonnegative values, and the data are homogeneous, it is usually performed for analyzing the association structure in a contingency table. CA is a principal axes method, mainly characterized by a symmetric processing of rows and columns; the use of a special weighted Euclidean distance, known as chi-square metrics; simple transition formulas allowing for a simultaneous representation of rows and columns, the so called joint plot (Lebart *et al.*, 1998).

In the case of textual data analysis, it is performed on a lexical table, and it is named Lexical Correspondence Analysis, where usually documents are aggregated with respect to some common characteristics. Dealing with a lexical table **T** (I, J) which cross-tabulates the I documents and the J terms of a corpus, CA is generally performed to identify the latent semantic structure in the corpus and graphically represent the latent lexical relations.

3.2 Network Analysis

Network Analysis (NA) is the analysis of a set of relations among objects. Relations are not the properties of objects, but of systems of objects, in a larger relational system (Scott, 2000). The relational system is represented by graphs. In the graphical representation of a network, objects (vertices) are the points, and relations are drawn as lines connecting pairs of vertices.

NA has found applications in many fields although its developments are often connected with the social sciences, where it is named Social Network Analysis. Attention to NA has extraordinarily increased in recent times, both from a methodological and applicative viewpoint.

(Popping, 2000) defined Network Text Analysis as a method for encoding the relationships between words in a text and constructing a network of the linked words. The assumption is that language and knowledge can be modeled as networks of words and the relations between them. Several ways to derive networks from textual data and an overview of applications is presented in Batagelj *et al*. (2002).

Here we are mainly interested in a particular network, the so called ego-network. An ego-network is a network formed by a focal node ("ego"), all the actors ("alters") connected to that node and all the connections among the alters. According to the graph theory, ego networks define a specific class of graphs, called centered graph (Freeman, 1982). Consider a graph, $\mathcal{G}$, consisting of a set, $\mathcal{W}$, of k vertices and a set, $\mathcal{E}$, of e symmetrical edges linking pairs of points. Now if $k > 2$ and there are $(k – 1)$ edges such that some one point, w^*, is adjacent to all of the others, G is defined a k-star**.** A centered graph, then, is any graph of k points that contains a k-star. Ego-networks are extracted from the whole network and illustrate its local areas. Here we explore the relations of each focal node with its alters but we do not pay attention to the connections among the alters.

4. Our proposal

In a previous paper (Balbi *et al*., 2012), we considered the effectiveness of jointly using Correspondence Analysis and Network Analysis in order to build lexical sources in a peculiar domain (economic and financial statements), by analyzing a small database (*training set*), chosen as highly representative of the vocabulary of a larger corpus.

We use CA for identifying the most "relevant" words in the *training set*, considering relevance with respect to contributions, and then we build the ego-networks for the selected words, as expression of different concepts.

However, when we draw the networks on the first factorial plane, sometimes we have unsatisfying representations due to the differences between manifest and latent relations between words (figure 2). The fact is strictly connected with the problem of ambiguity, as ambiguous words have different relational systems with respect to their different meanings, captured by the latent semantic components underlying the *corpus*.

In this work we propose to improve our strategy, in order to identify the different neighborhoods of an ambiguous word. The objective is to obtain a data driven sense-tagging of some selected, ambiguous terms, in order to refine lexical sources. This procedure can be performed without external dictionaries, by producing different sub-lists of alters, related with the different meanings of the word.

4.1 The disambiguation strategy

Let $\mathcal{W}=\{w_1,\ldots, w_N\}$ denote a set of relevant words identified by the strategy proposed by Balbi *et al.* (2012), on a selected part of the *corpus* to be analysed (*training set*).

Let $\mathcal{W}_A \subset \mathcal{W}$ be a set of ambiguous terms which are individuated, as mentioned above, from the unsatisfying representations of their ego-networks emerged from the first factorial plane of a Correspondence Analysis performed on the training set.

Let K be the cardinality of $\mathcal{W}_A$.

To obtain a data driven sense-tagging of the ambiguous terms in the entire corpus, we propose a semi-supervised strategy, where we consider a small training set manually labeled in order to produce resources useful for the semantic tagging of the ambiguous words of the entire corpus.

```
INPUT: the factorial coordinates (coming from the previous CA) of each $w_i \in \mathcal{W}_A$ and
       the factorial coordinates of the $w_i$'s alters in the ego-network;
for $i$ = 1 to K
STEP 1 Identification of the axes where $w_i$ has a high contribution;
STEP 2 Identification of the alters near $w_i$ on the different identified axes;
STEP 3 Identification of different subsets made up of the term $w_i$ and groups of its alters
       according different factorial planes;
OUTPUT: Different lists of words useful for the sense tagging of the entire corpus.
```

The lists enable the sense tagging of the term w_i in any further analysis performed on documents belonging to the entire *corpus* . It consists in annotating w_i with different labels, according to the presence in a sentence of the alters specific to a list. It can be performed by TALTAC (Bolasco, 2010), using the lexical analysis procedure, by building different lists of the nearest alters, and assigning a proper label to differentiate the meanings of each w_i.

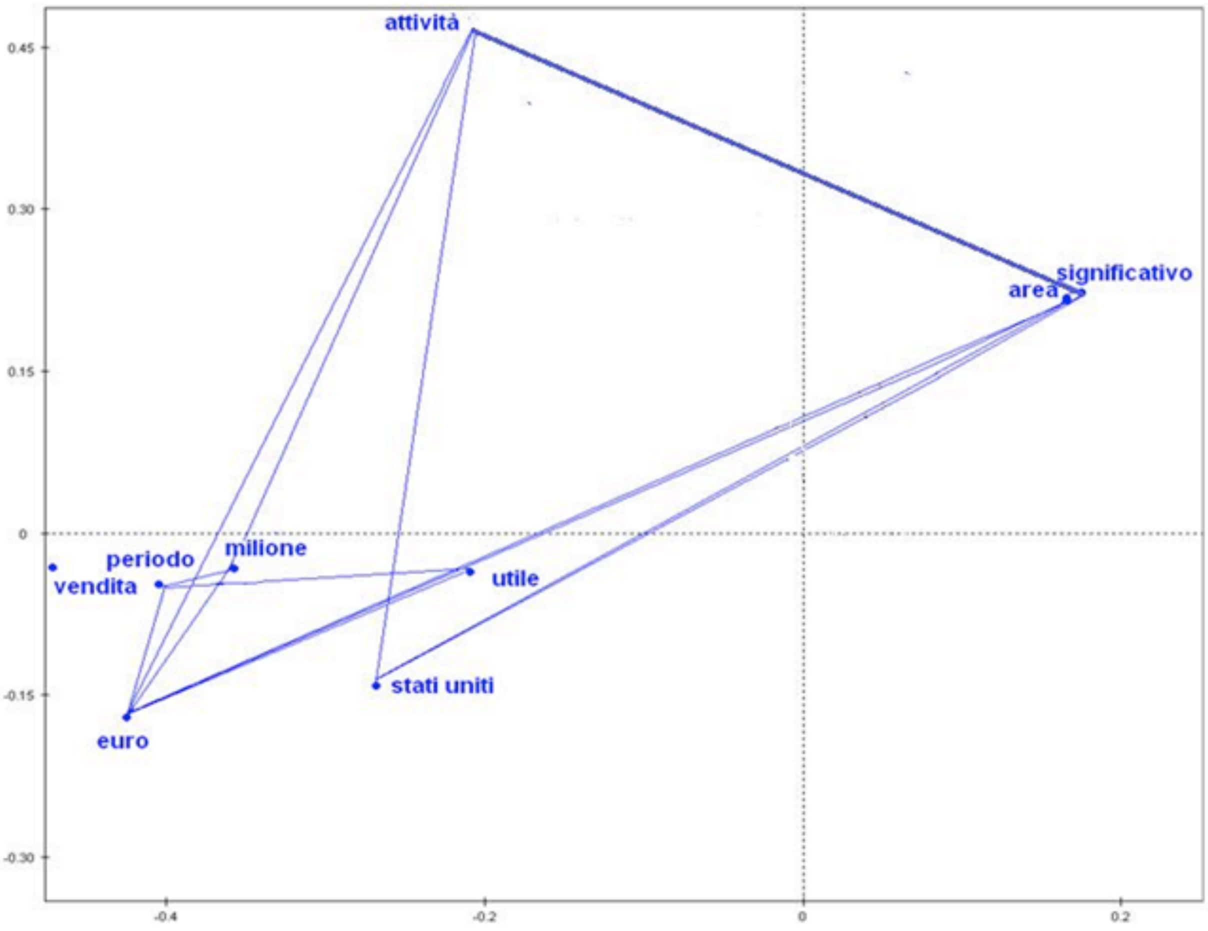

Figure 2. The focal node "area" and its "well represented" alters on the CA First Factorial Plane (from Balbi et al.*, 2012)*

Apart from the tool proposed by Bolasco (2010), it could be possible to improve the proposed strategy and define labels of identified groups of words by considering external knowledge about ambiguous terms.

5. Sense tagging in analyzing Italian management commentaries

In order to build lexical sources for the automatic analysis of Italian management commentaries, we have chosen the documents written by Luxottica group, as training set. Luxottica group, the world leader of luxury and sports eyewear is listed both on Italian and U.S. stock markets. In the U.S. market, the law thoroughly explains all information that the management commentary must include, while in Italy there are no specific rules. For this reason we considered Luxottica management commentary a good training set. The reference year of our analysis is 2009. For a detailed analysis, see (Balbi *et al.*, 2012).

Here we focus our attention on some non trivial cases of ambiguity, emerging during the analysis.

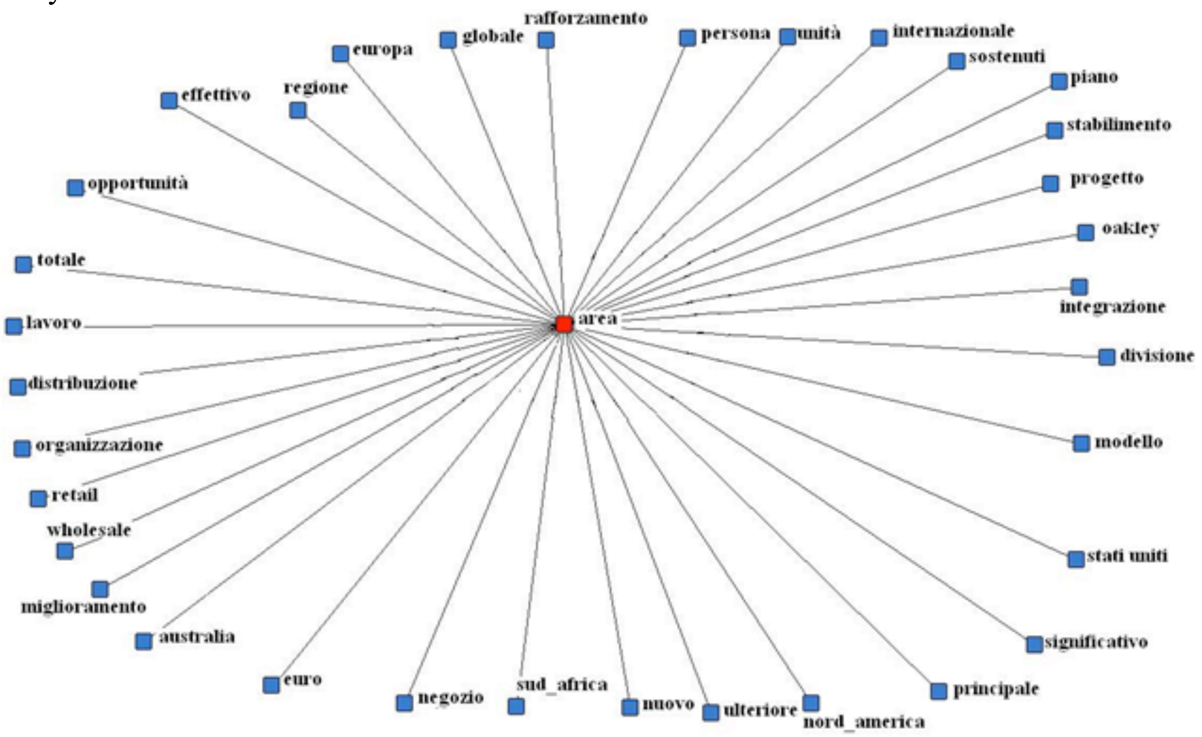

Figure 3. The ego network of "area"

5.1 The data

The textual database is Luxottica's management commentary which is composed of 44 sections and 22,621 tokens. The pre-processing procedure has been performed using the TalTac 2.0 software. During this phase the text is firstly normalized and cleaned up from empty words (conjunction, articles, adverbs, etc.). Then the corpus is grammatically tagged to select the lexical-part-of-speech (nouns, adjectives). Following this procedure, we manage to get the lexical matrix of size (44×375).

A CA on the lexical table is performed, using the software SPAD. Therefore, from a statistical viewpoint, we decide to study the basic latent structure which explains the 75% of the association in the table. From a practical view point this leads to the analysis of the first 10 axes where each one illustrates a percentage of the association higher than the average equal to 2.5%.

Here we go in depth into exploring the latent and manifest relations of the term "area", which is one of the most relevant words, identified by Correspondence Analysis. The joint use of

CA and NA individuates the word “area” as one of the terms where the problem of ambiguity should be investigated deeply. Figure 3 illustrates the ego network which represents the manifest relations among the term “area” and its alters. This ego network is extracted from the whole network where two terms are linked if the strength association between them is higher than 0.3, measured by Jaccard similarity index. In this way weak relations are avoided. It is worth noting that the manifest relations show the terms with which the word “area” occurs frequently enough to build strong relations but unfortunately they are not able to emphasize the various meanings this term can have in the *corpus*, e.g. referred to marketing (distribution chanels, geographical markets) or organisation (models, plans).

In order to investigate them the analysis of the different CA factorial axes is necessary. The first factorial plane explains the 17.1% with the first axis and 10.8 with the second axis of the total inertia. It reveals the use of the term “area” in the sense of **distribution channels** (Figure 4). We find “area” as reference markets (*Sud Africa*, *Europa*, *Australia*, *Nord America*). Only by the analysis of the further factorial planes do we manage to capture the whole meaning of the term “area”. Figure 5 shows the fourth factorial plane, based on the axes 7 (4.8%) and 8 (3.9%) where we can individuate the meaning of “area” in the sense of **company structure** (*piano*, *organizzazione*, *modello*, *effettivo*, *miglioramento*). By analysing the axes 9 (3.3%) and 10 (2.2%) the **financial aspects** related to the term “area” (*euro*) emerge. Note that "area" has uninteresting contributions to axes 3 (10.4%), 4 (8.5%), 5 (7.1%), 6 (6.3%).

The sense tagging of the term "area" in the entire *corpus* of the management commentaries of Italian companies is now possible by the obtained results. "Area" has to be tagged, according to the presence of the identified alters in each sentence, by TALTAC. We can identified an "area_mercato", when "area" appears together with the alters related to the **distribution channel**, i. e. with *Sud Africa*, *Australia*, *Europa*, *Nord America*, *wholesale*, *retail*, *negozio*. We identify an "area_struttura", when "area" appears together with the alters related to the **company structure**, i. e. with *piano*, *organizzazione*, *modello*, *effettivo*, *miglioramento,* and an "area_finanziaria", when "area" appears together with the alters related to the **financial aspects**, i. e. mainly with *euro*.

This practical example illustrates the effectiveness of the proposed strategy. The partition of the egonet of a selected (ambiguous) term in different components, by taking into account the alters which lie near the selected term on the different axes, gives the possibility to capture the different meanings of a word and deal with the problem of ambiguity. As a matter of fact, the procedure does not produce the entire sense tagging of the word in the whole *corpus*, but makes it possible to better identify nets of words representing different concepts in a semi-supervised, WSD data driven procedure.

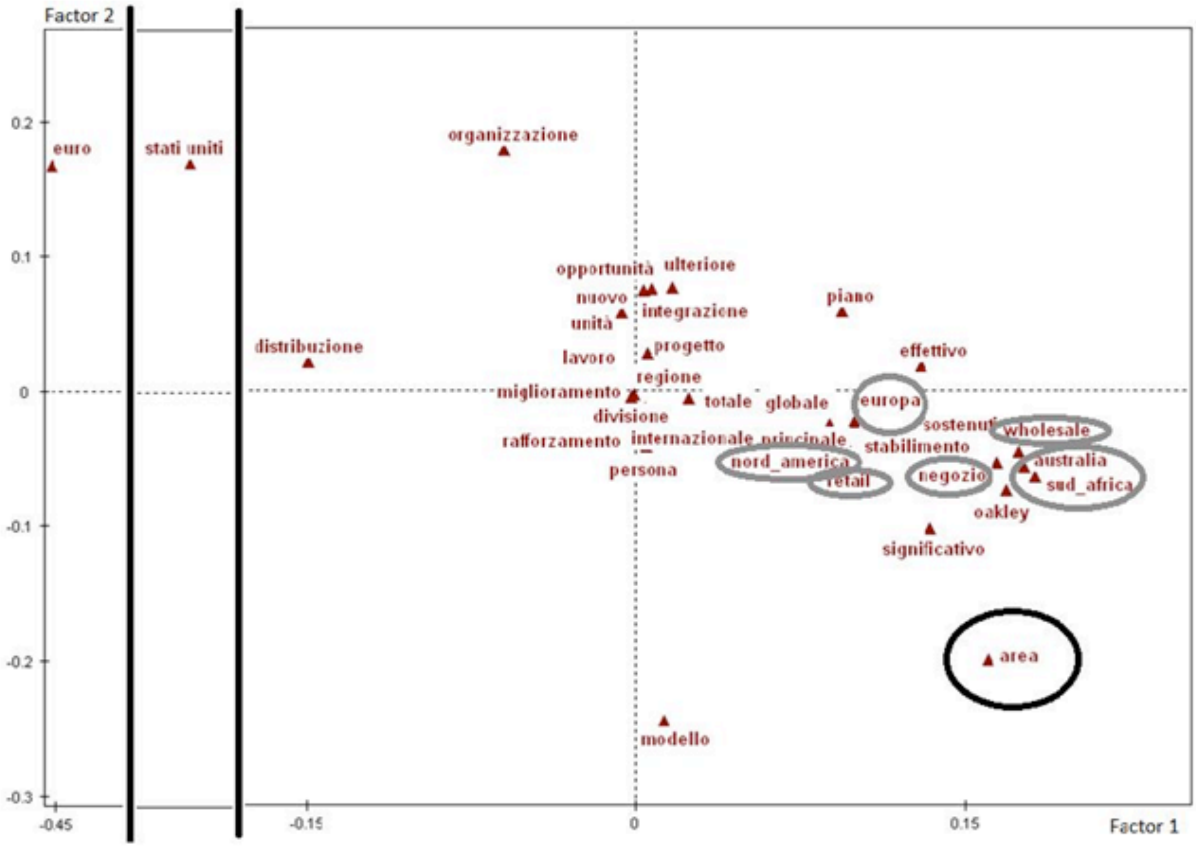

Figure 4. The alters of the focal node "area" on the CA First Factorial Plane

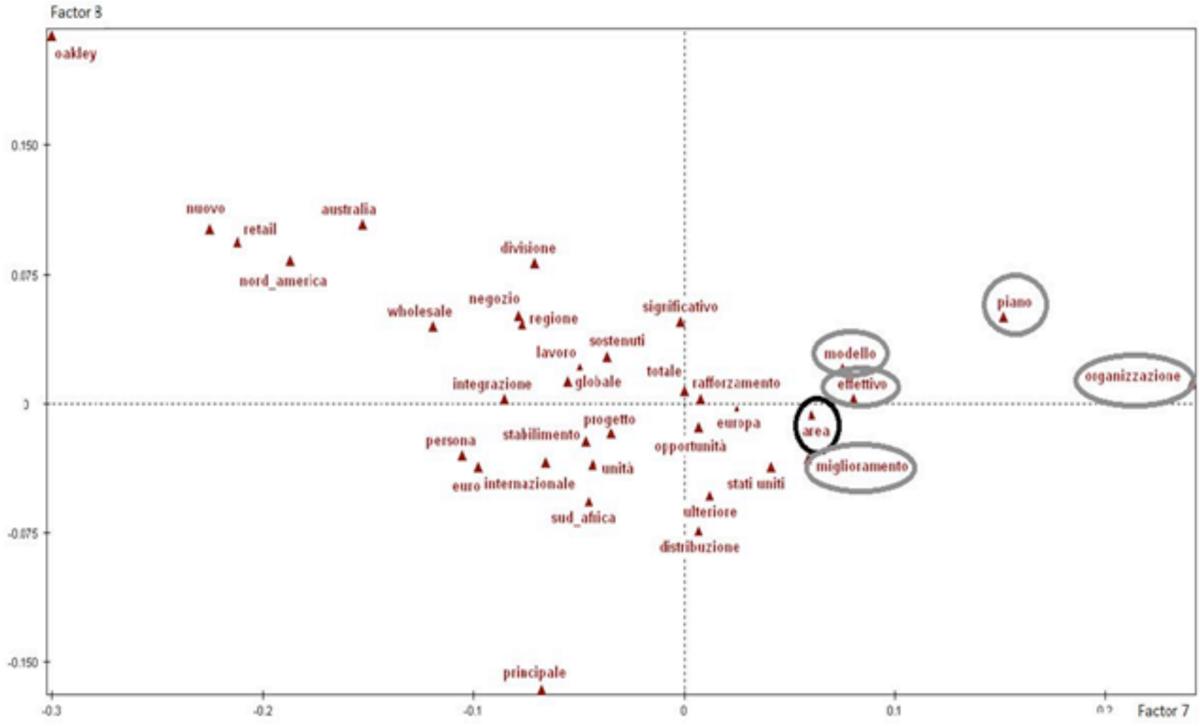

Figure 5. The alters of the focal node "area" on the CA forth factorial plane (factor 7and 8)

6. Concluding remarks and further developments

The factorial methods are a well-established approach for identifying latent features in a corpus (Biber, 1988; Lebart and Salem, 1988). Here we propose their use at a micro level, focusing our attention on the words and the variability in their use. Although the starting point consists in identifying the most relevant words with respect to the first principal components, we show that the different meanings of a word can be captured by the factors corresponding to smaller eigenvalues. Therefore it is possible to partition the egonet of a selected term in different components. Each component consists of the words which lie near the selected term on a different axis. The identified components can be useful for further analyses on corpora belonging to the same field, as they can make the identification of different contexts easier.

Further developments of the research activity will be devoted to choosing and comparing metrics for building the similarity matrix in order to fully explore relational structures among the terms in a corpus. As the proposed strategy helps to select a set of terms which can be viewed as new features prior to further analyses, in the future, we aim to put the strategy into the theoretical frame of feature selection.

Acknowledgements

This work is a result of the research activity relative to the project REPOS/DOTTORI DI RICERCA ESPERTI/ 1 and it is financially supported by the Project REPOS.

This paper derives by a strict and continuous collaboration among the authors. Anyway sections 1, 2, 5 may be mainly attributed to S. Balbi; sections 3, 4, 6 to A. Stawinoga. The Authors thank M. Spano for the helpful comments.

References

Balbi S., Stawinoga A. and Triunfo N. (2012). Text Mining tools for extracting knowledge from Firms Annual Reports. In Dister A., Longrée D., Purnelle G. (eds.) *JADT 2012: 11es Journées internationales d'Analyse statistique des Données Textuelles*, 67-79, LASLA-SESLA, Liège.

Batagelj V., Mrvar A. and Zaversnik M. (2002). Network analysis of texts.

Paper online : http://nl.ijs.si/isjt02/zbornik/sdjt02-24bbatagelj.pdf

Biber, D. (1988). *Variation across speech an*d *writing*. Cambridge: Cambridge University Press.

Bolasco S. (2010). *TaLTaC2.10, Sviluppo, esperienze ed elementi essenziali di analisi automatica dei testi*, LED on line.

Bourdieu P. (1991). Introduction in Bourdieu P., Chamboredon J-C.andPasseron J-C. *The Craft of Sociology*.Walter de Guyte r. Berlin.

deNooy W. (2003). Fields and networks: corresponding analysis and social network analysis in the framework of field theory. *Poetics*, 31: 305-327.

Freeman L.C. (1982). Centered graphs and the structure of ego networks. Mathematical Social Sciences 3: 291 -304.

Hanneman R. and Riddle M. (2005). *Introduction to social network methods*. Riverside, CA: University of California, Riverside. http://faculty.ucr.edu/~hanneman/.

Kasture N. R. and Agrawal A. (2012). A supervised Word Sense Disambiguation method using ontology and context knowledge, *Computer Engineering and Intelligent Systems*, Vol 3, No 8.

Lebart L. and Salem A. (1988). *Statistique Textuelle*. Dunod, Paris.

Lebart L., Salem A. and Berry L. (1998). *Exploring Textual Data*. Kluwer Academic Publishers. The Netherlands.

Popping R. (2000). *Computer-Assisted Text Analysis*. Sage. London.

Scott J. (2000). *Social Network Analysis: A Handbook*. Sage. London.

Stawinoga A. and Balbi S. (2012). The Use of Network Analysis Tools for Dimensionality Reduction in Text Mining, Symposium on Learning and Data Science SLDS, Florence, https://www.ceremade.dauphine.fr/SLDS2012/abstracts_slides_papers/5-CHALLENGES_TEXT_MINING/Slides_BALBI_STAWINOGA.pdf.

Stevenson M. and Wilks Y. (2003). Word Sense Disambiguation. In R. Mitkov (ed.) *The Oxford Handbook of Computational Linguistics*.

Visual Thesaurus.com