Semantic structure from Correspondence Analysis

Barbara McGillivray
Dipartimento di Linguistica
Università di Pisa
Pisa, Italy
barbara.mcgillivray@aksis.uib.no

Christer Johansson
Dept. of Linguistics
University of Bergen
Bergen, Norway
christer.johansson@uib.no

Daniel Apollon
Text Technology Lab.
Aksis, UNIFOB
Bergen, Norway
daniel.apollon@aksis.uib.no

Abstract

A common problem for clustering techniques is that clusters overlap, which makes graphing the statistical structure in the data difficult. A related problem is that we often want to see the distribution of factors (variables) as well as classes (objects). Correspondence Analysis (CA) offers a solution to both these problems. The structure that CA discovers may be an important step in representing similarity. We have performed an analysis for Italian verbs and nouns, and confirmed that similar structures are found for English.

1 Introduction

Over the past years, distributional methods have been used to explore the semantic behaviour of verbs, looking at their contexts in corpora (Landauer and Laham, 1998; Redington and Finch, 1998; Biemann, 2006, inter al.). We follow a general approach suggested already by Firth (1957), to associate distributional similarity with semantic similarity.

One question concerns the syntax-semantics interface. Results using distributions of verbs in context had an impact on verb classification (Levin, 1993), automatic verb clustering (Schulte im Walde, 2003), and selectional preference acquisition (Resnik, 1993; Li and Abe, 1995; McCarthy, 2001; Agirre and Martinez, 2001, inter al.).

In automatic verb clustering, verbs are represented by vectors of a multidimensional space whose dimensions (variables) are identified by some linguistic features, ranging, for example, from subcategorization frames to participation in diathesis alternations and lexical selectional preferences. The verbs cluster on co-occurrence with the features chosen, and such information provide a generalisation over the verbs with respect to the variables.

In the case of selectional preference acquisition, a verb (or a verb class) is associated to a class of nouns that can be the lexical fillers of a case frame slot for the verb. This allows us to calculate the association strength between the verb and its filler nouns. The generalisation step is performed for the case frame instances (observations) and produces more abstract noun classes that can be applied to unseen cases. This often utilizes hierarchies of existing thesauri or wordnets.

We propose a method that uses Correspondence Analysis (CA) to study the distribution (and associated semantic behaviour) of a list of verbs with nouns occurring in a particular syntactic relation, for example their subjects. This is collected from a corpus, and reflects usage in that corpus. Unlike clustering methods, this technique does not imply an exclusive choice between a) classifying verbs on the basis of the noun fillers in their syntactic frame, or b) associating noun classes to verbs (sometimes mediated by a semantic hierarchy). Instead, this approach yields a geometric representation of the relationships between the nouns and the verbs in a common dual space (biplot). CA aims to find an overall structure (if any) of the data. The method emphasizes unusual observations, as deviance from the expected is what creates the axes of the analysis. CA generalizes over the actual occurrences of verb-noun pairs in the corpus, and visualizes the shape of the correspondence space.

When associating verbs with nouns, CA takes as input a contingency table (here rows correspond to the verbs, and columns correspond to their subject fillers). Each verb is a row point in the multidimensional noun space, and each noun is a column point in the multidimensional verb space. The CA goals

Coling 2008: Proceedings of 3rd Textgraphs workshop on Graph-Based Algorithms in Natural Language Processing, pages 49–52
Manchester, August 2008

are to reduce the dimension of the dual original space, and to find an optimal subspace that is the closest to this cloud of points in the χ^2-metric. The best subspace is determined by finding the smallest number of orthogonal axes that describe the most variance from the original cloud.

Finally the coordinates of both row and column points of the χ^2 contingency table are projected onto this optimal subspace, simultaneously displaying row and column points. If we consider those points that are well represented, the closer they are in this geometric representation, the more similar their original distributions are. In this way, we can detect not only that there is a relationship between the verb (e.g. *explode*) and the noun (e.g. *bomb*), but also how each word relates to each *other* word.

2 Correspondence Analysis

CA is a data analytic technique developed by Benzécri in the 1960s, which has been widely used in describing large contingency tables and binary data. At the heart of CA is Singular Value Decomposition (SVD), from which many other methods were derived (Biplot, Classical Principal Component Analysis, PCA and more).

Compared to usual clustering methods, CA gives a more fine-grained view of the spread of the input points. Benzécri (1973) points out that CA is more efficient than clustering in terms of decomposition of variance. Secondly, CA represents possible regions in space with varying density, and produces a flexible "compound clustering" on both objects and variables. Verb-nouns association profiles may not cluster in distinct space regions, but may be evenly distributed, follow a gradient-like distribution, or show overlapping clusters. In such difficult cases for clustering, CA is able to offer a representation of the geometry of the input profiles. Finally, CA offers the possibility of reconstructing the original space from the output subspace.

Let us consider a data matrix M whose size is (r, c), the $(i, j)^{th}$ entry of M containing the number of occurrences of verb j with noun i as its subject in a corpus. We calculate the relative frequencies by dividing each entry $M(i, j)$ by the sum of row i, i. e. the frequency of noun i, to get the matrix of row profiles $R(i, j)$. Therefore, the more similar two row profiles i_1 and i_2 are, the more these two nouns can be considered as distributional synonyms.

The next step implies comparing the row profiles with the average distribution where each entry (i, j) is the product of the frequency of noun i by the frequency of verb j divided by the grand total N of the table. This comparison is calculated using the χ^2-distance (i.e. a weighted Euclidean distance), which eliminates effects of high frequency alone. The next formula shows calculations for rows. Calculations for columns are analogous.

$$\delta^2(i_1, i_2) = \sum_{j=1}^{c} \frac{(R(i_1, j) - R(i_2, j))^2}{\sum_{i=1}^{r} M(i, j)}$$

The χ^2-distance between a profile point and the average profile (barycentre) is called inertia of the profile point and the total inertia measures how the individual profiles p_i are spread around the barycentre:

$$Inertia = \frac{1}{N} \sum_{i=1}^{r} \sum_{j=1}^{c} M(i, j) \delta^2(p_i, \bar{p})$$

CA then searches for the optimal subspace S that minimises the distance from the profile points. Once specified its dimension $k \leq min(r - 1, c - 1)$, S is found by applying the Singular Value Decomposition (SVD) to matrix $R - 1\bar{p}$, which decomposes it as the product $N \cdot D \cdot M$: where D is a diagonal matrix with positive coefficients $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$ (singular values) and N and M are orthonormal matrices ($N^T N = M^T M = I$). The rows of M are the orthonormal basis vectors that define S (called principal axes of inertia) and the rows of matrix $F = N \cdot D$ are the projections of the row profiles onto S. For $k = 2$, this allows us to plot the new coordinates in a two-dimensional space and get the correspondence analysis of the row profiles.

The total inertia is decomposed into the direction of the principal axes of inertia. The first axis represents the direction where the inertia of the cloud is the maximum; the second axis maximises the inertia among all the directions orthogonal to the first axis, and so on.

The geometry of column profiles can be analysed similarly, because the two problems are directly linked and two transition formulae can be used to pass from one coordinate system to the other, explaining the French name "analyse des correspondances".

As a result, both analyses decompose the same inertia into the same system of principal axes. This allows us to merge the two representations in one single geometric display showing at the same time the projections of row and column points in the subspace.

In addition to this dual space representation, CA gives a system of diagnostic measures for each of the two dual spaces:

- contributions of the rows (and columns) to the axes, i. e. the inertia of the points projected onto the axes, which contributes to the principal inertia;
- contributions of the axes to the row (and column) points;
- quality of representation (cumulative sum of contributions of the axes for each point); this highlights *well represented* points.

3 Explorations

We performed a CA using the Matlab Analytica Toolbox developed by Daniel Apollon. We tested this technique first on the Italian newspaper corpus LA REPUBBLICA, which consists of 450 million word tokens. This corpus was syntactically parsed using the MALT dependency parser (Nivre, 2006). A list of 196 verbs was compiled following the list of German verbs contained in (Schulte im Walde, 2003) and adapting it to Italian. Looking at the syntactic analyses of the corpus where the verbs of the list showed a subcategorization frame containing a subject slot, their lexical subject fillers were automatically extracted. The matrix M, whose 2553 row entries correspond to the nouns extracted as subject fillers, was then used as input for the CA ($|M| = 196 \times 2553 = 500388$).

Starting from the quality of representation scores of this analysis, we isolated a set of points with increasingly good representation, ending with an extremely faithful and low dimensional representation. We called this method ”incremental pruning”. Figure 1 shows the dual display of the analysis for the Italian data in a two dimensional space, after filtering out those points showing a quality of representation below a threshold of 30%.

We can conceptualize the data set C after a CA as the cumulative effect of three different underlying phenomena: K, R and E.

K can be seen as a reduction of the latent structure of C; it contains its *core* structure as it has been underlined by the analysis and left after pruning.

R refers to the *residual* variance, not included in the core analysis. It contains the most predictable points[1], which are plotted near to origin (barycentre of the data cloud). These points give a small contribution to the inertia of the principal axes.

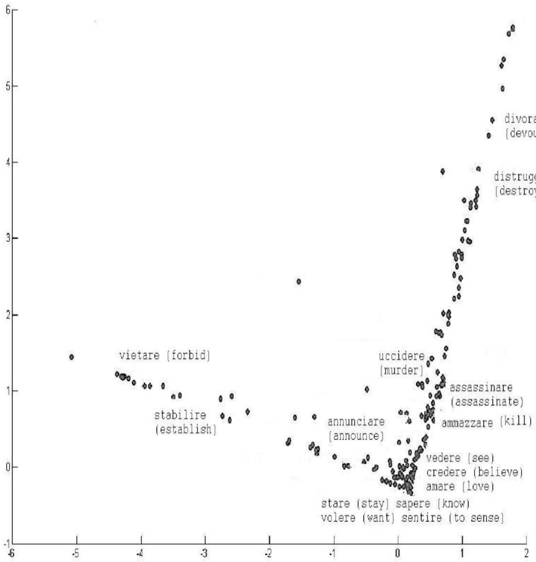

Figure 1: Correspondence graph for Italian

E contains the *error* in the representation, as well as badly represented points.

Points far from the origin display strong structure; they may correspond to rare words used in special contexts. Figure 1 shows that words related to destruction[2] are aligned in the same direction, whereas the second vector is mainly constituted by nouns and verbs that have to do with the political and legal area[3]. The first principal axis accounts for nearly 16% of the total inertia, whereas the second axis accounts for 12%. The first six axes accounts for over 70% of variation. Many words were not well represented, but contribute to variance.

We confirmed our method on English, using the British National Corpus[4]. A similar structure was found. We restrict ourselves to reproduce the graph for Italian.

[1] In our data: pronouns *she, I, he, every-, no-, some-body, who*, nouns with partly pronominal qualities *husband, wife, friend, sir, son, father, mother, fact, event*.

[2] Along the y-axis from top down to the middle, we find the nouns *flame, extend, stick of dynamite, excavator, chemotherapy, effusion, blaze, seism, demon, dynamite, fire, aviation, earthquake, artificer, explosive device, insect, gas, landslide, virus, rain, bulldozer, hurricane, wave, speculation, artillery, remorse, bomb, missile, violence, revolution*, etc.

[3] Along the x-axis we find, from left to the middle, the nouns *order, regulations, norm, code, legislation, rules, treaty, constitution, circular letter, system, directive, law, decree, article, judgement, amendment, court*, etc.

[4] via sketchengine http://www.sketchengine.co.uk

4 Conclusion

CA detects a structure for Italian verb-noun correspondences in LA REPUBBLICA ($\sim$ 450 million words). A similar structure was confirmed using BNC for English. Both global and local structures are found, which gives possibilities to represent lexical units with reference to both principal axes and word similarity. The main dimensions of the Italian corpus are topical (crime related vs. natural catastrophes, and laws vs. political institutions). Semantic relatedness were observed in closely mapped words. Both global and local structure is found, and we can speculate that this helps representing lexical units in semantic labeling (Giuglea and Moschitti, 2006) for machine learning tasks. We can conceptualize text graphs in two distinct usages: knowledge re-presenting (e.g. FrameNet) and visualizing relations in a data set. Our method belongs in the second category.

Acknowledgements

The first author is a MULTILINGUA fellow at Uni. Bergen, financially supported by a Marie Curie action (European Commission). We wish to thank the anonymous reviewers who gave important leads to future research.

References

Agirre, Eneko and David Martinez. 2001. Learning class-to-class selectional preferences. In *Proc. of the ACL/EACL Workshop on Computational Natural Language Learning*, pages 1–8, Toulouse, France.

Benzécri, Jean-Paul. 1973. *L'Analyse des Données*, volume 1. Dunod.

Biemann, Chris. 2006. Chinese Whispers – an Efficient Graph Clustering Algorithm and its Application to Natural Language Processing Problems. In *Proc. of the HLT-NAACL-06 Workshop on Textgraphs-06*, pages 73–80, New York, USA.

Firth, John R., 1957. *Studies in Linguistic Analysis*, chapter A synopsis of linguistic theory 1930-1955. Philological Society.

Giuglea, Ana-Maria and Alessandro Moschitti. 2006. Semantic Role Labeling via FrameNet, VerbNet and PropBank. In *Proc. of the 21st Int. Conf. on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 929–936, Sydney, Australia.

Landauer, Thomas K., Peter W. Foltz and Darrell Laham. 1998. Introduction to Latent Semantic Analysis. *Discourse Processes*, 25:259–284.

Levin, Beth. 1993. *English Verb Classes and Alternations*. The University of Chicago Press.

Li, Hang and Naoki Abe. 1995. Generalizing case frames using a thesaurus and the MDL principle. In *Proc. of Recent Advances in Natural Language Technology*, pages 239–248.

McCarthy, Diana. 2001. *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences*. Ph.D. thesis, University of Sussex.

Nivre, Joakim. 2006. *Inductive Dependency Parsing*. Springer.

Redington, Martin, Nick Chater and Steven Finch. 1998. Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22:425–469.

Resnik, Philip. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania.

Schulte im Walde, Sabine. 2003. *Experiments on the Automatic Induction of German Semantic Verb Classes*. Ph.D. thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.