Categorial Type Logics and Italian Corpora

Raffaella Bernardi[a], Andrea Bolognesi[b], Simone Romagnoli[b], Corrado Seidenari[b], Laura Surace[b], Fabio Tamburini[b]

[a] *Free University of Bolzano Bozen, Italy*

[b] *CILTA, University of Bologna, Italy*

Abstract

In this abstract we will present work in progress on the annotation of Italian Corpora carried out at the Interfaculty Center for Theoretical and Applied Linguistics (CILTA) - University of Bologna. The project aims at tagging the 100-million-words synchronic corpus of contemporary Italian, CORIS/CODIS, with syntactic information. In particular, we will focus attention on our first task, namely the definition of an empirical motivated part-to-speech (PoS) tagset for Italian. To achieve this goal, we plan to apply a clustering algorithm to Categorial Type Lexicon Assignments automatically induced by a dependency treebank, a simplified version of the Turin University Treebank (TUT).

Key words: Lambek syntactical calculus; Multimodal categorial grammars; Dependency Tree-bank; Corpus Annotation; Clustering.

1 Introduction

As for other languages, for Italian as well there exist guidelines regarding the definition of a proper Part-to-Speech (PoS) tagset. They have been developed under the EAGLES project (Expert Advisory Group on Language Engineering Standards) by Monachini [1]. Furthermore, there are several other research groups (Torino University, Xerox and Venice University) which worked on PoS annotation for Italian developing different classification strategies.

⋆ Our thanks go to FIRB 2001 project N. RBNE01H8RS coordinated by prof. R. Rossini Favretti. Thanks to M. Moortgat and R. Moot for the stimulating discussion during their staying in Bologna.

Email address: bernardi@inf.unibz.it (Raffaella Bernardi).

We have compared the PoS tag sets used by these groups with Monachini's guidelines. From this comparison, it results that though there is a general agreement on the part of speech to be used[1], the resulting classifications of Italian words is rather heterogeneous. This holds, particularly, for adjectives, determiners and adverbs.

In order to avoid the differences in definition inherited from the linguistic theory forming the background of any annotation schema, we propose to follow a distributional approach to PoS definition by making use only of linguistic information which is largely accepted, namely Head-Dependent (H-D) and Function-Argument (F-A) relations. In particular, we use a simplified treebank extracted from TUT [2] containing information covering only the most general dependency relations that play a role in the F-A structures, namely argument, modifier and auxiliary relations.

We encode this information into Categorial Types automatically induced from F-A structures by means of a type-resolution algorithm [2,3]. We then apply a clustering algorithm on the obtained types exploiting the expressivity of Categorial Type Logic formulas and its inference system [4]. In this way we expect to reach an empirically founded PoS tagset definition. In the remaining part of the abstract we sketch the main steps of our procedure from F-A structures to clusters.

2 Empirically Motivated PoS classification

Early approaches to the distributional study of lexicon were based on the hypothesis that if two words are syntactically and semantically different, they will appear in different contexts. In brief, these approaches [5,6] examine the distributional behaviour of some target words by comparing the lexical distribution of their respective collocates and using some quantitative measures of distributional similarity [7]. The main drawback of these techniques is the limited context of analysis. Collecting information from a defined context, typically 2 or 3 words, will invariably miss all the syntactic dependencies longer than the context interval. To overcome this problem we propose to exploit the

[1] The rather standard classification consists of nouns, verbs, adjectives, determiners, articles, adverbs, prepositions, conjunctions, numerals, interjections, punctuation and residuals which differ from project to project.

[2] Turin University Treebank (TUT) is a corpus of Italian sentences annotated by specifying relational structures augmented with morpho-syntactic information and semantic role in a single-layered dependency-based representation. The currently released tree-bank includes 38,653 words (1,500 sentences) from the Italian civil law code, national newspapers, reviews, novels, and academic papers.

expressivity of Categorial Type Assignments (CTAs) (with encoded core dependency relations) by applying clustering algorithms on them. Our next step is to define a notion of "distance" between CTAs. Currently, we are studying the application of proper distance measures considering types as trees and adapting the theoretical results on tree metrics to our problem. The algorithm for computing the tree-edit distance [8], designed for generic trees, appears to be a good candidate for clustering in categorial-type domain. More experiments have to be performed to test the method and fine-tune the metric parameters to our purpose.

3 Conclusion

We have described work in progress on a distributional approach to PoS tagset definition that exploits the logical power of Categorial Type Logic and the expressivity of its language. We still have to experiment the studied algorithm and fine-tune the clustering algorithm to serve our needs.

References

[1] M. Monachini, Elm-it: An italian incarnation of the eagles-ts. definition of lexicon specification and classification guidelines, Tech. rep., Pisa (1995).

[2] W. Buszkowski, G. Penn, Categorial grammars determined from linguistic data by unification, Studia Logica 29 (1990) 431–454.

[3] M. van Emden, Conditional answers for polymorfic type inference, in: Proceedings of the 5th International Conference on Logic Programming, 1988.

[4] M. Moortgat, Categorial type logics, in: J. van Benthem, A. ter Meulen (Eds.), Handbook of Logic and Language, The MIT Press, Cambridge and Massachusetts, 1997, pp. 93–178.

[5] E. Brill, M. Marcus, Tagging an unfamiliar text with minimal human supervision, in: Proceedings of the Fall Symposium on Probabilistic Approaches to Natural Language, MA: American Association for Artificial Intelligence, Cambridge, 1992, pp. 10–16.

[6] C. D. Santis, F. Tamburini, E. Zamuner, Identifying phrasal connectives in italian using quantitative methods, in: S. Nuccorini (Ed.), Phrases and Phraseology - Data and Description, Berlin: Peter Land., 2002.

[7] L. Lee, Measures of distributional similarity, in: Proceedings of the 37th ACL, College Park, MD, 1999, pp. 25–32.

[8] D. Shasha, D. Zhang, Approximate tree pattern matching, in: A. Apostolico, Z. Galig (Eds.), Pattern matching algorithms, Oxford University Press, 1997.