Stefano Federici (Cagliari, Italy)

Grammars, regularities, languages: the analogy-based support to hypergrammars

Abstract

Grammars are the classical tool to describe the structure of natural languages. In the field of language studies, grammars aim exactly to what other sciences aim to: generalize the phenomena of a world (in this case, the world of words and sentences of a human language) by creating a model of its underlying structure. The model will make then possible to check the correctness of a given word or sentence, or to generate correct words and sentences.

As in other sciences, when it comes to building a model that generalizes phenomena, grammars are not perfect, nor can they possibly be perfect. Indeed, if this is true in models of physics or medicine or other physical sciences (the complexity of the world to model is so high that the model can only take into account several of the important aspects of the world to be modelled), this is also true of models of human languages, i.e. grammars.

This does not mean that there are no “100% areas” in human languages. Just to give an oversimplified example, the selection of the right article in front of an Italian word follows very strict rules, without exceptions. But classical grammars are the mirror of the usage of the language as it is extracted from a source, i.e. a given collection of “correct” texts. The competence of the linguist then adds to the source, to get a final set of rules that will allow to check and generate correct linguistic objects with respect to those rules.

From one side, grammars derived from an extremely limited amount of sources have a common problem: large parts of texts (especially digital texts) cannot be considered as grammatical, even if they are perfectly understandable to humans. From the other side, collecting too many texts does not allow the linguist to derive a unitary, comprehensive, coherent corpus of grammatical rules.

This state of affairs does not imply that it is impossible to create such a corpus of rules. Grammars are not the wrong instrument. But the classical, “linear” grammar is not sufficient anymore. In a world in which linear texts are not sufficient anymore to get the necessary information, “hyper-grammars”, i.e. grammars in which dependent information is linked to each other, are necessarily the next evolutive stage. A hyper-grammar will be created by the interaction of different types of regularities: from simple grammatical rules to specific and exceptional cases, to generally valid regular behaviors. In a hyper-grammar each single context derived from a text will be a regularity, that is a rule that is valid until another, more “justified” rule, comes into action. The addition of new regularities will not change the status of previously recorded regularities, but will possibly change the response of the grammar to future generations or validation judgments.

In this talk we will propose Analogy-Based Systems as a good candidate to the role of hypergrammar. In the past years analogy-based systems have been successfully shown to be able to cope with linguistic tasks at different levels (phonological (Federici et al., 1996), morphological (Federici and Pirrelli, 1997), grammatical (Federici, 1998), semantic (Federici et al., 1999, 2000) of linguistic description.

A grammatical rule in an analogy-based system is the result of the interaction of a potentially unlimited set of analogy-based regularities. Each analogy-based regularity is derived from several actual similar texts contained in the corpora from which the hypergrammar is derived. Similarity is defined as contextual identity, modulated by the different type of linguistic level.

Hypergrammars based on analogy-based systems have a twofold objective: creating a tool for the automatic analysis of textual information and highlighting how the grammatical structure of a given word is really “implemented” in the source corpora (Federici and Wade, 2007). Examples of actual usage are automatically linked to the analogy-based rule.

In this talk we will show examples of how analogy-based hypergrammars can be built and used for several different purposes.

References

Federici S. (1998), “An efficient algorithm for the automatic building of a lexicon from textual corpora”, in Proceedings of EURALEX 98, Liège, Belgium.

Federici S., Pirrelli V. (1997), “Analogy, computation and linguistic theory”, New Methods In Language Processing, Taylor & Francis Ltd, Bristol, United States.

Federici, S., Wade J. (2007), “Letting in the light and working with the Web – A dynamic corpus development approach to interpreting metaphor”, in Proceedings of Corpus Linguistics 2007, University of Birmigham, Birmigham (UK).

Federici S., Montemagni S., Pirrelli V. (1999), “SENSE: an analogy-based Word Sense Disambiguation system”, Natural Language Engineering, pp. 207-218.

Federici S., Montemagni S., Pirrelli V. (2000), “ROMANSEVAL: Results for Italian by SENSE”, in Computers And The Humanities, 34, pp. 199-204.

Federici S., Pirrelli V., Yvon F. (1996), “A dynamic Approach to Paradigm-driven Analogy”, in Connectionist, Statistical, and Symbolic Approaches To Learning For Natural Language Processing, Springer-Verlag, Berlin, Germany.