Shape Analysis as an Aid for Grammar Induction[1]

Ife Adebara [a], Veronica Dahl [a]
[a] *Department of Computer Science, Simon Fraser University, 8888 University Drive, Burnaby, Canada*

Abstract.
Visual shapes inherent in different aspects of language processing have been manifesting themselves as important not only for enhancing that process itself, but also for helping solve open problems in ways that are more economical and more intuitive than the usual statistical-based, massive processing approaches. In this article we investigate an interesting use of gleaning shape from input sets, as an aid for mixed language grammar induction.

Keywords. shape implicit in errors, womb grammars, constraint-based parsing, multilingual text, grammar induction.

1. Introduction, Background, and The Main Problem

Imagine an automatic language processing system (for a language we shall call the source language) that can adjust its own grammar rules so that they become those of another language (which we shall call the target language). Imagine that for doing so, our system only needs access to a corpus of representative correct sentences of the target language, plus access to the target language's lexicon.

A computational methodology exists- Womb Grammars, or WG [6]- for solving precisely this grammar induction problem. It is implemented on top of CHRG [5], and evolved from Property Grammars [4,7]. WGs have been useful in various applications such as second language tutoring [3], language acquisition [8] and bio-inspired computation [2].

WGs expect a language's grammar to be formulated in terms of grammar constraints (properties) between pairs of constituents of a phrase. For instance we can define a (very simple) noun phrase pattern by saying that its allowable constituents are determiners and nouns (constituency constraint), that each can appear only once (unicity), that the noun is obligatory, that the determiner must precede the noun (precedence), and so on. WGs work by observing the list of violated properties that are output when correct sentences in the target language are fed to the source grammar, and "correcting" that grammar so that these properties are no longer violated.

[1]This research was supported by NSERC Discovery grant 31611024

With the spontaneous language mixes inherent in social media communications across countries, it has become important to automate the processing of mixed languages and jargons also. A proposal for using WGs in this sense was put forward in [1], but an open problem remained: that of determining how predominantly a main language's given constructs show up versus the secondary language's counterpart. There was only the suggestion that it might be solved by using some statistical analysis in a second round of parsing.

It is our thesis here that visual clues can adequately address this problem, together with an expert's small amount of time and interaction with the system. We shall present our ideas through the example of noun phrase's properties. Similar considerations apply to other types of phrases.

2. Our Proposed Solution

For illustration purposes, let us assume that English noun phrases consistently follow the linear precedence rule adj < noun. Concretely, we propose to line up our set of input phrases one below another and visually mark all the incorrectly (with respect to the source grammar) ordered nouns in this set (or conversely, all adjectives), say by writing them all in blue. If the input corpus is correct with respect to English, the resulting blue shape will be a straight line. For a Yòrùbá native speaker, the corpus may be tinted with alternative, Yòrùbá-inspired orderings, introducing visible "scoliosis" into the resulting blue shape.

Thus, a simple visual inspection of the coloured shape formed in the set of input phrases would give us a quick idea not only of whether the corpus exhibits violations of the main language's constraints, but also of how predominant the secondary language is, for the property in question, for the user at hand. In general terms, the more visual scoliosis, the more deviation from the norm- independently of how the input sentences are ordered. Similarly, we can visually mark failed properties that WG have found out along our input corpus, using different colours for each property: disallowed constituents (such as a verb as direct daughter of a noun phrase) can be marked in red, to bring them to the human expert's attention, who may then decide to include the "extraneous" category because ubiquitous.

Obligatory categories that are missing can be marked as labelled arcs across the phrase where they are missing. Violations of uniqueness can be highlighted in another colour, to quickly draw the expert's eye towards a decision of whether to delete the extra occurrence because of deeming it a typo, or to adjust the grammar in order to relax the uniqueness constraint. Visually marking two constituents that exclude each other could quickly call on the expert to modify the grammar so as to accept, e.g. the Yòrùbá-influenced coexistence of a determiner with a proper name, as in "the Veronica".

Should the main language in our system include a noun's strict requirement for a determiner, the correct action when our English-tinted input "Lions sleep tonight" shows up would be to relax the requirement under the stated circumstance. Again we must colour absence if we are to catch the expert's eye to solicit their input on whether to add a determiner, or adjust the grammar to include the said relaxation condition for plural generic nouns. And again, the marking must be done as a label on an arc that covers the entire phrase.

3. Concluding Remarks

We have shown how some Womb Grammar parsing results can be re-expressed in terms of shape, so that a human expert can quickly determine visually the relative strengths of competing properties of the grammar. With this work we hope to stimulate further research into extending grammars with visual interactive means for adjusting them. We believe that complementing logic-based grammars with visually driven interactions with an expert can become a very fruitful, while less expensive alternative to statistical parsing.

References

[1] Adebara I., Dahl V., T.S.: Completing mixed language grammars through womb grammars plus ontologies. In: In Proceedings of the International Conference on Agents and Artificial Intelligence, Lisbon, Portugal. pp. 292–297 (2015)

[2] Becerra, L., Dahl, V., Jiménez-López, M.D.: Womb grammars as a bio-inspired model for grammar induction. In: Trends in Practical Applications of Heterogeneous Multi-Agent Systems. The PAAMS Collection, pp. 79–86. Springer International Publishing (2014)

[3] Becerra Bonache, L., Dahl, V., Miralles, J.: On second language tutoring through womb grammars. In: Rojas, I., Joya, G., Gabestany, J. (eds.) Advances in Computational Intelligence. Lecture Notes in Computer Science, vol. 7902, pp. 189–197. Springer Berlin Heidelberg (2013), http://dx.doi.org/10.1007/978-3-642-38679-4_18

[4] Blache, P.: Property grammars: A fully constraint-based theory. In: Proceedings of the First International Conference on Constraint Solving and Language Processing. pp. 1–16. CSLP'04, Springer-Verlag, Berlin, Heidelberg (2005), http://dx.doi.org/10.1007/11424574_1

[5] Christiansen, H.: CHR grammars. TPLP 5(4-5), 467–501 (2005)

[6] Dahl, V., Miralles, J.: Womb grammars: Constraint solving for grammar induction. In: Sneyers, J., Frühwirth, T. (eds.) Proceedings of the 9th Workshop on Constraint Handling Rules. vol. Technical Report CW 624, pp. 32–40. Department of Computer Science, K.U. Leuven (2012)

[7] Dahl, V., Blache, P.: Directly executable constraint based grammars. In: Proc. Journees Francophones de Programmation en Logique avec Contraintes, JFPLC 2004 (2004)

[8] Dahl, V., Miralles, E., Becerra, L.: On language acquisition through womb grammars. In: 7th International Workshop on Constraint Solving and Language Processing. pp. 99–105 (2012)

[9] Wattenberg, M.: Arc diagrams: Visualizing structure in strings. In: Proceedings of the IEEE Symposium on Information Visualization. pp. 110–116 (2002)

[10] Strzalkowski(ed), T.: Reversible grammar in natural language processing. Springer Science + Business Media, B.V. (2012)