Modeling sentence processing difficulty with a conditional probability calculator

Zhong Chen (zc77@cornell.edu)
Department of Linguistics, Cornell University
Ithaca, NY, 14853 USA

Tim Hunter (timh@umn.edu)
Institute of Linguistics, University of Minnesota, Twin Cities
Minneapolis, MN 55455 USA

Jiwon Yun (jiwon.yun@stonybrook.edu)
Department of Linguistics, Stony Brook University
Stony Brook, NY 11794 USA

John Hale (jthale@cornell.edu)
Department of Linguistics, Cornell University
Ithaca, NY, 14853 USA

Abstract

We present the conditional probability calculator CCPC for predicting word-by-word processing difficulties in human sentence comprehension. This system, in conjunction with weighted grammars and the linking hypothesis Entropy Reduction (Hale, 2006), derives the subject-object asymmetry in Italian relative clauses, including the animacy effect of head nouns.

Keywords: sentence processing; computational modeling; Entropy Reduction; Italian; relative clauses

Introduction

The computational modeling of incremental sentence comprehension has attracted increasing attention in recent years. Proposals such as Surprisal (Hale, 2001; Levy, 2008) and Entropy Reduction (Hale, 2003, 2006) use probabilistic information about syntactic structures to derive word-by-word processing difficulty predictions, for instance regarding different types of relative clauses (RCs). However, the relationship between these comprehension difficulty metrics/predictions and any specific syntactic disambiguation decisions is often hard to visualize. In this paper, we demonstrate a freely available software system, CCPC[1], which allows psycholinguists to calculate the probability of alternative completions of initial substrings from user-supplied weighted grammars. A key feature of this system is the ability to enumerate syntactic alternatives that are in play at a given point, essentially visualizing the contents of a ranked parallel parser state. In addition to our earlier modeling efforts on East Asian languages (Yun et al., 2010; Chen et al., 2012; Yun et al., In press), this conditional probability calculator allows us to make correct predictions about the subject-object asymmetry in Italian RCs and model the effect of head-noun animacy.

Modeling Procedure

Our modeling procedure starts from a user-prepared grammar. This includes expressive Minimalist Grammars (MGs) in the style of Stabler (1997) that model the relationship between fillers and gaps in RCs. Using CCPC (Figure 1), we convert a MG to an equivalent Multiple Context Free Grammar (MCFG) (Seki et al., 1991). We weight the MCFG rules by attestation counts of relevant structures in the corpus. Using the resulting weighted grammar (WMCFG), the CCPC calculates the "remainder set" of syntactic alternatives and its Entropy on the model defined by the probabilistic rules. Changes in this Entropy value before and after a given word reflect the processing cost on disambiguation and can be graphed as in Figure 2. Section 3-5 of Yun et al. (In press) present this method in a tutorial fashion.

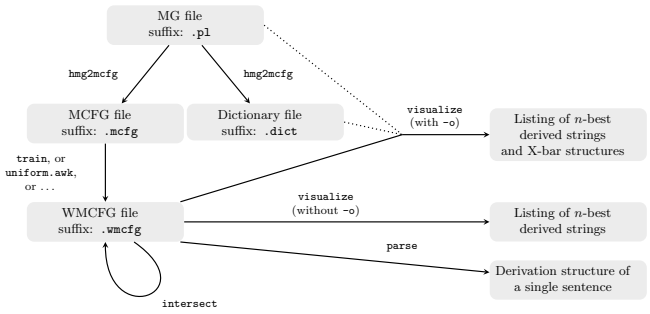

Figure 1: Overview of the system workflow and the dependencies among various kinds of inputs and outputs

Processing Predictions

The Subject-Object Asymmetry in Italian Relatives

Using an explicit grammar fragment for Italian RCs, the CCPC predicts that the widely-observed subject relative advantage holds in Italian (Di Domenico & Di Matteo, 2009). Figure 2 illustrates the predicted disambiguation costs at each word, quantified by reductions in Entropy. The comparison between subject relatives (SRs) and object relatives (ORs) suggests early processing difficulty in ORs at the embedded subject "il pagliaccio", as compared to the embedded verb "guarda" in SRs. This is roughly compatible with the English modeling results using Surprisal (Hale, 2001; Levy, 2008). However, CCPC makes it possible to understand the Entropy Reduction account in terms of specific syntactic disambiguation decisions. The Italian SR prefix is more ambiguous because the omitted phrase before the verb "guarda" can either be an extracted subject or a dropped *pro*. Therefore, less disambiguation work has been done in the SR than in the OR.

This Italian RC example illustrates a general method for deriving processing predictions from grammars. These predictions can reflect distributional factors. We leverage the Turin University Treebank (Bosco et al., 2000) to estimate a handful of parameters: the rate of noun phrase post-modification by relative clauses vs complement clauses, the

[1] http://conf.ling.cornell.edu/compling/software.htm

choice between transitive and intransitive verbs, and the rate of subject omission. The selection of this parameterization, rather than some other one, is a consequence of the grammar fragment itself.

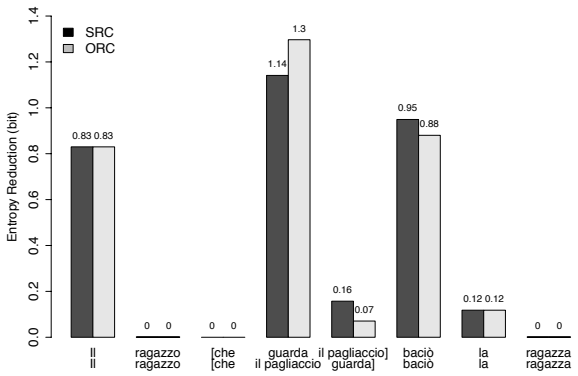

Figure 2: Predicted reading difficulty of Italian RCs

Modeling the effect of Head Noun Animacy

The generality of our approach allows a theorist to explore the role of "formalist" as well as "functionalist" features, such as animacy, in sentence processing. The animacy feature is interesting because of its differential influence on RC comprehension across languages. For example, Traxler et al. (2002) report that the difficulty of processing English ORs is reduced when the head noun of the OR is inanimate. However, Belletti and Chesi (2011) find that participants of a Italian sentence production task are immune to the change of head noun animacy in ORs. Using CCPC as discussed above, the Entropy Reduction metric correctly derives this observed result from a combination of grammatical assumptions and corpus distributions.

We rewrite the Italian grammar fragment so that the animacy information is encoded by subcategorized rules. The animacy-tagged Siena University Treebank (Chesi et al., 2011) allows us to obtain frequency distributions of noun phrase animacy both in matrix clauses and in RCs. Table 1 compares the ER predictions for the RC region, suggesting that the processing difficulty of ORs will not be reduced with an inanimate head noun phrase.

Table 1: Predicted processing difficulty for the RC region

	Head-NP	Embedded-NP	ER	Diff.
SR	+anim	−anim	1.39	**0.19**
OR	+anim	−anim	1.58	
SR	−anim	+anim	1.93	**0.22**
OR	−anim	+anim	2.15	

Conclusion

The conditional probability calculator CCPC, along with a complexity metric, can predict processing difficulty profiles in incremental sentence comprehension. We demonstrate this system by modeling the subject-object asymmetry in Italian relative clauses. Incorporating frequency information about functional features like animacy allows us to take an initial step toward rich conceptual structure.

Acknowledgments

We thank Cristiano Chesi for giving us access to the Siena University Treebank. This work was supported by the NSF CAREER Award 0741666.

References

Belletti, A., & Chesi, C. (2011). Relative clauses from the input: syntactic considerations on a corpus-based analysis of Italian. In *STiL-Studies in Linguistics* (Vol. 4).

Bosco, C., Lombardo, V., Vassallo, D., & Lesmo, L. (2000). Building a Treebank for Italian: a Data-driven Annotation Schema. In *Proceedings of the 2nd LREC* (pp. 99–105).

Chen, Z., Jäger, L., & Hale, J. (2012). Uncertainty reduction as a predictor of reading difficulty in Chinese relative clauses. In Y.-O. Biq & L. Chen (Eds.), *Proceedings of the 13th International Symposium on Chinese Languages and Linguistics (IsCLL-13)* (pp. 245–261). Taipei, Taiwan.

Chesi, C., Lebani, G., & Pallottino, M. (2011). A Bilingual Treebank (ITA-LIS) suitable for Machine Translation: what Cartography and Minimalism teach us. In *STiL-Studies in Linguistics* (Vol. 2, p. 165-185).

Di Domenico, A., & Di Matteo, R. (2009). Processing Italian relative clauses: working memory span and word order effects on RTs. *Journal of General Psychology*(136), 387-406.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the 2nd NAACL.*

Hale, J. (2003). The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, *32*(2), 101–123.

Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, *30*(4).

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126-1177.

Seki, H., Matsumura, T., Fujii, M., & Kasami, T. (1991). On multiple context-free grammars. *Theoretical Computer Science*, *88*(2), 191-229.

Stabler, E. (1997). Derivational minimalism. In C. Retoré (Ed.), *Logical aspects of computational linguistics.* Springer-Verlag.

Traxler, M., Morris, R. K., & Seely, R. E. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language*, *47*, 69-90.

Yun, J., Chen, Z., Hunter, T., Whitman, J., & Hale, J. (In press). Uncertainty in processing relative clauses across East Asian languages. *Journal of East Asian Linguistics*.

Yun, J., Whitman, J., & Hale, J. T. (2010). Subject-object asymmetries in Korean sentence comprehension. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society.*