A Cost Metric for Pronoun Resolution: Uncertainty Increases Processing Cost

Olga Seminck, Pascal Amsili, Université Catholique de Louvain, Université Paris 3
olga.seminck@cri-paris.org

Pronoun resolution is the process of finding the antecedent of an anaphoric pronoun. Whereas it is almost always successful — it seems even the case that the use of pronouns contributes to the fluidity of a discourse (1) — there are differences in the facility of resolution between different pronouns demonstrated by small differences in processing times. For example, in English, subject antecedents are processed quicker than object antecedents (2).

However, the list of contributing factors is long and it is difficult to combine all factors to make predictions about pronouns in corpora. Therefore, we investigated whether a broader hypothesis about processing cost induced by anaphoric pronouns can make relevant predictions when it is implemented as a cost metric.

Our cost metric is based on the following hypothesis: more uncertainty about the antecedent of a pronoun leads to higher processing cost. We use the notion of entropy (3) to estimate uncertainty: the antecedent of a pronoun is modeled as a random variable that can take the value of different discourse referents. Each of these referents can be attributed a probability that it is actually the antecedent and then the pronoun's entropy can be calculated (see the formula H(pro) at the bottom of this page).

An issue with this proposal is that entropy increases when the number of discourse referents rises and will thus systematically be higher further in the text. Therefore, we propose to make a small modification on the metric and take the relative entropy (3) in which the 'distance' between the actual entropy and the maximal entropy is measured (see the formula of $H_{relative}$ at the bottom of this page). So, when the relative entropy is low, we predict more processing cost.

We tested the uncertainty hypothesis on the English part of the Dundee Corpus (4), a corpus containing reading times of ten native speakers of English. For each anaphoric pronoun in the corpus, we estimated the probabilities of every discourse referent occurring in the text before the pronoun with a state of the art coreference resolution system (5). We used a Bayesian mixed model (6) to test whether the cost-metric contributed to the prediction of pronoun reading. We found that lower relative entropy lead to more participants fixating a pronoun: a result in line with the uncertainty hypothesis (95% credible interval).

This work illustrates that uncertainty about the antecedent influences pronoun resolution. It also illustrates how notions from Information Theory can make relevant predictions about human language processing and how NLP-systems can be used as robust tools to estimate probabilities of language.

$$H(pro) = -\sum_{a \in A} P(pro = a) \cdot \log_2(P(pro = a)) \qquad H_{relative}(P\|Q) = \sum_{i \in P \wedge i \in Q} P(i) \log \frac{P(i)}{Q(i)}$$

(1) Gordon, P. C., & Chan, D. (1995). Pronouns, passives, and discourse coherence. Journal of Memory and Language, 34(2), 216-231. (2) Crawley, R., Stevenson, R., & Kleinman, D. (1990). The use of heuristic strategies in the interpretation of pronouns. Journal of Psycholinguistic Research, 4, 245-264. (3) Cover, T. M., & Thomas, J. A. (2012). Elements of Information Theory. John Wiley & Sons. (4) Kennedy, A., Hill, R., & Pynte, J. (2003). The Dundee Corpus. In Proceedings of the 12th European conference on eye movement. (5) Lee, K., He, L., Lewis, M., & Zettlemoyer, L. (2017). End-to-end Neural Coreference Resolution. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 188-197. (6) Bürkner, Paul-Christian (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. In: Journal of Statistical Software 80(1), 1–28.