Constructing Distributions of Variation in Referring Expression Type from Corpora for Model Evaluation

T. Mark Ellison, Fahime Same
{t.m.ellison, f.same}@uni-koeln.de
University of Cologne

1. Referring Expression Generation (REG) in context

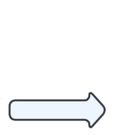
REG-in-context: Given an intended referent and a discourse context, how do we generate appropriate referring expressions (REs) to refer to the referent at different points in the discourse? (Belz and Varges 2007)

Rule-based & feature-based studies often approach REG in 2 steps:

1. Choosing the referring expression form (REF), one of: proper noun, definite noun phrase, or pronoun
2. Determining the content of that form

2. A REG-in-context example

Homer_Simpson (born May 12 1956) is the main protagonist and one of the five main characters of The Simpsons series (or show). **Homer_Simpson** is the spouse of Marge Simpson and father of Bart, Lisa and Maggie Simpson. **Homer_Simpson** is overweight (said to be ~240 pounds), lazy, and often ignorant to the world around **Homer_Simpson**.



Homer Jay Simpson (born May 12 1956) is the main protagonist and one of the five main characters of The Simpsons series (or show). **He** is the spouse of Marge Simpson and father of Bart, Lisa and Maggie Simpson. **Homer** is overweight (said to be ~240 pounds), lazy, and often ignorant to the world around **him**.

3. REG-in-context is a non-deterministic task

For many contexts, there is not a single correct REF.How do we know?

- Human choices vary, even for simple texts.
- Machine systems do not converge on singleton distributions, even when trained on big corpora.

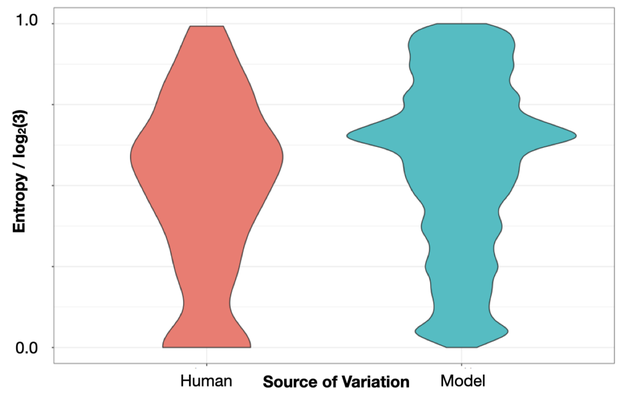
Algorithms for REG-in-context are generally evaluated against corpora of written texts, offering a single correct response in the given context.

4. Referring Expression Form Distributions (REFDs)

Violin plots of variation (relative entropy) in human REF choice and machine learning models trained on the WSJ.

Without variation, all density would be clustered at 0.0.

If all options were equally likely, the density would be clustered at 1.0.

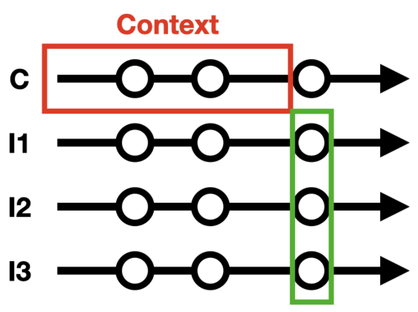


5. 2-Dimensional Corpora

To determine the distributions over REFs at a particular point, we must aggregate multiple RE form choices as the repeated measures of a single random variable. We can create two different kinds of corpora of variation:

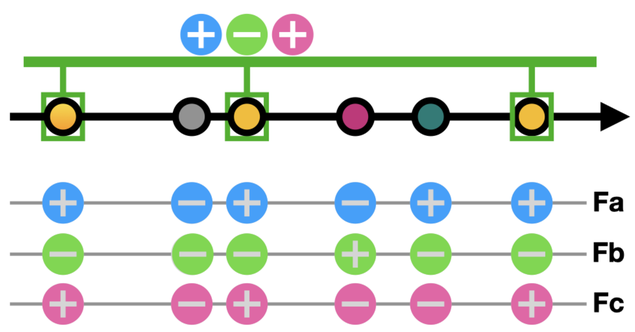
- **Parallel** Keep identical context and referent. Find REFDs by asking distinct (but similar) informants (I1, I2, I3) to choose RE forms. OR
- **Longitudinal** Generalise over contexts using features. Find REFDs by aggregating all REF choices with the same combinations of values for features (Fa,Fb,Fc).

6. Parallel vs Longitudinal Corpora



Participants all see the context from the original corpus and internalise it. They then choose an REF. These choices define the REFD for this RE.

This is a *parallel* corpus of variation as the judgements forming the REFDs come from independent choices for the same context.



RE contexts are categorised by contextual features, shown by binary values for features and colour on the REs. REFDs are formed across each distinct combination of feature values.

This is a *longitudinal* corpus of variation as the variation is organised sequentially.

7. VaREG corpus and studies

VaREG corpus (Castro Ferreira, Krahmer, and Wubben 2016a)

- 36 texts (563 REs) in 3 genres: news texts, reviews of commercial products, and Wikipedia texts
- Approximately 20 participants filled each RE gap. So it is a latitudinal corpus.

Problem: a lot of human time is required to build a corpus of parallel human judgements.

Their study

- showed substantial variation between participants in their REFD entropies,
- used Jensen-Shannon Divergence to evaluate how well model REFDs matched human REFDs from the parallel corpus (Castro Ferreira, Krahmer, and Wubben 2016b).

8. The current study

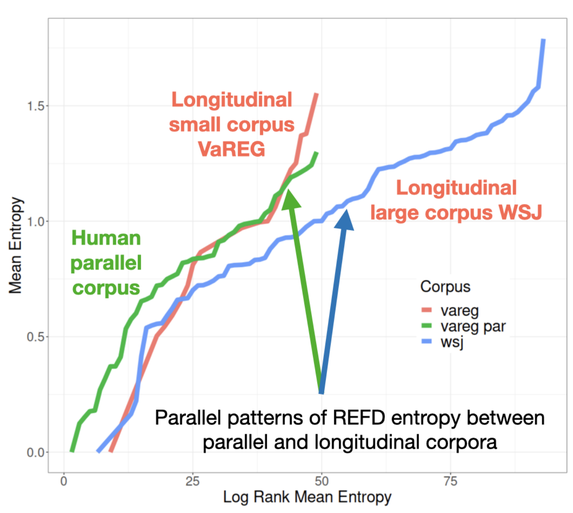
GOAL: generate REFDs of human free variation from standard corpora (without expensive parallel REF judgements).

Method: make longitudinal corpora of REFDs using feature-value combinations to aggregate REF choices into distributions

Corpora: (1) VaREG:long, (2) VaREG:lat, (3) WSJ

Our study
Learning algorithms: (1) Random Forest, (2) XGBoost, (3) CatBoost
Feature set: grammatical role, form of the antecedent, animacy, recency

9. Pattern of Entropies



10. Comparing Evaluations

VaREG	Parallel	Longitudinal
RF	0.094	0.065
XGBoost	0.086	0.061
CatBoost	**0.076**	**0.059**

JSD divergences between machine learning algorithms on parallel and longitudinal REFD corpora. Lower divergence values indicate more-similar distributions. Both corpora give the same ranking of algorithm accuracy.

11. Conclusion

Longitudinal corpora parallel structural properties and evaluative patterns of human parallel corpora.

Longitudinal corpora open the door to evaluating REG-in-context models by distribution, rather than using maximum *a posteriori* categorical choices.

References

Belz, A. and S. Varges (2007). "Generation of repeated references to discourse entities". In: *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)*, pp. 9–16. Castro Ferreira, T., E. Krahmer, and S. Wubben (June 2016a). "Individual Variation in the Choice of Referential Form". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, pp. 423–427. Castro Ferreira, T., E. Krahmer, and S. Wubben (Aug. 2016b). "Towards more variation in text generation: Developing and evaluating variation models for choice of referential form". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 568–577.