Harnessing the richness of the linguistic signal in predicting pragmatic inferences

Abstract

[jd: This is the abstract of my pre-XPrag workshop talk and will NOT be the abstract for this paper]

Theories of pragmatic inference have come a long way by building on introspective judgments and, more recently, judgment and processing data from naive participants in controlled experiments, as primary sources of data. Based on such data, common lore has it that scalar inferences are drawn very regularly and relatively independently of the context in which the scalar expression occurs. To the extent that variability in scalar inferences is investigated or acknowledged, it is typically attributed to properties of the particular scale under investigation or to working memory demands, rather than to features of the linguistic or extra-linguistic context. In this talk I will argue that a very different picture emerges once we take into account the natural distribution of scalar items, i.e., the rich complexity of the signal actually experienced by listeners in naturalistic interaction, a source of data that has received remarkably little attention to date. In particular, I will present two large-scale corpus investigations of the occurrence and interpretation of ”some” and ”or” in the Switchboard corpus of naturally occurring speech. Inference ratings were collected in web-based studies for each sentence containing a scalar expression. I will show for both ”some” and ”or” that they are much less likely to give rise to scalar inferences than commonly assumed, and that the strength of a scalar inference is systematically modulated by multiple syntactic, semantic, and pragmatic features of the context in which the scalar expression occurs. I will then go further and show that for ”some”, many effects of our carefully hand-mined contextual features can also be captured by distributed vector-based sentence representations based on recent developments in natural language processing. In particular, sentence embeddings based on deep contextualized word representations (ELMo), which model both complex characteristics of word use and how these uses vary across linguistic contexts, capture both most of the variance that the hand-mined features captured as well as additional variance. This excitingly suggests that we can use state-of-the-art neural network models of the representation of word and sentence meanings to reverse-engineer contextual information that listeners recruit in drawing pragmatic inferences, thereby further informing pragmatic theory.

Keywords: computational pragmatics; scalar implicature;

Introduction

The role of context has long been recognized as central to pragmatics. The field of experimental pragmatics has been instrumental in identifying features of context that listeners use in deriving pragmatic inferences.Recent Bayesian accounts of pragmatic inference within the Rational Speech Act framework (?, ?, ?, ?) have provided proof-of-concept demonstrations that pragmatic interpretation can be modeled as the result of listeners integrating their expectations about the utterances a pragmatic speaker with a particular meaning in mind is likely to produce, with their prior beliefs about likely meanings, via Bayes’ rule. Pragmatic speakers in turn are modeled as soft-optimal agents that attempt to produce utterances that are both contextually informative and cheap. These models have been successfully applied to many pragmatic phenomena, including scalar implicature (?, ?, ?), reference-based Quantity inferences (?, ?, ?, ?, ?), embedded implicatures (?, ?), figurative meaning (?, ?), the interpretation of gradable adjectives (?, ?), and generics (?, ?).

Especially for the case of scalar implicature, these models have been shown to be useful for formalizing the way in which extra-linguistic features of context affect the inference. For instance, [jd: list example sentence plus extra-linguistics features like QUD and speaker knowledge and world knowledge that matter for interpretation]

However, despite this success, criticisms of RSA include

- [jd: restricted toy domain problem, how can this ever work for ”real” language?]
- [jd: Bayesian inference is computationally intractable, so how could humans be doing it?]
- [jd: lack of compositionality in the models, reasoning usually based on the assumption that an entire sentence is just magically given]

[jd: i’m not sure we need to bash on rsa to get this paper off the ground, but it is a useful way of motivating some of this]

One possibility is that language users learn to use shortcuts to the inference (or lack thereof) by learning associations between surface-level cues present in the linguistic signal and the speaker’s intention, across many instances of encountering a scalar expression like *some*. While little is known about the extent to which listeners use information from the linguistic signal directly when drawing pragmatic inferences in naturally occurring speech, ?, ?, in a crowd-sourced annotation task of naturally occurring sentences with *some* in the Switchboard corpus, has shown that listeners draw stronger scalar inferences from *some* to *not all* as a function of several features present in the linguistic signal, including when the *some*-NP uses the partitive construction, when it occurs in subject position, and when the embedded NP-referent was previously mentioned. An issue with this study, however, is that the features explored, while motivated by the theoretical literature, are unlikely to constitute the full list of surface features that are likely to be relevant for listeners' inferences. Moreover, it is unclear how the armchair researcher could come by such a list.

Motivated by the problems laid out above – the need for [jd: something we can use on real, naturally occurring language] and the pragmatic theorist's exhaustive feature list problem – we decided to enlist the help of recent advances in word embedding models (?, ?, ?, ?). In particular, we ask:

1. How well do neural models that encode each of the sentences with *some* from the ?, ? dataset predict the obtained human inference judgments?

2. To what extent does the best neural model capture the qualitative effects of the hand-mined features reported by ?, ??

To address these questions, we first compare the performance of neural models that differed in the underlying word embedding model (GLoVe, ELMo) and in the sentence embedding model (mean, LSTM, LSTM+attention). We then analyze which, if any, of the original effects reported by ?, ? remain once the model predictions are included as a predictor in the original linear regression model. We then conduct a more qualitative analysis and demonstrate for two of the hand-mined features (partitive and subjecthood) that the best neural model qualitatively replicates those effects. [jd: do any of the attention visualizations add something useful? if so, we should include them]. We close with some remarks on how these models can be used for bootstrapping further relevant linguistic features in the future.

The dataset

[jd: yuxing, fill out this part (you can reuse some of what you already have in the thesis). should include the briefest of descriptions of the 3 highlighted features: partitive, subjecthood, previous mention, with examples from the corpus (see my pre-xprag conference slides 32-47 for examples of high and low-scoring cases for each feature)].

[jd: this section should also contain the original table of effects from the regression, with the new table (ie with the additional neural model indicator) appended on the right side, ie slides 101 and 102 together]

Model description

[jd: basic model architecture description. highlight the main ways in which models differ: *word embedding mechanism* and *sentence encoding mechanism*. refer to Figure 1, which is just a copy of what's on slide 80 and can be changed to make prettier]

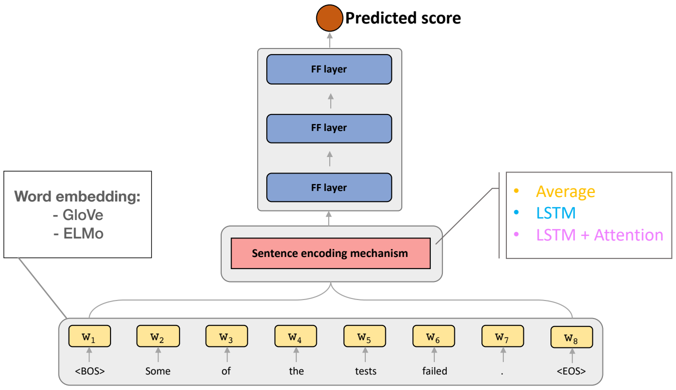

Figure 1: Model architecture.

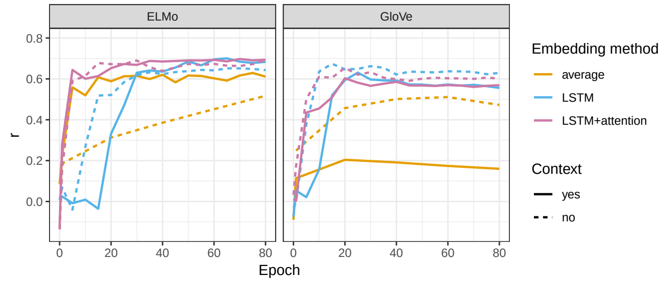

Figure 2: Correlation of each model's predictions with empirical means, by training epoch.

Word embedding model

[jd: GLoVe vs ELMo descriptionn (and BERT if we're using it?)]

Sentence embedding model

[jd: meann vs LSTM vs LSTM plus attention]

[jd: also mention here: manipulation of whether or not context was included in the sentence embedding (in which case, i suppose, it's no longer strictly a sentence embedding)]

Model evaluation

Model performance

[jd: Main points: 1. asymptotic performance very similar across ELMo based models (which are generally better than the GLoVe models). 2. Averaging is dumb (because it gets rid of signal). 3. Whether you use LSTM with or without attention doesn't matter for asymptotic performance, but it learns more quickly with attention. 4. Whether or not context is used doesn't matter for the best performing model. refer to Figure 2, which may need to be remade with BERT results (but only if they're better)]

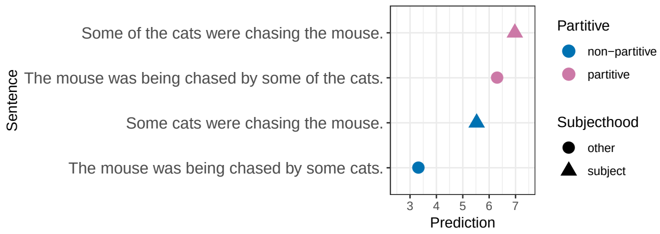

Figure 3: Model prediction for artificial dataset of four sentences that are minimal variations of each other where partitive and subjecthood are crossed.

Re-analysis of original dataset with neural model indicator

[jd: Main point: Including the best neural model predictions (in the talk that was ELMo - LSTM + attention - with context) as indicator in original regression model makes all other effects disappear]

[jd: refer back to the table from the dataset section to make the point that all effects disappear]

[jd: include scatterplots like on slide 105 to make point that variance explained is very similar]

While this is an impressive result, it is not clear whether the neural model captures the same amount of variance as the hand-mined features because it's encoding the same information, or for a different reason. We next provide a small qualitative analysis to address this question.

Qualitative model evaluation

General discussion

XXX