Towards unsupervised language models for QUD prediction

Matthijs Westera
⟨firstname⟩.⟨lastname⟩@gmail.com
Universitat Pompeu Fabra

To be presented at SemDial 2018, Aix-en-Provence.

Research goal

Central to explaining many linguistic phenomena is an understanding of what the goals of the given discourse are. This is made difficult however by the fact that goals are often left implicit in discourse. Much theoretical work in semantics and pragmatics assumes that discourse goals can be identified with implicit or explicit questions, or *Questions Under Discussion* (QUD; e.g., Ginzburg 1996; Roberts 1996). Semantic/pragmatic theories typically yield strong, falsifiable predictions *given* a certain QUD, but no comprehensive theory exists of what that QUD should be for any given piece of discourse. This limits the testability of these theories in practice, and it stands in the way of a proper understanding of results from experimental linguistics, where participants' judgments are due in part to their understandings of the implicit goals underlying the linguistic stimuli (e.g., Schwarz 1996; Westera and Brasoveanu 2014).

I propose to employ *language models* to help overcome this challenge, by using them to generate (or compute the probability of) a plausible QUD based on a discourse. To my awareness no quantitative, data-driven model of QUDs like this has been attempted. *This is work in progress*, and besides hoping to demonstrate the promise of this kind of approach and obtaining feedback, I foremost wish to draw attention to this important open issue for QUD-based theories, and the need for a tighter integration with computational modeling.

Language models

Language models are statistical models that can assign probabilities to sequences of words. For state of the art performance, language models are typically artificial neural networks (Mikolov et al. 2010 and much subsequent work). These are *generative* models: they generate natural-seeming language by sampling words from a probability distribution conditioned on the words generated before. Neural language models are trained in an *unsupervised* manner on large amounts of naturally occurring language, the training task being simply to always predict the next word.

For this work-in-progress presentation, I train a standard Recurrent Neural Network of the Long Short-Term Memory (LSTM) type (Hochreiter and Schmidhuber, 1997), which has become the de facto standard for language modeling. They have been shown to be able to acquire many aspects of syntax (including long-term dependencies) and lexical meanings (represented as high-dimensional vectors; cf. distributional semantics). However, discourse-level (inter-sentential) dependencies are still challenging (e.g., Paperno et al. 2016), which leads me to be modest in my expectations of a simple LSTM for the current task of question prediction, a typical discourse-level task. The current model will serve only as a first illustration, and I plan to apply more sophisticated models to this task in due course. Let the main contribution for now be merely to highlight the necessity of connecting pragmatic theory to computational models, and to bring attention to one possible way of doing so.

Once trained, a language model can start generating words from scratch, or from a writing prompt, e.g., one could give it "love" and it may generate "...is in the air", "...kills" or "...me please", and any of an open-ended range of continuations. We can also prompt it to *generate a question based on a prior discourse*, which I will pursue below, and/or based on a subsequent utterance, which is an option I will pursue in the near future. Indeed, to understand the QUD served by a given utterance, it will typically be necessary to combine both sources of information, i.e., about the preceding discourse and about the utterance itself – but the present work concentrates on the former.

Dataset used

Since my aim is to get language models to generate (and/or compute the probability of) questions, the training data must contain sufficiently many questions to learn from. Moreover, for these explicit questions to be able to teach the model about supposed implicit QUDs, the two types of 'questions' must have some correspondence. We will here assume such a correspondence, between explicit questions and implicit QUDs, as also assumed for instance in Roberts 1996 – but it is in certain respects a simplification.

The need for a dataset with sufficiently many explicit questions rules out non-fictional sources of data like newswire texts and Wikipedia, which have virtually none. Dialogue contains a lot of questions, but currently available dialogue datasets are comparatively small, whereas language models need a lot of data. Instead I will use literary text, which is a convenient middle way: much data is available, and it contains a reasonable number of questions (though of course results obtained may not be representative of other genres). More precisely, I use the raw data released as part of the LAMBADA dataset (Paperno et al., 2016). The training data consists of the full text of 2,662 novels, comprising more than 200M words; test data consists of 5,153 passages from 1,332 novels disjoint from the training data. As a rough indication of its question density: the training data contains 1.5M question marks (compared to 15M periods). It is worth noting that, in this genre, questions occur almost exclusively in reported speech – something quite like dialogue after all.

Since I want to be able to prompt the trained language model to generate a question, and to compute the probability of particular questions *given* that a question was to be produced, the training data must be minimally augmented to include such prompts. We do so automatically by inserting tags ⟨ask⟩, ⟨say⟩ and ⟨shout⟩ at the start of every sentence in the dataset, based on whether the sentence ends with a question mark ?, period . or exclamation mark ! – though this assumed alignment between punctuation and speech act type is a simplification. After training on the data with such tags, one can then prompt the model to generate a question based on a given discourse by first inputting the discourse, then inputting the tag ⟨ask⟩, and finally letting the model generate a sentence.

Outlook

At SemDial I will present some early results obtained on the above task, and a preliminary analysis. I will do so both by letting the model freely generate some questions for a piece of discourse, and by letting the model compute probabilities for a handful of plausible questions given a discourse. This may be the first exploration of using language models, trained on raw data, to ground QUD-based theories in natural data. As such, much remains to be seen, but I think that the current approach to leverage language models for generating questions holds some promise.

In the future I aim to combine the above type of model, which predicts a subsequent question given a prior discourse, with a 'backwards' model that predicts a prior question given an utterance. Ultimately I hope to apply the resulting models to stimuli used in experimental linguistics, and explain gradience in linguistic judgments in terms of gradience in QUD probability (Westera and Brasoveanu, 2014) – but this is not yet within reach.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 715154). This paper reflects the authors' view only, and the EU is not responsible for any use that may be made of the information it contains. I also gratefully acknowledge the support of NVIDIA Corporation with the donation of GPUs used for this research.

References

• Ginzburg, J. (1996). Dynamics and the semantics of dialogue. In Seligman, J. and Westerståhl, D., editors, *Language, Logic, and Computation*, volume 1.

• Hochreiter, S. & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

• Mikolov, T., Karafiát, M., Burget, L., Černockỳ, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In *11th Annual Conference of the International Speech Communication Association.*

• Paperno, D., Kruszewski, G., Lazaridou, A., Pham, Q. N., Bernardi, R., Pezzelle, S., Baroni, M., Boleda, G., and Fernández, R. (2016). The LAMBADA dataset: Word prediction requiring a broad discourse context. *Proceedings of ACL 2016, East Stroudsburg PA: pages 1525-1534.*

• Roberts, C. (1996). Information structure in discourse. In Yoon, J. and Kathol, A., editors, *OSU Working Papers in Linguistics*, volume 49, pages 91–136. Ohio State University.

• Schwarz, N. (1996). *Cognition and Communication: Judgmental Biases, Research Methods and the Logic of Conversation.* Erlbaum, Hillsdale, NJ.

• Westera, M. & Brasoveanu, A. (2014). Ignorance in context: The interaction of modified numerals and QUDs. In Snider, T. & Weigand, M., eds., *Semantics and Linguistic Theory (SALT) 24*, pages 414–431.