The Influence of Context on Sentence Acceptability Judgements

Jean-Philippe Bernardy[1], Shalom Lappin[1], and **Jey Han Lau**[2,3]

[1] University of Gothenburg
[2] IBM Research Australia
[3] The University of Melbourne

July 17, 2018

Introduction

- Sentence acceptability: the extent to which a sentence is natural to native speakers.
- It encompasses semantic, syntactic and pragmatic plausibility and other non-linguistic factors such as memory limitation.
- Grammaticality, by contrast, is a theoretical concept that measures the syntactic well-formedness of a sentence.
- Here we are interested in predicting acceptability judgements.

Motivation

- We previously explored using unsupervised probabilistic methods to predict sentence acceptability, and found some success.
- It provides evidence that linguistic knowledge can be represented as a probabilistic system, addressing foundational questions concerning the categorical nature of grammatical knowledge.

Acceptability in Context

- In previous experiments sentence acceptability was judged (by humans) or predicted (by models) independently of context.
- Here we extend the research to investigate the impact of context on acceptability.
- Context is defined as the full document environment surrounding a sentence.
- Specifically, we want to understand the influence of context on:
  - Human acceptability ratings
  - Model prediction of acceptability

Human Acceptability Ratings in Context

- We perform round-trip translation of sentences (e.g. EN→FR→EN) from English Wikipedia to generate a set of sentences with varying degrees of acceptability.
- We use MTurk to collect acceptability judgements (rated on a 4-point scale).
- Annotation task was run twice: first without context, and second within the document context.
- We collect multiple ratings for a sentence and take the mean.
- Human acceptability ratings:
  - without context $= \mathbf{h}^{-}$;
  - with context $= \mathbf{h}^{+}$

Instructions and guidelines:

To do this HIT you must be a native speaker of English. We have inserted controls to identify this. If your judgments are non-native then your HITs will most likely be rejected.

Please assess each of the following highlighted sentences (in bold) for naturalness using the four-point scale given after each sentence.

Each sentence occurs within an article. You are shown the sentences preceding and following the sentence to rate. You can expand the text in each direction, as you wish, by using the indicated buttons.

Description of scoring

The following scale should be used for scoring:

- 4 = Good
- 3 = Mostly good; a little odd
- 2 = Not very good, but I can understand the gist of the sentence
- 1 = Bad; can't interpret the sentence

Examples of '4':

- Mary had a little lamb

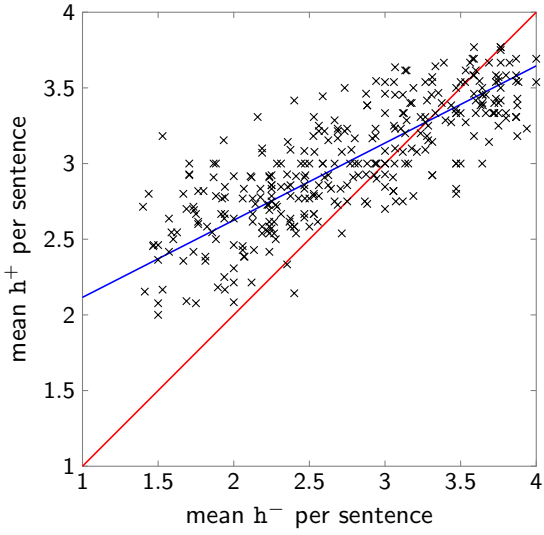
Examples of '1' (incoherent):

- lamb little Mary this had

Note: You must rate all sentences for the work to be approved.

1. Show/Hide preceding context Anthony Philip "Tony" Thirlwall (born 1941) is Professor of Applied Economics at the University of Kent.
He made a significant contribution to the regional economy. Analysis of unemployment and inflation; the theory of balance of payments, in particular the economics of growth and development in developing countries.
He is the author of the bestselling textbook "Economics of Development: Theory and Evidence" (Palgrave Macmillan) now in its ninth edition. Show/Hide following context
○ 1 = bad ○ 2 = not very good ○ 3 = mostly good ○ 4 = good
2. Show/Hide preceding context Anthony Philip "Tony" Thirlwall (born 1941) is Professor of Applied Economics at the University of Kent.
He made a significant contribution to the regional economy. Analysis of unemployment and inflation; the theory of balance of payments, in particular the economics of growth and development in developing countries.
He is the author of the bestselling textbook "Economics of Development: Theory and Evidence" (Palgrave Macmillan) now in its ninth edition. Show/Hide following context
○ 1 = bad ○ 2 = not very good ○ 3 = mostly good ○ 4 = good
3. Show/Hide preceding context Anthony Philip "Tony" Thirlwall (born 1941) is Professor of Applied Economics at the University of Kent.
He made a significant contribution to the regional economy. Analysis of unemployment and inflation; the theory of balance of payments, in particular the economics of growth and development in developing countries.
He is the author of the bestselling textbook "Economics of Development: Theory and Evidence" (Palgrave Macmillan) now in its ninth edition. Show/Hide following context
○ 1 = bad ○ 2 = not very good ○ 3 = mostly good ○ 4 = good
4. Show/Hide preceding context Anthony Philip "Tony" Thirlwall (born 1941) is Professor of Applied Economics at the University of Kent.
He made a significant contribution to the regional economy. Analysis of unemployment and inflation; the theory of balance of payments, in particular the economics of growth and development in developing countries.
He is the author of the bestselling textbook "Economics of Development: Theory and Evidence" (Palgrave Macmillan) now in its ninth edition. Show/Hide following context
○ 1 = bad ○ 2 = not very good ○ 3 = mostly good ○ 4 = good
5. Show/Hide preceding context Anthony Philip "Tony" Thirlwall (born 1941) is Professor of Applied Economics at the University of Kent.
He made a significant contribution to the regional economy. Analysis of unemployment and inflation; the theory of balance of payments, in particular the economics of growth and development in developing countries.
He is the author of the bestselling textbook "Economics of Development: Theory and Evidence" (Palgrave Macmillan) now in its ninth edition. Show/Hide following context
○ 1 = bad ○ 2 = not very good ○ 3 = mostly good ○ 4 = good

With-context $\mathtt{h}^+$ Against Without-context $\mathtt{h}^-$ Ratings



Observations

- Pearson's $r = 0.80$ between $\mathtt{h}^+$ and $\mathtt{h}^-$.
- Context boosts acceptability ratings most for ill-formed sentences.
- Surprisingly, context reduces acceptability for the most acceptable sentences.
- Context "compresses" distribution of ratings.
- One-vs-rest correlation, performance of a single annotator against the rest: 0.628 for $\mathtt{h}^-$ and 0.293 for $\mathtt{h}^+$.
- Low correlation is explained by the compression effect of context — good and bad sentences are now less separable.

Modelling Acceptability with Unsupervised Models

- `lstm`: standard LSTM language model
- `tdlm`: a topically driven language model; language model is driven by a topic vector automatically learnt on the document context.
- 4 variants at **test time**:
  - Use only the sentence as input: $\texttt{lstm}^{-}$ and $\texttt{tdlm}^{-}$;
  - Use both sentence and context as input: $\texttt{lstm}^{+}$ and $\texttt{tdlm}^{+}$.
- $\texttt{lstm}^{+}$ incorporates context by feeding it to the LSTM network and taking the final state as the initial state for the current sentence.
- Models trained on 100K English Wikipedia articles (40M tokens).

Acceptability Measures

To map sentence probability to acceptability, we compute several **acceptability measures**, which are designed to normalise sentence length and word frequency.

$$SLOR = \frac{\log P - \log U}{L}$$

- P = probability of the sentence given by a model;
- U = unigram probability of the sentence;
- L = sentence length

Results

	$\texttt{lstm}^-$	$\texttt{lstm}^+$	$\texttt{tdlm}^-$	$\texttt{tdlm}^+$
$\texttt{h}^-$	0.584	0.633	0.640	0.653
$\texttt{h}^+$	0.503	0.546	0.557	0.568

- Across all models (`lstm` or `tdlm`) and human ratings ($\texttt{h}^-$ or $\texttt{h}^+$), using context at test time improves performance.
- `tdlm` consistently outperforms `lstm` (even $\texttt{tdlm}^- > \texttt{lstm}^+$).
- Lower correlation when predicting sentence acceptability judged with context.
- It suggests $\texttt{h}^+$ ratings are more difficult to predict than $\texttt{h}^-$, which corresponds to the low one-vs-rest human performance.

Summary

- Context positively influences acceptability, particularly for ill-formed sentences.
- But it also has the reverse effect for well-formed sentences.
- Incorporating context (during training or testing) helps modelling acceptability.
- Prediction performance declines when tested on acceptability ratings judged with context, due to the "compression" effect of ratings.
- Future work: investigate why context reduces acceptability for highly acceptable sentences.

Questions?