Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Not All Dialogues are Created Equal:

Instance Weighting for Neural Conversational Models

Pierre Lison & Serge Bibauw

Norwegian Computing Center

KULeuven, imec
& UCLouvain

SIGDIAL 2017

Neural models of dialogue

- Increasing popularity of **neural architectures** for the development of conversational agents
  - ⊕ Can be learned from *raw dialogues,* without needing much domain knowledge or feature engineering
  - ⊖ Requires *large amounts of training data* to learn good conversation models (large parameter space)

↓

Typically large online resources such as Twitter discussions, technical web fora, online chat logs, movie scripts or subtitles, etc.

These resources are undeniably useful, but also face some limitations in terms of dialogue modelling

Some limitations

- Several dialogue corpora, most notably those extracted from movie & TV subtitles, do not include any explicit turn segmentation or speaker identification

1	If we wanted to kill you, Mr Holmes, we would have done it by now.	01:17:34.76	01:17:37.75
2	We just wanted to make you inquisitive.	01:17:37.80	01:17:40.59
3	Do you have it?	01:17:42.40	01:17:43.91
4	Do I have what?	01:17:43.91	01:17:45.43
5	The treasure.	01:17:45.48	01:17:46.43
6	I don't know what you're talking about.	01:17:46.43	01:17:48.91
7	I would prefer to make certain.	01:17:48.96	01:17:52.03
8	Everything in the West has its price.	01:17:57.00	01:17:59.63
9	And the price for her life - information.	01:17:59.68	01:18:04.55

Some limitations

- The dialogues may also contain references to named entities (in particular, fictional characters)

1	If we wanted to kill you, Mr Holmes, we would have done it by now.	01:17:34.76	01:17:37.75
2	We just wanted to make you inquisitive.	01:17:37.80	01:17:40.59
3	Do you have it?	01:17:42.40	01:17:43.91
4	Do I have what?	01:17:43.91	01:17:45.43
5	The treasure.	01:17:45.48	01:17:46.43
6	I don't know what you're talking about.	01:17:46.43	01:17:48.91
7	I would prefer to make certain.	01:17:48.96	01:17:52.03
8	Everything in the West has its price.	01:17:57.00	01:17:59.63
9	And the price for her life - information.	01:17:59.68	01:18:04.55

Key idea

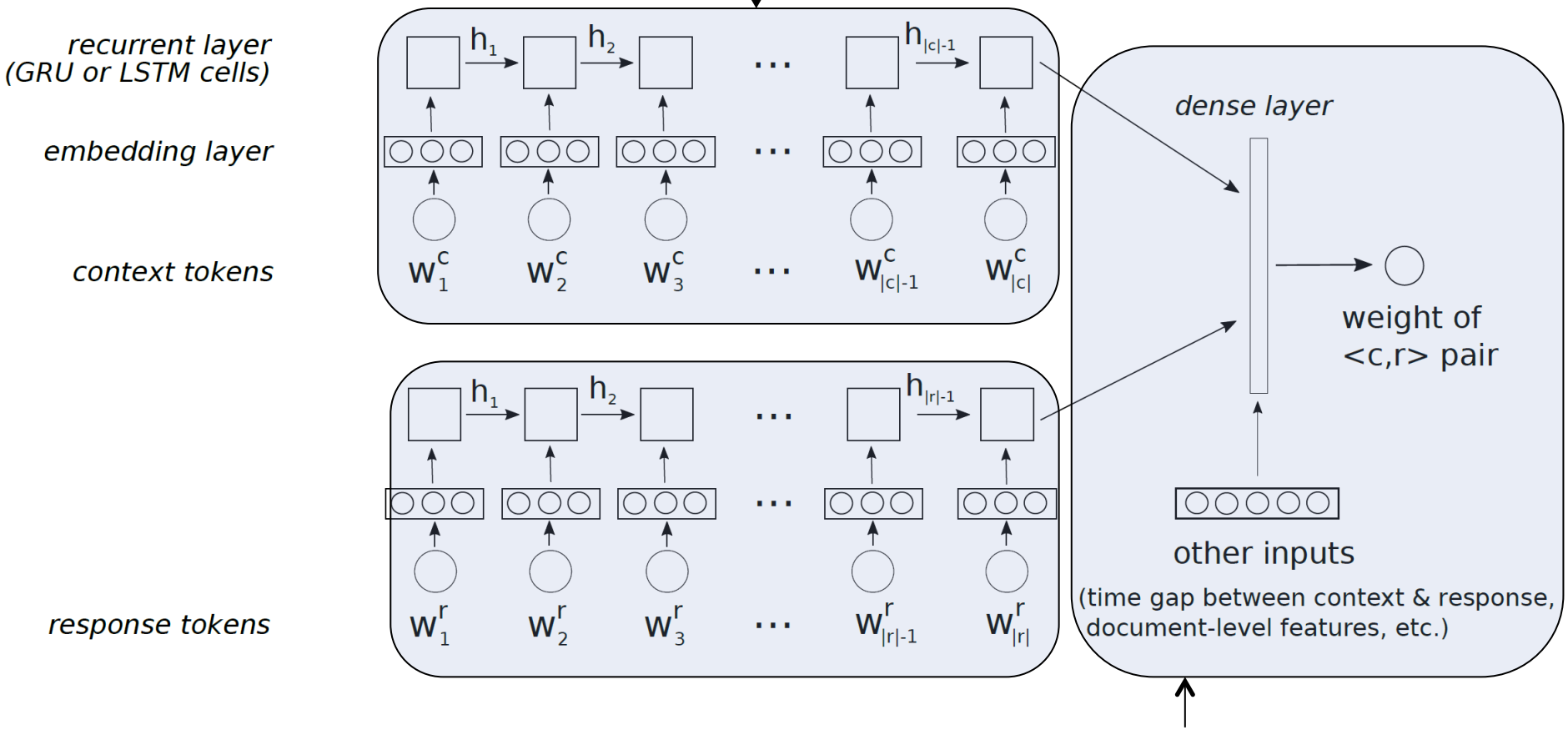
- The examples of <*context*, *response*> pairs are not all equally useful or relevant for the conversation models
  - Some might even be detrimental
- Can be viewed as a *domain adaptation problem*:
  - Discrepancy between the <*context*, *response*> pairs observed in the dialogue data and the ones we wish to encode in the neural conversation model
- Proposed solution: add a **weighting model**
  - Maps each<*context*, *response*> pair to a weight value
  - These weights are then used at learning time to quantify the importance of each training example

Weighting model

- How to assign weights to <*context*, *response*> pairs?
  - Annotating each pair manually is not feasible
  - Handcrafted rules are also difficult to apply, since the "quality" of examples may depend on multiple factors
- **Data-driven approach**: *learn* a weighting model from examples of high-quality responses
  - What constitutes a "high-quality response" may depend on the type of conversation model one wishes to build
  - The weighting model also uses a neural architecture

Weighting model

Two recurrent neural networks with shared weights



Dense layer combining the output vectors of the two sequences (+ additional features if available), and outputting a weight value for the pair

Instance weighting

- Once the weighting model is estimated, it can be used to assign each training example <c_i, r_i> to a weight w_i
- The weights are then included in the empirical loss to minimise when training the neural conversation model
  - For retrieval-based models:

$$\boldsymbol{\theta}^* = min_{\boldsymbol{\theta}} \sum_{1}^{n} w_i \, L(y_i, f(c_i, r_i; \boldsymbol{\theta}))$$

  - For generative models:

$$\boldsymbol{\theta}^* = min_{\boldsymbol{\theta}} \sum_{1}^{n} w_i \, L(r_i, f(c_i; \boldsymbol{\theta}))$$

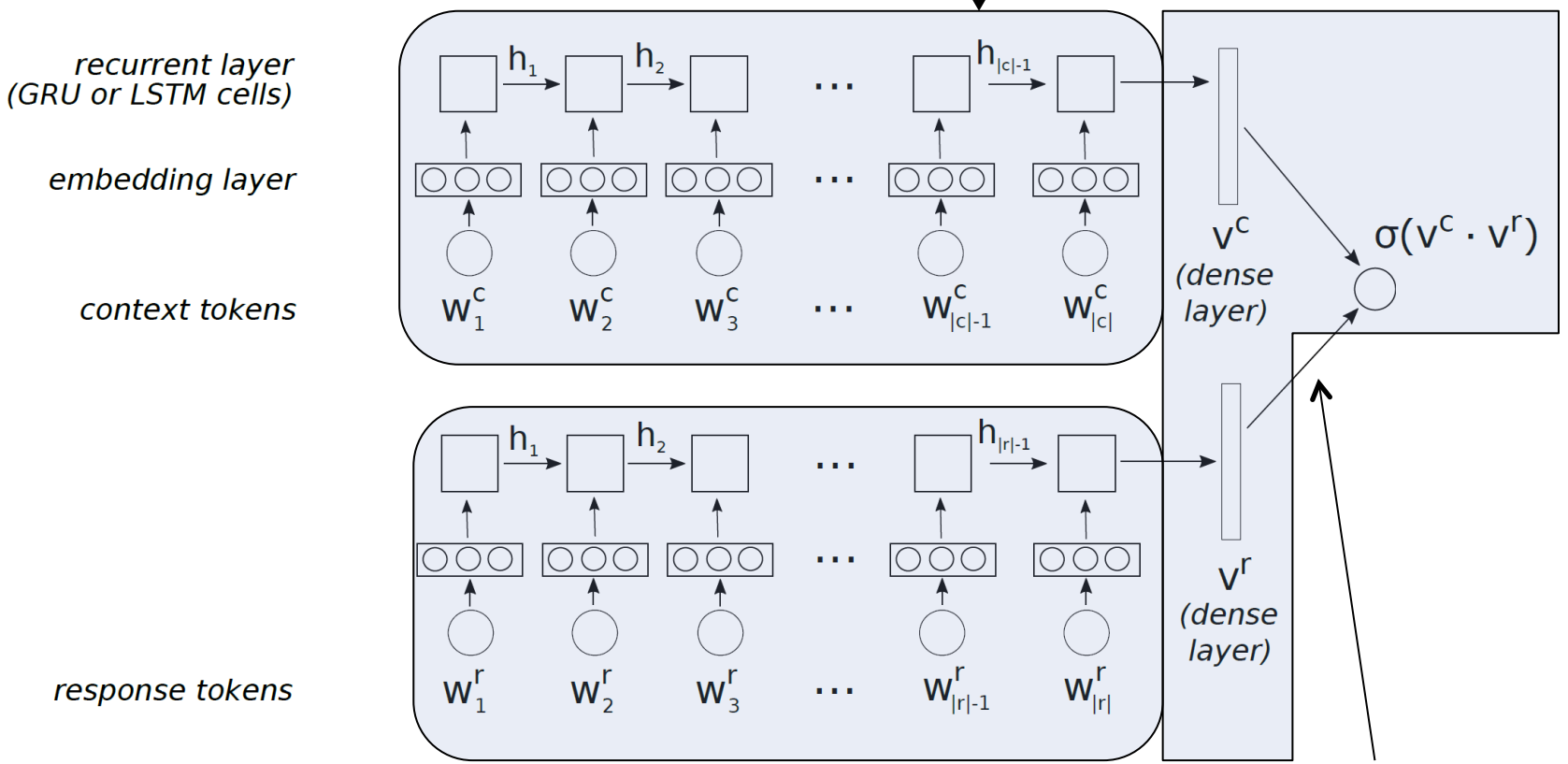
(in both cases, *L* is the loss function and *f* is the output of the conversation model)

Evaluation

- Evaluation with *retrieval-based* neural models
  - Selection of most relevant response for a given context amongst a set of possible ones (*score* each <c, r> pair)
  - Training data: English subtitles from OpenSubtitles
- Comparison of three conversation models:
  - TF-IDF model
  - Dual Encoder model with uniform weights
  - Dual Encoder model combined with a weighting model
- Both automatic evaluation (using the Recall_m@i metric) as well as human evaluation using crowdsourcing

Dual Encoder models

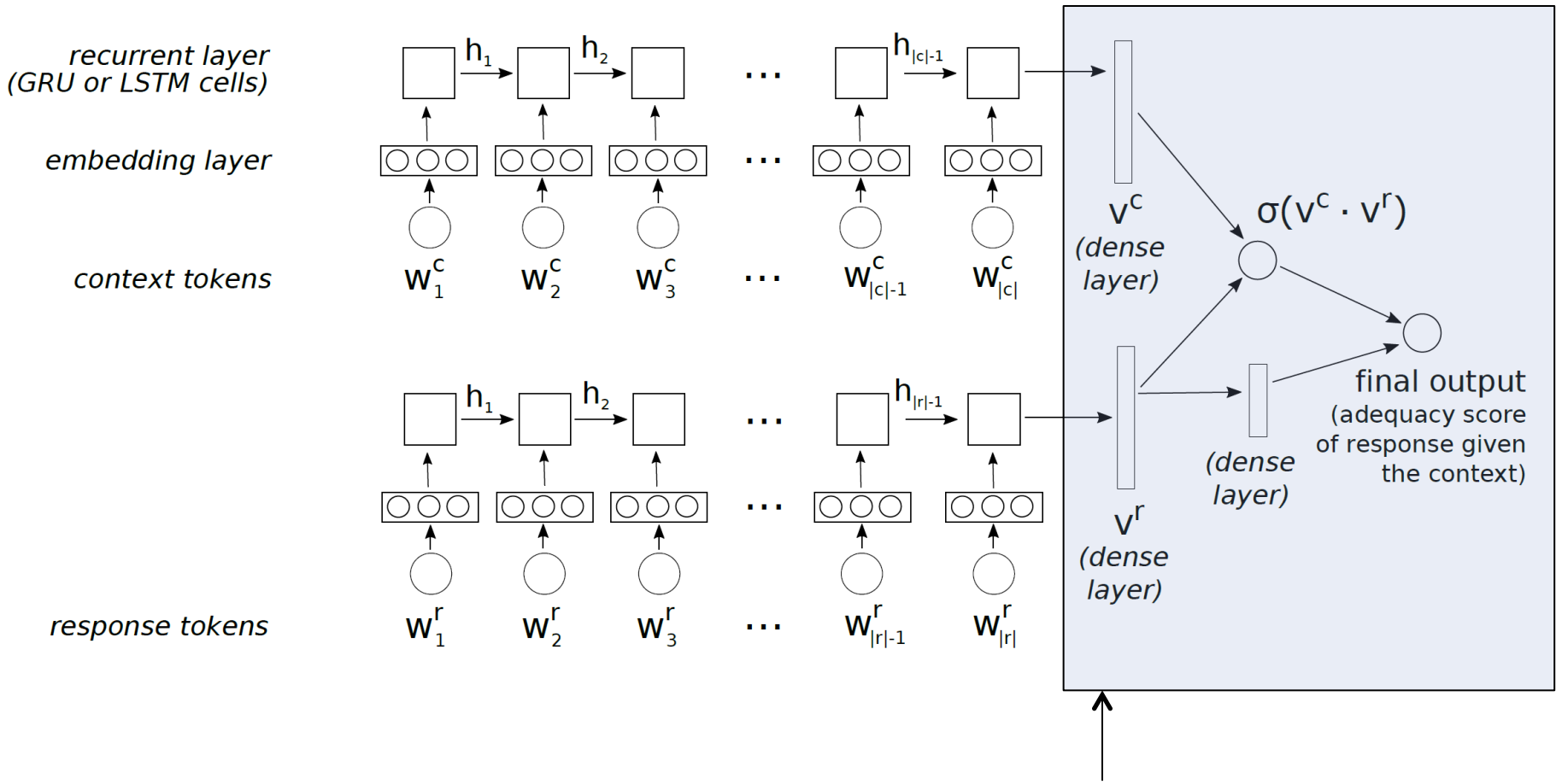
Again, two recurrent networks with shared weights



Dot product between the latent representations of the two sequences

[R. Lowe et al. 2017. Training end-to-end dialogue systems with the Ubuntu Dialogue Corpus. Dialogue & Discourse 8(1):31–65.]

Dual Encoder models



Small modification to original Dual Encoder model:
Final score also depends on the response itself

Evaluation – data sets

- **Training data**: English-language portion of the OpenSubtitles corpus of movie and TV subtitles, composed of 105 445 subtitles and 95.5 million utterances

[P. Lison and J. Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie & TV subtitles. In LREC 2016.]

Evaluation – data sets

- **Training data**: English-language portion of the OpenSubtitles corpus of movie and TV subtitles, composed of 105 K subtitles and 95.5 M utterances
- **High-quality responses** (for weighting model):
  - Extracted from subset of subtitles for which the turn structure is known (through alignment with movie scripts)
  - *Heuristic 1*: only keep responses that introduce a new dialogue turn and appear in two-party conversations
  - *Heuristic 2*: filter out responses containing fictional character names and out-of-vocabulary words
  - Resulting dataset: 96 K ⟨context, response⟩ pairs

[Pierre Lison and Raveesh Meena. 2016. Automatic turn segmentation of movie & TV subtitles. In SLT 2016.]

Evaluation – data sets

- **Test data**:
  - Cornell Movie Dialog Corpus, a collection of fictional conversations extracted from movie scripts (67 Kpairs)
  - Small corpus of 62 theatre plays from the web (3 K pairs)
- Preprocessing:
  - All sentences were tokenised and POS-tagged
  - Named entities were replaced by generic tags
  - Vocabulary capped to 25 000 words

[C. Danescu-Niculescu-Mizil and L. Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2016.]

Evaluation – experimental design

- Input and output:
  - Contexts limited to the last 10 utterances preceding the response and a maximum of 60 tokens.
  - Responses limited to a maximum of 5 utterances (in case of multi-utterance turns) and 30 tokens.
  - 1:1 ratio between positive examples (actual pairs observed in corpus) and negative examples drawn at random
- Training details:
  - Embedding layers of dim=300
  - GRU cells used for recurrent layer, with output dim=400
  - Batch size=256 and RMSProp used for the optimisation
  - Dropout of 0.2 applied to all layers

Evaluation results

Model name	Cornell Movie Dialogs			Theatre plays		
	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
TF-IDF	0.33	0.44	0.67	0.33	0.44	0.53
Dual Encoder	0.44	0.62	0.83	0.52	0.67	0.75
Dual Encoder + weighting	**0.47**	**0.63**	**0.85**	**0.56**	**0.70**	**0.80**

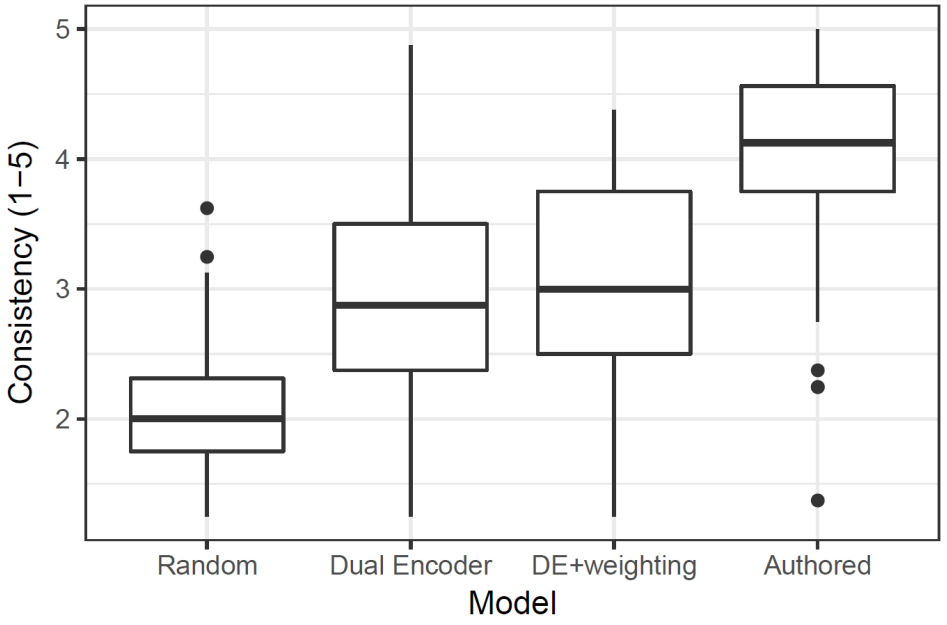
- The three models evaluated with the $Recall_m@i$ metric
- Dual Encoder model combined with weighting model outperforms the two baselines on both test sets.
- Weighting model gives more importance to ”cohesive” adjacency pairs between context and response

Human evaluation

- Human evaluation of the responses generated by the two Dual Encoder models,
  - 115 random contexts drawn from the Cornell Corpus
  - 4 possible responses: random, responses from the two Dual Encoder models, and expert response
- The resulting 460 pairs were each evaluated by 8 distinct human judges, who were asked to rate the *consistency* between context and response on a 5-points scale
  - 118 individuals participated in the evaluation through a crowdsourcing platform

Human evaluation

- **Inconclusive results**: no statistically significant difference between the two models (Wilcoxon rank-sum test)
  - Very low agreement between the evaluation participants (Krippendorff's α = 0.36).
- Difficulty for the raters to discriminate between responses
  - Probably due to the nature of the corpus, which is heavily dependent on an external context (the movie scenes)



Conclusion

- Large dialogue corpora can include many ”noisy examples”
  - Not all examples have the same quality/relevance when learning neural conversation models
- Possible remedy: include a **weighting model**
  - Can be seen as a form of domain adapation
  - The weighting model can itself be learned from examples (based on their adherence to certain quality criteria)
  - Can be applied to any data-driven conversation model
- Future work: extend it to generative neural models
  - Possible performance benefits?

Question time

- *Why not train on high-quality responses directly?*
  - This would restrict the training set to a subset for which we can explicitly determine a quality score
  - In the evaluation, we could do this for about 0.1% of the subtitles (for which the turn structure was known in advance thanks to alignments with movie scripts)
  - The weighting model enables us to continue using the full, noisy training set
  - But also assign higher weights to the ⟨c, r⟩ pairs whose latent representations are close to the high-quality examples, and a lower weight for those further away.