# Not All Dialogues are Created Equal: Instance Weighting for Neural Conversational Models

**Pierre Lison**
Norwegian Computing Center
Oslo, Norway
plison@nr.no

**Serge Bibauw***
KU Leuven, imec
Université catholique de Louvain
serge.bibauw@kuleuven.be



## Abstract

Neural conversational models require substantial amounts of dialogue data to estimate their parameters and are therefore usually learned on large corpora such as chat forums, Twitter discussions or movie subtitles. These corpora are, however, often challenging to work with, notably due to their frequent lack of turn segmentation and the presence of multiple references external to the dialogue itself. This paper shows that these challenges can be mitigated by adding a *weighting model* into the neural architecture. The weighting model, which is itself estimated from dialogue data, associates each training example to a numerical weight that reflects its intrinsic quality for dialogue modelling. At training time, these sample weights are included into the empirical loss to be minimised. Evaluation results on retrieval-based models trained on movie and TV subtitles demonstrate that the inclusion of such a weighting model improves the model performance on unsupervised metrics.

## 1 Introduction

The development of conversational agents (such as mobile assistants, chatbots or interactive robots) is increasingly based on data-driven methods aiming to infer conversational patterns from dialogue data. One major trend in the last recent years is the emergence of neural conversation models (Vinyals and Le, 2015; Sordoni et al., 2015; Shang et al., 2015; Serban et al., 2016; Lowe et al., 2017; Li et al., 2017). These neural models can be directly estimated from raw (non-annotated) dialogue corpora, allowing them to be deployed with a limited amount of domain-specific knowledge and feature engineering.

Due to their large parameter space, the estimation of neural conversation models requires considerable amounts of dialogue data. They are therefore often trained on conversations collected from various online resources, such as Twitter discussions (Ritter et al., 2010) online chat logs (Lowe et al., 2017), movie scripts (Danescu-Niculescu-Mizil and Lee, 2011) and movie and TV subtitles (Lison and Tiedemann, 2016).

Although these corpora are undeniably useful, they also face some limitations from a dialogue modelling perspective. First of all, several dialogue corpora, most notably those extracted from subtitles, do not include any explicit turn segmentation or speaker identification (Serban and Pineau, 2015; Lison and Meena, 2016). In other words, we do not know whether two consecutive sentences are part of the same dialogue turn or were uttered by different speakers. The neural conversation model may therefore inadvertently learn responses that remain within the same dialogue turn instead of starting a new turn.

Furthermore, these dialogues contain multiple references to named entities (in particular, person names such as fictional characters) that are specific to the dialogue in question. These named entities should ideally not be part of the conversation model, since they often draw on an external context that is absent from the inputs provided to the conversation model. For instance, the mention of character names in a movie is associated with a visual context (for instance, the characters appearing in a given scene) that is not captured in the training data. Finally, a substantial portion of the utterances observed in these corpora is made of neutral, commonplace responses ("*Perhaps*", "*I*

* Also affiliated with Universidad Central del Ecuador (Quito, Ecuador).

*don't know*", "*Err*", ...) that can be used in most conversational situations but fall short of creating meaningful and engaging conversations with human users (Li et al., 2016a).

The present paper addresses these limitations by adding a *weighting model* to the neural architecture. The purpose of this model is to associate each $\langle context, response \rangle$ example pair to a numerical *weight* that reflects the intrinsic "quality" of each example. The instance weights are then included in the empirical loss to minimise when learning the parameters of the neural conversation model. The weights are themselves computed via a neural model learned from dialogue data. Experimental results demonstrate that the use of instance weights improves the performance of neural conversation models on unsupervised metrics. Human evaluation results are, however, inconclusive.

The rest of this paper is as follows. The next section presents a brief overview of existing work on neural conversation models. Section 3 provides a description of the instance weighting approach. Section 4 details the experimental validation of the proposed model, using both unsupervised metrics and a human evaluation of the selected responses. Finally, Section 5 discusses the advantages and limitations of the approach, and Section 6 concludes this paper.

## 2 Related Work

Neural conversation models are a family of neural architectures (generally based on deep convolutional or recurrent networks) used to represent mappings between dialogue contexts (or queries) and possible responses. Compared to previous statistical approaches to dialogue modelling based on Markov processes (Levin et al., 2000; Rieser and Lemon, 2011; Young et al., 2013), one benefit of these neural models is their ability to be estimated from raw dialogue corpora, without having to rely on additional annotation layers for intermediate representations such as state variables or dialogue acts. Rather, neural conversation models automatically *derive* latent representations of the dialogue state based on the observed utterances.

Neural conversation models can be divided into two main categories, *retrieval models* and *generative models*. Retrieval models are used to select the most relevant response for a given context amongst a (possibly large) set of predefined responses, such as the set of utterances extracted from a corpus (Lowe et al., 2015; Prakash et al., 2016). Generative models, on the other hand, rely on sequence-to-sequence models (Sordoni et al., 2015) to generate new, possibly unseen responses given the provided context. These models are built by linking together two recurrent architectures: one encoder which maps the sequence of input tokens in the context utterance(s) to a fixed-sized vector, and one decoder that generates the response token by token given the context vector (Vinyals and Le, 2015; Sordoni et al., 2015). Recent papers have shown that the performance of these generative models can be improved by incorporating attentional mechanisms (Yao et al., 2016) and accounting for the structure of conversations through hierarchical networks (Serban et al., 2016). Neural conversation models can also be learned using adversarial learning (Li et al., 2017). In this setting, two neural models are jointly learned: a generative model producing the response, and a discriminator optimised to distinguish between human-generated responses and machine-generated ones. The discriminator outputs are then used to bias the generative model towards producing more human-like responses.

The linguistic coherence and diversity of the models can be enhanced by including speaker-addressee information (Li et al., 2016b) and by expressing the objective function in terms of Maximum Mutual Information to enhance the diversity of the generated responses (Li et al., 2016a). As demonstrated by (Ghazvininejad et al., 2017), neural conversation models can also be combined with external knowledge sources in the form of factual information or entity-grounded opinions, which is an important requirement for developing task-oriented dialogue systems that must ground their action in an external context.

Dialogue is a sequential decision-making process where the conversational actions of each participant influence not only the current turn but the long-term evolution of the dialogue (Levin et al., 2000). To incorporate the prediction of future outcomes in the generation process, several papers have explored the use of reinforcement learning techniques, using deep neural networks to model the expected future reward (Li et al., 2016c; Cuayáhuitl, 2017). In particular, the Hybrid Code Networks model of (Williams et al., 2017) demonstrate how a mixture of supervised learning, reinforcement learning and domain-specific knowl-

edge can be used to optimise dialogue strategies from limited amount of training data.

In contrast with the approaches outlined above, this paper does not present a new neural architecture for conversational models. Rather, it investigates how the performance of existing models can be improved "upstream", by adapting how these models can be trained on large, noisy corpora with varying levels of quality. It should be noted that, although the experiments presented in Section 4 focus on a limited range of neural models, the approach presented in this paper is designed to be model-independent and can be applied as a preprocessing step to any data-driven model of dialogue.

## 3 Approach

As mentioned in the introduction, the interactions extracted from large dialogue corpora do not all have the same intrinsic quality, due for instance to the frequent lack of turn segmentation or the presence of external, unresolvable references to person names. In other words, there is a discrepancy between the actual $\langle$context, response$\rangle$ pairs found in these corpora and the conversational patterns that should be accounted for in the neural model.

One way to address this discrepancy is by framing the problem as one of *domain adaptation*, the source domain being the original dialogue corpus and the target domain representing the dialogues we want our model to produce. The target domain is in this case not necessarily another dialogue domain, but simply reflects the fact that the distribution of responses in the raw corpus does not necessarily reflect the distribution of responses we ultimately wish to encode in the conversational model.

A popular strategy for domain adaptation in natural language processing, which has notably been used in POS-tagging, sentiment analysis, spam filtering and machine translation (Bickel et al., 2007; Jiang and Zhai, 2007; Foster et al., 2010; Xia et al., 2013), is to assign a higher weight to training instances whose properties are similar to the target domain. We present below such an instance weighting approach tailored for neural conversational models.

### 3.1 Weighting model

The quality of a particular $\langle$context, response$\rangle$ pair is difficult to determine using handcrafted rules – for instance, the probability of a turn boundary may depend on multiple factors such as the presence of turn-yielding cues or the time gap between the utterances (Lison and Meena, 2016). To overcome these limitations, we adopt a data-driven approach and automatically learn a weighting model from examples of "high-quality" responses. What constitutes a high-quality response depends in practice on the specific criteria we wish to uphold in the conversation model – for instance, favouring responses that are likely to form a new dialogue turn (rather than a continuation of the current turn), avoiding the use of dull, commonplace responses, or disfavouring the selection of responses that contain unresolved references to person names.

The weighting model can be expressed as a neural model which associates each $\langle$context, response$\rangle$ example pair to a numerical weight. The architecture of this neural network is depicted in Figure 1. It is composed of two recurrent sub-networks with shared weights, one for the context and one for the response. Each sub-network takes a sequence of tokens as input and pass them through an embedding layer and a recurrent layer with LSTM or GRU cells. The fixed-size vectors for the context and response are then fed to a regular densely-connected layer, and finally to the final weight value through a sigmoid activation function. Additional features can also be included whenever available – for instance, timing information for movie and TV subtitles (such as the duration gap between the context and its response, in milliseconds), or document-level features such as the dialogue genre or the total duration of the dialogue.

To estimate its parameters, the neural model is provided with positive examples of "high-quality" responses along with negative examples sampled at random from the corpus. Based on this training data, the network learns to assign higher weights to the $\langle$context, response$\rangle$ pairs whose output vectors (combined with the additional inputs) are close from the high-quality examples, and a lower weight for those further away. In practice, the selection of high-quality example pairs from a given corpus can be performed through a combination of simple heuristics, as detailed in Section 4.1.

### 3.2 Instance weighting

Once the weighting model is estimated, the next step is to run it on the entire dia-

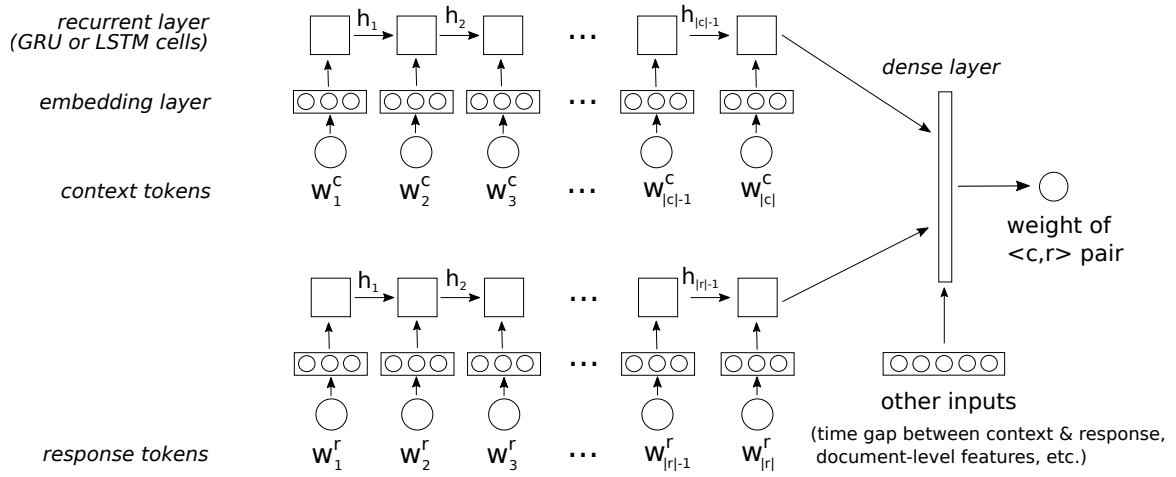


Figure 1: Neural weighting model, taking as input the ⟨context, response⟩ pairs, possibly along additional features (such as timing information for subtitles), and returning an associated weight value.

logue corpus to compute the expected weight of each ⟨context, response⟩ pair. These sample weights are then included in the empirical loss that is being minimised during training. Formally, assuming a set of context-response pairs $\{(c_1, r_1), (c_2, r_2), ...(c_n, r_n)\}$ with associated weights $\{w_1, ...w_n\}$, the estimation of the model parameters $\boldsymbol{\theta}$ is expressed as a minimisation problem. For retrieval models, this minimisation is expressed as:

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

where $L$ is a loss function (for instance, the cross-entropy loss), and $y_i$ is set to either 1 if $r_i$ is the response to $c_i$, and 0 otherwise (when $r_i$ is a negative example). For generative models, the minimisation is similarly expressed as:

$$\boldsymbol{\theta}^* = min_{\boldsymbol{\theta}} \sum_{1}^{n} w_i \, L(r_i, f(c_i; \boldsymbol{\theta})) \quad (2)$$

In both cases, the loss computed from each example pair is multiplied by the weight value determined by the weight model. Examples associated with a larger weight $w_i$ will therefore have a larger influence on the gradient update steps.

## 4 Evaluation

The approach is evaluated on the basis of retrieval-based neural models trained on English-language subtitles from (Lison and Tiedemann, 2016). Three alternative models are evaluated:

1. A traditional TF-IDF model,
2. A Dual Encoder model trained directly on the corpus examples,
3. A Dual Encoder model combined with the weighting model from Section 3.1.

### 4.1 Models

**TF-IDF model**

The TF-IDF (Term Frequency - Inverse Document Frequency) model computes the similarity between the context and its response using methods from information retrieval (Ramos, 2003). TF-IDF measures the importance of a word in a "document" (in this case the context or response) relative to the whole corpus. The model transforms the context and response (represented as bag-of-words) into TF-IDF-weighted vectors. These vectors are sparse vectors of a size equivalent to the vocabulary size, where each row corresponds, if the given word is present in the context or response, to its TF-IDF weight, and is 0 otherwise. The matching score between the context and its response is then determined as the cosine similarity between the two vectors:

$$similarity = \frac{v^c \cdot v^r}{\|v^c\|_2 \, \|v^r\|_2} \quad (3)$$

where $v^c$ and $v^r$ respectively denote the TF-IDF-weighted vectors for the context and response.

**Dual Encoder**

The Dual Encoder model (Lowe et al., 2017) consists of two recurrent networks, one for the context and one for the response. The tokens are first

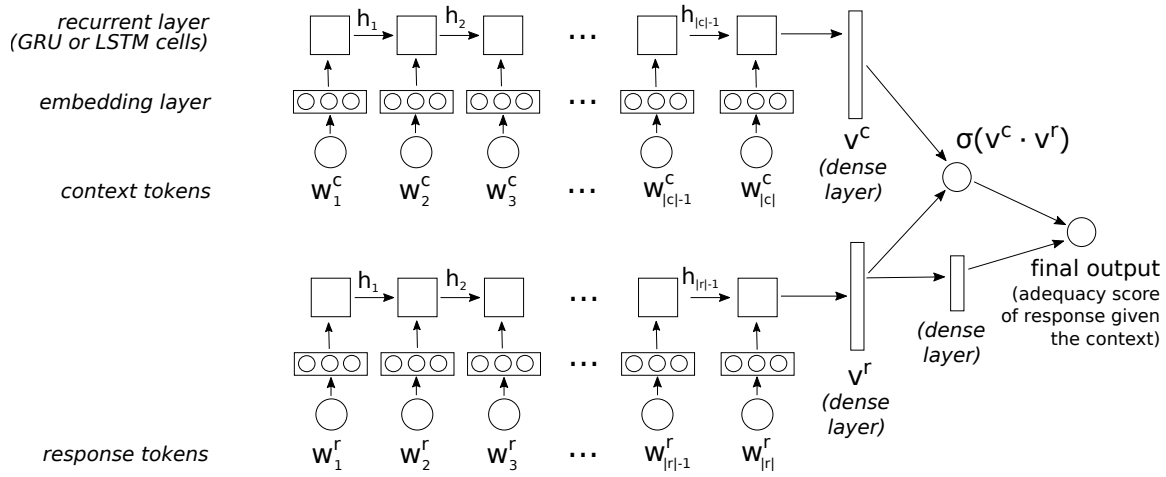


Figure 2: Dual encoder model, taking as input the ⟨context, response⟩ pairs and returning a score expressing the adequacy of the response given the context.

passed through an embedding layer and then to a recurrent layer with LSTM or GRU cells. In the original formalisation of this model (Lowe et al., 2015), the context vector is transformed through a dense layer of same dimension, representing the "predicted" response. The inner product of the predicted and actual responses is then calculated and normalised, yielding a similarity score. This model, however, only seeks to capture the semantic similarity between the two sequences, while the selection of the most adequate response in a given context may also need to account for other factors such as the grammaticality and coherence of the response. We therefore extend the Dual Encoder model in two ways. First, both the context and response vectors are transformed through a dense layer at the end of the recurrent layer (instead of just the context vector). Second, the final prediction is connected to both the inner product of the two vectors and to the response vector itself, as depicted in Figure 2.

**Dual Encoder with instance weighting**

Finally, the third model relies on the exact same Dual Encoder model as above, but applies the weighting model described in Section 3.1 prior to learning in order to assign weights to each training example. The weighting model is estimated on a subset of the movie and TV subtitles augmented with speaker information and filtered through heuristics to ensure a good cohesion between the context and its response. These heuristics are detailed in the next section.

Although the architecture of the Dual Encoder is superficially similar to the weighting model of Figure 1, the two models serve a different purpose: the weighting model returns the expected *quality* of a training example, while the Dual Encoder returns a score expressing the *adequacy* between the context and the response.

### 4.2 Datasets

**Training data for the conversation models**

The dataset used for training the three retrieval models is the English-language portion of the OpenSubtitles corpus of movie and TV subtitles (Lison and Tiedemann, 2016). The full dataset is composed of 105 445 subtitles and 95.5 million utterances, each utterance being associated with a start and end time (in milliseconds).

**Training data for the weighting model**

For training the weighting model, we extracted a small subset of the full corpus of subtitles corresponding to ⟨context, response⟩ pairs satisfying specific quality criteria. The first step was to align at the sentence level the subtitles with an online collection of movie and TV scripts (1 069 movies and 6 398 TV episodes), following the approach described in (Lison and Meena, 2016).

This alignment enabled us to annotate the subtitles with speaker names and turn boundaries. Based on these subtitles, we then selected example pairs with two heuristics:

1. To ensure the response constitutes an actual reply from another speaker and not simply a continuation of the current turn, the

subtitles were segmented into sub-dialogues. ⟨context, response⟩ pairs including a change of speaker from the context to the response were then extracted from these sub-dialogues. Since multi-party dialogues make it harder to determine who replies to whom, only sub-dialogues with two participants were considered in the subset.

2. To ensure the response is intelligible given the context (without drawing on unresolved references to e.g. fictional person names), we also filtered out from the subset the dialogue turns including mentions of fictional character names and out-of-vocabulary words.

A total of 95 624 ⟨context, response⟩ pairs can be extracted using these two heuristics. This corresponds to about 0.1 % of the total number of examples for the OpenSubtitles corpus. These pairs are used as positive examples for the weighting model, along with negative pairs sampled at random from the corpus.

**Test data**

Two distinct corpora are used as test sets for the evaluation. The first corpus, whose genre is relatively close to the training set, is the Cornell Movie Dialog Corpus (Danescu-Niculescu-Mizil and Lee, 2011), which is a collection of fictional conversations extracted from movie scripts (unrelated to the ones used for training the weighting model). The transcripts from this corpus are segmented into conversations. Each conversation is represented as a sequence of dialogue turns. As this paper concentrates on the selection of relevant responses in a given context, we limited the test pairs to the ones where the context ends with a question, which yields a total of 67 305 ⟨context, response⟩ pairs.

The second test set comes from a slightly different conversational genre, namely theatre plays. The scripts of 62 English-language theatre plays were downloaded from public websites. We also limited the test pairs to the pairs where the context ends with a question, for a total of 3 427 pairs.

#### 4.2.1 Experimental design

**Preprocessing**

The utterances from all datasets were tokenised, lemmatised and POS-tagged using the spaCy NLP library[1]. We also ran the named entity recogniser from the same library to extract named entities. Since the person names mentioned in movies and theatre plays typically refer to fictional characters, we replaced their occurrences by tags, one distinct tag per entity. For instance, the pair:

**Dana**: Frank, do you think you could give me a hand with these bags?
**Frank**: I'm not a doorman, Miss Barrett. I'm a building superintendent.

is simplified as:

**Dana**: <person1>, do you think you could give me a hand with these bags?
**Frank**: I'm not a doorman, <person2>. I'm a building superintendent.

Named entities of locations and numbers are also replaced by similar tags. To account for the turn structure, turn boundaries were annotated with a <newturn> tag. The vocabulary is capped to 25 000 words determined from their frequency in the training corpus. Tokens not covered in this vocabulary are replaced by <unknown>.

**Training details**

The dialogue contexts were limited to the last 10 utterances preceding the response and a maximum of 60 tokens. The responses were defined as the next dialogue turn after the context, and limited to a maximum of 5 utterances and 30 tokens.

The embedding layers of the Dual Encoders were initialised with Skip-gram embeddings trained on the OpenSubtitles corpus. For the recurrent layers, we tested the use of both GRU and LSTM cells, along with their bidirectional equivalents (Chung et al., 2014), without noticeable differences in accuracy. As GRU cells are faster to train than LSTM cells, we opted for the use of GRU-based recurrent layers. The dimensionality of the output vectors from the recurrent layers was 400. The neural networks are trained with a batch size of 256, binary cross-entropy as cost function and RMSProp as optimisation algorithm. To avoid overfitting issues, a dropout of 0.2 was applied at all layers of the neural model.

Both the weighting model and the Dual Encoder models were training with a 1:1 ratio between positive examples (actual ⟨ context, response ⟩ pairs) and negative examples with a response sampled at random from the training set.

[1] https://spacy.io/

| 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** |

Table 1: Performance of the 3 retrieval models on the two test sets, namely the Cornell Movie Dialogs Dataset and the smaller dataset of theatre plays, using the $\text{Recall}_{10}@i$ metric.

### 4.3 Results

The three models (the TF-IDF model, the baseline Dual Encoder and the Dual Encoder combined with the weighting model) are evaluated using the $\text{Recall}_m@i$ metric, which is the most common metric for the evaluation of retrieval-based models. Let $\{\langle c_i, r_i\rangle, 1 \leq i \leq n\}$ be the list of $m$ context-response pairs from the test set. For each context $c_i$, we create a set of $m$ alternative responses, one response being the actual response $r_i$, and the $m$–1 other responses being sampled at random from the same corpus. The $m$ alternative responses are then ranked based on the output from the conversational model, and the $\text{Recall}_m@i$ measures how often the correct response appears in the top $i$ results of this ranked list. The $\text{Recall}_m@i$ metric is often used for the evaluation of retrieval models as several responses may be equally "correct" given a particular context.

The experimental results are shown in Table 1. As detailed in the table, the Dual Encoder model combined with the weighting model outperforms the Dual Encoder baseline on both test sets (the Cornell Movie Dialogs corpus and the smaller corpus of theatre plays). Our hypothesis is that the weighting model biases the responses selected by the conversation model towards more cohesive adjacency pairs between context and response[2].

Figure 3 illustrates the learning curve for the two Dual Encoder models, where the accuracy is measured on a validation set composed of the high-quality example pairs described in the previous section along with randomly sampled alternative responses (using a 1:1 ratio of positive vs. negative examples). We can observe that the Dual Encoder with instance weights outperforms the baseline model on this validation set – which is not *per se* a surprising result, since the purpose of the weighting model is precisely to bias the conversation model to give more importance to these types of example pairs.

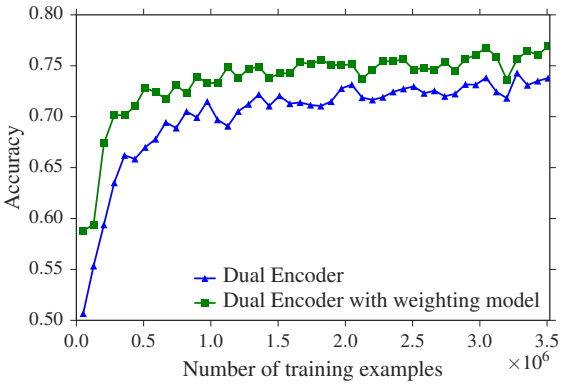


Figure 3: Learning curve for the two Dual Encoder models, showing the evolution of their accuracy on the validation set as a function of the number of observed training examples.

### 4.4 Human evaluation

To further investigate the potential of this weighting strategy for neural conversational models, we conducted a human evaluation of the responses generated by the two neural models included in the evaluation. We collected human judgements on $\langle$context, response$\rangle$ pairs using a crowdsourcing platform. We extracted 115 random contexts from the Cornell Movie Dialogs corpus and used four distinct strategies to generate dialogue responses: a random predictor (used to identify the lower bound), the two Dual Encoder models (both without and with instance weights), and expert responses (used to identify the upper bound). The expert responses were manually authored by two human annotators. The resulting 460 $\langle$context, response$\rangle$ pairs were evaluated by 8 distinct human judges each (920 ratings per model). The human judges were asked to rate the consistency between context and response on a 5-points scale, from *Inconsistent* to *Consistent*. In total,

[2]Contrary to the OpenSubtitles corpus which is made of subtitles with no turn segmentation, the Cornell Movie Dialogs corpus and the corpus of theatre plays are derived from scripts and are therefore segmented in dialogue turns.

118 individuals participated in the crowdsourced evaluation.

The results of this human evaluation are presented in Figure 4. There is unfortunately no statistically significant difference between the baseline Dual Encoder ($M = 2.97$, $SD = 1.27$) and the one combined with the weighting model ($M = 3.04$, $SD = 1.27$), as established by a Wilcoxon rank-sum test, $W(1838) = 410360$, $p = 0.23$. These inconclusive results are probably due to the very low agreement between the evaluation participants (Krippendorff's $\alpha$ for continuous variable $= 0.36$). The fact that the lower and upper bounds are only separated by 2 standard deviations confirms the difficulty for the raters to discriminate between responses. We hypothesise that the nature of the corpus, which is heavily dependent on an external context (the movie scenes), makes it particularly difficult to assess the consistency of the responses.

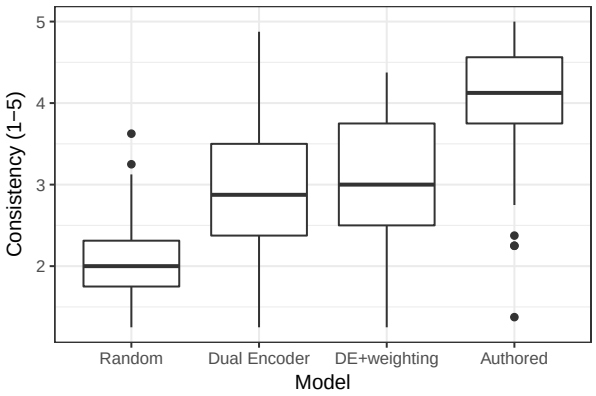


Figure 4: Distribution of human ratings of the responses generated by the four models tested.

Some examples of responses produced by the two Dual Encoder models illustrate the improvements brought by the weighting model. In (1), the baseline Dual Encoder selected a turn continuation rather than a reply, while the second model avoids this pitfall. Both (1) and (2) also show that the dual encoder with instance weighting tends to select utterances with fewer named entities.

(1) *Context of conversation:*
– This is General Ripper speaking.
– Yes, sir.
– Do you recognize my voice?"
⇒ *Response of Dual Encoder:*
– This is General Nikolas Pherides, Commander of the Third Army. I'm Oliver Davis.
⇒ *Response of Dual Encoder + weighting:*
– Yes, sir. I'm Gideon.

(2) *Context of conversation:*
– Let me finish dinner before you eat it... Chop the peppers...
– Are you all right?
⇒ *Response of Dual Encoder:*
– No thanks, not hungry. Harry Dunne.
⇒ *Response of Dual Encoder + weighting:*
– Yes I'm fine. Everything is ok.

## 5 Discussion

The limitations of neural conversational models trained on large, noisy dialogue corpora such as movie and TV subtitles have been discussed in several papers. Some of the issues raised in previous papers are the absence of turn segmentation in subtitling corpus (Vinyals and Le, 2015; Serban and Pineau, 2015; Lison and Meena, 2016), the lack of long-term consistency and "personality" in the generated responses (Li et al., 2016b), and the ubiquity of dull, commonplace responses when training generative models (Li et al., 2016a). To the best of our knowledge, this paper is the first to propose an instance weighting approach to address some of these limitations. One related approach is described in (Zhang et al., 2017) which also relies on domain adaptation for neural response generation, using a combination of online and offline human judgement. Their focus is, however, on the construction of personalised conversation models and not on instance weighting.

The empirical results corroborate the hypothesis that assigning weights to the training examples of "noisy" dialogue corpora can boost the performance of neural conversation models. In essence, the proposed approach replaces a one-pass training regime with a two-pass procedure: the first pass to determine the quality of each example pair, and a second pass to update the model based on the observed pair and its associated weight. We also showed that these weights can be determined in a data-driven manner with a neural model trained on example pairs selected for their adherence to specific quality criteria.

Instead of this two-pass procedure, an alternative approach is to directly learn a conversation model on the subset of example pairs that are known to be of high-quality. However, one major shortcoming of this approach is that it consider-

ably limits the size of the training set that can be exploited. For instance, the data used to estimate the weighting model in Section 4.2 corresponds to a mere 0.1 % of the total English-language part of the OpenSubtitles corpus (since the utterances had to be associated with speaker names derived from aligned scripts in order to apply the heuristics). In contract, the proposed two-pass procedure can scale to datasets of any size.

The results from Section 4 are limited to retrieval-based models. One important question for future work is to investigate whether the results carry over to generative, sequence-to-sequence models. As generative models are more computationally intensive to train than retrieval models, the presented approach may bring another important benefit, namely the ability to filter out part of the training data to concentrate the training time on "interesting" examples with a high cohesion between the context and its response.

## 6 Conclusion

Dialogue corpora such as chat logs or movie subtitles are very useful resources for developing open-domain conversation models. They do, however, also raise a number of challenges for conversation modelling. Two notable challenges are the lack of segmentation in dialogue turns (at least for the movie subtitles) and the presence of external context that is not captured in the dialogue transcripts themselves (leading to mentions of person names and unresolvable named entities).

This paper showed how to mitigate these challenges through the use of a *weighting model* applied on the training examples. This weighting model can be estimated in a data-driven manner, by providing example of "high-quality" training pairs along with random pairs extracted from the same corpus. The criteria that determine how these training pairs should be selected depend in practice on the type of conversational model one wishes to learn. This instance weighting approach can be viewed as a form of *domain adaptation*, where the data points from the source domain (in this case, the original corpus) are re-weighted to improve the model performance in a target domain (in this case, the interactions in which the conversation model will be deployed).

Evaluation results on retrieval-based neural models demonstrate the potential of this approach. The weighting model is essentially a preprocessing step and can therefore be combined with any type of conversational model.

Future work will focus on two directions. The first is to extend the weighting model to account for other criteria, such as ensuring diversity of responses and coherence across turns. The second is to evaluate the approach on other types of neural conversational models, and more particularly on generative models.

## References


Steffen Bickel, Michael Brückner, and Tobias Scheffer. 2007. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th International Conference on Machine Learning*. ACM, New York, NY, USA, ICML '07, pages 81–88.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR* abs/1412.3555.

Heriberto Cuayáhuitl. 2017. SimpleDS: A simple deep reinforcement learning dialogue system. In Kristiina Jokinen and Graham Wilcock, editors, *Dialogues with Social Robots: Enablements, Analyses, and Evaluation*, Springer, Singapore, Lecture Notes in Electrical Engineering, pages 109–118.

Cristian Danescu-Niculescu-Mizil and Lillian 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 2011*. Association for Computational Linguistics.

George Foster, Cyril Goutte, and Roland Kuhn. 2010. Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, EMNLP '10, pages 451–459.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2017. A knowledge-grounded neural conversation model. *CoRR* abs/1702.01932.

Jing Jiang and Chengxiang Zhai. 2007. Instance weighting for domain adaptation in NLP. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, Prague, Czech Republic, pages 264–271.

E. Levin, R. Pieraccini, and W. Eckert. 2000. A stochastic model of human-machine interaction for learning dialog strategies. *IEEE Transactions on Speech and Audio Processing* 8(1):11–23.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 110–119.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 994–1003.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016c. Deep reinforcement learning for dialogue generation. *CoRR* abs/1606.01541.

Jiwei Li, Will Monroe, Tianlin Shi, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. *CoRR* abs/1701.06547.

Pierre Lison and Raveesh Meena. 2016. Automatic turn segmentation of movie & TV subtitles. In *Proceedings of the 2016 Spoken Language Technology Workshop*. IEEE, San Diego, CA, USA, pages 245–252.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting on Discourse and Dialogue (SIGDIAL 2015)*. pages 285–294.

Ryan Lowe, Nissan Pow, Iulian V. Serban, Laurent Charlin, Chia-Wei Liu, and Joelle Pineau. 2017. Training end-to-end dialogue systems with the Ubuntu Dialogue Corpus. *Dialogue & Discourse* 8(1):31–65.

Abhay Prakash, Chris Brockett, and Puneet Agrawal. 2016. Emulating human conversations using convolutional neural network-based IR. *CoRR* abs/1606.07056.

Juan Ramos. 2003. Using TF-IDF to Determine Word Relevance in Document Queries. In *Proceedings of the First Instructional Conference on Machine Learning*. Rutgers University, New Brunswick, NJ, USA.

V. Rieser and O. Lemon. 2011. *Reinforcement Learning for Adaptive Dialogue Systems*. Springer, Berlin, Heidelberg.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '10, pages 172–180.

Iulian V Serban and Joelle Pineau. 2015. Text-based speaker identification for multi-participant open-domain dialogue systems. In *NIPS Workshop on Machine Learning for Spoken Language Understanding*. Montreal, Canada.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI Press, AAAI'16, pages 3776–3783.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, pages 1577–1586.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, CO, USA, pages 196–205.

Oriol Vinyals and Quoc Le. 2015. A Neural Conversational Model. *CoRR* abs/1506.05869.

Jason D. Williams, Kavosh Asadi, and Geoffrey Zweig. 2017. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. *CoRR* abs/1702.03274.

Rui Xia, Xuelei Hu, Jianfeng Lu, Jian Yang, and Chengqing Zong. 2013. Instance selection and instance weighting for cross-domain sentiment classification via pu learning. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*. AAAI Press, IJCAI '13, pages 2176–2182.

Kaisheng Yao, Baolin Peng, Geoffrey Zweig, and Kam-Fai Wong. 2016. An attentional neural conversation model with improved specificity. *CoRR* abs/1606.01292.

S. Young, M. Gai, B. Thomson, and J. D. Williams. 2013. POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE* 101(5):1160–1179.

Weinan Zhang, Ting Liu, Yifa Wang, and Qingfu Zhu. 2017. Neural personalized response generation as domain adaptation. *CoRR* abs/1701.02073.