Redundancy Localization for the Conversationalization of Unstructured Responses

Sebastian Krause[1,*], **Mikhail Kozhevnikov**[2], **Eric Malmi**[3,*], **Daniele Pighin**[2]

[1]DFKI Language Technology Lab, Berlin, Germany
sebastian.krause@dfki.de
[2]Google, Zürich, Switzerland
{qnan,biondo}@google.com
[3]Aalto University, Espoo, Finland
eric.malmi@aalto.fi

Abstract

Conversational agents offer users a natural-language interface to accomplish tasks, entertain themselves, or access information. Informational dialogue is particularly challenging in that the agent has to hold a conversation on an open topic, and to achieve a reasonable coverage it generally needs to digest and present unstructured information from textual sources. Making responses based on such sources sound natural and fit appropriately into the conversation context is a topic of ongoing research, one of the key issues of which is preventing the agent's responses from sounding repetitive. Targeting this issue, we propose a new task, known as redundancy localization, which aims to pinpoint semantic overlap between text passages. To help address it systematically, we formalize the task, prepare a public dataset with fine-grained redundancy labels, and propose a model utilizing a weak training signal defined over the results of a passage-retrieval system on web texts. The proposed model demonstrates superior performance compared to a state-of-the-art entailment model and yields encouraging results when applied to a real-world dialogue.

1 Introduction

Recent years have seen a growing interest in research on conversational agents. Several strands of dialogue systems have emerged which differ in underlying goals and methods. Some systems focus on data-driven learning of models which can autonomously hold conversations with humans or one another, potentially even on open domains (Vinyals and Le, 2015; Sordoni et al., 2015; Li et al., 2016). Other works deal with task-oriented dialogues, which offer natural-language interfaces to real-world services like restaurant booking (Bordes and Weston, 2016; Dhingra et al., 2016; Crook et al., 2016). We focus in this paper on a third dialogue setting where the goal is to have a natural conversation with a user, during which the user's information needs are satisfied in an iterative manner. Such a setting is common in question-answering experiences implemented in personal digital assistants (Sarikaya et al., 2016).

User: *What is Malaria?*

Agent: *A disease caused by a plasmodium parasite, transmitted by the bite of infected mosquitoes.*

User: *Is it a virus?*

Agent: *Malaria is a parasitic infection spread by Anopheles mosquitoes. The Plasmodium parasite that causes Malaria is neither a virus nor a bacterium – it is a single-celled parasite that multiplies in red blood cells of humans as well as in the mosquito intestine.*

Figure 1: Informational-dialogue example between a human and a conversational agent. The second agent utterance is partially redundant (the underlined text).

We call this setting *informational dialogues*. They start with the user posing a fact-seeking question, e.g., to learn about current events or to explore unknown terms and concepts. Consider the example dialogue in Fig. 1, which is initiated by the user requesting a definition of a specific disease and which also features a subsequent question on the same topic. Many approaches have been proposed which can produce suitable replies to such questions. Examples include techniques which find pertinent passages or short text chunks in collections of documents (Hermann et al., 2015; Miller et al., 2016; Trischler et al., 2016) or find rele-

*Work performed during an internship at Google.

vant entries in structured knowledge bases (Bordes et al., 2014, 2015; Yin et al., 2016a,b). Generation techniques can then be employed to generate well-formed natural-language utterances from the candidate replies (Wen et al., 2015, 2016a,b; Zhou et al., 2016; Dušek and Jurcicek, 2016). In the dialogue in Fig. 1, both agent replies are coherent wrt. the questions. However, they sound strange when occurring together in a single dialogue context because information is partially reiterated (see the underlined part in the second agent reply). It is this very problem that we focus on in this work, i.e., the *localization of redundancy* in conversation. Information on the location of non-novel portions of a passage could either be fed back to the retrieval model, so that only text passages with new information would be selected, or alternatively this localized redundancy might be used as input to a summarization model (Rush et al., 2015).

The specific contributions of this work are as follows:

- We propose a new task, motivated by practical issues that dialogue applications face (Sec. 3).
- We release a new dataset with manual annotations for this task, which allows to evaluate and compare competing approaches (Sec. 4).
- Due to the insufficient amount of annotated data for training purposes, we report on a weak supervision signal over a large collection of passages with partially redundant content (Sec. 5).
- We augment a recently introduced entailment model (Parikh et al., 2016) with means for representing local similarities in passages in a unidirectional way (Sec. 6) and find that this extension outperforms the original model (Sec. 8).
- Furthermore, we briefly discuss an experiment on real-world dialogue data (Sec. 9), which gives insights on the application-relevance of the proposed task and model.

2 Related Work

A lot of work has been presented on reasoning with short texts for tasks on similarity and entailment. Knowledge-rich approaches define lexical and syntactic inference rules over phrase pairs and employ decision algorithms that rely on matches of these rules in input texts (Magnini et al., 2014). Other approaches generate structured representations of the input to enable sophisticated alignment of the texts with now available rich lexical, syntactic, and semantic information (Liang et al., 2016). The use of kernel methods for similarity tasks has also been reported (Filice et al., 2015). In contrast to these approaches, neither do we use external knowledge nor do we build explicit syntactic representations of input texts.

Sentence fusion (Barzilay and McKeown, 2005; Filippova and Strube, 2008) is a technique that is related to the overall problem setting of this paper. This technique is used in the context of abstractive multi-document summarization, where a particular challenge is to identify shared content in a cluster of sentences and to subsequently produce a single sentence that covers all information fragments. In our work, we focus on a similar but different problem formulation, in which we fix one text fragment and want to find reiterations of its content in other texts. Furthermore, we focus on identifying and localizing redundancy and leave the generation of low-redundancy text mostly as future work.

Neural approaches are common for bi-sequence classification problems (Laha and Raykar, 2016). Yin and Schütze (2015), He et al. (2015), and He and Lin (2016) use convolutional networks to represent input texts on multiple granularity levels and model the interactions of these. We also aim to find fine-granular interactions in texts, but in addition to their models, we aim to make these interactions explicit rather than latent intermediate results. Another line of research has proposed recurrent networks for modeling phrases/sentences, including various forms of neural attention (Bowman et al., 2015; Rocktäschel et al., 2015; Zhao et al., 2016). These approaches come with high computational cost during training and inference, in contrast we rely on cheaper feed-forward connections.

3 Problem Definition

We focus in this work on the problem of *redundancy localization* in a passage with respect to another text, i.e., we aim to understand when a sub-passage is redundant with what is mentioned in the context.[1] Consider the following example with a context passage $\mathbf{c}$ and a follow-up passage $\mathbf{p}$ with sub-sequences $\mathbf{s}_0$–$\mathbf{s}_3$, which need to be ranked according to the extent to which their semantics are covered by $\mathbf{c}$. In this case, one may expect the

[1] Note that the problem definition is not limited to the dialogue scenario used as motivation in the introduction.

order to be $(\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3, \mathbf{s}_0)$:

$\mathbf{c}$: *The Allianz Arena is a football stadium in Munich, Bavaria, Germany, with a seating capacity of more than 70,000.*

— — — — — — — — — — — —

$\mathbf{s}_0$: *Bayern to increase stadium capacity.*

$\mathbf{s}_1$: *Bayern Munich have revealed plans to increase the capacity of Allianz Arena to 75,000,*

$\mathbf{s}_2$: *which would make it the second largest stadium in Germany.*

$\mathbf{s}_3$: *The Allianz Arena is currently the third largest stadium in Germany.*

More formally, let $\mathbf{p}$ be a sequence of n tokens. Let $\mathbf{S} = \{\mathbf{s}_k\}_{k=0}^{m-1}$ be a set of m sub-sequences of $\mathbf{p}$ such that for integers $s_0, s_1, \ldots, s_m$ with $s_0 = 0 < s_1 < \ldots < s_{m-1} < s_m = n$, each subsequence $\mathbf{s}_k \in \mathbf{S}$ is ranging from tokens s_k to $(s_{k+1} - 1)$, inclusive. Given a context sequence $\mathbf{c}$, the task of redundancy localization is to produce a ranking function $rank(\mathbf{s}_k) \in \{1, \ldots, m\}$ that induces an ordering of the subsequences $\mathbf{s}_k \in \mathbf{S}$ of $\mathbf{p}$ which corresponds to the degree of information in $\mathbf{s}_k$ that is semantically covered by $\mathbf{c}$. Here, a low rank corresponds to a high semantic overlap of a subsequence with $\mathbf{c}$, where segments are allowed to have equal ranks.

We formulate this task as a ranking problem instead of a more expressive yet also more complex regression setting in order to pose less restrictions on the collection of data for training and evaluation. The design decision to rank sub-sequences rather than individual tokens is intended to keep manual annotation feasible and cost-effective.

Relation to Other Tasks The problem we pose here is related to bi-sequence problems like *semantic textual similarity* (STS) (Agirre et al., 2016a) and *recognizing textual entailment* (RTE) (Bowman et al., 2015). In contrast to these tasks, we are not interested in determining the overall relation between sequences, but aim to generate more fine-grained sub-passage-level information. The task of *interpretable semantic textual similarity* (Agirre et al., 2016b) requires systems to provide human-understandable explanations for STS ratings of *sentence* pairs. Chunks from both sentences need to be paired and for each such pairing, similarity and relation type need to be assessed. While this type of annotation is richer than what we propose, it is also harder to produce, likely requiring specially-trained raters, and would likely be impossible to predict accurately using a surrogate supervision signal like we rely on. Besides, it does not scale well beyond single sentences, since the number of ratings per sequence pair grows proportionally to the multiple of their lengths, while the model we present can handle longer, multi-sentence passages. The setting proposed in the next section is more restricted, but easier to learn and directly applicable in downstream applications.

4 A Testbed for Redundancy Localization

The evaluation dataset (EVAL) is constructed from pairs of *potentially* redundant passages from Wikipedia, which were segmented into sub-passages and presented to human raters for manual redundancy assessment. The collection of passages was guided by a need for text pairs with various degrees of semantic overlap; we employed a *passage-retrieval* system for the purpose of text selection. Passage retrieval (Khalid and Verberne, 2008; Aktolga et al., 2011; Xu et al., 2011) is a common intermediate step in information-retrieval and question-answering settings, the goal of which is to return a passage containing the answer to a given query. Most systems generate a list of candidate passages, rank them by relevance and return the top one.

We picked a random set of 1200 fact-seeking questions and retrieved corresponding passages from Wikipedia. The questions were then discarded, as they are not relevant to our task. We selected the top-scoring passage as the context $\mathbf{c}$ and paired it with a low-scoring one from further down the result list ($\mathbf{p}$). $\mathbf{p}$ was then heuristically split into chunks $\mathbf{s}_k$, corresponding to verb-governed phrases. The example shown in the last section is an instance of such a pair ($\mathbf{c}$, $\mathbf{p}$).

We asked three raters per item to select for each segment $\mathbf{s}_k$ of $\mathbf{p}$ one out of three labels: NOTREDUNDANT, PARTIALLYREDUNDANT, and FULLYREDUNDANT, depending on the degree of which the content of a sub-passage is covered by the context $\mathbf{c}$. The annotators fully/partially agreed[2] on 64%/96% of examples, their annotation has an intra-class correlation of .55. We aggregated the rating by mapping the categorical labels to a numeric scale (0, 1, 2) and averaging the scores. We used 200 examples as a development

[2]Full: 3/3 annotators agreed on a label. Partial: At least 2/3 annotators agreed on a label.

$\mathbf{c}$:	*Brewer's yeast is made from a one-celled fungus called Saccharomyces cerevisiae.*
$\mathbf{p}^+$:	*Brewer's yeast is named so because it comes from the same fungus that's used to ferment and make beer - Saccharomyces cerevisiae.*
$\mathbf{p}^-$:	*Because brewer's yeast is a rich source of chromium, scientists think it may help treat high blood sugar.*
$\mathbf{c}$:	*The height of the net in men's volleyball is 7 feet 11 5/8 inches, and in women's volleyball, it is 7 feet 4 1/8 inches.*
$\mathbf{p}^+$:	*Outdoor volleyball, played on grass, will use the standard net heights of 7 feet, 4 1/8 inches for women, with men and co-ed teams using the height of 7 feet, 11 5/8 inches.*
$\mathbf{p}^-$:	*The first volleyball net was borrowed from a tennis court and was set at 6 feet 6 inches high.*
$\mathbf{c}$:	*The world's tallest artificial structure is the 829.8 m tall Burj Khalifa in Dubai, United Arab Emirates.*
$\mathbf{p}^+$:	*The 828-metre tall Burj Khalifa in Dubai has been the tallest building in the world since 2008.*
$\mathbf{p}^-$:	*Burj Khalifa broke the height record in all four categories for completed buildings.*

Table 1: Three weakly-labeled examples (Sec. 5). Underlining used to indicate overlapping/distinct information between items.

Label	DEV #	DEV %	TEST #	TEST %
REDUNDANT	95	15.83	495	16.50
PARTIALLYREDUNDANT	81	13.50	541	18.03
NOTREDUNDANT	424	70.67	1964	65.47

Table 2: Distribution of sub-passage labels in EVAL.

dataset for the experiments in this paper (DEV), and the remaining 1000 items as a test dataset (TEST). Tab. 2 reports the label distribution in both parts of the dataset. We make the dataset publicly available at https://github.com/kraseb/redundancy-localization.

5 Training with a Proxy Signal

While the annotation required for our task is comparatively simple and can be performed by raters without special training, a workable fully-supervised model would require a very considerable amount of data and is likely to prove costly.[3] Suppose, however, we were supplied with a large number of short texts with varying degrees of similarity and relatedness to one another and we had a means of assessing at the coarse level of text pairs whether or not they were similar. Our hypothesis is that given appropriate model capacity and structure, a model trained to predict the passage-level similarity would learn to compare smaller units of text to make an appropriate high-level decision.

We derive a proxy signal from passage-level retrieval scores which allows to bootstrap the redundancy-localization model described in Sec. 6. The model is presented with passage triples, where two passages are very closely related and the third one is on the same general topic, but less similar to the other two and hence likely contains less redundancy. The model is then trained to rank the more closely related passage pairs above the less closely related ones.

We retrieve lists of relevant passages from the web using the same passage-retrieval system that we utilized to collect data for manual annotation. Through manual inspection of a small subset of candidate passage lists, we identified a range of passage scores, where candidate passages are topically close to the top-scoring one, but sufficiently different in factual content. To ensure that the top-scoring passage and the lower-scoring one are on the same topic, we further require that they be extracted from the same webpage.

From each of the queries' passage lists we extract three passages, the top-scoring passage $\mathbf{c}$, the second-highest ranking passage $\mathbf{p}^+$, and a lower-scoring passage $\mathbf{p}^-$ from the score corridor described above. The stream of passage triples $(\mathbf{c}, \mathbf{p}^+, \mathbf{p}^-)$ generated in this way allows to train a model with a margin-based ranking objective. This objective enforces that the similarity score of the two high-scoring passages $\mathbf{c}, \mathbf{p}^+$ is greater than the similarity of the low-scoring passage $\mathbf{p}^-$ and the top-scoring one, plus a margin; see Sec. 6.3. This pushes a model to find what differentiates two given text sequences, so that it can assign a higher similarity to the near-paraphrases.

Tab. 1 shows three example passage triples constructed with this signal. Here, underlining is a means of visualizing the overlapping/disjoint content between triple elements. Note that we do not

[3] Among other things, to accurately identify redundancy the model needs to have at least some notion of paraphrasing.

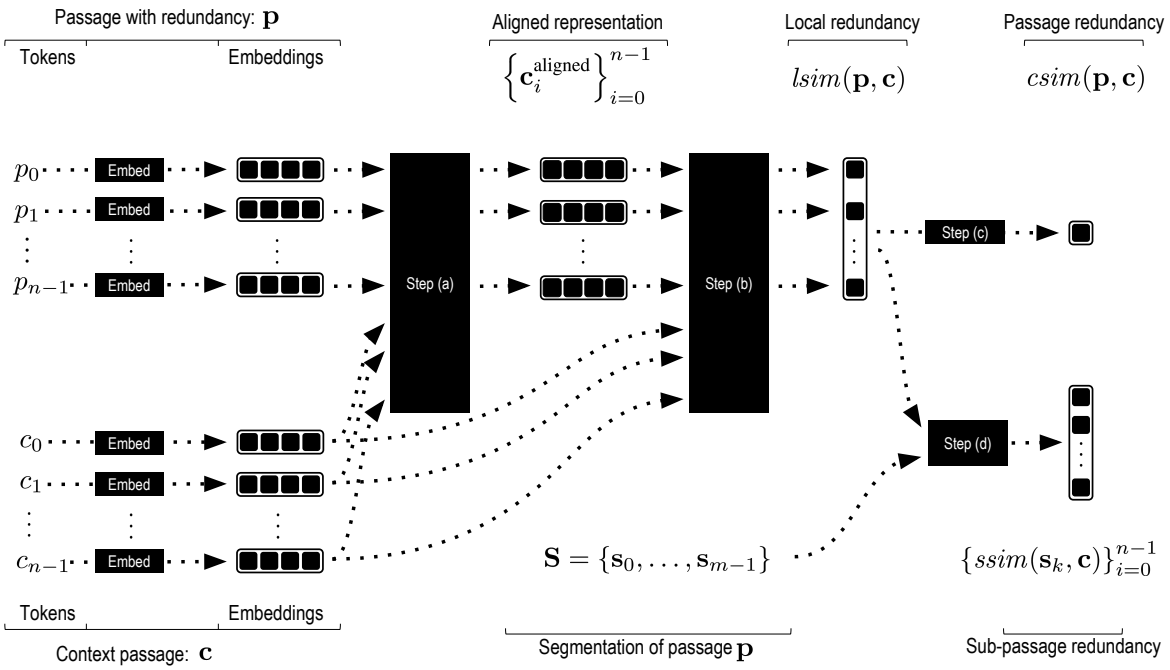

Figure 2: Overview of the model architecture.

make this information available to a model during training. In the interest of brevity, we selected short, single-sentence passages for this example.

6 Model Design

This section first gives a brief overview of the proposed model, before going into details of its architecture and use during training and inference time.

Architecture Overview Existing models for bi-sequence tasks (Bahdanau et al., 2014; Rush et al., 2015; He and Lin, 2016) often learn to align texts as an intermediate step, i.e., reasoning is done with pairs of short text units, which allows to build a task-specific output for whole sequences on top of local decisions. A particular example for RTE is the three-layer model of Parikh et al. (2016). The first layer produces a bi-directional alignment between input sentences, which is utilized in the second component to perform local comparisons, which in turn are fed to the top layer to make the final entailment decision. We follow the same pattern in the design of our model.

We implement a multi-component neural-network that takes two passages as input. It first (a) learns a *uni*-directional alignment between the passages, which is utilized to produce a customized representation of the context passage, specific to each token of the potentially redundant passage. Next, (b) token-level redundancy scores are produced via local comparison operations. During training, (c) an additional layer aggregates the local scores and produces a passage-level similarity score on top of which a ranking objective is applied. At inference time, (d) the local scores from (b) serve as the basis for the ranking of the sub-passage elements as described in Sec. 3. Fig. 2 outlines steps (a) – (d).

6.1 Step (a): Alignment

Input to the model are two sequences of n tokens each, $\mathbf{p} = (p_0, \ldots, p_{n-1})$ and $\mathbf{c} = (c_0, \ldots, c_{n-1})$, with shorter sequences being padded to this length. The goal of this step is to generate for each $p_i \in \mathbf{p}$ a fixed-length representation $\mathbf{c}_i^{\text{aligned}}$ of $\mathbf{c}$, which captures the meaning aspects of $\mathbf{c}$ specifically relevant for p_i.

The tokens p_i, c_j are represented via word embeddings of size d_w, which are updated during model training and are stored in a matrix $W_w \in \mathbb{R}^{d_w \times |V|}$, with V being the vocabulary. For ease of notation, we use $\mathbf{p}$, p_i, $\mathbf{c}$, c_i to refer to both the original tokens and their embedding representation.

We create a soft alignment of $\mathbf{c}$ to the tokens of $\mathbf{p}$ via the decomposed attention mechanism described by Parikh et al. (2016). At its core is the application of the attention function $\mathrm{f1}$ to each token of the input sequences, which is implemented as a feed-forward neural network with h_{f1} layers of

d_{f1} rectified linear units (Glorot et al., 2011, ReLu) each. Using this function, unnormalized attention weights are produced:

$$\alpha_{ij} = \text{f1}(p_i) \cdot \text{f1}(c_j), \quad (1)$$

then normalized per token in $\mathbf{p}$ via

$$\alpha'_{ij} = \exp(\alpha_{ij}) / \sum_k \exp(\alpha_{ik}). \quad (2)$$

The customized (*aligned*) representation of $\mathbf{c}$ is then calculated as

$$\mathbf{c}_i^{\text{aligned}} = \sum_{j=0}^{n-1} \alpha'_{ij} c_j. \quad (3)$$

6.2 Step (b): Learning Local Redundancy

Each token p_i from $\mathbf{p}$ is compared to the corresponding representation $\mathbf{c}_i^{\text{aligned}}$ of the context sequence via a single-layer feed-forward network f2 with a ReLu:

$$lsim(p_i, \mathbf{c}) := \text{f2}\left(\left[p_i, \mathbf{c}_i^{\text{aligned}}\right]\right) \quad (4)$$

$$lsim(\mathbf{p}, \mathbf{c}) := [lsim(p_i, \mathbf{c})]_{i=0}^{n-1} \quad (5)$$

with $[\,]$ being the concatenation operator and $lsim(\mathbf{p}, \mathbf{c}) \in \mathbb{R}^n$. This local similarity score measures for each token the degree with which its meaning is covered by $\mathbf{c}$.

6.3 Step (c): Learning to Aggregate Local Redundancy Scores

As described in Sec. 5, supervised training with local redundancy labels is costly, which is why we add another layer on top which learns to calculate a coarse passage-level similarity score $csim(\mathbf{p}, \mathbf{c})$ from the local redundancy information. Given a passage triple $(\mathbf{c}, \mathbf{p}^+, \mathbf{p}^-)$ (Sec. 5), two such coarse scores are calculated and used to determine a loss which allows to train steps (a–c) of the network in Fig. 2 in a weakly supervised way.

The passage-level score is computed by another feed-forward network f3 with h_{f3} layers of d_{f3} ReLus, followed by another hidden layer with a logistic activation function that projects to a scalar value in $(0, 1)$:

$$csim(\mathbf{p}, \mathbf{c}) := \text{f3}(lsim(\mathbf{p}, \mathbf{c})). \quad (6)$$

Then, for a given passage triple $(\mathbf{c}, \mathbf{p}^+, \mathbf{p}^-)$, the loss is defined as:

$$\mathcal{L} = max\{0, 0.5 - csim(\mathbf{p}^+, \mathbf{c}) + csim(\mathbf{p}^-, \mathbf{c})\} \quad (7)$$

This ranking criterion is similar to what has been used by Collobert et al. (2011) and Bordes et al. (2013). It is intended to push the model to assign a higher coarse similarity score to the more similar sequences from the triple, and in doing so, ideally forces the model to learn to detect local redundancies.

6.4 Step (d): Generation of Sub-sequence Redundancy Scores

During inference time, the goal of this model is to rank a set of given sub-sequences $\mathbf{S}$ of $\mathbf{p}$ with respect to their redundancy with $\mathbf{c}$; note that during inference time the model is presented with pairs of passages in contrast to the triples it sees in the training phase.

We calculate a redundancy score for a subsequence $\mathbf{s}_k \in \mathbf{S}$ as follows:

$$ssim(\mathbf{s}_k, \mathbf{c}) := \frac{1}{s_{k+1} - s_k} \sum_{l=s_k}^{s_{k+1}-1} (lsim(p_l, \mathbf{c})), \quad (8)$$

where $\mathbf{s}_k$ is the subsequence running from positions s_k to $s_{k+1} - 1$ (see Sec. 3). A ranking of the subsequences is then given by:

$$rank(\mathbf{s}_k) := |\{\mathbf{s}_l \mid ssim(\mathbf{s}_l, \mathbf{c}) \geq ssim(\mathbf{s}_k, \mathbf{c})\}| \quad (9)$$

In other words, sub-passages are ranked by comparing the mean of their local redundancy scores. In the evaluation of Sec. 8, we refer to the model that uses this way of ranking sub-passages as UA (short for uni-directional alignment). We compare this against a number of other variants of processing internal activations of the model to extract information about local redundancy, see Sec. 8.

6.5 Baseline Ranking Method

The bi-directional alignment model (BA) of Parikh et al. (2016) can be trained in a similar fashion as our proposed model, i.e., with triples of passages and the loss from Eq. (7). Although it has not been developed with the localization of redundancy in mind, its native problem formulation (RTE) is structurally related to the problem at hand by requiring models to assess to what degree the semantic content of one passage is embedded in a second one. We believe BA constitutes a strong baseline because it has been shown to achieve state-of-the-art performance on RTE and because it has the means to decompose coarse inference decisions on two text sequences into local comparison operations,

d_w	100	η	0.01
d_{f1}	200	$\|V\|$	$10k$
d_{f3}	100	p_{f1}	0.21
h_{f1}	1	p_{f2}	0.46
h_{f3}	1	p_{f3}	0.05
batch size	256	epochs	≈ 200

Table 3: Hyperparameter settings for UA.

Dataset	Model	ρ	Model	ρ
DEV	UA	.5298	BA′	.1384
	UA_Σ	.4169	BA'_Σ	.2232
	UA′	.3862	BA″	.2817
	UA'_Σ	.4071	BA''_Σ	.2923
TEST	UA	.5544	BA''_Σ	.2688

Table 4: Comparison of alternative strategies for step (d) (Sec. 6.4) on DEV and results of optimal strategies on TEST.

a key requisite to successfully utilize the training signal from Sec. 5.

However, in contrast to our model, the results of comparing the aligned sequences $\mathbf{c}_i^{\text{aligned}}$ with individual tokens from $\mathbf{p}$ are not directly interpretable as redundancy scores, also the architecture is designed for a bi-directional alignment of the input sequences. In order to produce $lsim$ values for the tokens of $\mathbf{p}$, we use the alignment matrices as a basis for a max-based aggregation, i.e., we take the row-wise maximum value and use this as the localized redundancy value for the corresponding token. Sub-sequence similarity is then determined either via Eq. (8) or alternatively via summation.

7 Experimental Setting, Model Training

We implemented both UA and BA in the TensorFlow framework (Abadi et al., 2015) and trained them with the signal from Sec. 5. As input to the passage-retrieval system we used a set of 1.5 million queries, resulting in the same amount of passage triples; 80% were used for training, 10% were used as a separate validation set for hyperparameter optimization, and the final 10% were held out and served as the basis for the smaller dataset with manually annotated labels (EVAL, Sec. 4)[4].

The hyperparameters of UA ($h_{\text{f1}}, d_{\text{f1}}, h_{\text{f3}}, d_{\text{f3}}$) and BA (like our model, plus a few additional ones) were optimized separately. We also experimented with Dropout (Srivastava et al., 2014) for the feed-forward networks in step (a–c) ($p_{\text{f1}}, p_{\text{f2}}, p_{\text{f3}}$), with different initial learning rates (η) for Adagrad (Duchi et al., 2011), with different batch sizes, and with different vocabulary sizes ($|V|$). The final settings for UA used in the reported experiments are shown in Tab. 3. Word embeddings were initialized with pre-trained embeddings (Mikolov et al., 2013), the other model parameters were randomly initialized; out-of-vocabulary words were hashed into 100 buckets. The models were trained for 1 million steps.

[4]We only annotated a subset of the passages in this part of the data.

8 Evaluation on EVAL

We first compare the performance of different variants of generating the redundancy scores for sub-passage ranking, for both UA and BA, on DEV. We then pick the respective best-performing model variant and compare the systems on TEST. The model variants we test are the following:

- **UA**: The uni-directional alignment model described in Sec. 6.
- **UA_Σ**: Summation instead of averaging in Eq. (8), which gives higher weight to long sub-sequences with redundancy.
- **UA′**: Calculation of $lsim$ in analogous fashion as BA (see below).
- **UA'_Σ**: Combination of two variants above.
- **BA′/BA″**: Models with bi-directional alignment of input texts. $lsim$ values for tokens of $\mathbf{p}$ are produced by using the first/second one of the two alignment matrices as a basis for the max-based aggregation of the normalized attention weights described in Sec. 6.5.
- **BA'_Σ / BA''_Σ**: Like above, but sub-sequence similarity is determined via summation rather than calculating the mean in Eq. (8).

We measure performance by calculating the Spearman correlation of the raw passage scores with the gold redundancy for all segments in the respective partition of the dataset. The top of Tab. 4 reports results of the different model variants. For UA, making direct use of the local redundancy scores calculated in step (b) of the model yields slightly better results than post-processing the alignments

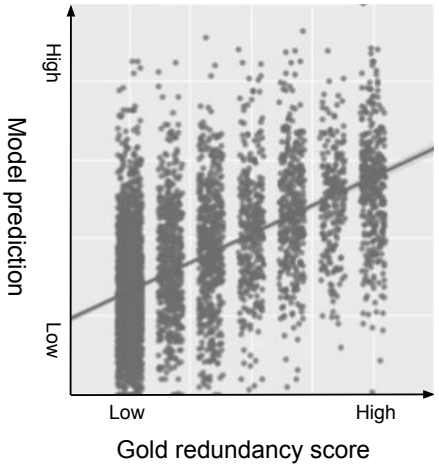

Figure 3: Plot of predictions of UA on TEST against annotated redundancy.

from step (a) of the model. The best overall results for UA are achieved when this is combined with the strategy that represents sub-sequence redundancy as the arithmetic mean of the contained tokens' local scores, meaning sub-sequence length needs to be taken into account.

For the baseline BA, exploiting the reverse alignment matrix and summing over the alignment scores without correction for sub-sequence length gives the best results. The bottom of the table reports the results of applying both models with the respective best strategy on the test partition of the dataset. The proposed uni-directional model clearly outperforms the bi-directional baseline. This indicates that the direct modeling of uni-directional redundancy during both training and inference time allows a model to better learn to compare a subsequence to another full passage, in comparison to the case where both passages are analyzed in a fine-granular way.

Fig. 3 depicts a scatter plot of the segments in TEST, with the x-axis corresponding to the gold redundancy scores (Sec. 4) and the y-axis showing the redundancy assessment by UA. While actually redundant segments tend to be handled correctly by the model, a certain amount of non-redundant segments get assigned a relatively high absolute redundancy value, which is not problematic as long as the actually redundant segments of the same passage are rated even higher. The next section elaborates on an experiment that looks into the quality of this internal ranking of segments for given passages, and how this ranking could potentially be utilized in an application.

9 Redundancy Localization for Passage Compression

This section briefly discusses an experiment in a dialogue setting, in which redundancy information is used for the compression of passages. Consider again the example from Fig. 1, where a conversational agent engages a human user in an informational dialogue whose quality suffers from repetition of information on the agent side. In this experiment, we asked human raters to assess whether the removal of redundancy improves the dialogue flow. Note, however, that given the small scale of the experiment, results are only indicative and not conclusive.

We selected 50 passage pairs from the held-out portion of the training data where the second passage consisted of at least three sentences. We then fed the passages to UA and removed the sentence from the second passage which had the largest semantic overlap with the context (the first passage). We asked three human raters, (a) whether the two original passages are coherent at all (as the following questions assume this), (b) whether the compressed passage sounds more or less natural (due to the dropped redundant sentence), and (c) whether the modified passage is equally informative as the original passage.

For comparison, we implemented a baseline which always dropped the first sentence of a passage, as well as one that removed the sentence with the highest term overlap. For the following example, dropping the underlined sentence from the passage would result in a more natural and equally informative text:

c : *The 1966 FIFA World Cup was won by the England national football team.*

p : *The day England won the World Cup. Long-suffering fans of the England football team can always look back with nostalgia on one year: 1966. This was the year Bobby Moore's team defeated West Germany 4-2 in the World Cup final on 30 July, after a nail-biting and controversial match.*

Among the 50 uncompressed passage pairs, only one third was rated as being coherent (question a; independent of the model). For these pairs, UA tended to produce more natural compressions (question b) compared to the baselines. This might be explained by the term-overlap baseline's restriction to only look at the level of individual words, which results in erroneously removing sentences that are essential for discourse coherence but do not

repeat facts. Similarly, always dropping the first sentence can leave a passage with dangling backward references, e.g., in the case of anaphors. In terms of the informativeness dimension (question c), all approaches resulted in slightly less informative compressed passages, which is expected. However, UA's score on this metric is slightly worse than the one of the baselines.

10 Contributions and Outlook

In this paper, we described the problem of localizing redundancy in pairs of passages. We proposed a model based on a uni-directional alignment from one passage to the context passage, which can be efficiently trained using a novel weak supervision signal defined over the output of common passage-retrieval systems. We applied this signal in a one-off process to train our model and a reasonable baseline; from a held-out part of the retrieved passages we created a publicly available dataset which allows to compare and evaluate models on this task and enables other researchers to reproduce the evaluation setting of this work. The conducted evaluation showed that the proposed uni-directional alignment model is indeed capable of finding the redundant sub-segments in texts.

In future work, we would like to represent and model more facets of the naturalness and coherence of dialogues. For instance in dialogue settings, a certain amount of redundancy between the utterances of participants may actually tie the dialogue turns together, i.e., may be beneficial in terms of discourse coherence and naturalness. Incorporating this consideration into the structure of a model can potentially improve the results of passage compression techniques in settings similar to Sec. 9.

Acknowledgments

The first author was partially supported by the German Federal Ministry of Education and Research, project ALL SIDES (contract 01IW14002).

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016a. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, pages 497–511. https://doi.org/10.18653/v1/S16-1081.

Eneko Agirre, Aitor Gonzalez-Agirre, Inigo Lopez-Gazpio, Montse Maritxalar, German Rigau, and Larraitz Uria. 2016b. SemEval-2016 task 2: Interpretable semantic textual similarity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, pages 512–524. https://doi.org/10.18653/v1/S16-1082.

Elif Aktolga, James Allan, and David A. Smith. 2011. Passage reranking for question answering using syntactic structures and answer types. In *Advances in Information Retrieval - Proceedings of the 33rd European Conference on IR Research (ECIR)*. Springer, Dublin, Ireland, volume 6611 of *Lecture Notes in Computer Science*, pages 617–628. https://doi.org/10.1007/978-3-642-20161-5_62.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR* abs/1409.0473. http://arxiv.org/abs/1409.0473.

Regina Barzilay and Kathleen McKeown. 2005. Sentence fusion for multidocument news summarization. *Computational Linguistics* 31(3):297–328. https://doi.org/10.1162/089120105774321091.

Antoine Bordes, Sumit Chopra, and Jason Weston. 2014. Question answering with subgraph embeddings. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 615–620. https://doi.org/10.3115/v1/D14-1067.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *CoRR* abs/1506.02075. http://arxiv.org/abs/1506.02075.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26 (NIPS)*. pages 2787–2795. http://papers.nips.cc/paper/5071-translating-embeddings-for-modeling-multi-relational-data.

Antoine Bordes and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. *CoRR* abs/1605.07683. http://arxiv.org/abs/1605.07683.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 632–642. https://doi.org/10.18653/v1/D15-1075.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12:2493–2537. http://dl.acm.org/citation.cfm?id=2078186.

Paul Crook, Alex Marin, Vipul Agarwal, Khushboo Aggarwal, Tasos Anastasakos, Ravi Bikkula, Daniel Boies, Asli Celikyilmaz, Senthilkumar Chandramohan, Zhaleh Feizollahi, Roman Holenstein, Minwoo Jeong, Omar Khan, Young-Bum Kim, Elizabeth Krawczyk, Xiaohu Liu, Danko Panic, Vasiliy Radostev, Nikhil Ramesh, Jean-Phillipe Robichaud, Alexandre Rochette, Logan Stromberg, and Ruhi Sarikaya. 2016. Task completion platform: A self-serve multi-domain goal oriented dialogue platform. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. Association for Computational Linguistics, San Diego, California, pages 47–51. https://doi.org/10.18653/v1/N16-3010.

Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng. 2016. End-to-end reinforcement learning of dialogue agents for information access. *CoRR* abs/1609.00777. http://arxiv.org/abs/1609.00777.

John C. Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* 12:2121–2159. http://dl.acm.org/citation.cfm?id=2021068.

Ondřej Dušek and Filip Jurcicek. 2016. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 45–51. https://doi.org/10.18653/v1/P16-2008.

Simone Filice, Giovanni Da San Martino, and Alessandro Moschitti. 2015. Structural representations for learning relations between pairs of texts. 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 1003–1013. https://doi.org/10.3115/v1/P15-1097.

Katja Filippova and Michael Strube. 2008. Sentence fusion via dependency graph compression. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 177–185. http://aclweb.org/anthology/D08-1019.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*. JMLR.org, volume 15 of *JMLR Proceedings*, pages 315–323.

Hua He, Kevin Gimpel, and Jimmy Lin. 2015. Multi-perspective sentence similarity modeling with convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1576–1586. https://doi.org/10.18653/v1/D15-1181.

Hua He and Jimmy Lin. 2016. Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 937–948. https://doi.org/10.18653/v1/N16-1108.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28 (NIPS)*. pages 1693–1701. http://papers.nips.cc/paper/5945-teaching-machines-to-read-and-comprehend.

Mahboob Khalid and Suzan Verberne. 2008. Passage retrieval for question answering using sliding windows. In *Coling 2008: Proceedings of the 2nd Workshop on Information Retrieval for Question Answering*. Coling 2008 Organizing Committee, pages 26–33. http://aclweb.org/anthology/W08-1804.

Anirban Laha and Vikas Raykar. 2016. An empirical evaluation of various deep learning architectures for bi-sequence classification tasks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 2762–2773. http://aclweb.org/anthology/C16-1260.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 1192–1202. https://doi.org/10.18653/v1/D16-1127.

Chen Liang, Praveen K. Paritosh, Vinodh Rajendran, and Kenneth D. Forbus. 2016. Learning paraphrase identification with structural alignment. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI)*. IJCAI/AAAI Press, pages 2859–2865. http://www.ijcai.org/Abstract/16/406.

Bernardo Magnini, Roberto Zanoli, Ido Dagan, Kathrin Eichler, Guenter Neumann, Tae-Gil Noh, Sebastian Padó, Asher Stern, and Omer Levy. 2014. The excitement open platform for textual inferences. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, pages 43–48. https://doi.org/10.3115/v1/P14-5008.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26 (NIPS)*. pages 3111–3119. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 1400–1409. https://doi.org/10.18653/v1/D16-1147.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 2249–2255. https://doi.org/10.18653/v1/D16-1244.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomás Kociský, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *CoRR* abs/1509.06664. http://arxiv.org/abs/1509.06664.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 379–389. https://doi.org/10.18653/v1/D15-1044.

Ruhi Sarikaya, Paul A. Crook, Alex Marin, Minwoo Jeong, Jean-Philippe Robichaud, Asli Çelikyilmaz, Young-Bum Kim, Alexandre Rochette, Omar Zia Khan, Xiaohu Liu, Daniel Boies, Tasos Anastasakos, Zhaleh Feizollahi, Nikhil Ramesh, H. Suzuki, Roman Holenstein, Elizabeth Krawczyk, and Vasiliy Radostev. 2016. An overview of end-to-end language understanding and dialog management for personal digital assistants. In *2016 IEEE Spoken Language Technology Workshop (SLT 2016)*. pages 391–397. https://doi.org/10.1109/SLT.2016.7846294.

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*. Association for Computational Linguistics, Denver, Colorado, pages 196–205. https://doi.org/10.3115/v1/N15-1020.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15(1):1929–1958. http://dl.acm.org/citation.cfm?id=2670313.

Adam Trischler, Zheng Ye, Xingdi Yuan, Philip Bachman, Alessandro Sordoni, and Kaheer Suleman. 2016. Natural language comprehension with the EpiReader. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 128–137. https://doi.org/10.18653/v1/D16-1013.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. *CoRR* abs/1506.05869. http://arxiv.org/abs/1506.05869.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Lina M. Rojas Barahona, Pei-Hao Su, Stefan Ultes, David Vandyke, and Steve Young. 2016a. Conditional generation and snapshot learning in neural dialogue systems. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 2153–2162. https://doi.org/10.18653/v1/D16-1233.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 1711–1721. https://doi.org/10.18653/v1/D15-1199.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, and Steve Young. 2016b. Multi-domain neural network language generation for spoken dialogue systems. 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 120–129. https://doi.org/10.18653/v1/N16-1015.

Wei Xu, Ralph Grishman, and Le Zhao. 2011. Passage retrieval for information extraction using distant supervision. In *Proceedings of 5th International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, pages 1046–1054. http://aclweb.org/anthology/I11-1117.

Jun Yin, Xin Jiang, Zhengdong Lu, Lifeng Shang, Hang Li, and Xiaoming Li. 2016a. Neural generative question answering. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI)*. IJCAI/AAAI Press, pages 2972–2978. http://www.ijcai.org/Abstract/16/422.

Pengcheng Yin, Zhengdong Lu, Hang Li, and Ben Kao. 2016b. Neural enquirer: Learning to query tables in natural language. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI)*. IJCAI/AAAI Press, pages 2308–2314. http://www.ijcai.org/Abstract/16/329.

Wenpeng Yin and Hinrich Schütze. 2015. Convolutional neural network for paraphrase identification. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 901–911. https://doi.org/10.3115/v1/N15-1091.

Kai Zhao, Liang Huang, and Mingbo Ma. 2016. Textual entailment with structured attentions and composition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, pages 2248–2258. http://aclweb.org/anthology/C16-1212.

Hao Zhou, Minlie Huang, and Xiaoyan Zhu. 2016. Context-aware natural language generation for spoken dialogue systems. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 2032–2041. http://aclweb.org/anthology/C16-1191.