# Summary Cloze: A New Task for Content Selection in Topic-Focused Summarization

**Daniel Deutsch and Dan Roth**
Department of Computer and Information Science
University of Pennsylvania
{ddeutsch,danroth}@seas.upenn.edu

## Abstract

A key challenge in topic-focused summarization is determining what information should be included in the summary, a problem known as content selection. In this work, we propose a new method for studying content selection in topic-focused summarization called the *summary cloze* task. The goal of the summary cloze task is to generate the next sentence of a summary conditioned on the beginning of the summary, a topic, and a reference document(s). The main challenge is deciding what information in the references is relevant to the topic and partial summary and should be included in the summary. Although the cloze task does not address all aspects of the traditional summarization problem, the more narrow scope of the task allows us to collect a large-scale datset of nearly 500k summary cloze instances from Wikipedia. We report experimental results on this new dataset using various extractive models and a two-step abstractive model that first extractively selects a small number of sentences and then abstractively summarizes them. Our results show that the topic and partial summary help the models identify relevant content, but the task remains a significant challenge.

## 1 Introduction

Topic-focused multi-document summarization (MDS) has long been a goal for natural language processing (Dang, 2005, 2006). In contrast to generic summarization, topic-focused summarization systems attempt to summarize a set of reference documents with respect to a specific topic or information need.

Recent research on summarization has mostly focused on generic summarization, in part due to the size of the available datasets. Work on topic-focused summarization (Lin and Bilmes, 2011; Ma et al., 2016; Feigenblat et al., 2017) is largely based on the DUC 2005 and 2006 datasets (Dang, 2005, 2006), which are orders of magnitude smaller than comparable generic summarization datasets (Nallapati et al., 2016; Grusky et al., 2018). Because the available corpora are so small, it is difficult to use them to train recent state-of-the-art summarization systems, which require large amounts of training data.

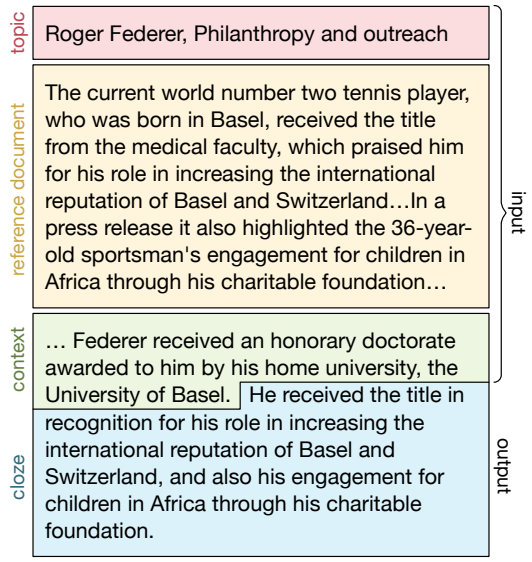


Figure 1: Given a topic, reference document, and a partial summary (the context), the objective of the summary cloze task is to predict the next sentence of the summary, known as the cloze.

Instead of focusing on the full topic-focused MDS problem, we address one aspect of the task known as *content selection*, the problem of deciding what information should be included in the summary (Nenkova, 2006b). Narrowing the scope of the problem makes it easier to collect a large-scale dataset that is tailored to the specific task and thus build upon recent work on summarization.

To that end, we formulate a new task for content selection in topic-focused summarization which

we call the *summary cloze* task (§2). Given a topic, a partial summary, and a reference document(s), the goal of the summary cloze task is to produce the next sentence in the summary (referred to as the *cloze*). The underlying assumption of the task is that the content of the cloze is known to come from the reference document, and thus the primary challenge of the task is to decide what information in the reference document(s) is relevant to the topic and partial summary. An example instance of the summary cloze task is presented in Figure 1.

Then, we collected a new large-scale summary cloze dataset from Wikipedia, called the WIKICITE dataset (§3). Each paragraph in Wikipedia can be viewed as a topic-focused summary of the references cited within the paragraph, where the topic is defined as the article title and section headings. The citations provide supervision at the sentence-level that indicates where the content of the preceding sentence came from. We scraped hundreds of thousands of Wikipedia articles and corresponding references to collect nearly 500k summary cloze instances.[1]

With a new task definition and corresponding large-scale dataset, we then propose an extractive model and a two-step abstractive model, both of which are based on recent successful work on generic summarization (§4). The extractive model combines representations of the topic and partial summary with representations of the document sentences through an attention mechanism to extract one reference sentence. The two-step model first reduces the length of the input data by extractively selecting a small number of sentences, then abstractively summarizes them using a decoder that has an initial hidden state which depends on the partial summary. Our experimental results (§6) show that the topic and context help the models identify relevant information, but the task remains a significant challenge.

The contributions of this work are three-fold: (1) We formalize the summary cloze task as a method to develop content selection models for topic-focused summarization; (2) we release the WIKICITE dataset, which contains nearly 500k summary cloze instances collected from Wikipedia; and (3) we propose sensible baseline extractive and abstractive models for content selection in topic-focused summarization.

[1]Available for download at https://github.com/danieldeutsch/wikicite

## 2 The Summary Cloze Task

One of the main challenges for a topic-focused MDS system is selecting what content from the reference documents to include in the summary. Intuitively, this decision largely depends on the topic itself and the content of the summary generated thus far. Therefore, being able to predict the next sentence of a summary is an important component of a full topic-focused MDS system and a worthwhile task on its own. With this motivation in mind, we now formally define the summary cloze task.

**Task Definition** Given a partial summary (the *context*) and a reference document or documents (the *references*), the goal of the summary cloze task is to generate the next sentence of the summary (the cloze). It is assumed that the information in the cloze comes from the provided reference documents. This is the most general definition of the summary cloze task. However, because we are interested in topic-focused summarization, in this work, it is also assumed that a high-level topic is provided as input, such as in Figure 1.

One advantage that the summary cloze task has over the traditional topic-focused MDS problem is the availability of data. Topic-focused MDS requires collecting example summaries of a set of documents about a particular topic. This data is hard to collect from human annotators and is difficult to find occurring naturally. The cloze task requires sentence-level supervision for which we were able to collect a large-scale dataset using Wikipedia (§3). By conditioning the cloze generation on a partial summary and working at the sentence-level, we are able to get around the problem of data scarcity.

**Evaluation** Evaluating summary cloze models is challenging; The difficulties of evaluating summarization systems (Nenkova, 2006a) still persist for the summary cloze task. We propose for the main evaluation metric to be ROUGE (Lin, 2004), which has been shown to be an effective metric for summarization (Owczarzak et al., 2012). Additionally, for abstractive models we use perplexity to quantify how likely the ground-truth cloze is according to the model.

Because the cloze is only one sentence long, the chance of the ground-truth and model clozes being equally valid but very different to each other is likely higher than for other summarization tasks.

It is desirable to have a good automatic evaluation metric that could capture the coherence of the model's cloze with the context in such situations. However, developing such an automatic metric for measuring summary coherence that correlates well with human judgments is an open problem for summarization in general, and thus we leave it for future work.

## 3 The WikiCite Dataset

In order to build content selection models, we collected a large-scale dataset of summary cloze instances using Wikipedia.

**Wikipedia as a Summary** Each paragraph in a Wikipedia article can be viewed as a topic-focused multi-document summary. Many Wikipedia articles have citations to external sources that provide supporting evidence for the information within the article. The articles' paragraphs can be viewed as summarizing the documents cited within the paragraph with respect to the page entity and section headings, which together form the topic of the summary. Since the citations are at the sentence-level, they provide supervision that identifies where the information in the preceding sentence came from, and thus can be used to generate summary cloze instances.

For example, consider a paragraph from the "Family and personal life" section of Barack Obama's Wikipedia page, presented in Figure 2. Nearly all of the sentences in the paragraph are supported with citations, and these citations contain the same information written in the paragraph. Each sentence with a citation can be used to produce a summary cloze instance where the preceding sentences are the context and the topic is "Barack Obama, Family and personal life." Although not all paragraphs on Wikipedia are as well-cited as the example in Figure 2, only one citation is necessary to create a cloze instance.

**Dataset Collection** To collect a summary cloze dataset, we first parsed the articles and citations from the January 2019 Wikipedia dump. Only articles marked with the category "Living people" were kept because these articles are more likely to have reference documents which can be scraped automatically. The reference document HTMLs were collected from 12 months of Common Crawl data.[2] Then, the text of each document was extracted from the HTML using unfluff.[3]

| |
|---|
| In June 1989, Obama met Michelle Robinson when he was employed as a summer associate at the Chicago law firm of Sidley Austin.[62] Robinson was assigned for three months as Obama's adviser at the firm, and she joined him at several group social functions but declined his initial requests to date.[63] They began dating later that summer, became engaged in 1991, and were married on October 3, 1992.[64] The couple's first daughter, Malia Ann, was born in 1998,[65] followed by a second daughter, Natasha ("Sasha"), in 2001.[66] The Obama daughters attended the University of Chicago Laboratory Schools. When they moved to Washington, D.C., in January 2009, the girls started at the Sidwell Friends School.[67] The Obamas have two Portuguese Water Dogs; the first, a male named Bo, was a gift from Senator Ted Kennedy.[68] In 2013, Bo was joined by Sunny, a female.[69] |
| 62: [Michelle] Obama was assigned to mentor Barack, who was a summer intern at Sidley & Austin… |
| 67: … Barack Obama and his wife have chosen Sidwell Friends School for their two daughters |
| 68: … the Obamas have chosen a 6-month-old Portuguese water dog — a gift from Senator Edward M. Kennedy… The girls have named the dog Bo… |

Figure 2: A paragraph from the "Family and personal life" section of Barack Obama's Wikipedia page and selected excerpts from the cited documents which provide supporting evidence.

In order to filter low-quality instances (e.g., HTML was a 404 page, the body text was incorrectly extracted, etc.), we labeled around 500 instances based on whether or not the reference document provided supporting evidence for the corresponding cloze. We trained a simple linear classifier with length and BM25 features that achieved an $F_1$ score of 85 to filter the data. The remaining data was split between training, 10k validation, and 10k testing instances, ensuring that no reference document or Wikipedia article appears in more than one set.

The dataset statistics compared to those of the DUC 2005 and 2006 datasets are presented in Table 1. Other than the number of instances, the biggest differences between the two datasets are the length of the reference documents and topics. The WIKICITE documents are significantly longer, over double the length DUC documents at one standard deviation above the mean. The DUC topics, which are generally text that describe an information need, are much longer than the WIKICITE topics, which are high-level topic keywords.

Examples from the dataset are provided in Appendix A.

[2] http://commoncrawl.org/

[3] https://github.com/ageitgey/node-unfluff

| Statistic | WikiCite | DUC 2005 + DUC 2006 |
|---|---|---|
| #Instances | 476,193 | 100 |
| #Document Tokens | 1081.47 (2103.1) | 723.3 (629.0) |
| #Document Sents | 51.3 (137.3) | 30.7 (30.1) |
| #Topic Tokens | 5.7 (3.4) | 29.5 (11.3) |
| #Context Tokens | 65.6 (56.1) | - |
| #Context Sents | 2.7 (2.3) | - |
| #Cloze/Summ. Tokens | 24.4 (11.9) | 277.8 (19.9) |
| #Cloze/Summ. Sents | 1 | 13.5 (3.4) |

Table 1: The WIKICITE and DUC 2005 and 2006 dataset statistics. The average values are reported with standard deviations in parentheses.

## 4 Content Selection Models

We propose two sensible extensions to successful single-document summarization systems for the summary cloze task, an extractive model and a two-step abstractive model.

### 4.1 Extractive Model

The extractive model is based on those presented in Kedzie et al. (2018). The base model works as follows. For every sentence $\mathbf{x}_i$ in the reference document, a sentence-level representation $\mathbf{h}_i$ is created by a sentence encoder (e.g., by averaging word vectors, by using a recurrent neural network, etc.). Then, a sentence extractor creates a document-level representation for each sentence $\mathbf{d}_i$, for example, by running a second RNN over the sentence-level representations. The document-level representation is passed through a multilayer perceptron and a sigmoid unit to calculate the probability of extracting a particular sentence under the model.

**Adding Topics and Context** We extended the standard extractive model by incorporating the topic and context information through an attention mechanism. First, each of the individual topics is encoded into a vector representation $\mathbf{t}_j$ by averaging the word vectors for each word in topic $j$. Then, a bidirectional RNN encodes the context tokens into a vector representation $\mathbf{c}_k$ for each token.

The topic and context token representations are combined with the sentence-level sentence representations through an attention mechanism. Specifically, an attention score is computed between each $\mathbf{h}_i$ and the topic and context representations. Then, the subsequent attention scores are used to create a weighted sum of the topic and context representations, $\mathbf{a}_i$.

Finally, the sentence representation $\tilde{\mathbf{h}}_i$ that includes the attention mechanism is computed as

$$\tilde{\mathbf{h}}_i = \mathbf{W}_a[\mathbf{h}_i; \mathbf{a}_i] \quad (1)$$

where $\mathbf{W}_a$ is a matrix of learned parameters.

In this extended version of the model, $\tilde{\mathbf{h}}_i$ is used to compute the extraction probabilities instead of $\mathbf{h}_i$. A graphical representation of the full extractive model is presented in Figure 3.

**Training & Inference** To convert the abstractive clozes into extractive reference sentence labels, we follow the procedure of Nallapati et al. (2017) and greedily select reference sentences that maximize ROUGE-1 recall until no more sentences improve the score. The training objective is the weighted negative log-likelihood of the labels with the same weighting as Kedzie et al. (2018). At inference, the sentence with the highest probability of extraction is selected.

### 4.2 Two-Step Abstractive Model

Since the reference documents in the WIKICITE dataset are rather long (see Table 1), we chose to use a two-step approach to build an abstractive system for the summary cloze task.

**Extractive Step** The first step uses the same extractive model from the previous subsection to significantly reduce the amount of input text. Instead of selecting just one sentence at inference, the extractive model repeatedly selects sentences until a pre-specified number of words has been met. Then, only the sentences which were extracted are passed as input to the abstractive step.

**Abstractive Step** For the abstractive step, we extended the Pointer-Generator + Coverage network (See et al., 2017) for single-document summarization to include the context.

The Pointer-Generator network is built on a sequence-to-sequence model with attention. The reference document is encoded using an RNN, and the summary is produced using a second RNN. The model is augmented with a copy mechanism by including a soft switch that allows the attention distribution to influence the decoder's probability distribution over the vocabulary, thus making it easier to copy words from the input.

Then, the coverage mechanism discourages the attention weights from repeatedly assigning high values to the same input tokens across decoding time steps. This is accomplished by adding a new

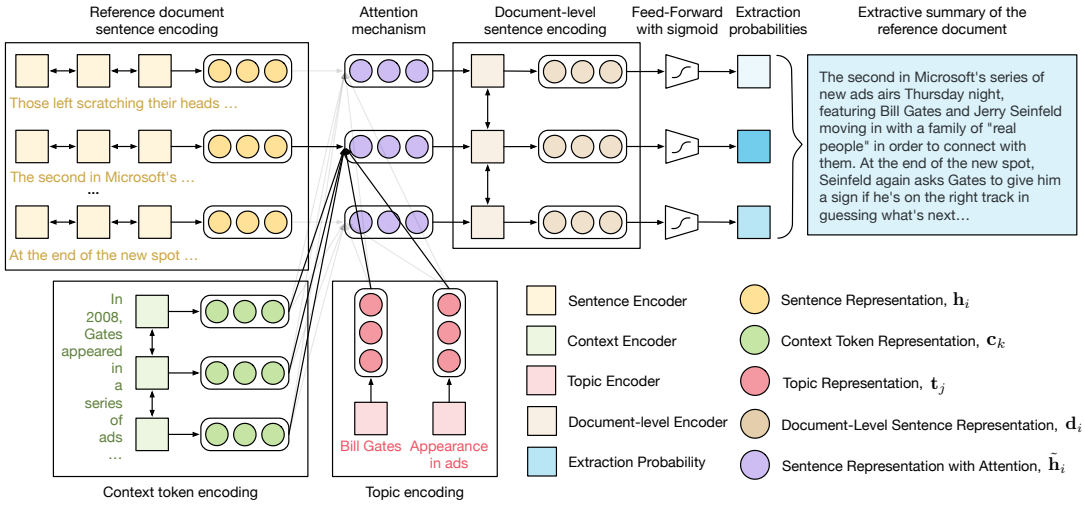


Figure 3: The extractive model uses three separate encoders create representations for the reference document sentences, context tokens, and topics. These are combined through an attention mechanism, encoded at a document-level, and passed through a feed-forward layer to compute an extraction probability for each reference sentence.

term to the loss function which penalizes such behavior and a new term to the attention score calculation that informs the model about the previous steps' attention weights. The coverage mechanism has the effect of decreasing the amount of redundancy in the generated summary.

**Conditioning on Context** In order to condition the generation of the cloze on the context, we use the context tokens to initialize the decoder's hidden state representation by first forcing the tokens through the decoder.

Concretely, let $\mathbf{s}_t$ be the decoder's hidden state for decoding step $t$ (with $\mathbf{s}_0$ initialized to the final encoder hidden state), $d_\theta(\cdot)$ be the decoder function, and $c_1, \ldots, c_T$ be the context tokens. The context tokens are passed through the decoder to create a new decoder hidden state representation using the following recursive formula

$$\mathbf{s}_t = d_\theta(c_t, \mathbf{s}_{t-1}) \qquad \forall t = 1, \ldots, T \qquad (2)$$

Then, $\mathbf{s}_T$ is used as the initial hidden state that the decoder uses to generate the cloze tokens.

Priming the decoder with the context tokens has the benefit of allowing the model to condition the cloze generation on a representation of the context. Additionally, because the context will impact the decoder's coverage calculation, it will discourage the cloze from repeating information already in the context.

**Training & Inference** Following See et al. (2017), the training objective is the negative log-likelihood of the ground-truth cloze tokens with the added coverage penalty. Inference is done using beam search.

## 5 Experimental Setup

Our experiments aim to better understand the WIKICITE dataset, evaluation metric, and performance of the extractive and abstractive models. Numbers in bold indicate statistical significance based on bootstrap resampling with $p = 0.05$.

### 5.1 Dataset Ablation & Human Performance

In order to better understand the WIKICITE dataset, we estimate how well models can perform without the reference documents and context and establish human performance on the task.

**No Reference Document** We used the publicly available large-scale language model from Radford et al. (2019) (the 345M parameter version) with the default parameters to generate the cloze. The language model (which we call NO REFERENCE) was conditioned on the context tokens and one full sentence was generated from the model without access to the reference document.

**No Context** We trained an abstractive model (§4.2) to summarize the first 200 tokens of the reference documents in one sentence, using the cloze as ground-truth, without access to the context. This is equivalent to training a single-document summarization system on the input text. This model is referred to as NO CONTEXT.

**Human Performance** We manually wrote clozes for 50 different instances with and without access to the topic and context to establish human performance on this task. We also had each cloze judged by 5 crowd workers to rate from 1 to 5 how coherent the topic, context, and cloze are. We report this average quality score alongside ROUGE with the HUMAN label.

### 5.2 Extractive Models

In addition to an ablation study on the extractive model (referred to as CONTENTSELECTOR; §4.1) where we remove the topic and context, we also evaluate several different extractive models on this task, described below.

**Lead & Oracle** The LEAD-1 chooses the first sentence from the reference documents as the cloze, a common baseline for summarization.

The ORACLE-1 model selects the one sentence from the reference set which maximizes the ROUGE score using the ground-truth. A different oracle model was created for each individual metric. The performance of the oracle represents the upper-bound ROUGE score that a system which extracts exactly one sentence can achieve.

**BM25** BM25 (Robertson et al., 1995) is an information retrieval-based score used to rank documents in search results for an input query. We selected the reference document sentence with the highest BM25 score for the context, treating the context as the query, the sentences as the documents, and computing the document frequency scores over the corpus of reference documents.

**SumFocus** SumFocus (Vanderwende et al., 2007) is an unsupervised model for topic-focused extractive summarization. The model constructs a unigram distribution over tokens by interpolating the unigram distributions of the documents and topic. Then, the model iteratively selects a sentence that has the highest probability token, discounts the probabilities of tokens in the selected sentence, and repeats. We extend SumFocus by additionally interpolating between the context unigram distribution in addition to the document and topic distributions.

### 5.3 Abstractive Models

Then, we measured how well the two-step abstractive system performs with and without access to the context and with different extractive models.

**Extractive Step Evaluation** We use the extractive models above to preprocess the reference set to select the top scoring sentences, up to 200 tokens. The preprocessed documents are evaluated against the ground-truth cloze by calculating the recall variant of ROUGE. This will measure what percent of the cloze $n$-grams are contained within the text input to the abstractive step.

In addition to the extractive models, we also report lead and extractive oracle scores. The LEAD-200 extractive step selects the first 200 tokens of the reference document. The extractive oracle model preprocesses the reference sentences by selecting only those which were labeled positively in the heuristically-generated extractive labels, limited to 200 tokens. This model, called HEURISTIC LABELS, serves to demonstrate how well the abstractive model would perform with a perfect extractive step.

**Abstractive Step Evaluation** Using the preprocessed documents from various extractive models, we trained abstractive models (§4.2) to produce the ground-truth cloze, both with and without access to the context. These systems were evaluated with the $F_1$ variant of ROUGE and perplexity.

### 5.4 Implementation Details

Our models were implemented using PyTorch (Paszke et al., 2017) in the AllenNLP framework (Gardner et al., 2017).[4] We used fixed 200-dimensional GloVe embeddings (Pennington et al., 2014). The RNNs were implemented with LSTMs (Hochreiter and Schmidhuber, 1997) of size 256. The models were trained with Adam (Kingma and Ba, 2014) using a learning rate of 1e-4 and 1e-3 and batch sizes of 32 and 16 for 290k and 62.5k iterations for the extractive and abstractive models, respectively. Following See et al. (2017), we trained the abstractive models without the coverage loss until convergence, then by 3k iterations with a coverage loss weight of 1.

## 6 Experimental Results

### 6.1 Human Performance & Evaluation

Table 2 shows the ROUGE and quality scores for the human-written clozes with and without access to the topic and context. For an additional comparison, we included a random human baseline that randomly shuffled the human-written clozes.

[4] All code available at https://github.com/danieldeutsch/summarize

| Model | R1 | R2 | RL | Qual. |
|---|---|---|---|---|
| GROUND-TRUTH | 100 | 100 | 100 | 4.0 |
| HUMAN | 31.78 | 16.94 | 28.32 | 3.9 |
| -TOPIC,-CONTEXT | 24.49 | 7.55 | 20.11 | 3.7 |
| RANDOM HUMAN | 9.96 | 0.08 | 8.24 | 2.3 |

Table 2: The ROUGE $F_1$ scores for the human clozes with and without access to the topic and context and crowdsourced quality judgments (on a 1-5 scale) of how well the cloze continues the corresponding text. The standard deviations for the quality judgments (omitted for space) were each around 0.9.

| Model | R1 | R2 | RL |
|---|---|---|---|
| NO REFERENCE | 14.47 | 1.43 | 11.41 |
| NO CONTEXT | 21.90 | 6.22 | 17.89 |

Table 3: The ROUGE $F_1$ results of the baseline models that do not have access to the references or context.

The ROUGE scores for when the human cloze writers had access to the topic and context are higher by 7.3 R1 points, a significant margin. From this observation, we can infer that, when provided with the topic and context, humans do a better job at identifying the same information present in the ground-truth cloze. This is indicative that the topic and context should also be necessary for a system to do the same.

However, the crowd worker quality judgments are roughly equal for the ground-truth and both human-written clozes, and human performance is at 31.78 R1, which is somewhat low on an absolute scale. We believe this reflects the difficulty of evaluating this task and summarization in general. Because there are many possible valid clozes (or valid summaries of documents in the more general case) and only one ground-truth cloze, evaluation with automatic metrics is difficult. Although ROUGE is imperfect, it does positively correlate with the crowd worker quality scores, with a Pearson's coefficient of 0.44.

### 6.2 Dataset Ablation

The two baseline models that do not have access to the references or context, respectively, are evaluated in Table 3. Although the NO REFERENCE model is a high-quality language model, it performs significantly worse than the NO CONTEXT model. This is likely due to the fact that the NO CONTEXT model was trained for this task.

| Model | R1 | R2 | RL |
|---|---|---|---|
| ORACLE-1 | 44.47 | 26.35 | 37.79 |
| LEAD-1 | 17.49 | 4.69 | 13.44 |
| BM25 | 21.29 | 5.75 | 16.17 |
| SUMFOCUS | 17.72 | 4.59 | 13.83 |
| CONTENTSELECTOR | **25.22** | **9.56** | **19.78** |
| -TOPIC | **25.18** | **9.59** | **19.74** |
| -CONTEXT | 22.06 | 6.47 | 16.63 |
| -TOPIC,-CONTEXT | 21.31 | 6.04 | 16.10 |

Table 4: The ROUGE $F_1$ scores for the extractive models, all of which are significantly lower than the oracle model, indicating there is room for improvement.

### 6.3 Extractive Models

The performance of the different extractive models on the summary cloze task is presented in Table 4.

The CONTENTSELECTOR model achieves the best score overall with 25.2 R1. Providing CONTENTSELECTOR with the topic and context improves the R1 score by over 4 points, however, this improvement is mostly due to the context. The topic does not change the model's performance by a statistically significant amount, which is an indication that it is not being properly utilized.

Among the symbolic models, BM25 outperforms SUMFOCUS by 3.57 R1. One reason for this could be that BM25 uses a token weighting to score each sentence based on how frequently the token appears in the entire reference corpora, whereas the standard SUMFOCUS implementation treats all tokens equally.

Although the models perform well, they are still significantly below the oracle model by at least 19 R1 and 16 R2 points. Such a large gap indicates that there is some signal in the data that is not yet captured by the extractive models.

### 6.4 Abstractive Models

**Extractive Step** The ROUGE recall scores for the different extractive models used to preprocess the input to the abstractive model are shown in Table 5. The results are largely consistent with the extractive model scores, with CONTENTSELECTOR performing the best, the topic causing no statistically significant difference, and a large gap remaining between current models and a perfect system (the heuristic labels).

Table 5 also contains an ablation study for the SUMFOCUS in which the topic did make a significant difference. This suggests that it may be easier to use the topic in a symbolic model and a model

| Model | R1-R | R2-R | RL-R |
|---|---|---|---|
| HEURISTIC LABELS | 80.52 | 36.02 | 55.41 |
| LEAD-200 | 60.97 | 22.39 | 45.09 |
| BM25 | 64.31 | 25.17 | 56.40 |
| SUMFOCUS | 61.91 | 22.84 | 46.02 |
| -TOPIC | 61.02 | 22.45 | 45.51 |
| -CONTEXT | 61.13 | 22.19 | 45.17 |
| -TOPIC,-CONTEXT | 59.18 | 20.72 | 43.49 |
| CONTENTSELECTOR | **66.43** | **27.74** | **58.42** |
| -TOPIC | **66.38** | **27.78** | **58.39** |
| -CONTEXT | 64.01 | 24.62 | 55.91 |
| -TOPIC,-CONTEXT | 63.14 | 23.82 | 55.14 |

Table 5: An evaluation using the recall variant of ROUGE of the different extractive preprocessing steps.

| Ext. Model | Abs. | R1 | R2 | RL | PPL |
|---|---|---|---|---|---|
| HEUR. LAB. | +C | 34.18 | 16.39 | 28.67 | 17.61 |
| | –C | 33.98 | 16.14 | 28.23 | 18.03 |
| LEAD-200 | +C | 23.06 | 7.18 | 19.05 | 38.76 |
| | –C | 21.90 | 6.22 | 17.89 | 57.31 |
| CONTSEL. | +C | **24.50** | **8.50** | **20.37** | 32.28 |
| | –C | 23.07 | 7.34 | 18.84 | 33.39 |

Table 6: The ROUGE $F_1$ and perplexity results for abstractive models with and without the context (+/–C) with heuristic labels, lead, and CONTENTSELECTOR extractive preprocessing steps.

may benefit from combining symbolic and continuous representations.

**Abstractive Step** The ROUGE scores and perplexities for the abstractive models that use three different extractive steps are presented in Table 6. We observe that the CONTENTSELECTOR extractive step outperforms LEAD-200 in both ROUGE and perplexity. Additionally, including the context improves performance across all models, with nearly a 1.5 R1 gain with the CONTENTSELECTOR. This result is a clear signal that priming the decoder with the context produces a better cloze.

Example output from the CONTENTSELECTOR model that uses the context can been seen in Figure 4. In general, the model tends to copy long sequences from the reference document, often preferring somewhat formulaic sentences that begin with phrases such as "he became," "she graduated," or "she received" and are similar in style to Wikipedia articles.

Similar to the extractive results, there is a sizable gap between the performance of the abstractive model when it uses the heuristic labels versus the extractive model for preprocessing. This indicates that improving the extractive model will provide large downstream abstractive improvements.

## 7 Related Work

**Summarization** Recent work on generic summarization has focused on single-document models on the CNN/DailyMail dataset (Nallapati et al., 2016), focusing on different neural architectures and loss functions for extractive and abstractive summarization (Cheng and Lapata, 2016; Nallapati et al., 2017; Chopra et al., 2016; Nallapati et al., 2016; See et al., 2017, inter alia). Our models were based on the successful approaches of See et al. (2017) and those described in Kedzie et al. (2018).

Most work on topic-focused (also referred to as "query-focused") summarization uses the DUC 2005 and 2006 datasets (Dang, 2005, 2006). Approaches are mostly extractive, with some work selecting sentences based on token frequency/salience (Otterbacher et al., 2005; Vanderwende et al., 2007), diversity with submodular functions (Lin and Bilmes, 2011), or using Bayesian networks (Daumé III and Marcu, 2006). To the best of our knowledge, ours is the first work which builds end-to-end models on a large-scale dataset for topic-focused summarization.

**Generating Wikipedia** In contrast to our work focusing on content selection for topic-focused summaries, there have been previous work interested in generating Wikipedia articles. Sauper and Barzilay (2009) try to generate articles from a small set of domains by learning to generate templates from similar articles using an integer linear program. Similarly, Banerjee and Mitra (2016) build topic classifiers based on sections on Wikipedia to identify content in search results for topic keywords as part of a pipeline to generate full articles. Like us, Liu et al. (2018) also employ a two-step abstractive approach, but their goal is to generate the lead Wikipedia paragraph from all of the citations and web search results.

## 8 Conclusion

In this work, we propose the summary cloze task, a new task for studying content selection in topic-focused summarization. By narrowing the scope from the full topic-focused summarization task to one at the sentence-level, we are able to collect a large-scale summary cloze dataset from Wikipedia. Our experimental results demonstrate

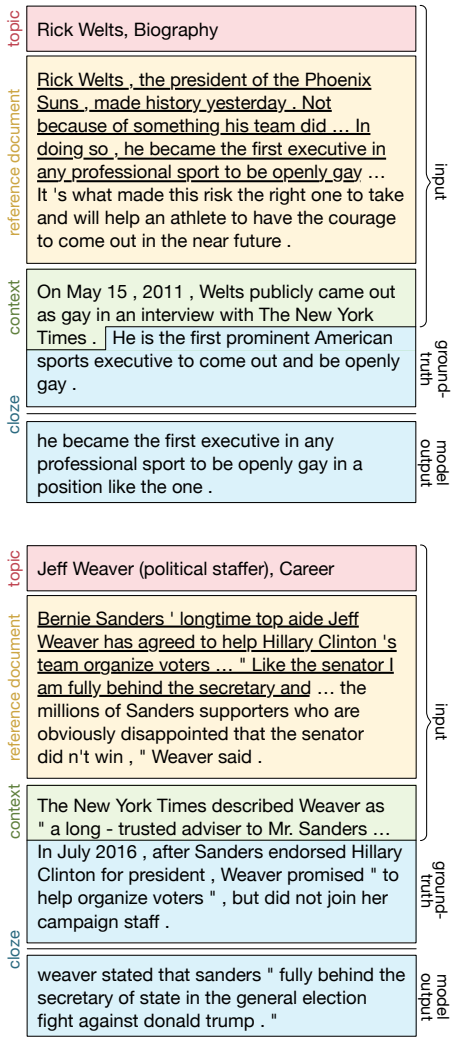


Figure 4: Example outputs from the abstractive model that uses the context. The model often copies sequences from the references which are sometimes correct (top) or incorrect but sensible (bottom), highlighting the difficulty of automatic evaluation. (Documents shortened for space. Sentences which are underlined were selected by the extraction step.)

that the topic and partial summary help the extractive and abstractive models, but the task remains a significant challenge with room for improvement in future work.

## Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful feedback and suggestions.

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053 (the ARL Network Science CTA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

## References

Siddhartha Banerjee and Prasenjit Mitra. 2016. WikiWrite: Generating Wikipedia Articles Automatically. In *IJCAI*, pages 2740–2746.

Jianpeng Cheng and Mirella Lapata. 2016. Neural Summarization by Extracting Sentences and Words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494. Association for Computational Linguistics.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive Sentence Summarization with Attentive Recurrent Neural Networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98. Association for Computational Linguistics.

Hoa Trang Dang. 2005. Overview of DUC 2005. In *Proceedings of the document understanding conference*, volume 2005, pages 1–12. Citeseer.

Hoa Trang Dang. 2006. Overview of DUC 2006. In *Document Understanding Conference*.

Hal Daumé III and Daniel Marcu. 2006. Bayesian Query-Focused Summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 305–312, Sydney, Australia. Association for Computational Linguistics.

Guy Feigenblat, Haggai Roitman, Odellia Boni, and David Konopnicki. 2017. Unsupervised Query-Focused Multi-Document Summarization Using the Cross Entropy Method. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 961–964, New York, NY, USA. ACM.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. AllenNLP: A Deep Semantic Natural Language Processing Platform.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. NEWSROOM: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational*

*Linguistics: Human Language Technologies*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Chris Kedzie, Kathleen McKeown, and Hal Daume III. 2018. Content Selection in Deep Learning Models of Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828, Brussels, Belgium. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Hui Lin and Jeff Bilmes. 2011. A Class of Submodular Functions for Document Summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 510–520, Portland, Oregon, USA. Association for Computational Linguistics.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating Wikipedia by Summarizing Long Sequences. *International Conference on Learning Representations 2018*.

Shulei Ma, Zhi-Hong Deng, and Yunlun Yang. 2016. An Unsupervised Multi-Document Summarization Framework Based on Neural Document Model. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1514–1523, Osaka, Japan. The COLING 2016 Organizing Committee.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A Recurrent Neural Network based Sequence Model for Extractive Summarization of Documents. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Ani Nenkova. 2006a. Summarization evaluation for text and speech: issues and approaches. In *INTERSPEECH*.

Ani Nenkova. 2006b. *Understanding the process of multi-document summarization: content selection, rewriting and evaluation*. Columbia University.

Jahna Otterbacher, Gunes Erkan, and Dragomir Radev. 2005. Using Random Walks for Question-focused Sentence Retrieval. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 915–922, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Karolina Owczarzak, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2012. An Assessment of the Accuracy of Automatic Evaluation in Summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 1–9, Montréal, Canada. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS-W*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at TREC-3. *Nist Special Publication Sp*, 109:109.

Christina Sauper and Regina Barzilay. 2009. Automatically Generating Wikipedia Articles: A Structure-Aware Approach. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 208–216, Suntec, Singapore. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. 2007. Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing & Management*, 43(6):1606–1618.

## A Additional WikiCite Examples

Here we include example instances from WIKICITE dataset in Figures 5, 6, and 7.

Cypress Grove (musician), Biography

Cypress Grove was born in London into a musical family. **His father was a professional jazz drummer** who at one time played with The Chris Barber band. In the late-1950s and early-1960s, Chris Barber was mainly responsible for arranging the first UK tours of seminal blues artists such as Big Bill Broonzy, Sonny Terry & Brownie McGhee and Muddy Waters. **Cypress was taught by his father to play drums at an early age. But by the time he turned 16, his interest had switched to guitar.**

His infatuation with pre-war acoustic blues continued to develop, which led to a chance encounter with Jeffrey Lee Pierce from The Gun Club. Jeffrey had a real passion for roots music of any kind, so the two hit it off immediately. Jeffrey and Cypress started to regularly jam together on acoustic guitars in Cypress' bedroom. Jeffrey eventually invited Cypress to collaborate with him on his long planned album of roots material - an album Jeffrey had been putting off for many years, as he could not find the right person to do it with - until he met Cypress. The album was released in 1992 - "Ramblin' Jeffrey Lee and Cypress Grove with Willie Love". Jeffrey and Cypress went on to tour the album on a handful of occasions, both with a band and as an acoustic duo.

October 1994. Jeffrey and Cypress are filmed for Henri-Jean Debon's "Hard Times Killin' Floor Blues" film- featuring Nick Cave. The film was eventually released in 2008.

Jeffrey and Cypress started to rehearse material for a follow up album. Jeffrey invited Cypress to join The Gun Club. Both events were destined not to happen. In mid 1995 Jeffrey returned to the States. Cypress takes him to the airport. It is the last time they will ever see each other. Jeffrey died in March 1996.

After Jeffrey's death, Cypress moved to France for some years. Traveling the entire country, playing various gigs, festivals and radio sessions.

In 2006, having returned to London, Cypress discovered a tape of Jeffrey and himself rehearsing some songs that Jeffrey had written.", These were plainly works in progress, and while the quality was deemed too poor for release, Cypress started to record the songs properly and invited a number of musicians to help him complete them, including Nick Cave, Iggy Pop, Thurston Moore, Debbie Harry, Chris Stein, Warren Ellis, Jim Sclavunos, Mark Lanegan, Isobel Campbell, Primal Scream, Mark Stewart and Lydia Lunch to name but a few. This series of collaborations became known as The Jeffrey Lee Pierce Sessions Project and resulted in what will eventually be a sequence of four albums - volume three will be released in May 2014. Isobel Campbell was so delighted with her duet with Mark Lanegan on volume one that she added the song to her live set, and invited Cypress to perform it live on stage with her and Mark at her London shows in 2008/9. It was around this time that Cypress and Lydia Lunch decided to expand on their highly successful pairing during the JLP sessions and started recording what would eventually become "A Fistful Of Desert Blues".

Cypress Grove was born into a musical family. He was taught by his father, a professional jazz drummer, to play drums from a very early age, but by the time he turned 16, his passion became the guitar.

Dharmajan Bolgatty, Biography

Vishnu Unnikrishnan-starrer Nithyaharitha Nayakan is all set for release on November 16. **The AR Binuraj directorial is Dharmajan Bolgatty's maiden production venture**. Dharmajan also has a starring role. Both Vishnu and Dharmajan have lent their voice to two of the songs.

The rest of the cast comprises Indrans, Basil Joseph, Jafar Idukki, Bijukuttan, Sunil Sukadha, Saju Navodaya, AK Sajan, Sajan Pallurithi, Robin Machan, Muhammed Prasad, Manju Pillai, Sruthi Jayan, Anju Aravind and Gayathri.

Binuraj directs from a script penend by Jayagopal. The editing is by Noufal Abdulla. Pavi K Pavan is handling the camera. Music is by Ranjin Raj. Dharmajan and Manu Thachett are bankrolling the film under the banner of Aadithya Creations.

After his full length role in Pappy appacha, his roles in Kattappanayile Rithwik Roshan, Aadu 1 & aadu 2 established him as a comedy actor in Malayalam film Industry. He is also announced to become a playback singer. He is also a businessman and has established fish stores himself. He is also a film producer and produced a film named Nithyaharitha Nayakan.

Figure 5: Addtional example instances from the WIKICITE dataset. The topic is in red, reference document in yellow, context in green, and cloze in blue. Relevant reference document information in bold.

Dustin Fowler, Career, New York Yankees

CHICAGO -- Joe Girardi stood near Dustin Fowler with his hands covering his face. The manager was absolutely gutted, heartbroken as Fowler sat on the ground, waiting for a cart to take him off the field. Fowler, a 22-year-old prospect who was selected by the Yankees in the 18th round of the 2013 MLB Draft, was making his MLB debut on Thursday night, living out his childhood dream. But with two outs in the bottom of the first inning, disaster struck. Fowler hustled to try to catch a foul ball off the bat of White Sox first baseman Jose Abreu down the right-field line. He collided with the short-wall, and his right leg slammed into it as he nearly plunged into the stands. It all happened so fast. Soon, Fowler was down and Girardi and trainer Steve Donohue were quickly out to check on him. His teammates surrounded him. Aaron Judge looked sick from the dugout. It was bad. A brace was placed on Fowler's right leg, and soon he was being picked up and carted off the field to applause. The Yankees eventually announced that Fowler had an open rupture of the right patella tendon in his right knee, and underwent surgery at Rush University Medical Center in Chicago. Surely, it is season-ending. Dustin Fowler leaves on a cart after crashing into right field wall. (Nam Y. Huh/AP) The whole scene was awful. **Fowler would've led off the second and had his first career at-bat.** He played but didn't bat in his MLB debut -- exactly 112 years to the day that Moonlight Graham did the same (June 29, 1905). Fowler had absolutely hit the cover off the ball -- both during the spring for New York and during the season for Triple-A Scranton. He displayed power. He displayed an ability to drive in runs. He was a gamer. The front office always raved about his makeup. "I'm overwhelmed right now," Fowler had said a few hours earlier in the Yankees' clubhouse. "I'm very excited, glad to be here." His promotion had caught him by surprise. But the Yankees have been decimated by injuries of late, leading to the promotions of fellow RailRiders Miguel Andujar (sent down to Triple-A on Thursday) and Fowler's close friend, Tyler Wade, in recent days. Clint Frazier could be the next Baby Bomber on his way up as a result of yet another injury -- this one to Fowler. Frazier has been a lightning rod for criticism -- much of it self-inflicted -- but is this the right time? Is he mature enough? And from a baseball perspective, do the Yankees want to bring him up if he isn't going to play every day? Sending our thoughts to Dustin Fowler of the @Yankees who left tonight's game following a collision with the right field foul wall.-- Chicago White Sox (@whitesox) June 30, 2017 After all, they already have superstar Aaron Judge and veterans Jacoby Ellsbury and Brett Gardner. Is he ready? Outfield prospect Jake Cave could be another option. Or perhaps the Yankees decide to bring Andujar back. They just designated Mason Williams for assignment. On Wednesday night, Fowler had been pulled from Triple-A Scranton's lineup. His mind raced. What's going on? Am I being called up? Am I being traded? He soon learned he was needed in Chicago. He flew from Syracuse to Chicago on Thursday afternoon, and didn't arrive until 4:30 p.m. CT. as typical traffic on I-90 from O'Hare International Airport halted his commute. He then had to wait out a 2-hour, 50-minute rain delay. And shortly after that, an awful injury. This wasn't how Dustin Fowler's career was supposed to start. And you can't help but feel for him.

Fowler made his major league debut on June 29, 2017, after a nearly three-hour rain delay, with the Yankees facing the Chicago White Sox on the road. During the first inning, he ran into a rail while chasing a fly ball, hitting his knee on a sharp edge of an electrical box. He collapsed to the ground and was carted off the field, before being diagnosed with an open rupture of the right patellar tendon. He was ruled out for the season. He underwent surgery that night at Rush University Medical Center. Fowler would have led off the next inning for his first major league plate appearance

Figure 6: Addtional example instances from the WIKICITE dataset. The topic is in red, reference document in yellow, context in green, and cloze in blue. Relevant reference document information in bold.

Caitlin Beevers, Refereeing career

Local rugby player Caitlin Beevers has certainly had a week to remember - scoring a try last Saturday as she won the Women's Challenge Cup, and learning that she will officiate in a fortnight's time at Wembley.

At just 16 years of age, the St John Fisher pupil was part of the Leeds Rhinos side that defeated Castleford Tigers 20-14 to win the trophy in Warrington, and she said the experience was "absolutely amazing".

"It was great," Caitlin said. "This is my first season at open age so I'm privileged to play a final as big as that, and it was amazing to experience that with all my team-mates."

"She described the elation of scoring a try in such a big game, saying that "it was absolutely amazing getting everyone to run and jump on you with excitement."

She added: "But in the end I wouldn't have got there without my teammates getting me up the field and giving me the opportunity to do things like that."

"It was a very tough game and we knew Cas were going to bring it all out as they always do, but I think we worked hard enough in the end."

She started playing rugby at the age of six, and says that she never expected to get so far and so quickly.

"I played at Birstall Victoria boys' under 7s. As soon as I got to under 12s to play for a girls' team, that's when it started really."

"I got into the 19s [at Leeds] and I was the youngest there, so I was absolutely over the moon with the opportunity I was given."

"I never expected to get into the first team, and it was just an amazing feeling that in my first year in open age I got that big an opportunity coming my way, and I was one of the first three to be chosen."

Caitlin says her natural position is at full-back, but that she has been asked to play in a variety of positions with the Rhinos this year.

"I've been moved about quite a lot! I've had hooker, winger, centre, loose forward, full-back. I never came into the team expecting to play full-back as we've got such an amazing one, Charlotte Booth, but I'm fine with any position I get put at."

"At the end of the day I'm lucky to get in the team, and if they think that might be the best position for me then I'll take it."

"She is very encouraged by the progress that women's rugby league is making and she said that is the case at youth levels as well."

"We've got loads of girls teams coming through now," she said. "When I first started, as you can imagine it was very small, and I only found one girl playing in a boys' team like I did."

But now you see girls' games and you see them playing from a young age and it's amazing to see them starting so young, and I hope that it does become a lot bigger than it is already."

"Winning the Challenge Cup wasn't the only highlight of her week though - **she also learnt that she will appear at Wembley, refereeing the Year 7 National Schools Final.**"

She said that she was "absolutely over the moon" to be chosen, although she waited a few days to share the news.

"I found out last Thursday I think, but I decided to wait until after the final to promote it as I wanted to focus on that and then focus on the refereeing."

Caitlin explained how she first got into refereeing at the age of 13.

"Originally I just wanted to do it to help out, because I took a gap year from playing and I wanted to stay in the game and still have a great deal to do with it.

"So I decided to go into refereeing, not expecting to do it that much but now I do it every weekend and I love it.

"I do community games for my society, the NCL every week, under 19s, academy games, and I got my first reserves the other week."

She had already officiated a national final for Year 8 girls, **but will now become the first female to referee any rugby league match at Wembley Stadium on 25th August, in the curtain-raiser to the men's Challenge Cup final.**

Beevers started refereeing aged 13 and is a member of Dewsbury & Batley Referees Society. In 2018 she became the first women to referee a rugby league game at Wembley Stadium when she refereed the Year 7 Boys National Schools Final (the curtain raiser to the Challenge Cup Final).

Figure 7: Addtional example instances from the WIKICITE dataset. The topic is in red, reference document in yellow, context in green, and cloze in blue. Relevant reference document information in bold.