QA-Align: Representing Cross-Text Content Overlap by Aligning Question-Answer Propositions

Daniela Brook Weiss [1] **Paul Roit** [1] **Ayal Klein** [1] **Ori Ernst** [1] **Ido Dagan** [1]
[1]Computer Science Department, Bar-Ilan University
{dani.b.weiss,plroit,ayal.s.klein,oriern}@gmail.com dagan@cs.biu.ac.il

Abstract

Multi-text applications, such as multi-document summarization, are typically required to model redundancies across related texts. Current methods confronting consolidation struggle to fuse overlapping information. In order to explicitly represent content overlap, we propose to align predicate-argument relations across texts, providing a potential scaffold for information consolidation. We go beyond clustering coreferring mentions, and instead model overlap with respect to redundancy at a propositional level, rather than merely detecting shared referents. Our setting exploits QA-SRL, utilizing question-answer pairs to capture predicate-argument relations, facilitating laymen annotation of cross-text alignments. We employ crowd-workers for constructing a dataset of QA-based alignments, and present a baseline QA alignment model trained over our dataset. Analyses show that our new task is semantically challenging, capturing content overlap beyond lexical similarity and complements cross-document coreference with proposition-level links, offering potential use for downstream tasks.

1 Introduction

End-to-end neural methods have become the de-facto standard for natural language understanding models. While these often work well for single document tasks, tasks that consider multiple textual inputs remain more challenging. A key difficulty concerns consolidating information from different, possibly redundant texts, which is crucial for applications such as multi-document summarization (MDS), sentence fusion (McKeown et al., 2010; Thadani and McKeown, 2013) or multi-hop question-answering (Welbl et al., 2018; Feldman and El-Yaniv, 2019). Previous works show that MDS methods for example, often just concatenate rather than merge inputs (Lebanoff et al., 2019a), or erroneously consolidate on non-coreferring elements (Lebanoff et al., 2020b). Recognizing such challenges, a few earlier MDS works (Liu et al., 2015; Liao et al., 2018; Shapira et al., 2017) attempted at leveraging semantic structures for consolidation, such as Abstract Meaning Representation (Banarescu et al., 2013) or Open Knowledge Graph (Wities et al., 2017), however, these initial means were found either too fine-grained (AMR) or immature (OKR) for efficient downstream consolidation.

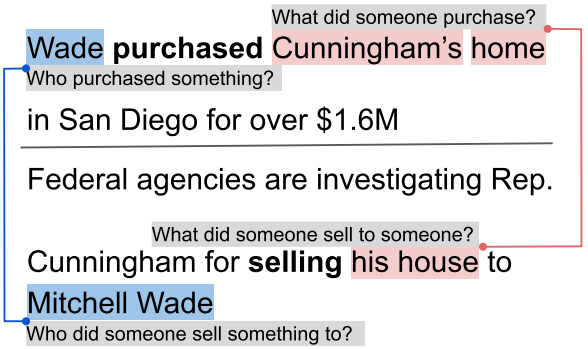


Figure 1: An example of two alignments expressing the same propositions captured by QASRL question-answers.

In this paper, we propose that a useful step toward effective consolidation is to detect and align *minimal* propositions that refer to the same information. Therefore, we choose to capture propositions at the fine-grained level of predicate-argument relations, each representing a single predication involving two sentence elements, along with the semantic role that relates them. Identifying these types of alignments at the propositional level would facilitate recognizing information redundancies and salience. Aligning propositions across documents may also prove useful to a variety of cross-text tasks, such as knowledge-base population, Multi-hop QA and cross document event extraction.

Consider the example in Figure 1; our alignments represent that both sentences express similar propositional information — sentence A talks

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 9879–9894
November 7–11, 2021.

about a "buying" event while sentence B is framing the same event through the "seller" perspective, essentially capturing the same reversed roles. In previous predicate-argument alignment works (Roth and Frank, 2012; Wolfe et al., 2013), predicates and arguments were aligned individually, rather than aligning the predicate argument relation as a whole, along with its semantic role. Our approach can be seen as a conceptual extension over the established Coreference Resolution framework, while handling content overlap at the more complex level of semantic relations (conveying stated information) rather than entity and event mentions (denoting referents). This difference in alignment scope also pertains to the cross-document coreference dataset used in these prior works, as we analyze and compare to in §5.2.

Unlike earlier work (Roth and Frank, 2012) that leveraged structured SRL to align predicate-argument-structures, we leverage the Question-Answer driven Semantic Role Labeling paradigm (QA-SRL) (He et al., 2015). QA-SRL captures predicate-argument relations using a naturally phrased question-answer pair, where the question type (who, what, where, etc.) reflects the role being captured, while the answer denotes the argument. For example, in Figure 1 sentence A, the Agent role (A0) of the predicate "purchase" is captured through the question-answer *Who purchased something? — Wade*. Once such relations are captured by QAs, we can align these QAs to capture similar propositional information expressed in a pair of sentences (see Figure 1). Since QAs are naturally intelligible and do not require a pre-defined schema, they can be easily annotated and used by non-experts.

Our contributions are outlined as follows: We introduce the QA-Align task, modeling cross-document proposition-level content overlap as QA-based alignments (§3). We show that this task, though semantically challenging, is attainable via crowd annotations, and publish our guidelines and crowdsourcing methodology. In addition, we compile and release our crowdsourced QA-Align dataset (accompanied by new QA-SRL annotations), over semantically similar paired texts (§4) that we collect. [1] We further analyze the quality of our data and compare it to an established cross-document (CD) coreference benchmark, ECB+ (Cybulska and Vossen, 2014) (§5). Finally, we implement a baseline modeling approach for QA-Align (§6), and present an analysis of appealing potential downstream use, over a sentence fusion task (§7).

[1] Our Code and data can be found here: https://github.com/DanielaBWeiss/QA-ALIGN

2 Related Work

Proposition alignment is closely related to the Cross Document Coreference Resolution (CDCR) task (Gooi and Allan, 2004; Mayfield et al., 2009). This task concerns clustering together entity or event mentions across topically related documents that refer to the same "real world" element, (entity or event). It has drawn substantial recent attention (Barhom et al., 2019; Zeng et al., 2020; Cattan et al., 2020; Yu et al., 2020), as its considered a fundamental intermediate task for cross-text language understanding. While the concept of coreference is essential for defining our alignment criteria (§3), we consider matching predicate–argument relations as expressing alignments of propositional information, and accordingly as a useful component in capturing information correspondence between texts, unlike prior approaches which match individual event and entity mentions disjointly.

Our work is largely inspired by the earlier predicate-argument alignment task (Roth and Frank, 2012, 2015; Wolfe et al., 2013, 2015). These works used either a semantic role labeler to align predicates to predicates (Roth and Frank, 2012), or annotated coreference chains to align both predicates and arguments, though disjointly evaluated (Wolfe et al., 2013). Unlike our work, these works did not take into consideration the relationship between the predicates and arguments, as expressed by the semantic role relating them. In this sense, these prior works are closer to the CDCR setting, whereas we directly align on propositional relations as the core units of interest, which, as we show later, captures more exhaustively information overlap across related texts (§5.2).

As potential uses for downstream tasks, various prior works presume that effectively handling multiple text sources, such as for multi-document summarization (MDS), requires intermediate explicit representations for cross-text content consolidation (Liao et al., 2018; Wities et al., 2017). Further, modeling redundant or overlapping text has been tackled through the sentence fusion task (Barzilay and McKeown, 2005; Marsi and Krahmer, 2005; McKeown et al., 2010; Thadani and McKeown,

2013), and recently as "disparate" sentence fusion, targeting related sentences in a single document (Nayeem et al., 2018; Geva et al., 2019; Lebanoff et al., 2019a,b, 2020b). In a recent series of investigations, Lebanoff et al. (2019a) highlight sentence fusion as a necessary step for improving summarization. In particular, pairs of sentences to be fused were empirically shown as a better source for generating summary sentences than single sentences (Lebanoff et al., 2019b). In Section 7, we analyze the potential utility of our QA alignments as a redundancy signal in a sentence fusion model.

QA-SRL QA-SRL has been shown to attain high quality annotation of predicate-argument structure for verbs, via crowdsourcing (FitzGerald et al., 2018), achieving above 90% coverage of PropBank arguments (Roit et al., 2020). In addition, Roit et al. (2020) showed that it captures much implicit information that is often missed by traditional SRL schemes. For these reasons, we employ the QA-SRL scheme for crowdsourcing propositional alignments and then modeling them in our QA-Align parser (§6).

3 Task Definition

Setting the task of aligning semantic content across text requires defining the informational units that are to be aligned. Moving from coreference resolution onto the propositional level, it is desired to pinpoint the minimal unit of propositional information, at which information overlap may be flexibly identified. Inspired by the seminal Neo-Davidsonian approach (Parsons, 1990) and following previous works (Roth and Frank, 2012, 2015; Wolfe et al., 2013; Liu et al., 2015), we view a single predicate-argument relation as an atomic propositional unit. As mentioned, we capture these using QA-SRL (He et al., 2015) Question-Answer (QA) pairs, as was shown in Figure 1.

The current scope of our work considers alignments between two sentences. We define the task as follows: given a pair of texts concerning (partly) overlapping scenarios, along with their predicate-argument structure representation — consisting of QA-SRL QAs in our setting, we want to find all cross-text alignments between QAs that refer to the same fact within the context of their sentences.

More concretely, our annotation guidelines, targeted for trained crowd workers (see §4), require aligned QAs to correspond on the following elements:

1. The main verbs in both questions refer to the same event.
2. Answers refer to the same referent or meaning.
3. Questions convey the same relation between the two, i.e. ask about same role of information regarding the predicate.

Figure 1 is an example of our crowd-sourced QA-alignments. In this example, our crowd-workers intuitively aligned the corresponding two roles for the predicates "buy" and "sell", even though their syntactic structure is reversed, as it is clear that all verbs, answers and questions correspond.

The vast majority of the alignments we observe are 1-to-1, that is, aligning a single QA from each sentence. However, we also allow for many-to-many alignments in our annotation protocol, which constitute about 4% of our data. These are required, for instance, to handle light-verb constructions, as in the following example:

1. *The owner hopes to* ***display*** *the painting*
2. *He hopes the Picasso painting will* ***go on display*** *tomorrow*

where the QA *What might someone display? — the painting* is aligned to the set of two QAs *What will something go on? — display*, and *What will go on something? — the Picasso painting*. Such alignments must also be minimal, meaning that taking out any QA from the alignment would posit it invalid.

Leveraging verbal QA-SRL, our task formulation targets verbal predicates, leaving the coverage of other predicate types, including nominalizations (Klein et al., 2020), for future work.

4 Dataset

4.1 Data Sources

Annotating cross-text propositional alignments requires a collection of semantically related sentences. Previous works (Roth and Frank, 2012; Wolfe et al., 2013; Lebanoff et al., 2019b) collected such sentence pairs based on computed similarity scores, which biases the data toward sentences that the utilized model already recognizes as similar. To avoid such bias, we leverage available human annotations that identified information overlap from various sources. Such sources yield naturally occurring texts, representative for multi-text applications,

Sentence Contributor	The **alleged bomb-builder was arrested in Egypt.**
Sentence Contributor	ABC News reported that **the alleged bomb maker behind the London attacks was arrested in Egypt.**
SCU Label	Bomb-maker arrested in Egypt

Table 1: Source summary sentences with their SCU contributing spans (in bold), and their given SCU Label.

	Train	Dev	Test
Num. Paired Texts	**1374**	**400**	**605**
Avg Num. Alignments	2.3	3.0	2.5
Num. QASRL questions	19191*	4299	8048
Total alignments	**3162**	**1205**	**1508**
No alignments	29%	16%	24%
Many-to-many	3%	4%	4%

Table 2: Statistics on the dataset collected. *Questions are produced using the QASRL parser. Over 94% of our many-to-many alignments are 2-to-1.

while challenging models with realistic lexical and content diversity. We next describe the concrete sources for our dataset.

ECB+ The Event Coreference Bank (ECB+, an extension over ECB) (Cybulska and Vossen, 2014) provides annotations for coreferring event and entity mentions across a "topic" — a set of documents related to the same event. We use this source to collect related pairs of sentences that share at least one coreferring verbal event mention. To control for data diversity, we take at most 6 sentence-pairs from each ECB+ topic, those with varying degrees of similarity based on a shared number of coreferring mentions.

DUC The Document Understanding Conference (DUC)[2] and the Text Analysis Conference (TAC)[3] both provide multi-document summarization evaluation datasets. In particular, we leveraged available manual Pyramid annotations over these datasets (Nenkova and Passonneau, 2004). Under this scheme, annotators extract from reference summaries small units of information, termed "summary content units" (SCUs), and then match them with corresponding information in system summaries to evaluate their quality. SCUs that repeat in several reference summaries create a cluster of similar *SCU contributors*, as can be seen in Table 1 in bold, and are given a manually written SCU label describing the main content of the cluster.

[2]https://www-nlpir.nist.gov/projects/duc/data.html, years used 2005-2008

[3]https://tac.nist.gov/, years used 2009-2011

Inspired by Thadani and McKeown (2013), we collect clusters of summary sentences that include matching SCU contributors. Then, we take each pair of summary sentences that appear in the same cluster to be a sentence-pair instance for our dataset.

	ECB+	**DUC**	**MN**
Paired Sentences	611	1028	740
Unique Sentences	704	1563	1265
Topics	86	347	503
Avg ROUGE 2 between pairs	0.21	0.12	0.13

Table 3: Distribution of our crowdsourced train, dev, and test data from the three different multi-text sources.

MultiNews MultiNews (MN) (Fabbri et al., 2019) is a recent Multi Document Summarization dataset, containing clusters of news articles along with human-written summaries. Recently, Ernst et al. (2020) crowdsourced alignments between semantically matching proposition spans, of a similar nature to Pyramid SCUs, across documents and summaries. To generate our data, we collected from their gold data pairs of sentences that include aligned spans.

Table 3 details the source distributions of our data. The average ROUGE-2 similarity score across paired sentences indicates that we indeed achieve a dataset of semantically similar sentences, yet exhibiting a limited degree of lexical overlap, providing challenging alignment cases. We maintained original train/dev/test splits for each corpora where suitably available and created our own otherwise. For more details regarding our dataset creation, see Appendix B.

4.2 Crowdsourcing

Aiming at high quality annotation, we applied a controlled crowdsourcing methodology (Roit et al., 2020), over the Amazon Mechanical Turk platform. Crowd workers were selected and trained, ensuring a reliable, scalable, cheap and rapid process.

4.2.1 QA-SRL Annotation

We begin with crowdsourcing the prerequisite QA-SRL annotation over our data. We followed the guidelines and protocol of Roit et al. (2020), utilizing their released annotation tools. For our dev and test sets, we collected a single worker's annotation per predicate. Evaluating their performance against an expert, we found its quality to match that reported in Roit et al. (2020). For the larger training set, we used the available QA-SRL parser[4] (FitzGerald et al., 2018) to produce question-answer pairs, thus reducing costs and demonstrating the feasibility of automatically generating question-answer pairs to be subsequently aligned by our proposed approach.[5]

4.2.2 Annotating QA-Alignments

Following the controlled crowdsourcing methodology, we first publish a "trap" task, for the purpose of identifying an initial pool of workers that perform adequately on the QA alignment task. We then begin a training process, where workers read detailed annotation guidelines and answer a follow-up survey.[6] Next, they go through 3 practice rounds, during which they receive detailed feedback. Subsequently, we selected 6 of the best performing workers for dataset annotation.

For our training set, each sentence pair was annotated by a single worker, who detected all QA alignments for that pair. For the development and test sets, each pair was annotated independently by two workers (*aligners*), yielding two QA alignment sets. A third worker then arbitrated any disagreements, yielding the final set of QA alignments for the sentence pair. We qualitatively observed that most disagreements arise from recall misses, while the rest arise from true semantic ambiguity. Section 5.1 presents inter-annotator agreement analysis, for both individual workers as well as for the arbitrated annotations.

Table 2 presents the details of our final dataset. As shown, the average number of QA alignments per sentence pair in the arbitrated Dev and Test sets ranges from 2.5 to 3. As each alignment corresponds to a matched (propositional) predicate-argument relation, this indicates a substantial level of information overlap captured in our data. We also note that our training set contains a high percentage of "no alignments", meaning instances where workers deemed nothing to align. This fact is likely attributed to using the QA-SRL parser to produce the QA pairs for each sentence (versus using human QA-SRL annotations), which is known to struggle with coverage (FitzGerald et al., 2018).[7]

Compensation Initially, workers are compensated for the time and effort invested during their training process. In data production, in addition to a fixed compensation per annotation instance (i.e. sentence pair), we use bonuses to account for instance complexity, which is measured by the size of QA sets for both sentences. Overall, aligners receive on average **$12**/hour, or **30¢** per instance, while arbitrators are paid **22¢** per instance, given it is an easier task. To conclude, a final instance in our gold dataset costs **82¢** (two *aligners* and one *arbitrator*), while a "silver" instance in our training set costs **30¢**.

5 Dataset Quality Evaluation

5.1 Inter-Annotator Agreement

To estimate the reliability of our annotated alignments, we measure inter-annotator agreement (IAA), both between single workers as well as between "teams" that produce our final annotation (i.e. two aligners and an arbitrator). Agreement between two alignment sets for an input sentence pair is measured simply by an exact match criterion. A QA-alignment is regarded as true positive if it fully matches a reference QA-alignment, while measuring agreement by the yielded F1 score.

The mean agreement between two single annotators (the two *aligners*) for the sentence pairs in our dataset is **71** F1. In 44% of the instances, the two aligners fully agreed in their annotation, where the remaining 56% instances are passed to a third worker for arbitration. Hence, since final annotations are collected using a team of workers, evaluating the overall consistency of our data requires assessing team-vs-team agreement. For this evaluation, we performed 5 experiments, each including two teams (disjoint worker triplets) which annotate the same 20 sentence pairs (for a total of 100 instances over the 5 experiments). Averaging the agreement score across the 5 experiments results

[4] http://github.com/nafitzgerald/nrl-qasrl

[5] A quality account for the automatically produced QAs and subsequent alignments can be found in Appendix D.

[6] Our task interfaces and our detailed guidelines can be found in Appendix C and at the end of the Appendix respectively.

[7] For more examples of our crowdsourced QA-alignments, see Appendix A.

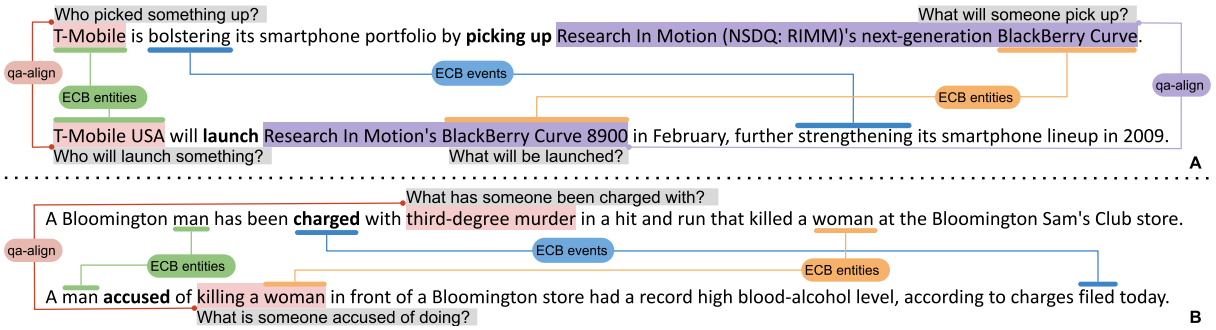


Figure 2: Examples of comparing ECB+ coreference annotations to QA-alignments. **A. Missed ECB+ event alignment** ECB+ fails to align the main event (in **bold**) while correctly aligning its core participants. In comparison, QA-Align annotates the arguments using propositions that depict the main event. QA-Align also aligns more propositions that include other ECB+ corefering events (blue underline), not shown for brevity. **B. Argument out of scope for ECB+** QA-Align aligns the theme of the charged/accused event mentions (in pink), while such type of participants are out of scope for ECB+, which considers only entities as arguments.

in **84** F1, expressing a substantial data consistency for a challenging semantic task.

Notably, the arbitration procedure reduced disagreements by almost 50%, and indicated that a major part of worker-vs-worker disagreements are not a matter of controversy. Rather, they stem from the difficulty of a single annotator to exhaustively identify all valid alignments between the two sets of QAs, justifying our annotation protocol.

5.2 Comparison to ECB+ Coreference Annotations

In this section, we assess and analyze the additional scope of information correspondence captured by our proposition-level annotation scheme, relative to the common paradigm in prior work which builds on event and entity coreference. To this end, we leverage the prominent ECB+ cross-document coreference dataset as a reference point over the corresponding subset of our data.

To compare our QA-alignments dataset with ECB+, we automatically induce one-to-one proposition level *ECB-based alignments* from event and entity mentions. We re-use gold QA-SRL data as the underlying propositions, and consider two QA pairs as aligned if their predicates corefer as mentions of the same event, and if their answers have any overlap with coreferring entity mentions. This procedure assumes that, in most cases, a match of both the question predicates and their answers would correspond to a match of the two QA pairs as a whole, including a match of their semantic role and compositional meaning.

We found that **70.3**% of the ECB-based alignments are covered by our QA-alignments. To understand the gap, we manually analyzed a random sample of 66 out of the 302 ECB-based alignments not covered by the QA-alignments in our dataset. The majority of these (77%) were erroneous or redundant, due to artifacts in their automatic generation process.[8] 20% of the non-matched ECB-based alignments correspond to proposition alignments that were effectively captured in our data, using different expressions in the sentence (see Example 4 in Table 7 in Appendix A, where a QA-alignment captures more information than an ECB+ coreference). Importantly, only two alignments (3%) reflect true misses of our annotations, evaluating the interpolated effective recall of our data relative to ECB+ at **99%**.

On the other hand, **37.3**% of our gold QA-alignments are not covered by the ECB-based alignments. Manually inspecting these illuminates the additional scope gained from addressing information overlap at the level of predicate-argument relations, as captured by QA pairs, rather than at the level of individual event and entity mentions. First, entity coreference often falls short in accounting for corresponding propositions, since many verbal arguments (answers) — e.g. those comprised of full clauses — are not entity mentions. In Figure 2-**B**, for example, although ECB+ aligns the events *charged / filed*, their corresponding "Theme" arguments (highlighted in pink) cannot be aligned us-

[8] Redundancies commonly derive from within-document entity coreference, where two mentions (e.g. *the man* and *he*) yield two redundant QAs (*Who came? — the man* and *Who came? — he*). Erroneous ECB-based alignments mostly consist of QAs whose answers do not corefer, but do encompass coreferring entity mentions, e.g. the misleading **he** in "When did someone come? – after **he** finished working".

ing entity coreference. More generally, identifying overlapping information may fail using naive coreference criteria. Consider the *picking up* / *launch* events in Figure 2-**A**. The two verbs are quite distinct in terms of their basic lexical meaning, which is probably why ECB+ refrains from annotating them as coreferring events. Nevertheless, as manifested by the predicate-argument relations involving these mentions, it is clear that in the given context, these two event mentions convey the same information.

All in all, these analyses show that proposition-level alignments are required in order to properly capture information overlap, complementing the more elementary task of traditional event and entity coreference resolution.

6 Baseline QA-Align Model

6.1 Model Description

Taking a simplistic modeling approach as an initial baseline, we reduce the QA alignment prediction problem into a binary classification task. Let (S_1, S_2) be a sentence pair, where each sentence is provided with a set R of QA-SRL relations (QAs):

$$r_i^s \in R_s\ ,\ 1 \geq i \geq |R_s|\ ,\ s \in \{1, 2\}$$

Our baseline considers only 1:1 alignments (recall this covers 96% of the cases in our data). Given a single pair of QAs (r_i^1, r_j^2) as a candidate alignment along with their contexts, each comprising of the sentence itself and its predecessor sentence, the model predicts the probability that the pair of QAs is aligned. We serialize the full information of a candidate QA pair into the input sequence, and feed it into a pre-trained language model with a binary classifier on top, as typical for sequence classification. An example input for our model can be seen in Table 4. For each QA, we concatenate the question and the context sentences, while denoting the predicate and the answer span with special markup tokens, attuning the model toward the correspondences within and across candidates (Baldini Soares et al., 2019; Lebanoff et al., 2020a). More details about the training procedure appear in Appendix E. During inference, we treat each QA-to-QA candidate alignment as an edge and weigh it with the probability score produced by the model, filtering edges scored under a threshold $\tau = 0.5$. The final alignment set is then decoded using maximal bipartite matching.

We experimented with BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and CorefRoBERTa (Ye et al., 2020)[9] as the pretrained models. Following previous works on CD coreference (Cybulska and Vossen, 2014; Barhom et al., 2019; Zeng et al., 2020) and predicate-argument alignment (Wolfe et al., 2013; Roth and Frank, 2012), we compare our models to a lemma-based baseline method. Similarly to the alignment criterion applied for the ECB-based analysis of our crowdsourced data (§5.2), our lemma baseline model aligns QA pairs in which the two predicates, as well as the head words of the answer spans, share a lemma.

6.2 Model Performance

Results are shown in Table 5. [10] Notably, the lemma baseline performance is relatively low, reinforcing that aligning predicate-argument propositions across texts is challenging.[11] Applying pre-trained language models yields a modest yet clear improvement. The best performance is obtained using CorefRoBERTa (Ye et al., 2020), which is specialized on coreferential reasoning by further pre-training on predicting masked entity mentions repeating in a single document. Overall, we suggest that QA alignment provides a challenging task for future research, with much room for improvement, particularly considering the high inter-annotator agreement of our laymen annotators (§5.1).

Analysis shows that the lemma baseline, while exhibiting high precision (as expected), misses many alignments where the predicates or arguments are paraphrastic rather than lexically identical. On the other hand, we find that the alignment model often misses alignments of similar questions that include different "Wh" words. For instance, the QA alignment containing *Why was someone pulled over?* and *What was someone pulled over for?* is missed by our model. Indeed, recognizing "same-role" paraphrastic questions is a known challenge in the context of QA-SRL evaluation (FitzGerald et al., 2018; Roit et al., 2020), with

[9] The paper presented CorefBERT, while we use a released CorefRoBERTa version.

[10] Skipped many-to-many alignments are counted as recall misses.

[11] As a rough non-comparable reference point, it is interesting to note that the "same lemma" baseline for CD event coreference achieves 76.5 CoNLL F1 on ECB+, providing a strong baseline which recent state-of-the-art models only surpass by 4 to 8 F1 points. (Cattan et al., 2020; Zeng et al., 2020).

Example QA-Align Model Input	
Input A	Who did someone [P] fire [/P] ? [Q] The Philadelphia 76ers [P] fired [/P] [A] coach Maurice Cheeks [/A] on Saturday, one day after the team continued its slide with a season-worst offensive effort, dpa reported.
Input B	Who was [P] fired [/P] ? [Q] If you don't know by now: you disappoint in the NBA, you get canned. Today, [A] Maurice Cheeks [/A] became the fifth coach [P] fired [/P] within the first quarter of the season.

Table 4: Example input encoding for our baseline models (§6.1).

Method / Model	Dev			Test		
	P	R	F1	P	R	F1
Lemma	89	35	50	**89**	30	45
Bert-base-cased	72	51	60	62	41	49
Roberta-base	66	50	57	58	44	50
CorefRoberta	71	59	64	60	**48**	**53**

Table 5: Precision, Recall, and F1 results for multiple QA-Align baselines.

potential value also for QA-SRL downstream tasks, and is left for future work.

7 Analysis of Potential Downstream Use

Aiming at an initial analysis of the potential extrinsic utility of our alignments, we experiment with using them as a signal for the sentence fusion task (Barzilay and McKeown, 2005; Marsi and Krahmer, 2005). Sentence fusion is a summarizing task applied for the single sentence level, thus providing a suitable use-case for exploring the extrinsic utility of our sentence-level alignments. Indeed, it was shown that fusion models tend to generate naive extractive summaries that are copied from a single source (Lebanoff et al., 2019a), rather than merging information from multiple sources. In that vein, we aim to examine whether explicitly incorporating alignments into the fusion task would encourage a model to attend to corresponding information originated in multiple sentences.

7.1 The Sentence Fusion Experiment

Barzilay and McKeown (2005) formulated sentence fusion as a generation task that consolidates the information across multiple input sentences into a single sentence. Here, we adopt and reproduce the fusion dataset constructed by Thadani and McKeown (2013) (referred in §4), which is the most recent source for *multi-document* sentence fusion (vs. the different type of fusion within a single document; see Appendix F for more details regarding the dataset). In this dataset, the input for fusion instances consists of clusters of two to four source sentences, originating in different documents, which need to be summarized.

We create a new modern baseline for the dataset of Thadani and McKeown (2013), which outperforms their pre-neural one, evaluated using bigram-F1. As a baseline end-to-end fusion model, we employ the pre-trained auto-encoder BART (Lewis et al., 2020), which has achieved state-of-the-art results on summarization tasks. In comparison, we predict QA alignments for the fusion data using our best reported model, and then incorporate them into the fusion model (termed Fuse-Align) using an input-augmentation technique (resembling the one in §6.1). As shown in Table 6, the input encompasses the alignment information by attaching indexed markup tokens around aligned predicates and arguments.[12]

7.2 Results and Analysis

We create 20 different variations of the dataset by randomly shuffling the order of the input sentences in each input, and take the average ROUGE 2 (R2) across all runs. We find that the baseline and Fuse-Align models achieve a similar R2 of 41 and 40 respectively. Although ROUGE scores are similar, upon closer examination we find the outputs of both models to be of different nature.

In order to quantify these differences, we automatically classify fusion outputs stemming from single versus multiple source sentences. Based on lexical overlap, we link every word in the fused output to its contributor source sentences, if exist. An output is considered multi-sourced or *consolidating* if it consists of different words exclusively originating from separate sources (at least two). This classification is related to the distinction of abstractive vs. extractive summarization, where merging multiple sources in a summary sentence falls under the abstractive summarization regime, and is not possible in extractive summarization.

Based on this classification, we find that Fuse-Align significantly creates substantially more consolidating outputs (**30**%) compared to the baseline

[12] In this initial trial, we do not incorporate the questions in our alignments.

Example Fuse-Align Input	Law enforcement agencies [P1] use [\P1] [A1] dogs [\A1] worldwide. </s> Dogs perform many different law-enforcement tasks around the world. </s> City and county police agencies, customs departments, fire departments, the Secret Service, highway patrol, border patrol, military bases and some prisons in the US and many other countries [P1] use [\P1] [A1] dogs [\A1] to help in law enforcement work.
Baseline Output	Dogs perform many different law-enforcement tasks around the world
Fuse-Align output	Law enforcement agencies use dogs to help in law enforcement

Table 6: An example input for the Fuse-Align model marked with predicted alignments (the baseline does not include these tokens). The indices in the special tokens indicate which spans are aligned across sentences, </s> are sentence separators. The orange span is the one contributing to the baseline's output, while the blue spans contribute to Fuse-Align's output, showcasing a merge across redundant and complementary information that our alignment model identifies.

model (**20**%), as illustrated in Table 6 (see Appendix G for more examples).[13]

To further investigate qualitative differences, we analyze a sample of 50 multi-source outputs from each model. We find that both models consistently produce grammatical and faithful outputs for these short input/output pairs, and exhibit a similar level of compression. The key difference attributed to our Fuse-Align input augmentation is thus encouraging the BART seq-to-seq model to merge multi-source information 50% more frequently than over the raw input, in the baseline configuration, and also triggering merges of more than two input sentences, which never happens for the baseline.

To conclude, though further investigation of this trend is worthwhile, this analysis demonstrates a promising potential utilization of propositional alignments. The fact that adding alignment information to BART's input encourages it merge information sources much more frequently is notable, especially as sequence-to-sequence MDS models often tend to produce mostly extractive outputs (Lebanoff et al., 2019a).

8 Conclusion

We present a new task for aligning propositional predicate-argument information, captured via the question-answer pairs of QA-SRL. We compile and publish a QA-Align dataset and present our crowdsourcing methodology, showing that a high inter-annotator agreement is achievable for this challenging semantic task even for non-expert annotators. We compare our annotation scheme to that of ECB+ with respect to aligning information across related texts, and show that alignments at the level of predicate-argument relations capture substantially more information correspondences than aligning individual referents. In addition, we present a baseline QA-Align model, and utilize its predicted alignments to increase information consolidation within a high-redundancy sentence fusion task. Our exploratory findings warrant further future investigations concerning the potential of predicate-argument alignments for modeling cross-document consolidation.

[13] In both models the majority is still single-sourced, probably due to the "extractive" nature of the this data (Thadani and McKeown, 2013).

Acknowledgments

We would like to thank the anonymous reviewers for their thorough and insightful comments. The work described herein was supported in part by grants from Intel Labs, Facebook, and the Israel Science Foundation grant 1951/17.

References

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Shany Barhom, Vered Shwartz, Alon Eirew, Michael Bugert, Nils Reimers, and Ido Dagan. 2019. Revisiting joint modeling of cross-document entity and event coreference resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4179–4189.

Regina Barzilay and Kathleen McKeown. 2005. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31:297–328.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2020. Streamlining cross-document coreference resolution: Evaluation and modeling. *arXiv preprint arXiv:2009.11032*.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4545–4552, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ori Ernst, Ori Shapira, Ramakanth Pasunuru, Michael Lepioshkin, Jacob Goldberger, Mohit Bansal, and Ido Dagan. 2020. SuperPAL: Supervised proposition alignment for multi-document summarization and derivative sub-tasks. *CoRR*, abs/2009.00590.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Yair Feldman and Ran El-Yaniv. 2019. Multi-hop paragraph retrieval for open-domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2296–2309.

Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer. 2018. Large-scale QA-SRL parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2051–2060, Melbourne, Australia. Association for Computational Linguistics.

M. Geva, E. Malmi, I. Szpektor, and J. Berant. 2019. DiscoFuse: A large-scale dataset for discourse-based sentence fusion. In *North American Association for Computational Linguistics (NAACL)*.

Chung Heong Gooi and James Allan. 2004. Cross-document coreference on a large scale corpus. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 9–16.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 643–653.

Ayal Klein, Jonathan Mamou, Valentina Pyatkin, Daniela Stepanov, Hangfeng He, Dan Roth, Luke Zettlemoyer, and Ido Dagan. 2020. QANom: Question-answer driven SRL for nominalizations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3069–3083, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Logan Lebanoff, Franck Dernoncourt, Doo Soon Kim, Lidan Wang, Walter Chang, and Fei Liu. 2020a. Learning to fuse sentences with transformers for summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4136–4142, Online. Association for Computational Linguistics.

Logan Lebanoff, John Muchovej, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019a. Analyzing sentence fusion in abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 104–110, Hong Kong, China. Association for Computational Linguistics.

Logan Lebanoff, John Muchovej, Franck Dernoncourt, Doo Soon Kim, Lidan Wang, Walter Chang, and Fei Liu. 2020b. Understanding points of correspondence between sentences for abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 191–198, Online. Association for Computational Linguistics.

Logan Lebanoff, Kaiqiang Song, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019b. Scoring sentence singletons and pairs for abstractive summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2175–2189, Florence, Italy. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Kexin Liao, Logan Lebanoff, and Fei Liu. 2018. Abstract meaning representation for multi-document summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1178–1190, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Fei Liu, Jeffrey Flanigan, Sam Thomson, Norman Sadeh, and Noah A. Smith. 2015. Toward abstractive summarization using semantic representations.

In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1077–1086, Denver, Colorado. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Erwin Marsi and Emiel Krahmer. 2005. Explorations in sentence fusion. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*.

James Mayfield, David Alexander, Bonnie J Dorr, Jason Eisner, Tamer Elsayed, Tim Finin, Clayton Fink, Marjorie Freedman, Nikesh Garera, Paul McNamee, et al. 2009. Cross-document coreference resolution: A key technology for learning by reading. In *AAAI Spring Symposium: Learning by Reading and Learning to Read*, volume 9, pages 65–70.

Kathleen McKeown, Sara Rosenthal, Kapil Thadani, and Coleman Moore. 2010. Time-efficient Creation of an Accurate Sentence Fusion Corpus. pages 317–320.

Mir Tafseer Nayeem, Tanvir Ahmed Fuad, and Yllias Chali. 2018. Abstractive unsupervised multi-document summarization using paraphrastic sentence fusion. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1191–1204, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

Terence Parsons. 1990. Events in the semantics of english: A study in subatomic semantics.

Paul Roit, Ayal Klein, Daniela Stepanov, Jonathan Mamou, Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer, and Ido Dagan. 2020. Controlled crowdsourcing for high-quality QA-SRL annotation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7008–7013, Online. Association for Computational Linguistics.

Michael Roth and Anette Frank. 2012. Aligning predicate argument structures in monolingual comparable texts: A new corpus for a new task. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 218–227, Montréal, Canada. Association for Computational Linguistics.

Michael Roth and Anette Frank. 2015. Inducing implicit arguments from comparable texts: A framework and its applications. *Computational Linguistics*, 41(4):625–664.

Ori Shapira, Hadar Ronen, Meni Adler, Yael Amsterdamer, Judit Bar-Ilan, and Ido Dagan. 2017. Interactive abstractive summarization for event news tweets. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 109–114.

Kapil Thadani and Kathleen McKeown. 2013. Supervised sentence fusion with single-stage inference. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1410–1418, Nagoya, Japan. Asian Federation of Natural Language Processing.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Rachel Wities, Vered Shwartz, Gabriel Stanovsky, Meni Adler, Ori Shapira, Shyam Upadhyay, Dan Roth, Eugenio Martínez-Cámara, Iryna Gurevych, and Ido Dagan. 2017. A consolidated open knowledge representation for multiple texts. In *Proceedings of the 2nd workshop on linking models of lexical, sentential and discourse-level semantics*, pages 12–24.

Travis Wolfe, Mark Dredze, and Benjamin Van Durme. 2015. Predicate argument alignment using a global coherence model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 11–20, Denver, Colorado. Association for Computational Linguistics.

Travis Wolfe, Benjamin Van Durme, Mark Dredze, Nicholas Andrews, Charley Beller, Chris Callison-Burch, Jay DeYoung, Justin Snyder, Jonathan Weese, Tan Xu, and Xuchen Yao. 2013. PARMA: A predicate argument aligner. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 63–68, Sofia, Bulgaria. Association for Computational Linguistics.

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential Reasoning Learning for Language Representation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7170–7186, Online. Association for Computational Linguistics.

Xiaodong Yu, Wenpeng Yin, and Dan Roth. 2020. Paired representation learning for event and entity coreference. *arXiv preprint arXiv:2010.12808*.

Yutao Zeng, Xiaolong Jin, Saiping Guan, Jiafeng Guo, and Xueqi Cheng. 2020. Event coreference resolution with their paraphrases and argument-aware embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3084–3094, Barcelona, Spain (Online). International Committee on Computational Linguistics.

A Examples of Crowdsourced QA-Alignments

Table 7 presents a few examples collected by our crowd-workers following our crowdsourcing methodology. Examples 1 & 3 present an example of lexically different entities being captured and aligned using our proposition alignments. Example 2 is a light-verb construction case talked about in §3. Example 4 (Align2) presents an instance where QA-SRL captures a *causality* discourse relation (in both sentences), not explicitly manifested in syntactic structure (and indeed missed by an SRL parser trained on PropBank). This enables our workers to align semantically rich propositions that go beyond syntactic analysis. Consequently, workers align on these relations rather then the more narrow alignment captured by the ECB-based alignments to sentence B's QA: *What was someone charged in? — her murder*, resulting from a coreference link between *Woman* and *her*.

B Dataset Construction – Further Details

In this section we further explain the details regarding our dataset creation corpus. Given the smaller number of topics annotated in ECB+ and DUC, we don't want to take too many pairs of sentences that overlap over the same content, therefore we filter out any pair of sentences with ROUGE 2 above 0.9. In addition, we attach previous sentences to each instance for context, used when presenting the instance to crowd-workers, and in our QA-Alignment model. In terms of train, dev, and test splits, we use official splits where available, and create our own where needed. For ECB+, we take the official split published; for MultiNews, we also use the official split provided, and in addition, supplement with 5 extra topics from the train and add it to the dev and test (each gets unique 5 topics), given a very small amount of sentence pairs were released with gold annotations (312 and 345 for dev and test respectively). For DUC, an official split is not available, therefore we follow Thadani and McKeown (2013) official sentence fusion data split, based on DUC and TAC's published years. Table 8 presents the sources distributions by training splits.

B.1 ECB+

Due to the relative small number of topics annotated in ECB+ (Cybulska and Vossen, 2014), with usually only a handful of shared events per topic and on average only 2 annotated sentences per document, the resulting pairs are not well varied in content and are highly repetitive. Because of this, we rank the sentence pairs in ECB+ by their number of shared coreferring mentions, and limit to top 6 pairs per topic. In addition, we collect sentence pairs from the lowest levels of shared coreferring mentions, in order to maintain a lexical variety and negative examples in our data. We select only verbal predicates from ECB+ when creating semantically related pairs.

B.2 DUC

In addition to the summary sentence pairs we collect using their shared SCU label, we also collect document to document, and document to summary pairs of sentences using the SCU Marked Corpus (Copeck and Szpakowicz, 2005), which links SCU labels to document source sentences. We don't perform any filtering on DUC, given it's varied in content and lexicality due to the large amount of topics and years available.

B.3 MultiNews

For MultiNews (MN), we also extract document to document aligned sentences, using a shared aligned summary sentence as an anchor. Because two different document sentences can be aligned to the same summary sentence, but with different non-overlapping spans, we also filter out any document pairs that have a span IOU (intersection over union) of lower than 0.1. We measure this IOU based on how much overlap the two document sentences have in their aligned spans to the summary sentence they share.

C Crowdsourcing

C.1 User-Interface

Figures 3 and 4 exhibit Step 1 in our crowdsourcing task on Amazon Mechanical Turk. Workers are instructed to read a pair of sentences carefully, noting any overlap of information they find, and in

	Sentence A: An earthquake measuring 4 . 6 **rattled** Sonoma and Lake counties early Thursday, according to the U.S. Geological Survey. Sentence B: The temblor **struck** about 26 miles north of Santa Rosa in the Geysers area.		
Ex 1	**Aligned QAs from A**		**Aligned QAs from B**
	What **rattled** something? — An earthquake measuring 4 . 6	–Align1–	Who **struck** somewhere? — The temblor
	What did something **rattle**? — Lake counties	–Align2–	Where did someone **strike**? — about 26 miles north of Santa Rosa in the Geysers area
	What did something **rattle**? — Sonoma		
	Sentence A: A jury in eastern Oklahoma has **convicted** a 27 - year - old man for killing his girlfriend. Sentence B: Man **found** GUILTY of shooting girlfriend		
Ex 2	**Aligned QAs from A**		**Aligned QAs from B**
	Who did someone **convict**? — a 27 - year - old man	–Align1–	Who was **found** as something? — Man
			What was someone **found** as? — GUILTY
	Sentence A: Piccard **got** Chinese airspace clearance before starting but was limited to a narrow strip, forbidden north of the 26th parallel. Sentence B: They gained vital time by **obtaining** flight permission from China before taking off.		
Ex 3	**Aligned QAs from A**		**Aligned QAs from B**
	What did someone **get**? — airspace clearance	–Align1–	What did someone **obtain** from someone? — flight permission
	Sentence A: Woman Killed In Queens Hit - And - Run , Driver **Charged**		
Ex 4	Sentence B: An allegedly intoxicated driver who tried to flee after striking and fatally injuring a woman in Queens has been **charged** in her murder, according to police.		
	Aligned QAs from A		**Aligned QAs from B**
	Who was **charged** with something? — Driver	–Align1–	Who was **charged** in something? — An allegedly intoxicated driver
	Why was someone **charged** with something? — Woman Killed	–Align2–	Why was someone **charged** in something? — tried to flee after striking and fatally injuring a woman in Queens

Table 7: A table showcasing our crowdsourced alignments.

		Train	Dev	Test
Num. Instances	**ECB+**	355	116	140
	DUC	534	182	312
	MN	485	102	153

Table 8: Distribution of our train/dev/test splits. Using Official splits of ECB+ and MN, while splitting our DUC in accordance with Thadani and McKeown (2013)'s fusion dataset split, in order not to compromise our training when testing on our extrinsic task.

Step 2 they carefully select QAs from the presented table and create the alignments.

D Quality Assessment of QAs Predicted by QASRL Parser

We make use of an updated version of the QASRL parser published by FitzGerald et al. (2018), which according to developer's comment in the repository, reaches 85 F1 on span detection and 47.6% question prediction accuracy. Nevertheless, accurately evaluating produced question is difficult, since various questions can validly describe the same semantic role, depending on the predicate's lexical semantics and on context. This results in a underestimation of question quality by the question prediction accuracy measure. A manual assessment of the QASRL parser based on an improved evaluation protocol (Roit et al., 2020) suggest that its role precision is relatively good (83.9), but its role recall is mediocre (64.3), as has been previously acknowledged (FitzGerald et al., 2018), Nevertheless, we opt for having a more sparse training set while showing it is possible for future endeavors to scale up training set annotation using an off-the-shelve QASRL parser.

Limited QA coverage might also result in lower QA alignment quality. In order to assess the quality of alignments based on parser produced QAs, we give one worker two sets of the same 30 instances, once with the parser-generated QAs and once with gold collected QAs. We find that the worker achieved **79** F1 agreement with herself, whereas the average number of alignments went from 2.3 with gold QAs to 1.9 with predicted QAs. To conclude, using the parser results in a modest alignment coverage decrease, while saving the costs of collecting gold QASRL annotations.

E QA-Alignment Model Training

For training, we apply negative sampling with a 1/10 positive/negative candidate alignment ratio, compared to an approximate 1/50 naturalistic ratio. Specifically, we sample negative alignments that include verbs or answers contained in the positive set, which we find leading to better performance, presumably because it leaves the "harder" cases for the model to distinguish from. As for training parameters, we use a learning rate of 3e-5 for both BERT and CorefRoBERTa, and 2e-5 for RoBERTa; for all we use the Adam optimizer with an epsilon of 1e-8, and max length tokens of 256, in addition

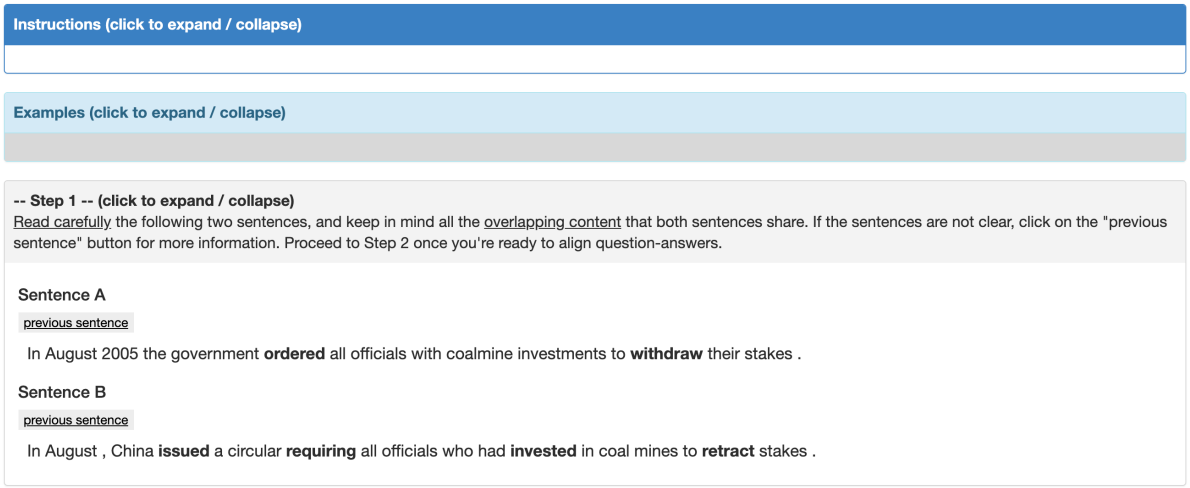
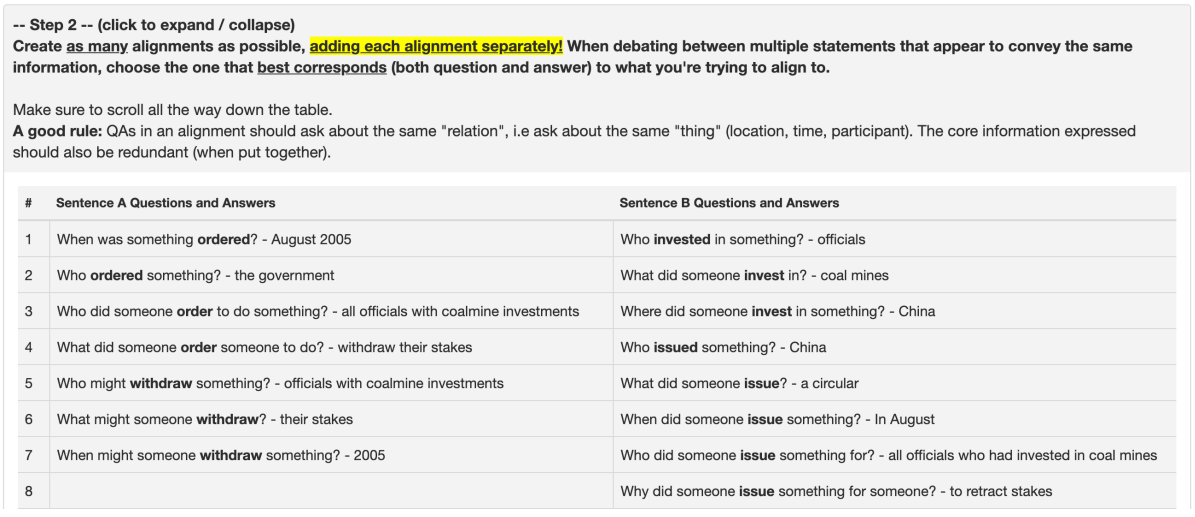


Figure 3: QA-Alignment Task Interface Step 1



#	Sentence A Questions and Answers	Sentence B Questions and Answers
1	When was something **ordered**? - August 2005	Who **invested** in something? - officials
2	Who **ordered** something? - the government	What did someone **invest** in? - coal mines
3	Who did someone **order** to do something? - all officials with coalmine investments	Where did someone **invest** in something? - China
4	What did someone **order** someone to do? - withdraw their stakes	Who **issued** something? - China
5	Who might **withdraw** something? - officials with coalmine investments	What did someone **issue**? - a circular
6	What might someone **withdraw**? - their stakes	When did someone **issue** something? - In August
7	When might someone **withdraw** something? - 2005	Who did someone **issue** something for? - all officials who had invested in coal mines
8		Why did someone **issue** something for someone? - to retract stakes

Figure 4: QA-Alignment Task Interface Step 2

to adding the special tokens used in our input to the model's vocabulary. We finetune BERT and CorefRoBERTa for 5 epochs, and 4 epochs for RoBERTa.

F Sentence Fusion Dataset Creation

For the extrinsic evaluation of our collected alignments on sentence fusion (§7) we reproduce the dataset constructed by Thadani and McKeown (2013). We note that while more recent datasets for the sentence fusion task exist (Geva et al.,2019; Lebanoff et al., 2020b), these works concern disparate sentences coming from a single document, which have minor content overlap (Lebanoff et al., 2019a) and thus diverge from the multi-doc setting.

As mentioned in §4. the fusion data is generated using SCU labels from the DUC datasets, where SCU labels function as the target fusion outputs, and the sentence contributors as input sentences. Although the original dataset is not available, they published detailed instructions that enabled us to recreate the data. We note also that the nature of this fusion dataset is not entirely comparable to the one we extracted from DUC explained in §4. The fusion dataset that was described in Thadani and McKeown (2013) contained post-processed sentence clusters (which we use as is only for the fusion experiment, we don't do any processing in our QA-Alignment data). This was because the authors in the paper applied a number of heuristics on the pairs created by the SCU labels in DUC, in order to remove noisy and longer sentences with small SCU contributors, discarding shorter and maintaining only highly overlapping sentences. This means that although the resulting dataset is highly relevant for cross-text information overlap, it also means that this fusion dataset created might not entirely reflect real document and summary sentences found "in the wild", however, it is the only sentence-level fusion dataset currently available.

Although Thadani and McKeown (2013) originally reported 1858 instances, they informed us they were using an unreleased version of DUC 2005. Our regenerated dataset thus consists of 1705 fusion instances, distributed into the author's original 70/30 train/test split, using DUC years 2005-2007 for test, TAC 2011 for dev, and TAC 2008-2011 for train.

G Fusion Output Examples

In Table 9 we display further fusion input examples that are fed to our baseline models, marked with special tokens as the predicted alignments. Fuse-Align's outputs in examples 1 and 3 display a behavior that is more abstractive than the baseline's (as discussed in §7), able to identify corresponding parts across source sentences, and "fuse" them into a single output. On the other hand, example 2 is an instance where our model also merges information, however, doing so on the incorrect entities (the 150,000 are not the survivors that live in tent cities, as the model predicts). Since our QA-Align model did not predict any alignments on this pair, this error could also result from not having any aligned propositions to attend to.

Fusion Input	[A2] She [\A2] called it " [P5] Change [\P5] [A5] Your Life TV [\A5] " ... </s> Recently [A4] she [\A4] has been [P4] presenting [\P4] a " [P5] change [\P5] [A5] your life [\A5] " theme [P3] focusing [\P3] on [A3] who you are [\A3] , and a truth - will - set - you - free attitude , influenced by her own traumatic childhood . [A1] Her show [\A1] is all about [P1] learning [\P1] who YOU are. </s> [A42] *The Oprah Winfrey Show* [\A42] [P3] features [\P3] [A3] [P4] talking [\P4] cures [\A3] and [P12] learning [\P12] [A1] who you are [\A1].
Baseline Output	The Oprah Winfrey Show features talking cures and learning who you are
Fuse-Align Output	The Oprah Winfrey Show is all about learning who YOU are
Fusion Input	A year later 150,000 were in prefab homes and only 26,000 still in tent cities. </s> A year later about 26,000 survivors were living in tent cities that had initially held some 120,000; and some 150,00 people were still in prefabricated houses.
Baseline Output	A year later 150,000 were in prefab homes
Fuse-Align Output	Some 150,000 survivors were living in tent cities
Fusion Input	In December 1994, [A1] Tony Blair [\A1], Labour leader, and probable prime minister after an upcoming election, signaled a constitutional battle with the Tories by [P312] setting [\P312] out [A2] [A3] proposals [\A3] for devolution for Scotland and Wales [\A2] </s> Labour has committed to the creation of a Welsh assembly, and [A1] party leader Tony Blair [\A1] [P12] set [\P12] out [A2] proposals for devolution [\A2].
Baseline Output	Labour has committed to the creation of a Welsh assembly
Fuse-Align output	In December 1994, Tony Blair set out proposals for devolution

Table 9: Example outputs for both fusion models. The input represents what our Fuse-Align models accepts, while the baseline takes the same input just without the special tokens.