A Models Details

S-Reader The model architecture of S-Reader is divided into the encoder module and the decoder module. The encoder module is identical to that of our sentence selector. It first takes the document and the question as inputs, obtains document embeddings $D \in \mathbb{R}^{L_d \times h_d}$, question embeddings $Q \in \mathbb{R}^{L_q \times h_d}$ and question-aware document embeddings $D^q \in \mathbb{R}^{L_d \times h_d}$, where D^q is defined as Equation 1, and finally obtains document encodings D^{enc} and question encodings Q^{enc} as Equation 3. The decoder module obtains the scores for start and end position of the answer span by calculating bilinear similarities between document encodings and question encodings as follows.

$$\beta = \text{softmax}(w_1^T Q^{enc}) \in \mathbb{R}^{L_q} \quad (10)$$

$$\tilde{q^{enc}} = \sum_{j=1}^{L_q} (\beta_j Q_j^{enc}) \in \mathbb{R}^h \quad (11)$$

$$score^{start} = D^{enc} W_{start} \tilde{q}^{enc} \in \mathbb{R}^{L_d} \quad (12)$$

$$score^{end} = D^{enc} W_{end} \tilde{q}^{enc} \in \mathbb{R} L_d \quad (13)$$

Here, $w_1 \in \mathbb{R}^h, W_{start}, W_{end} \in \mathbb{R}^{h \times h}$ are trainable weight matrices.

The overall architecture is similar to Document Reader in DrQA (Chen et al., 2017), except they are different in obtaining embeddings and use different hyperparameters. As shown in Table 5, our S-Reader obtains F1 score of 79.9 on SQuAD development data, while Document Reader in DrQA achieves 78.8.

Training details We implement all of our models using PyTorch. First, the corpus is tokenized using Stanford CoreNLP toolkit (Manning et al., 2014). We obtain the embeddings of the document and the question by concatenating 300-dimensional Glove embeddings pretrained on the 840B Common Crawl corpus (Pennington et al., 2014), 100-dimensional character n-gram embeddings by Hashimoto et al. (2017), and 300-dimensional contextualized embeddings pretrained on WMT (McCann et al., 2017). We do not use handcraft word features such as POS and NER tagging, which is different from Document Reader in DrQA. Hence, the dimension of the embedding (d_h) is 600. We use the hidden size (h) of 200. We apply dropout with 0.2 drop rate (Srivastava et al., 2014) to encodings and LSTMs for regularization. We train the models using ADAM optimizer (Kingma and Ba, 2014) with default hyperparameters. When we train and evaluate the model on the dataset, the document is truncated to the maximum length of $min(2000, max(1000, L_{th}))$ words, where L_{th} is the length which covers 90% of documents in the whole examples.

Selection details Here, we describe how to dynamically select sentences using `Dyn` method. Given the sentences $S_{all} = \{s_1, s_2, s_3, ..., s_n\}$, ordered by scores from the sentence selector in descending order, the selected sentences $S_{selected}$ is as follows.

$$S_{candidate} = \{s_i \in S_{all} | score(s_i) >= 1 - th\} \quad (14)$$

$$S_{selected} = \begin{cases} S_{candidate} & if\, S_{candidate} = \emptyset \\ \{s_1\} & o.w. \end{cases} \quad (15)$$

Here, $score(s_i)$ is the score of sentence s_i from the sentence selector, and th is a hyperparameter between 0 and 1.

The number of sentences to select can be dynamically controlled during inference by adjusting th, so that proper number of sentences can be selected depending on the needs of accuracy and speed. Figure 4 shows the trade-off between the number of sentences and accuracy, as well as the number of selected sentences depending on the threshold th.

B More Analyses

Human studies on TriviaQA We randomly sample 50 examples from the TriviaQA (Wikipedia) development (verified) set, and analyze the minimum number of sentences to answer the question. Despite TriviaQA having longer documents (488 sentences per question), most examples are answerable with one or two sentences, as shown in Table 11. While 88% of examples are answerable given the full document, 95% of them can be answered with one or two sentences.

Error analyses We compare the error cases (in exact match (EM)) of FULL and MINIMAL. The left-most Venn diagramin Figure 5 shows that MINIMAL is able to answer correctly to more than 97% of the questions answered correctly by FULL. The other two diagrams in Figure 5 shows the error cases of each model, broken down by the sentence where the model's prediction is from.

Table 12 shows error cases on SQuAD, which MINIMAL fails to answer correctly. In the first

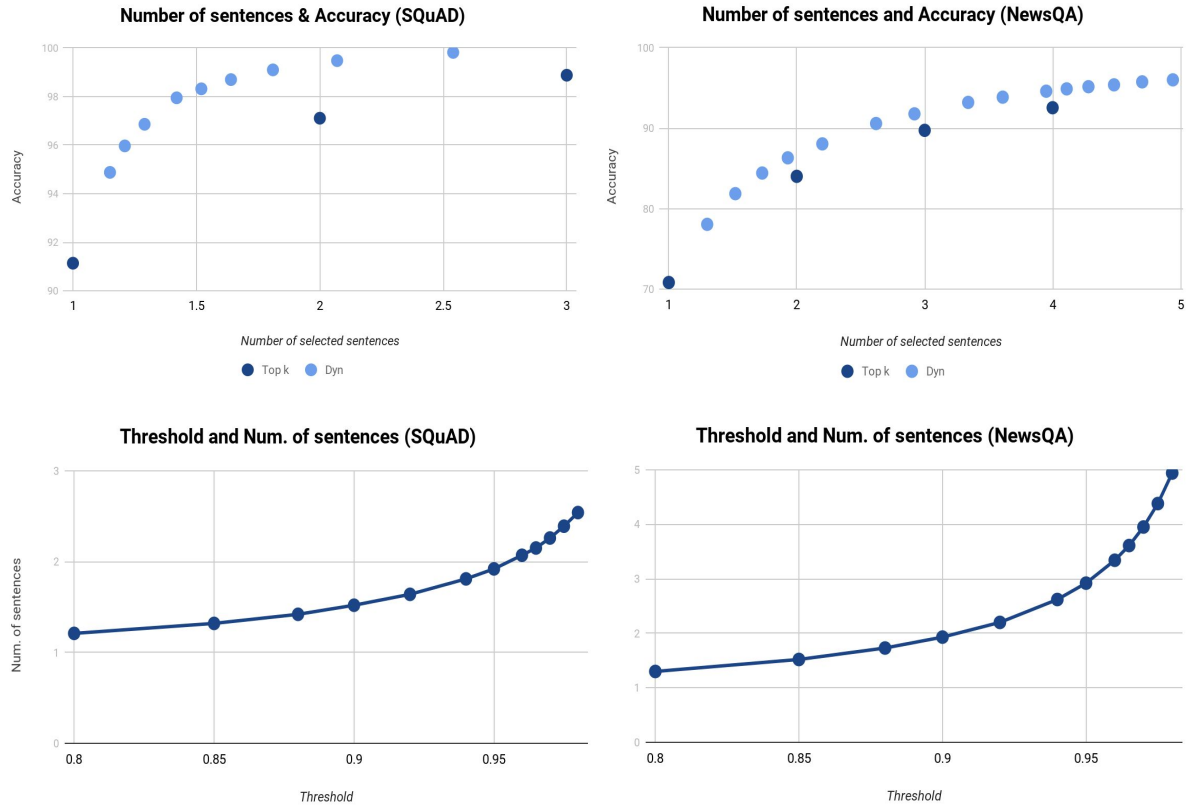

Figure 4: (Top) The trade-off between the number of selected sentence and accuracy on SQuAD and NewsQA. `Dyn` outperforms `Top k` in accuracy with similar number of sentences. (Bottom) Number of selected sentences depending on threshold.

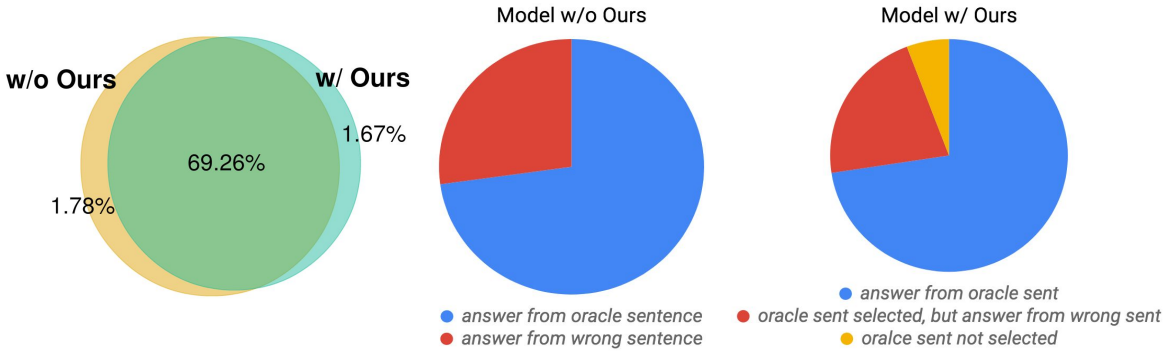

Figure 5: (Left) Venn diagramof the questions answered correctly by FULL and with MINIMAL. (Middle and Right) Error cases from FULL (Middle) and MINIMAL (Right), broken down by which sentence the model's prediction comes from.

N sent	%	Paragraph	Question
1	56	Chicago O'Hare International Airport, also known as O'Hare Airport, Chicago International Airport, Chicago O'Hare or simply O'Hare, is an international airport located on the far northwest side of Chicago, Illinois.	In which city would you find O'Hare International Airport?
		In 1994, Wet Wet Wet had their biggest hit, a cover version of the troggs' single "Love is All Around", which was used on the soundtrack to the film Four Weddings and A Funeral.	The song "Love is All Around" by Wet Wet Wet featured on the soundtrack for which 1994 film?
2	28	Cry Freedom is a 1987 British epic drama film directed by Richard Attenborough, set in late-1970s apartheid era South Africa. (...) The film centres on the real-life events involving black activist Steve Biko and (...)	The 1987 film 'Cry Freedom' is a biographical drama about which South Aftrican civil rights leader?
		Helen Adams Keller was an American author, political activist, and lecturer. () The story of how Kellers teacher, Anne Sullivan, broke through the isolation imposed by a near complete lack of language, allowing the girl to blossom as she learned to communicate, has become widely known through (...)	Which teacher taught Helen Keller to communicate?
3 ↑	4	(...) The equation shows that, as volume increases, the pressure of the gas decreases in proportion. Similarly, as volume decreases, the pressure of the gas increases. The law was named after chemist and physicist Robert Boyle, who published the original law. (...)	Who gave his name to the scientific law that states that the pressure of a gas is inversely proportional to its volume at constant temperature?
		The Buffalo six (known primarily as Lackawanna Six) is a group of six Yemeni-American friends who were convicted of providing material support to Al Qaeda in December 2003, () In the late summer of 2002, one of the members, Mukhtar Al-Bakri, sent () Yahya Goba and Mukhtar Al-Bakri received 10-year prison sentences. Yaseinn Taher and Shafal Mosed received 8-year prison sentences. Sahim Alwan received a 9.5-year sentence. Faisal Galab received a 7-year sentence.	Mukhtar Al-Bakri, Sahim Alsan, Faysal Galan, Shafal Mosed, Yaseinn Taher and Yahya Goba were collectively known as the Lackawanna Six and by what other name?
N/A	12	(...) A commuter rail operation, the New Mexico Rail Runner Express, connects the state's capital, its and largest city, and other communities. (...)	Which US state is nicknamed both 'the Colourful State' and 'the Land of Enchantment?'
		Smith also arranged for the publication of a series of etchings of Capricci in his vedette ideal, but the returns were not high enough, and in 1746 Canaletto moved to London, to be closer to his market.	Canaletto is famous for his landscapes of Venice and which other city?

Table 11: Human analysis of the context required to answer questions on TriviaQA (Wikipedia). 50 examples are sampled randomly. 'N sent' indicates the number of sentences required to answer the question, and 'N/A' indicates the question is not answerable even given all sentences in the document. The groundtruth answer text is in red text. Note that the span is not given as the groundtruth. In the first example classified into 'N/A', the question is not answerable even given whole documents, because there is no word 'corlourful' or 'enchantment' in the given documents. In the next example, the question is also not answerable even given whole documents, because all sentences containing 'London' does not contain any information about Canaletto's landscapes.

In On the Abrogation of the Private Mass, he condemned as idolatry the idea that the mass is a sacrifice, asserting instead that it is a gift, to be received with thanksgiving by the whole congregation. ✓ *What did Luther call the mass instead of sacrifice?*
Veteran receiver Demaryius Thomas led the team with 105 receptions for 1,304 yards and six touchdowns, while Emmanuel Sanders caught (...) ✓ Running back Ronnie Hillman also made a big impact with 720 yards, five touchdowns, 24 receptions, and a 4.7 yards per carry average. ✓ *Who had the most receptions out of all players for the year?*
In 1211, after the conquest of Western Xia, Genghis Kahn planned again to conquer the Jin dynasty. ✓ Instead, the Jin commander sent a messenger, Ming-Tan, to the Mongol side, who defected and told the Mongols that the Jin army was waiting on the other side of the pass. The Jin dynasty collapsed in 1234, after the siege of Caizhou. ✓ *Who was the Jin dynasty defector who betrayed the location of the Jin army?*

Table 12: Examples on SQuAD, which MINIMAL predicts the wrong answer. Grountruth span is in underlined text, the prediction from MINIMAL is in red text. Sentences selected by our selector is denoted with ✓. In the first example, the model predicts the wrong answer from the oracle sentence. In the second example, the model predicts the answer from the wrong sentence, although it selects the oracle sentence. In the last example, the model fails to select the oracle sentence.

TriviaQA		Inference			Dev-verified		Dev-full	
		n sent	Acc	Sp	F1	EM	F1	EM
FULL		69	95.9	x1.0	66.1	61.6	59.6	53.5
MINIMAL	TF-IDF	5	73.0	x13.8	60.4	54.1	51.9	45.8
		10	79.9	x6.9	64.8	59.8	57.2	51.5
		20	85.5	x3.5	67.3	62.9	60.4	54.8
	Our Selector	5.0	84.9	**x13.8**	65.0	61.0	59.5	54.0
		10.5	90.9	x6.6	**67.0**	**63.8**	**60.5**	**54.9**
		20.4	95.3	x3.4	**67.7**	**63.8**	**61.3**	**55.6**
MEMEN		-	-	-	55.8	49.3	46.9	43.2
Mnemonic Reader		-	-	-	59.5[a]	54.5[a]	52.9[a]	46.9[a]
Reading Twice		-	-	-	59.9[a]	53.4[a]	55.1[a]	48.6[a]
Neural Cascades		-	-	-	62.5[a]	58.9[a]	56.0[a]	51.6[a]

Table 13: Results on the dev-verified set and the dev-full set of TriviaQA (Wikipedia). We compare the results from the sentences selected by TF-IDF and our selector (Dyn). We also compare with MEMEN (Pan et al., 2017), Mnemonic Reader (Hu et al., 2017), Reading Twice for Natural Language Understanding (Weissenborn, 2017) and Neural Casecades (Swayamdipta et al., 2018), the published state-of-the-art.

[a]Numbers on the test set.

SQuAD-Open		Inference			Dev	
		n sent	Acc	Sp	F1	EM
FULL		124	76.9	x1.0	41.0	33.1
MINIMAL	TF-IDF	5	46.1	x12.4	36.6	29.6
		10	54.3	x6.2	39.8	32.5
		20	62.4	x3.1	41.7	34.1
		40	65.8	x1.6	42.5	34.6
	Our Selector	5.3	58.9	**x11.7**	**42.3**	**34.6**
		10.7	64.0	x5.8	**42.5**	**34.7**
		20.4	68.1	x3.0	**42.6**	**34.7**
		40.0	71.4	x1.5	42.6	34.7
R^3		-	-	-	37.5	29.1
DrQA		2376[a]	77.8	-	-	28.4
DrQA (Multitask)		2376[a]	77.8	-	-	29.8

Table 14: Results on the dev set of SQuAD-Open. We compare with the results from the sentences selected by TF-IDF method and our selector (Dyn). We also compare with R^3 (Wang et al., 2018) and DrQA (Chen et al., 2017).

[a]Approximated based on there are 475.2 sentences per document, and they use 5 documents per question

two examples, our sentence selector choose the oracle sentence, but the QA model fails to answer correctly, either predicting the wrong answer from the oracle sentence, or predicting the answer from the wrong sentence. In the last example, our sentence selector fails to choose the oracle sentence. We conjecture that the selector rather chooses the sentences containing the word 'the Jin dynasty', which leads to the failure in selection.

C Full Results on TriviaQA and SQuAD-Open

Table 13 and Table 14 show full results on TriviaQA (Wikipedia) and SQuAD-Open, respectively.

MINIMAL obtains higher F1 and EM over FULL, with the inference speedup of up to 13.8×. In addition, outperforms the published state-of-the-art on both TriviaQA (Wikipedia) and SQuAD-Open, by 5.2 F1 and 4.9 EM, respectively.

D Samples on SQuAD, TriviaQA and SQuAD-Adversarial

Table 15 shows the full index of samples used for human studies and analyses.

Analysis	Table	Dataset	Ids
Context Analysis	1	SQuAD	56f7eba8a6d7ea1400e172cf, 56e0bab7231d4119001ac35c, 56dfa2c54a1a83140091ebf6, 56e11d8ecd28a01900c675f4, 572ff7ab04bcaa1900d76f53, 57274118dd62a815002e9a1d, 5728742cff5b5019007da247, 572748745951b619008f87b2, 5730626624611fd1900a9cdf7, 56e1efa0e3433e140042321a, 57115f0a50c2381900b54aa9, 57286f373acd2414000df9db, 57300f8504bcaa1900d770d3, 5728192ff5b5019007da1e0, 571cd11add7acb1400e4c16f, 57094ca7efce8f15003a7dd7, 57300761947a6a140053cf9c, 571144d1a58dae1900cd6d6f, 572813b52ca10214002d9d68, 572969f51d046914007793e0, 56e0d6cf231d4119001ac423, 572754cd5951b619008f8867, 570d4a6bfed7b91900d45e13, 57284b904b864d19001648e5, 5726cc11dd62a815002e9086, 572966ebaf94a219006aa392, 5726c3da708984140094d0d9, 57277bfc708984140094dedd, 572747dd5951b619008f87aa, 57107c24a58dae1900cd69ea, 571cdcb85efbb31900334e0d, 56e10e73cd28a01900c674ec, 5726c0c5dd62a815002e8f79, 5725f39638643c19005acefb, 5726bcde708984140094cfc2, 56e74bf937bdd419002c3e36, 56d997cddc89441400fdb586, 5728349dff5b5019007d9f01, 573011de04bcaa1900d770fc, 57274f49f1498d1400e8f620, 57376df3c3c5551400e51ed7, 5726bd655951b619008f7ca3, 5733266d4776f41900660714, 5725bc0338643c19005acc12, 572ff760b2c2fd1400568679, 572fbfa504bcaa1900d76c73, 5726938af1498d1400e8e448, 5728ef8d2ca10214002daac3, 5728f3724b864d1900165119, 56f85bb8aef2371900626011
Oracle Error Analysis	2	SQuAD	57376df3c3c5551400e51eda, 5726a00cf1498d1400e8e551, 5725f00938643c19005aceda, 573361404776f4190066093c, 571bb2269499d21900609cac, 571cebc05efbb31900334e4c, 56d7096b0d65d214001982fd, 5732b6b5328d981900602025, 56beb6533aeaaa14008c928e, 5729e1101d04691400779641, 56d601e41c8504140094 6ecf, 57115b8b50c2381900b54a8b, 56e74d1f00c9c71400d76f70, 5728245b2ca10214002d9ed6, 5725c2a038643c19005acc6f, 57376828c3c5551400e51eba, 573403947 76f419006616df, 5728d7c54b864d1900164f50, 57265aaf5951b619008f706e, 5728151b4b864d1900164429, 57060cc352bb89140068980e, 5726e08e5951b619008f8110, 57266cc9f1498d1400e8df52, 57273455f1498d1400e8f48e, 572972f46aef051400154ef3, 5727482bf1498d1400e8f5a6, 57293f8a6aef051400154bde, 5726f8abf1498d1400e8f166, 5737a9afc3c5551400e51f63, 570614ff52bb89140068988b, 56bebd713aeaaa14008c9331, 57060a1175f01819005e78d3, 5737a9afc3c5551400e51f62, 57284618ff5b5019007da0a9, 570960cf200fba1400367f03, 572822233acd2414000df556, 5727b0892ca10214002d93ea, 57268525dd62a815002e8809, 57274b35f1498d1400e8f5d6, 56d98c53dc89441400fdb545, 5727ec062ca10214002d99b8, 57274e975951b619008f87fa, 572686fc708984140094c8e8, 572929d56aef051400154b0c, 570d30fdfed7b91900d45ce3, 5726b1d95951b619008f7ad0, 56de41504396321400ee2714, 5726472bdd62a815002e8046, 5727d3843acd2414000ded6b, 5726e9c65951b619008f8247
`Top k` vs `Dyn`	7	SQuAD	56e7504437bdd419002c3e5b
FULL vs MINIMAL	10	SQuAD-Adversarial	56bf53e73aeaaa14008c95cc-high-conf-turk0, 56dfac8e231d4119001abc5b-high-conf-turk0
Context Analysis	11	TriviaQA	qb 4446, wh 1933, qw 3445, qw 169, qz 2430, sfq 25261, qb 8010, qb 2880, qb 370, sfq 8018, sfq 4789, qz 1032, qz 603, sfq 7091, odql 10315, dpql 3949, odql 921, qb 6073, sfq 13685, bt 4547 sfq 23524, qw 446, jp 3302, jp 2305, tb 1951, qw 10268, bt 189, qw 14470, jp 3059, qw 12135, qb 7921, sfq 2723, odql 2243, qw 7457, dpql 4590, sfq 3509, bt 2065, qf 2092, qb 10019, sfq 14351, bb 4422, jp 3321, sfq 12682, sfq 13224, sfq 4027, qw 12518, qz 2135, qw 1983, sfq 26249, sfq 19992
Error Analysis	12	SQuAD	56f84485aef2371900625f74, 56bf38383aeaaa14008c956e, 5726bb64591b619008f7c3c

Table 15: QuestionIDs of samples used for human studies and analyses.