EMERGENT PREDICATION STRUCTURE IN VECTOR REPRESENTATIONS OF NEURAL READERS

Hai Wang* **Takeshi Onishi*** **Kevin Gimpel** **David McAllester**
Toyota Technological Institute at Chicago
6045 S. Kenwood Ave. Chicago, Illinois 60637. USA
{haiwang,tonishi,kgimpel,mcallester}@ttic.edu

ABSTRACT

Reading comprehension is a question answering task where the answer is to be found in a given passage about entities and events not mentioned in general knowledge sources. A significant number of neural architectures for this task (neural readers) have recently been developed and evaluated on large cloze-style datasets. We present experiments supporting the emergence of "predication structure" in the hidden state vectors of a class of neural readers including the Attentive Reader and Stanford Reader. We posits that the hidden state vectors can be viewed as (a representation of) a concatenation $[P, c]$ of a "predicate vector" P and a "constant symbol vector" c and that the hidden state represents the atomic formula $P(c)$. This predication structure plays a conceptual role in relating "aggregation readers" such as the Attentive Reader and the Stanford Reader to "explicit reference readers" such as the Attention-Sum Reader, the Gated-Attention Reader and the Attention-over-Attention Reader. In an independent contribution, we show that the addition of linguistics features to the input to existing neural readers significantly boosts performance yielding the best results to date on the Who-did-What dataset.[1]

1 INTRODUCTION AND OVERVIEW

Reading comprehension is a type of question answering task where the answer is to be found in a passage about particular entities and events not otherwise familiar to the reader. In particular, the entities and events should not be mentioned in structured databases of general knowledge. Reading comprehension problems are intended to measure a systems ability to extract semantic information about entities and relations directly from unstructured text. Several large scale reading comprehension datasets have been introduced recently. In particular the CNN & DailyMail datasets (Hermann et al., 2015), the Children's Book Test (CBT) (Hill et al., 2016), and the Who-did-What dataset (Onishi et al., 2016). The large sizes of these datasets enable the application of deep learning. These are all cloze-style datasets where a question is constructed by deleting a word or phrase from an article summary (in CNN/DailyMail), from a sentence in a Children's story (in CBT), or by deleting a person from the first sentence of a different news article on the same entities and events (in Who-did-What).

In this paper we present empirical evidence for the emergence of predication structure in a certain class of neural readers. To understand predication structure is it helful to review the anonymization performed in the CNN/DailyMail dataset. In this dataset named entities are replaced by anonymous entity identifiers such as "entity37". The passage might contain "entity52 gave entity24 a rousing applause" and the question might be "X received a rounding applause from entity52". The task is to fill in X from a given multiple choice list of candidate entity identifiers. A fixed relatively small set of the same entity identifiers are used over all the problems and the same problem is presented many times with the entity identifiers shuffled. This prevents a given entity identifier from having any semantically meaningful vector embedding. The embeddings of the entity identifiers are

*Authors contributed equally

[1] code will be available: https://github.com/sohuren

presumably just pointers to semantics-free tokens. We will write entity identifiers as logical constant symbols such as c rather than strings such as "entity37".

Aggregation readers, including Memory Networks (Weston et al.; Sukhbaatar et al., 2015), the Attentive Reader (Hermann et al., 2015) and the Stanford Reader (Chen et al., 2016), use bidirectional LSTMs or GRUs to construct a contextual embedding h_t of each position t in the passage and also an embedding q of the question. They then select and answer c using a criterion similar to

$$\operatorname*{argmax}_c \sum_t \; < h_t, q > \; < h_t, e(c) > \tag{1}$$

where $e(c)$ is the vector embedding of the constant symbol (entity identifier) c. In practice the inner-product $< h_t, q >$ is normalized over t using a softmax to yield an attention α_t over t and (1) becomes.

$$\operatorname*{argmax}_c \; < e(c), \sum_t \alpha_t h_t > . \tag{2}$$

Here $\sum_t \; \alpha_t h_t$ is viewed as a vector representation of the passage.

We argue that for aggregation readers, roughly defined by (2), the hidden state h_t of the passage at position (or word) t can be viewed as a vector concatenation $h_t = [e(\Phi_t), e'(c_t)]$ where Φ_t is a property (or statement or predicate) being stated of a particular constant symbol c_t. A logician might write this as $h_t = \Phi_t[c_t]$. Furthermore, the question can be interpreted as having the form $\Psi[x]$ where the problem is to find a constant symbol c such that the passage implies $\Psi[c]$. Assuming $h_t = [e(\Phi_t), e'(c_t)]$ and $q = [e(\Psi), 0]$ and $e(c) = [0, e'(c)]$ we can rewrite (1) as

$$\operatorname*{argmax}_c \sum_t \; < e(\Phi_t), e(\Psi) > \; < e'(c_t), e'(c) > . \tag{3}$$

The first inner product in (3) is interpreted as measuring the extent to which $\Phi_t[x]$ implies $\Psi[x]$ for any x. The second inner product is interpreted as restricting t to positions talking about the constant symbol c.

Note that the posited decomposition of h_t is not explicit in (2) but instead must emerge during training. We present empirical evidence that this structure does emerge. The empirical evidence is somewhat tricky as the direct sum structure that divides h_t into its two parts need not be axis aligned and therefore need not literally correspond to vector concatenation.

We also consider a second class of neural readers that we call explicit reference readers. Explicit reference readers avoid (2) and instead use

$$\operatorname*{argmax}_c \sum_{t \in R(c)} \alpha_t \tag{4}$$

where $R(c)$ is the subset of the positions where the constant symbol (entity identifier) c occurs. Note that if we identify α_t with $< e(\Phi_t), e(\Psi) >$ and assume that $< e'(c), e'(c_t) >$ is either 0 or 1 depending on whether $c = c_t$, then (3) and (4) agree. In explicit reference readers the hidden state h_t need not carry a pointer to c_t as the restriction on t is independent of learned representations. Explicit reference readers include the Attention Sum Reader (Kadlec et al., 2016), the Gated Attention Reader (Dhingra et al., 2016), the Attention-over-Attention Reader (Cui et al., 2016) and others (a list can be found in section 6).

So far we have only considered anonymized datasets that require the handling of semantics-free constant symbols. However, even for non-anonymized datasets such as Who-Did-What, it is helpful to add features which indicate which positions in the passage are referring to which candidate answers. This indicates, not surprisingly, that reference is important in question answering. The fact that explicit reference features are needed in aggregation readers on non-anonymized data indicates that reference is not being solved by the aggregation readers. However, as reference seems to be important for cloze-style question answering, these problems may ultimately provide training data from which reference resolution can be learned.

Sections 2 and 3 review various existing datasets and models respectively. Section 4 presents the logical structure interpretation of aggregation readers in more detail and the empirical evidence supporting it. Section 5 proposes new models that enforce the direct sum structure of the hidden

state vectors. It is shown that these new models perform well on the Who-did-What dataset provided that reference annotations are added as input features. Section 5 also describes additional linguistic features that can be added to the input embeddings and show that these improve the performance of existing models resulting in the best single-model performance to date on the Who-did-What dataset.

2 A BRIEF SURVEY OF DATASETS

Before presenting various models for machine comprehension we give a general formulation of the machine comprehension task. We take an instance of the task be a four tuple $(q, p, a, \mathcal{A})$, where q is a question given as sequence of words containing a special taken for a "blank" to be filled in, p is a document consisting of a sequence of words, $\mathcal{A}$ is a set of possible answers and $a \in \mathcal{A}$ is the ground truth answer. All words are drawn from a vocabulary $\mathcal{V}$. We assume that all possible answers are words from the vocabulary, that is $\mathcal{A} \subseteq \mathcal{V}$, and that the ground truth answer appears in the document, that is $a \in p$. The problem can be described as that of selecting the answer $a \in \mathcal{A}$ that answers question q based on information from p.

We will now briefly summarize important features of the related datasets in reading comprehension.

CNN & DailyMail: Hermann et al. (2015) constructed these datasets from a large number of news articles from the CNN and Daily Mail news websites. The main article is used as the context, while the cloze style question is formed from one short highlight sentence appearing in conjunction with the published article. To avoid the model using external world knowledge when answering the question, the named entities in the entire dataset were replaced by anonymous entity IDs which were then further shuffled for each example. This forces models to rely on the context document to answer each question. In this anonymized corpus the entity identifiers are taken to be a part of the vocabulary and the answer set $\mathcal{A}$ consists of the entity identifiers occurring in the passage.

Who-did-What (WDW): The Who-did-What dataset (Onishi et al., 2016) contains 127,000 multiple choice cloze questions constructed from the LDC English Gigaword newswire corpus (David & Cieri, 2003). In contrast with CNN and Daily Mail, it avoids using article summaries for question formation. Instead, each problem is formed from two independent articles: one is given as the passage to be read and a different article on the same entities and events is used to form the question. Further, Who-did-What avoids anonymization, as each choice is a person named entity. In this dataset the answer set $\mathcal{A}$ consists of the person named entities occurring in the passage. Finally, the problems have been filtered to remove a fraction that are easily solved by simple baselines. It has two training sets. The larger training set ("relaxed") is created using less baseline filtering, while the smaller training set ("strict") uses the same filtering as the validation and test sets.

Children's Book Test (CBT) Hill et al. (2016) developed the CBT dataset in a slightly different fashion to the CNN/DailyMail datasets. They take any sequence of 21 consecutive sentences from a children's book: the first 20 sentences are used as the passage, and the goal is to infer a missing word in the 21st sentence. The task complexity varies with the type of the omitted word (verb, preposition, named entity, or common noun). According to the original study on this dataset (Hill et al., 2016), n-gram and recurrent neural network language models are sufficient for predicting verbs or prepositions. However, for named entities and common nouns, current solvers are still far from human performance.

Other Related Datasets. It is also worth mentioning several related datasets. The MCTest dataset (Richardson et al., 2013) consists of children's stories and questions written by crowdsourced workers. The dataset only contains 660 documents and is too small to train deep models. The bAbI dataset (Weston et al., 2016) is constructed automatically using synthetic text generation and can be perfectly answered by hand-written algorithms (Lee et al., 2016). The SQuAD dataset (Rajpurkar et al., 2016) consists passage-question pairs where the passage is a wikipedia article and the questions are written by crowdsourced workers. Although crowdsourcing is involved, the dataset contains over 200,000 problems. But the answer is often a word sequence which is dificult to handle with the reader models considered here. The LAMBADA dataset (Denis et al., 2016) is a word prediction dataset which requires a broad discourse context and the correct answer might not in the context. Nonetheless, when the correct answer is in the context, neural readers can be applied effectively (Chu et al., 2016).

3 AGGREGATION READERS AND EXPLICIT REFERENCE READERS

Here we classify readers into aggregation readers and explicit reference readers. Aggregation readers appeared first in the literature and include Memory Networks (Weston et al.; Sukhbaatar et al., 2015), the Attentive Reader (Hermann et al., 2015), and the Stanford Reader (Chen et al., 2016). Aggregation readers are defined by equations (8) and (10) below. Explicit reference readers incluce the Attention-Sum Reader (Kadlec et al., 2016), the Gated-Attention Reader (Dhingra et al., 2016), and the Attention-over-Attention Reader (Cui et al., 2016). Explicit reference readers are defined by equation (14) below. We first present the Stanford Reader as a paradigmatic aggregation Reader and the Attention-Sum Reader as a paradigmatic explicit reference reader.

3.1 AGGREGATION READERS

Stanford Reader. The the Stanford Reader (Chen et al., 2016) computes a bi-directional LSTM representation of both the passage and the question.

$$h = \text{biLSTM}(e(p)) \tag{5}$$

$$q = [\text{fLSTM}(e(q))_{|q|}, \text{bLSTM}(e(q))_1] \tag{6}$$

In equations (5) and (6) we have that $e(p)$ is the sequence of word embeddings $e(w_i)$ for $w_i \in p$ and similarly for $e(q)$. The expression biLSTM(s) denotes the sequence of hidden state vectors resulting from running a bi-directional LSTM on the vector sequence s. We write biLSTM$(s)_i$ for the ith vector in this sequence. Similarly fLSTM(s) and bLSTM(s) denote the sequence of vectors resulting from running a forward LSTM and a backward LSTM respectively and $[\cdot, \cdot]$ denotes vector concatenation. The Stanford Reader, and various other readers, then compute a bilinear attention over the passage which is then used to construct a single weighted vector representation of the passage.

$$\alpha_t = \underset{t}{\text{softmax}}\ h_t^\top W_\alpha\ q \tag{7}$$

$$o = \sum_t \alpha_t h_t \tag{8}$$

Finally, they compute a probability distribution over the answers $P(a|p, q, \mathcal{A})$.

$$p(a|d, q, \mathcal{A}) = \underset{a \in \mathcal{A}}{\text{softmax}}\ e_o(a)^\top o \tag{9}$$

$$\hat{a} = \underset{a \in \mathcal{A}}{\text{argmax}}\ e_o(a)^\top o \tag{10}$$

Here $e_o(a)$ is an "output embedding" of the answer a. On the CNN dataset the Stanford Reader trains an output embedding for each the roughly 500 entity identifiers used in the dataset. In cases where the answer might be any word in $\mathcal{V}$ an output embedding must be trained for the entire vocabulary.

The reader is trained with log-loss $\ln 1/P(a|p, q, \mathcal{A})$ where a is the correct answer. At test time the reader is scored on the percentage of problems where $\hat{a} = a$.

Memory Networks. Memory Networks (Weston et al.; Sukhbaatar et al., 2015) use (8) and (10) but have more elaborate methods of constructing "memory vectors" h_t not involve LSTMs. Memory networks use (8) and (10) but replace (9) with

$$P(w|p, q, \mathcal{A}) = P(w|p, q) = \underset{w \in \mathcal{V}}{\text{softmax}}\, e_o(w)^T o. \tag{11}$$

It should be noted that (11) trains output vectors over the whole vocabulary rather than just those items occurring in the choice set $\mathcal{A}$. This is empirically significant in non-anonymized datasets such as CBT and Who-did-What where choices at test time may never have occurred as choices in the training data.

Attentive Reader. The Stanford Reader was derived from the Attentive Reader (Hermann et al., 2015). The Attentive Reader uses $\alpha_t = \text{softmax}_t\, \text{MLP}([h_t, q])$ instead of (7). Here MLP(x) is the output of a multi layer perceptron (MLP) given input x. Also, the answer distribution in the attentive reader is defined over the full vocabulary rather than just the candidate answer set $\mathcal{A}$.

$$P(w|p, q, \mathcal{A}) = P(w|p, q) = \underset{w \in \mathcal{V}}{\text{softmax}}\, e_o(w)^T \text{MLP}([o, q]) \tag{12}$$

Equation (12) is similar to (11) in that it leads to the training of output vectors for the full vocabulary rather than just those items appearing in choice sets in the training data. As in memory networks, this leads to improved performance on non-anonymized data sets.

3.2 EXPLICIT REFERENCE READERS

Attention-Sum Reader. In the Attention-Sum Reader (Kadlec et al., 2016) h and q are computed with equations (5) and (6) as in the Stanford Reader but using GRUs rather than LSTMs. The attention α_t is computed similarly to (7) but using a simple inner product $\alpha_t = \mathrm{softmax}_t \ h_t^\top q$ rather than a trained bilinear form. Most significanlty, however, equations (9) and (10) are replaced by the following where $t \in R(a,p)$ indicates that a reference to candidate answer a occurs at position t in p.

$$P(a|p,q,\mathcal{A}) = \sum_{t \in R(a,p)} \alpha_t \tag{13}$$

$$\hat{a} = \operatorname*{argmax}_a \sum_{t \in R(a,p)} \alpha_t \tag{14}$$

Here we think of $R(a,p)$ as the set of references to a in the passage p. It is important to note that (13) is an equality and that $P(a|p,q,\mathcal{A})$ is not normalized to the members of $R(a,p)$. When training with the log-loss objective this drives the attention α_t to be normalized — to have support only on the positions t with $t \in R(a,p)$ for some a. See the heat maps in the appendix.

Gated-Attention Reader. The Gated Attention Reader Dhingra et al. (2016) involves a K-layer biGRU architecture defined by the following equations.

$$\begin{aligned} q^\ell &= [\text{fGRU}(e(q))_{|q|}, \text{bGRU}(e(q))_1] \ \ 1 \le \ell \le K \\ h^1 &= \text{biGRU}(e(p)) \\ h^\ell &= \text{biGRU}(h^{\ell-1} \odot q^{\ell-1}) \ \ 2 \le \ell \le K \end{aligned}$$

Here the question embeddings q^ℓ for different values of ℓ are computed with different GRU model parameters. Here $h \odot q$ abbreviates the sequence $h_1 \odot q, h_2 \odot q, \ldots h_{|p|} \odot q$. Note that for $K = 1$ we have only q^1 and h^1 as in the attention-sum reader. An attention is then computed over the final layer h^K with $\alpha_t = \mathrm{softmax}_t \ (h_t^K)^\top \ q^K$ in the attention-sum reader. This reader uses (13) and (14).

Attention-over-Attention Reader, The Attention-over-Attention Reader (Cui et al., 2016) uses a more elaborate method to compute the attention α_t. We will use t to range over positions in the passage and j to range over positions in the question. The model is then defined by the following equations.

$$h = \text{biGRU}(e(p)) \qquad q = \text{biGRU}(e(q))$$

$$\alpha_{t,j} = \mathrm{softmax}_t \, h_t^\top q_j \quad \beta_{t,j} = \mathrm{softmax}_j \, h_t^\top q_j$$

$$\beta_j = \tfrac{1}{|p|} \textstyle\sum_t \beta_{t,j} \qquad \alpha_t = \textstyle\sum_j \beta_j \alpha_{t,j}$$

Note that the final equation defining α_t can be interpreted as applying the attention β_j to the attentions $\alpha_{t,j}$. This reader uses (13) and (14).

4 EMERGENT PREDICATION STRUCTURE

As discussed in the introduction the entity identifiers such as "entity37" introduced in the CNN/DailyMail dataset cannot be assigned any semantics other than their identity. We should think of them as pointers or semantics-free constant symbols. Despite this undermining of semantics, aggregation readers using (8) and (10) are able to perform well. Here we posit that this is due to an emergent predication structure in the hidden vectors h_t. Intuitively we want to think of the hidden state vector h_t as a concatenation $[e(\Phi_t), e'_o(a_t)]$ where Φ_t carries semantic information true of a_t. We think of h_t as representing $\Phi_t[a_t]$ for semantic statement $\Phi_t[x]$ asserted of the constant symbol

a_t. We also think of the vector representation q of the question as having the form $[e(\Psi), 0]$ and the vector embedding $e_o(a)$ as having the form $[0, e'_o(a)]$.

Unfortunately, the decomposition of h_t into this predication structure need not be axis aligned. Rather than posit an axis-aligned concatenation we posit that the hidden vector space H is a possibly non-aligned direct sum

$$H = S \oplus E \tag{15}$$

where S is a subspace of "statement vectors" and E is an orthogonal subspace of "entity pointers". Each hidden state vector $h \in H$ then has a unique decomposition as $h = \Psi + e$ for $\Psi \in S$ and $e \in E$. This is equivalent to saying that the hidden vector space H is some rotation of a concatenation of the vector spaces S and E.

We now present empirical evidence for this decomposition structure. We first note that the predication decomposition implies that $e_o(a)^\top h_t$ equals $e_o(a)^\top e_o(a_t)$. This suggests the following for some fixed positive constant c.

$$e_o(a)^\top h_t = \begin{cases} c & \text{if } t \in R(a,p) \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

Assuming the predication structure we have $c = ||e_o(a)||^2$. We note that if different entity constants had different norms then answers would be biased toward occurrences of the constant symbol of larger norm. But we need to have that all constant symbols are equivalent. We note that (**??**) gives

$$\begin{aligned} \operatorname*{argmax}_a \; e_o(a)^\top o &= \operatorname*{argmax}_a \; e_o(a)^\top \sum_t \alpha_t h_t \\ &= \operatorname*{argmax}_a \; \sum_t \alpha_t \, e_o(a)^\top h_t = \operatorname*{argmax}_a \sum_{t \in R(a,p)} \alpha_t \end{aligned}$$

and hence (10) and (14) agree — the aggregation readers and the explicit reference readers are using essentially the same answer selection criterion.

Empirical evidence for (16) is given in the first three rows of table 1. The first row empirically measures the "constant" c in (16) by measuring $e_0(a)^\top h_t$ for those cases where $t \in R(a,p)$. The second row measures "0" in (16) by measuring $e_o(a)^\top h_t$ in those cases where $t \notin R(a,p)$. Additional evidence for (16) is given in figure 1 showing that the output vectors $e_o(a)$ for different entity identifiers a are nearly orthogonal. Orthogonality of the output vectors is required by (16) provided that each output vector $e_o(a)$ is in the span of the hidden state vectors $h_{t,p}$ for which $t \in R(a,p)$. Intuitively, the mean of all vectors $h_{t,p}$ with $t \in R(a,p)$ should be approximately equal to $e_o(a)$. Of course empirically this will only be approximately true.

Equation (16) would suggest that the vector embedding of the constant symbols should have dimension at least as large as the number of distinct constants. However, in practice is sufficient that $e(a)^\top e(a')$ is small for $a \neq a'$. This allows the vector embeddings of the constants to have dimension much smaller than the number of constants. We have experimented with two-sparse constant symbol embeddings where the number of embedding vectors in dimention d is $2d(d-1)$ (d choose 2 times the four ways of setting the signs of the non-zero coordinates). Although we do not report results here, these designed and untrained constant embeddings worked reasonably well.

Table 1: Statistics to support (16) and (17). These statistics are computed for the Stanford Reader.

	CNN Dev			CNN Test		
	samples	mean	variance	samples	mean	variance
$e_o(a)^\top h_t, \quad t \in R(a,p)$	222,001	10.66	2.26	164,746	10.70	2.45
$e_o(a)^\top h_t, \quad t \notin R(a,p)$	93,072,682	-0.57	1.59	68,451,660	-0.58	1.65
$e_o(a)^\top h_{t\pm1}, \quad t \in R(a,p)$	443,878	2.32	1.79	329,366	2.25	1.84
$\text{Cosine}(q, h_t), \quad \exists a \; t \in R(a,p)$	222,001	0.22	0.11	164,746	0.22	0.12
$\text{Cosine}(q, e_o(a)), \quad \forall a$	103,909	-0.03	0.04	78,411	-0.03	0.04

As further support for (16) we give heat maps for $e_o(a)h_t$ in the appendix for different identifiers a and heat maps for α_t for different readers in the appendix.

As another testable predication we note that the posited decomposition of the hidden state vectors implies

$$q^\top(h_i + e_o(a)) = q^\top h_i. \tag{17}$$

This equation is equivalent to $q^\top e_o(a) = 0$. Experimentally, however, we cannot expect $q^\top e_o(a)$ to be exactly zero and (17) seems to provides a more experimentally meaningful test. Empirical evidence for (17) is given in the fourth and fifth row of table 1. The fourth row measures the cosine of the angle between the question vector q and the hidden state h_t averaged over passage positions t at which some entity identifier occurs. The fifth row measures the cosine of the angle between q and $e_o(a)$ averaged over the entity identifiers a.

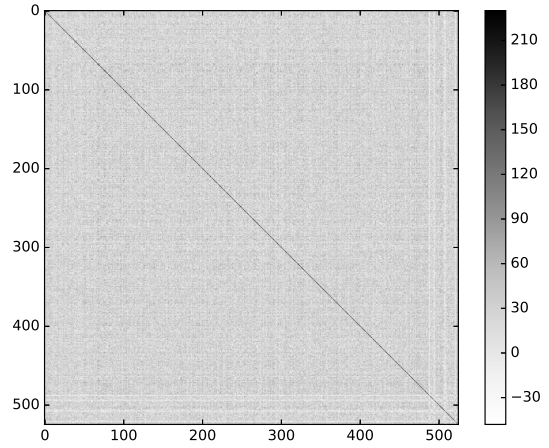

Figure 1: Plot of $e_o(a_i)^\top e_o(a_j)$ from Stanford Reader trained on CNN dataset. Off-diagonal values have mean 25.6 and variance 17.2 while diagonal values have mean 169 and variance 17.3.

A question asks for a value of x such that a statement $\Psi[x]$ is implied by the passage. For a question Ψ we might even suggest the following vectorial interpretation of entailment.

$$\Phi[x] \text{ implies } \Psi[x] \quad \text{iff} \quad \Phi^\top \Psi \geq ||\Psi||_1.$$

This interpretation is exactly correct if some of the dimensions of the vector space correspond to predicates, Ψ is a 0-1 vector representing a conjunction predicates, and Φ is also 0-1 on these dimensions indicating whether a predicate is implied by the context. Of course in practice one expects the dimension to be smaller than the number of possible predicates.

5 Pointer Annotation Readers

It is of course important to note that anonymization provides reference information — anonymization assumes that one can determine coreference so as to replace coreferent phrases with the same entity identifier. Anonymization allows the reference set $R(a, p)$ to be directly read off of the passage. Still, an aggregation reader must learn to recover this explicit reference structure.

Aggregation readers can have difficulty when anonymization is not done. The Stanford Reader achieves just better than 45% on Who-did-What dataset while Attention Sum Reader can get near 60%. But if we anonymize the Who-did-What dataset and then re-train the Stanford Reader, the accuracy jumps to near 65%. Anonymization has two effects. First, it greatly reduces the number of output word $e_o(a)$ to be learned — we need only learn output embeddings for the relatively small number of entity identifiers needed. Second, anonymization suppresses the semantics of the reference phrases and leaves only a semantics-free entity identifier. This suppression of semantics may facilitate the separation of the hidden state vector space H into a direct sum $S \oplus E$ with $q \in S$ and $e_o(a) \in E$.

We can think of anonymization as providing additional linguistic input for the reader — it explicitly marks positions of candidate answers and establishes coreference. A natural question is whether

Table 2: Accuracy on WDW dataset. All these results are based on single model. Results for neural readers other than NSE are based on replications of those systems. All models were trained on the relaxed training set which uniformly yields better performance than the restricted training set. The first group of models are explicit reference models and the second group are aggregation models. + indicates anonymization with better reference identifier.

Who did What	Val	Test
Attention Sum Reader (Onishi et al., 2016)	59.8	58.8
Gated Attention Reader (Onishi et al., 2016)	60.3	59.6
NSE (Munkhdalai & Yu, 2016)	66.5	66.2
Gated Attention + Linguistic Features$^+$	72.2	**72.8**
Stanford Reader	46.1	45.8
Attentive Reader with Anonymization	55.7	55.5
Stanford Reader with Anonymization	64.8	64.5
One-Hot Pointer Reader	65.1	64.4
One-Hot Pointer Reader + Linguistic Features$^+$	69.3	68.7
Stanford with Anonymization + Linguistic Features$^+$	69.7	**69.2**
Human Performance	-	84

this information can be provided without anonymization by simply adding additional coreference features to the input. Here we evaluate two architectures inspired by this question. This evaluation is done on the Who-did-What dataset which is not anonymized. In each architecture we add features to the input to mark the occurrences of candidate answers. These models are simpler than the Stanford reader but perform comparably. This comparable performance in table 2 further supports our analysis of logical structure in aggregation readers.

One-Hot Pointer Annotation: The Stanford Reader involves both input embeddings of words and output embeddings of entity identifiers. In the Who-did-What dataset each problem has at most five choices in the multiple choice answer list. This means that we need only five entity identifiers and we can use a five dimensional one-hot vector representation for answer identifiers. If an answer choice exists at position t in the passage let i_t be the index of that choice on the choice list. If no choice occurs t take i_t to be zero. Take $e'(i)$ to be the zero vector if $i = 0$ and otherwise to be the one-hot vector for i. We defined pointer annotation to be the result of adding $e'(i_t)$ as additional features to the input embedding.

$$e(w_t) = [e(w_t), e'(i_t)] \tag{18}$$

We then define a one-hot pointer reader by designates five dimensions of the hidden state as indicators of the answer and take the probability of choice i to be defined as

$$p(i|d, q) = \operatorname*{softmax}_i \; o_i \tag{19}$$

where o is computed by (8).

General Pointer Annotation: In the CNN dataset there are roughly 500 entity identifier and a one-hot representation is not desirable. Instead we can let $e'(i)$ be a fixed set of "pointers vectors" — vectors distributed widely on the unit sphere so that for $i \neq j$ we have that $e'(i)^\top e'(j)$ is small. We again use (18) but replace (19) with

$$p(i|d, q) = \operatorname*{softmax}_i \; [0, e'(i)]^\top o \tag{20}$$

In the general pointer reader the pointer embeddings $e'(i)$ are held fixed and not trained.

Linguistic Features. Each model can be modified to include additional input features for each input token in the question and passage. More specifically we can add the following features to the word embeddings.

- Binary feature: whether current token occurs in the question.
- Real value feature: the frequency of current token in the passage.

- Real value feature: position of the token's first occurrence in the passage as a percentage of the passage length.
- Binary feature: whether the text surrounding token match the text surrounding the placeholder in the question. We only have features for matching both left and right one word.
- One hot vector: Part-of-speech (POS) tagging. We only use such feature on CBT dataset.
- One hot vector: Name Entity Recognition (NER). We only use such feature on CBT dataset.

6 A SURVEY OF RECENT RESULTS

The performance of various recent readers on CNN, DailyMail and CBTest are summarized in Table 3. For purposes of comparison we only present results on single models. Model ensembles generally perform better than single models but are require more computation to train making comparisons more difficult. More experimental details can be found in appendix.

Table 3: Accuracy on CNN, DailyMail, CBTest NE and CBTest CN. All results are based on a single model. Results other than those involving pointer or linguistic feature annotations are taken from the original publications. Readers in the first group are explicit reference readers. Readers in the second group are aggregation readers. The final reader defies this classification.

	CNN		DailyMail		CBT NE		CBT CN	
	valid	test	valid	test	valid	test	valid	test
Human(context+query)	-	-	-	-	-	81.6	-	81.6
Attention Sum (Kadlec et al., 2016)	68.6	69.5	75.0	73.9	73.8	68.6	68.8	63.4
Gated Attention (Dhingra et al., 2016)	73.0	73.8	76.7	75.7	74.9	69.0	69.0	63.9
AoA Reader (Cui et al., 2016)	73.1	74.4	-	-	77.8	72.0	72.2	69.4
NSE (Munkhdalai & Yu, 2016)	-	-	-	-	78.2	**73.2**	74.2	**71.4**
DER Network (Kobayashi et al., 2016)	71.3	72.9	-	-	-	-	-	-
Epi Reader (Trischler et al., 2016)	73.4	74.0	-	-	75.3	69.7	71.5	67.4
Iterative Reader (Sordonif et al., 2016)	72.6	73.3	-	-	75.2	68.6	72.1	69.2
QANN (Weissenborn, 2016)	-	73.6	-	**77.2**	-	70.6	-	-
Gated Attention with linguistic features	74.7	**75.4**	78.6	**78.3**	75.7	**72.2**	73.3	**70.1**
MemNets (Sukhbaatar et al., 2015)	63.4	66.8	-	-	70.4	66.6	64.2	63.0
Attentive Reader (Hermann et al., 2015)	61.6	63.0	70.5	69.0	-	-	-	-
Stanford Reader (Chen et al., 2016)	72.5	72.7	76.9	76.0	-	-	-	-
Stanford Reader with linguistic features	75.7	**76.0**	-	-	-	-	-	-
ReasoNet (Shen et al., 2016)	72.9	74.7	77.6	76.6	-	-	-	-

In table 3, all the high-performance approaches are proposed very recently. Blue color represents the second highest accuracy and bold font indicates the state-of-the-art accuracy. Note that the result of Stanford Reader we report here is the one without relabeling since relabeling procedure doesn't follow the protocol used in Hermann et al. (2015).

7 DISCUSSION

Explicit reference architectures rely on reference resolution — a specification of which phrases in the given passage refer to candidate answers. Our experiments indicate that all existing readers benefit greatly from this externally provided information. Aggregation readers seem to demonstrate a stronger learning ability in that they essentially learn to mimic explicit reference readers by identifying reference annotation and using it appropriately. This is done most clearly in the pointer reader architectures. Furthermore, we have argued for, and given experimental evidence for, an interpretation of aggregation readers as learning emergent logical structure — a factoring of neural representations into a direct sum of a statement (predicate) representation and an entity (argument) representation.

At a very high level our analysis and experiments support a central role for reference resolution in reading comprehension. Automating reference resolution in neural models, and demonstrating its value on appropriate datasets, would seem to be an important area for future research.

Of course there is great interest in "learning representations". The current state of the art in reading comprehension is such that systems still benefit from externally provided linguistic features including externally annotated reference resolution. It would seem desirable to develop fully automated neural readers that perform as well as readers using externally provided annotations. It is of course important to avoid straw man baselines when making any such claim.

We are hesitant to make any more detailed comments on the differences between the architectural details of the readers discussed in this paper. The differences in scores between the leading readers are comparable to differences in scores that can be achieved by aggressive search over meta parameters or the statistical fluctuations in the quality of models learned by noisy statistical training procedures. More careful experiments over a longer period of time are needed. More dramatic improvements in performance would of course provide better support for particular innovations.

ACKNOWLEDGMENTS

We thanks the support of NVIDIA Corporation with the donation of GPUs used for this work.

REFERENCES

Danqi Chen, Jason Bolton, and Christopher D Manning. A thorough examination of the cnn/daily mail reading comprehension task. *In Proceedings of the ACL*, 2016.

Zewei Chu, Hai Wang, Kevin Gimpel, and David McAllester. Broad context language modeling as reading comprehension. *Arxiv*, 2016.

Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. Attention-over-attention neural networks for reading comprehension. *Arxiv*, 2016.

Graff David and Christopher Cieri. English gigaword ldc2003t05. *Philadelphia: Linguistic Data Consortium*, 2003.

Paperno. Denis, Germn Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernndez. The lambada dataset: Word prediction requiring a broad discourse context. *In Proceedings of the ACL*, 2016.

Bhuwan Dhingra, Hanxiao Liu, William W. Cohen, and Ruslan Salakhutdinov. Gated-attention readers for text comprehension. *Arxiv*, 2016.

Karm Moritz Hermann, Tom Kocisk, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. *In Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2015.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. The goldilocks principle: Reading childrens books with explicit memory representations. *In Proceedings of the 4th International Conference on Learning Representations*, 2016.

Pennington Jeffrey, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. *In Proceedings of the Conference on Empirical Methods on Natural Language Processing*, 14:1532–1543, 2014.

Rudolf Kadlec, Martin Schmid, Ondrej Bajgar, and Jan Kleindienst. Text understanding with the attention sum reader network. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 1:908–918, 2016.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *In Proceedings of the 3rd International Conference on Learning Representations*, 2015.

Sosuke Kobayashi, Ran Tian, Naoaki Okazaki, and Kentaro Inui. Dynamic entity representation with max-pooling improves machine reading. *In Proceedings of the North American Chapter of the Association for Computational Linguistics and Human Language Technologies (NAACL-HLT)*, 2016.

Moontae Lee, Xiaodong He, Scott Wen tau Yih, Jianfeng Gao, Li Deng, and Paul Smolensky. Reasoning in vector space: An exploratory study of question answering. *Proceedings of the 4th International Conference on Learning Representations*, 2016.

Tsendsuren Munkhdalai and Hong Yu. Reasoning with memory augmented neural networks for language comprehension. *Arxiv*, 2016.

Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. Who did what: A large-scale person-centered cloze dataset. *In Proceedings of the EMNLP*, 2016.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *In Proceedings of International Conference on Empirical Methods in Natural Language Processing*, 2016.

Pascanu Razvan, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. *In Proceedings of ICML*, pp. 1310–1318, 2013.

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. Mctest: A challenge dataset for the open-domain machine comprehension of text. *In Proceedings of the Conference on Empirical Methods on Natural Language Processing*, 3:4–10, 2013.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *Arxiv*, 2013.

Yelong Shen, Po-Sen Huang, Jianfeng Gao, and Weizhu Chen. Reasonet: Learning to stop reading in machine comprehension. *Arxiv*, 2016.

Alessandro Sordonif, Phillip Bachmanf, and Yoshua Bengio. Iterative alternating neural attention for machine reading. *Arxiv*, 2016.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. *In Advances in neural information processing systems*, pp. 2440–2448, 2015.

Adam Trischler, Zheng Ye, Xingdi Yuan, and Kaheer Suleman. Natural language comprehension with the epireader. *Arxiv*, 2016.

Bart van Merrienboer, Dzmitry Bahdanau, Vincent Dumoulin, Dmitriy Serdyuk, David Warde-farley, Jan Chorowski, and Yoshua Bengio. Blocks and fuel : Frameworks for deep learning. *Arxiv*, 2015.

Dirk Weissenborn. Separating answers from queries for neural reading comprehension. *Arxiv*, 2016.

Jason Weston, Sumit Chopra, and Antoine Bordes.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merrinboer, Armand Joulin, and Tomas Mikolov. Towards ai complete question answering: A set of prerequisite toy tasks. *In Proceedings of the 4th International Conference on Learning Representations*, 2016.

Fre de ric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian J. Goodfellow, Arnaud Bergeron, Nicolas Bouchard, and Yoshua Bengio. Theano: new features and speed improvements. *NIPS Workshop Deep Learning and Unsupervised Feature Learning*, 2012.

8 APPENDIX

8.1 EXPERIMENT DETAILS

We implemented the neural readers using Theano (de ric Bastien et al., 2012) and Blocks (van Merrienboer et al., 2015) and train them on single Nvidia Tesla K40 GPU. Negative log-likelihood is employed as training criterion. We used stochastic gradient descent (SGD) with the ADAM update rule (Kingma & Ba, 2015) and set the learning rate 0.0005.

For Stanford Reader and One-Hot Pointer Reader, we simply follows the Stanford Reader's setting and didn't tune it on each dataset. For gated attention reader, the lookup table was randomly initialized with uniform distribution from the interval [-0.2, 0.2] on CBT dataset, but on CNN&DailyMail, the lookup table was initialized by Glove vector (Jeffrey et al., 2014) trained on the train&validatation set (we found that the pre-trained word vector doesn't improve the accuracy but will accelerate the training) on CNN&DailyMail. On WDW dataset, the lookup table was initialized by pre-trained Glove vector[2]. It should be noticed that if we initialize the lookup table with pre-trained Glove vector from //nlp.stanford.edu/data/glove.6B.zip, it will slightly boost the accuracy compared with using the Glove vector trained on train&validation set. Input to hidden state weights were initialized by random orthogonal matrices (Saxe et al., 2013) and biases were initialized to zero. Hidden to hidden state weights were initialized by identity matrices to force the model can remember longer information. To compute the attention weight, we $\alpha_t = h_t \top W_\alpha q$ and initialize W_α with random uniform distribution. We also used the gradient clipping (Razvan et al., 2013) with threshold of 10 and batches of size 32.

During training we randomly shuffled all examples within each epoch. To speedup training, we always pre-fetched 10 batches worth of examples and sorted them according to document length as did by Kadlec et al. (2016). When trained on CNN, DailyMail and WDW (anonymization case) dataset, we randomly reshuffled the entity identifier to match the procedure proposed in Hermann et al. (2015).

During training we evaluated the accuracy after each epoch and stopped the training when the accuracy on the validation set started decreasing. We tried limiting the vocabulary to the most frequent tokens but didn't observed any performance improvement compared with using all the distinct tokens as vocabulary. Since part of our experiments need to check the word embedding assignment issues, finally we use all the distinct tokens as vocabulary. To find the optimal embedding and hidden state dimension, we tried several groups of different combinations, the optimal value and corresponding training statistics in Gated Attention readers are summarized in Table. 4. When anonymize the Who-did-What dataset, we can either use simple string match to replace answer in question and story with entity identifier, or we can use Name Entity Recognition(NER) tools[3] to detect name entities and then replace the answer name entities in question and story with entity identifier, we found the later one generally will bring 2 % improvement compared with simple string match. More experimental details can be found in code.

Table 4: Training Details on Different Datasets

Dataset	Embedding	Hidden State	Time Per Epoch	Trained Epochs	K
CNN	128	256	18 hours	5	3
DailyMail	128	256	2 days	5	3
WDW Relaxed	200	384	2.5 hours	8	1
CBT NE	384	384	1 hour	8	1
CBT CN	384	256	1 hour	7	1

8.2 HEAT MAP OF STANFORD READER FOR DIFFERENT ANSWER CANDIDATES

We randomly choose one article from CNN dataset and show softmax($e_o(a)h_t$) for $t \in [0, |p|]$ for each answer candidate a in figure.2, figure.3, figure.4, figure.5 and figure.6. Red color indicates

[2]http://nlp.stanford.edu/data/glove.6B.zip

[3]http://nlp.stanford.edu/software/CRF-NER.shtml

larger probability and orange indicates smaller probability and the remaining indicates very low probability that can be ignored. From those figures, we can see that our assumption that $e_o(a)$ is used to pick up its occurrence is reasonable.

@entity0 (@entity1) six survivors of the @entity0 kosher supermarket siege in january are suing a @entity5 media outlet for what they call dangerous live broadcasting during the hostage - taking . according to @entity0 prosecutor 's spokeswoman @entity10 , the lawsuit was filed march 27 and a preliminary investigation was opened by the prosecutor 's office wednesday . the media outlet , @entity1 affiliate @entity16 , is accused of endangering the lives of the hostages , who were hiding in a cold room during the attack , by broadcasting their location live during the siege . @entity23 in a statement friday said one of its journalists " mentioned only once the presence of a woman hidden inside the @entity27 , on the basis of police sources on the ground . " " immediately , the chief editor felt that this information should not be released . it therefore has subsequently never been repeated on air or posted on - screen . @entity16 regrets that the mention of this information could cause concern to the hostages , as well as their relatives , that their lives were in danger , " the statement said . gunman @entity47 , also suspected in the slaying of a police officer , stormed the @entity27 @entity51 supermarket on january 9 , killing four people and taking others hostage . he was killed in the police operation to end the siege . a 24 - year - old supermarket employee , @entity57 - born @entity56 , was hailed as a hero afterward when it emerged that he had risked his life to hide 15 customers from @entity47 in the cold room . the hostage - taking was the culmination of three days of terror in @entity0 that began with the january 7 shooting of 12 people at the offices of @entity5 satirical magazine @entity69 . the two brothers blamed for that attack , @entity72 and @entity73 , were killed on january 9 after a violent standoff at an industrial site . the terror attacks claimed the lives of 17 people and put @entity5 on a heightened state of alert . @entity1 's @entity80 reported from @entity0 , and @entity81 wrote from @entity82 . @entity1 's @entity83 contributed to this report .
query: they hid in a cold room during the attack in @entity0 by gunman @placeholder

Figure 2: Heat map of softmax($e_o(a)h_t$) when a = entity0.

8.3 Heat map of different Readers

We randomly choose one article from CNN dataset and show the attention map $\alpha_t = \text{softmax}(q^\top W_a h_t)$ for different readers (in Attention Sum and Gated Attention Reader, W_α is identity matrix). From figure 7, figure 8 and figure 9, we can see that different readers essential put the weights on the entity identifiers.

@entity0 (@entity1) six survivors of the @entity0 kosher supermarket siege in january are suing a @entity5 media outlet for what they call dangerous live broadcasting during the hostage - taking . according to @entity0 prosecutor 's spokeswoman @entity10 , the lawsuit was filed march 27 and a preliminary investigation was opened by the prosecutor 's office wednesday . the media outlet , @entity1 affiliate @entity16 , is accused of endangering the lives of the hostages , who were hiding in a cold room during the attack , by broadcasting their location live during the siege . @entity23 in a statement friday said one of its journalists " mentioned only once the presence of a woman hidden inside the @entity27 , on the basis of police sources on the ground . " " immediately , the chief editor felt that this information should not be released . it therefore has subsequently never been repeated on air or posted on - screen . @entity16 regrets that the mention of this information could cause concern to the hostages , as well as their relatives , that their lives were in danger , " the statement said . gunman @entity47 , also suspected in the slaying of a police officer , stormed the @entity27 @entity51 supermarket on january 9 , killing four people and taking others hostage . he was killed in the police operation to end the siege . a 24 - year - old supermarket employee , @entity57 - born @entity56 , was hailed as a hero afterward when it emerged that he had risked his life to hide 15 customers from @entity47 in the cold room . the hostage - taking was the culmination of three days of terror in @entity0 that began with the january 7 shooting of 12 people at the offices of @entity5 satirical magazine @entity69 . the two brothers blamed for that attack , @entity72 and @entity73 , were killed on january 9 after a violent standoff at an industrial site . the terror attacks claimed the lives of 17 people and put @entity5 on a heightened state of alert . @entity1 's @entity80 reported from @entity0 , and @entity81 wrote from @entity82 . @entity1 's @entity83 contributed to this report .
query: they hid in a cold room during the attack in @entity0 by gunman @placeholder

Figure 3: Heat map of softmax($e_o(a)h_t$) when a = entity1.

@entity0 (@entity1) six survivors of the @entity0 kosher supermarket siege in january are suing a @entity5 media outlet for what they call dangerous live broadcasting during the hostage - taking . according to @entity0 prosecutor 's spokeswoman @entity10 , the lawsuit was filed march 27 and a preliminary investigation was opened by the prosecutor 's office wednesday . the media outlet , @entity1 affiliate @entity16 , is accused of endangering the lives of the hostages , who were hiding in a cold room during the attack , by broadcasting their location live during the siege . @entity23 in a statement friday said one of its journalists " mentioned only once the presence of a woman hidden inside the @entity27 , on the basis of police sources on the ground . " " immediately , the chief editor felt that this information should not be released . it therefore has subsequently never been repeated on air or posted on - screen . @entity16 regrets that the mention of this information could cause concern to the hostages , as well as their relatives , that their lives were in danger , " the statement said . gunman @entity47 , also suspected in the slaying of a police officer , stormed the @entity27 @entity51 supermarket on january 9 , killing four people and taking others hostage . he was killed in the police operation to end the siege . a 24 - year - old supermarket employee , @entity57 - born @entity56 , was hailed as a hero afterward when it emerged that he had risked his life to hide 15 customers from @entity47 in the cold room . the hostage - taking was the culmination of three days of terror in @entity0 that began with the january 7 shooting of 12 people at the offices of @entity5 satirical magazine @entity69 . the two brothers blamed for that attack , @entity72 and @entity73 , were killed on january 9 after a violent standoff at an industrial site . the terror attacks claimed the lives of 17 people and put @entity5 on a heightened state of alert . @entity1 's @entity80 reported from @entity0 , and @entity81 wrote from @entity82 . @entity1 's @entity83 contributed to this report .
query: they hid in a cold room during the attack in @entity0 by gunman @placeholder

Figure 4: Heat map of softmax($e_o(a)h_t$) when a = entity16.

@entity0 (@entity1) six survivors of the @entity0 kosher supermarket siege in january are suing a @entity5 media outlet for what they call dangerous live broadcasting during the hostage - taking . according to @entity0 prosecutor 's spokeswoman @entity10 , the lawsuit was filed march 27 and a preliminary investigation was opened by the prosecutor 's office wednesday . the media outlet , @entity1 affiliate @entity16 , is accused of endangering the lives of the hostages , who were hiding in a cold room during the attack , by broadcasting their location live during the siege . @entity23 in a statement friday said one of its journalists " mentioned only once the presence of a woman hidden inside the @entity27 , on the basis of police sources on the ground . " " immediately , the chief editor felt that this information should not be released . it therefore has subsequently never been repeated on air or posted on - screen . @entity16 regrets that the mention of this information could cause concern to the hostages , as well as their relatives , that their lives were in danger , " the statement said . gunman @entity47 , also suspected in the slaying of a police officer , stormed the @entity27 @entity51 supermarket on january 9 , killing four people and taking others hostage . he was killed in the police operation to end the siege . a 24 - year - old supermarket employee , @entity57 - born @entity56 , was hailed as a hero afterward when it emerged that he had risked his life to hide 15 customers from @entity47 in the cold room . the hostage - taking was the culmination of three days of terror in @entity0 that began with the january 7 shooting of 12 people at the offices of @entity5 satirical magazine @entity69 . the two brothers blamed for that attack , @entity72 and @entity73 , were killed on january 9 after a violent standoff at an industrial site . the terror attacks claimed the lives of 17 people and put @entity5 on a heightened state of alert . @entity1 's @entity80 reported from @entity0 , and @entity81 wrote from @entity82 . @entity1 's @entity83 contributed to this report .
query: they hid in a cold room during the attack in @entity0 by gunman @placeholder

Figure 5: Heat map of softmax($e_o(a)h_t$) when a = entity27.

@entity0 (@entity1) six survivors of the @entity0 kosher supermarket siege in january are suing a @entity5 media outlet for what they call dangerous live broadcasting during the hostage - taking . according to @entity0 prosecutor 's spokeswoman @entity10 , the lawsuit was filed march 27 and a preliminary investigation was opened by the prosecutor 's office wednesday . the media outlet , @entity1 affiliate @entity16 , is accused of endangering the lives of the hostages , who were hiding in a cold room during the attack , by broadcasting their location live during the siege . @entity23 in a statement friday said one of its journalists " mentioned only once the presence of a woman hidden inside the @entity27 , on the basis of police sources on the ground . " " immediately , the chief editor felt that this information should not be released . it therefore has subsequently never been repeated on air or posted on - screen . @entity16 regrets that the mention of this information could cause concern to the hostages , as well as their relatives , that their lives were in danger , " the statement said . gunman @entity47 , also suspected in the slaying of a police officer , stormed the @entity27 @entity51 supermarket on january 9 , killing four people and taking others hostage . he was killed in the police operation to end the siege . a 24 - year - old supermarket employee , @entity57 - born @entity56 , was hailed as a hero afterward when it emerged that he had risked his life to hide 15 customers from @entity47 in the cold room . the hostage - taking was the culmination of three days of terror in @entity0 that began with the january 7 shooting of 12 people at the offices of @entity5 satirical magazine @entity69 . the two brothers blamed for that attack , @entity72 and @entity73 , were killed on january 9 after a violent standoff at an industrial site . the terror attacks claimed the lives of 17 people and put @entity5 on a heightened state of alert . @entity1 's @entity80 reported from @entity0 , and @entity81 wrote from @entity82 . @entity1 's @entity83 contributed to this report .
query: they hid in a cold room during the attack in @entity0 by gunman @placeholder

Figure 6: Heat map of softmax($e_o(a)h_t$) when a = entity47.

(@entity3) suspected @entity2 militants this week attacked civilians inside @entity5 for the first time in a month , killing at least 16 villagers , a military spokesman told @entity3 saturday . six attackers were killed by @entity5 forces , said maj. @entity10 , an operations officer with a special military unit set up to fight @entity2 . the attackers came thursday " in the hundreds ... torched @entity14 village in the @entity15 , " he said . @entity14 is a village that borders @entity17 and has been identified as a recruiting ground for @entity2 . regional gov. @entity19 said the insurgents have been attacking border villages in @entity5 in search of supplies . @entity5 troops retook cattle that was stolen by the attackers in @entity14 , @entity10 said . the last attack in @entity5 by the @entity29 - based militants was march 10 , when the assailants struck the locality of @entity32 in a failed attempt to overrun a military base . @entity2 , whose name translates as " @entity44 education is sin , " has been waging a years - long campaign of terror aimed at instituting its extreme version of @entity42 law in @entity29 . @entity2 's tactics have intensified in recent years , from battling @entity29 government soldiers to acts disproportionately affecting civilians -- such as raids on villages , mass kidnappings , assassinations , market bombings and attacks on churches and unaffiliated mosques . much of this violence has taken place in @entity29 , but neighboring countries -- @entity5 included -- have also been hit increasingly hard . journalist @entity61 in @entity63 , @entity5 , contributed to this report .

query: @placeholder is based in @entity29 but has attacked across the border of several neighbors

Figure 7: Heat map α_t for Stanford Reader

(@entity3) suspected @entity2 militants this week attacked civilians inside @entity5 for the first time in a month , killing at least 16 villagers , a military spokesman told @entity3 saturday . six attackers were killed by @entity5 forces , said maj. @entity10 , an operations officer with a special military unit set up to fight @entity2 . the attackers came thursday " in the hundreds ... torched @entity14 village in the @entity15 , " he said . @entity14 is a village that borders @entity17 and has been identified as a recruiting ground for @entity2 . regional gov. @entity19 said the insurgents have been attacking border villages in @entity5 in search of supplies . @entity5 troops retook cattle that was stolen by the attackers in @entity14 , @entity10 said . the last attack in @entity5 by the @entity29 - based militants was march 10 , when the assailants struck the locality of @entity32 in a failed attempt to overrun a military base . @entity2 , whose name translates as " @entity44 education is sin , " has been waging a years - long campaign of terror aimed at instituting its extreme version of @entity42 law in @entity29 . @entity2 's tactics have intensified in recent years , from battling @entity29 government soldiers to acts disproportionately affecting civilians -- such as raids on villages , mass kidnappings , assassinations , market bombings and attacks on churches and unaffiliated mosques . much of this violence has taken place in @entity29 , but neighboring countries -- @entity5 included -- have also been hit increasingly hard . journalist @entity61 in @entity63 , @entity5 , contributed to this report .

query: @placeholder is based in @entity29 but has attacked across the border of several neighbors

Figure 8: Heat map α_t for Gated Attention Reader

(@entity3) suspected @entity2 militants this week attacked civilians inside @entity5 for the first time in a month , killing at least 16 villagers , a military spokesman told @entity3 saturday . six attackers were killed by @entity5 forces , said maj. @entity10 , an operations officer with a special military unit set up to fight @entity2 . the attackers came thursday " in the hundreds ... torched @entity14 village in the @entity15 , " he said . @entity14 is a village that borders @entity17 and has been identified as a recruiting ground for @entity2 . regional gov. @entity19 said the insurgents have been attacking border villages in @entity5 in search of supplies . @entity5 troops retook cattle that was stolen by the attackers in @entity14 , @entity10 said . the last attack in @entity5 by the @entity29 - based militants was march 10 , when the assailants struck the locality of @entity32 in a failed attempt to overrun a military base . @entity2 , whose name translates as " @entity44 education is sin , " has been waging a years - long campaign of terror aimed at instituting its extreme version of @entity42 law in @entity29 . @entity2 's tactics have intensified in recent years , from battling @entity29 government soldiers to acts disproportionately affecting civilians -- such as raids on villages , mass kidnappings , assassinations , market bombings and attacks on churches and unaffiliated mosques . much of this violence has taken place in @entity29 , but neighboring countries -- @entity5 included -- have also been hit increasingly hard . journalist @entity61 in @entity63 , @entity5 , contributed to this report .

query: @placeholder is based in @entity29 but has attacked across the border of several neighbors

Figure 9: Heat map α_t for Attention Sum Reader