Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement

Nina Poerner, Benjamin Roth & Hinrich Schütze
Center for Information and Language Processing
LMU Munich, Germany
poerner@cis.lmu.de

Abstract

The behavior of deep neural networks (DNNs) is hard to understand. This makes it necessary to explore post hoc explanation methods. We conduct the first comprehensive evaluation of explanation methods for NLP. To this end, we design two novel evaluation paradigms that cover two important classes of NLP problems: small context and large context problems. Both paradigms require no manual annotation and are therefore broadly applicable. We also introduce LIMSSE, an explanation method inspired by LIME that is designed for NLP. We show empirically that LIMSSE, LRP and DeepLIFT are the most effective explanation methods and recommend them for explaining DNNs in NLP.

1 Introduction

DNNs are complex models that combine linear transformations with different types of nonlinearities. If the model is deep, i.e., has many layers, then its behavior during training and inference is notoriously hard to understand.

This is a problem for both scientific methodology and real-world deployment. Scientific methodology demands that we understand our models. In the real world, a decision (e.g., "your blog post is offensive and has been removed") by itself is often insufficient; in addition, an explanation of the decision may be required (e.g., "our system flagged the following words as offensive"). The European Union plans to mandate that intelligent systems used for sensitive applications provide such explanations (European General Data Protection Regulation, expected 2018, cf. Goodman and Flaxman (2016)).

A number of post hoc explanation methods for DNNs have been proposed. Due to the complexity of the DNNs they explain, these methods are necessarily approximations and come with their own sources of error. At this point, it is not clear which of these methods to use when reliable explanations for a specific DNN architecture are needed.

Definitions. (i) A *task method* solves an NLP problem, e.g., a GRU that predicts sentiment.

(ii) An *explanation method* explains the behavior of a task method on a specific input. For our purpose, it is a function $\phi(t, k, \mathbf{X})$ that assigns real-valued relevance scores for a target class k (e.g., positive) to positions t in an input text $\mathbf{X}$ (e.g., "great food"). For this example, an explanation method might assign: $\phi(1, k, \mathbf{X}) > \phi(2, k, \mathbf{X})$.

(iii) An *(explanation) evaluation paradigm* quantitatively evaluates explanation methods for a task method, e.g., by assigning them accuracies.

Contributions. (i) We present novel evaluation paradigms for explanation methods for two classes of common NLP tasks (see §2). Crucially, *neither paradigm requires manual annotations* and our methodology is therefore broadly applicable.

(ii) Using these paradigms, we perform a comprehensive evaluation of explanation methods for NLP (§3). We cover the most important classes of task methods, RNNs and CNNs, as well as the recently proposed Quasi-RNNs.

(iii) We introduce LIMSSE (§3.6), an explanation method inspired by LIME (Ribeiro et al.,

tasks	sentiment analysis, morphological prediction, ...
task methods	CNN, GRU, LSTM, ...
explanation methods	LIMSSE, LRP, DeepLIFT, ...
evaluation paradigms	hybrid document, morphosyntactic agreement

Table 1: Terminology with examples.

lrp	From : *kolstad* @ cae.wisc.edu (Joel **Kolstad**) Subject : Re : Can Radio Freq . Be Used To Measure Distance ? [...] What is the difference between vertical and horizontal ? Gravity ? Does n't gravity pull down the photons and cause a *doppler* shift or something ? (Just kidding !)
$\text{grad}^{\text{L2}}_{1p}$	If you find faith to be honest , show me how . David The whole **denominational** mindset only causes more problems , sadly . (See section 7 for details .) Thank you . 'The Armenians just shot and shot . Maybe *coz* they 're *'quality'* cars ; -) 200 *posts/day* . [...]
$\text{limsse}^{\text{ms}}_{s}$	If you find faith to be honest , show me how . David The whole denominational mindset only causes more problems , sadly . (See section 7 for details .) Thank you . 'The **Armenians** just shot and shot . Maybe *coz* they 're *'quality'* cars ; -) 200 *posts/day* . [...]

Figure 1: **Top:** sci.electronics post (not hybrid). Underlined: Manual relevance ground truth. Green: evidence for sci.electronics. Task method: CNN. **Bottom:** hybrid newsgroup post, classified talk.politics.mideast. Green: evidence for talk.politics.mideast. Underlined: talk.politics.mideast fragment. Task method: QGRU. Italics: OOV. Bold: rmax position. See supplementary for full texts.

2016) that is designed for word-order sensitive task methods (e.g., RNNs, CNNs). We show empirically that LIMSSE, LRP (Bach et al., 2015) and DeepLIFT (Shrikumar et al., 2017) are the most effective explanation methods (§4): LRP and DeepLIFT are the most consistent methods, while LIMSSE wins the hybrid document experiment.

2 Evaluation paradigms

In this section, we introduce two novel evaluation paradigms for explanation methods on two types of common NLP tasks, *small context* tasks and *large context* tasks. Small context tasks are defined as those that can be solved by finding short, self-contained indicators, such as words and phrases, and weighing them up (i.e., tasks where CNNs with pooling can be expected to perform well). We design the *hybrid document paradigm* for evaluating explanation methods on small context tasks. Large context tasks require the correct handling of long-distance dependencies, such as subject-verb agreement.[1] We design the *morphosyntactic agreement paradigm* for evaluating explanation methods on large context tasks.

We could also use **human judgments** for evaluation. While we use Mohseni and Ragan (2018)'s manual relevance benchmark for comparison, there are two issues with it: (i) Due to the *cost of human labor*, it is limited in size and domain. (ii) More importantly, *a good explanation method should not reflect what humans attend to, but what task methods attend to.* For instance, the family name "Kolstad" has 11 out of its 13 appearances in the 20 newsgroups corpus in sci.electronics posts. Thus, task methods probably learn it as a sci.electronics indicator. Indeed, the explanation method in Fig 1 (top) marks "Kolstad" as relevant, but the human annotator does not.

2.1 Small context: Hybrid document paradigm

Given a collection of documents, hybrid documents are created by randomly concatenating document fragments. We assume that, on average, the most relevant input for a class k in a hybrid document is located in a fragment that stems from a document with gold label k. Hence, an explanation method succeeds if it places maximal relevance for k inside the correct fragment.

Formally, let x_t be a word inside hybrid document $\mathbf{X}$ that originates from a document $\mathbf{X}'$ with gold label $y(\mathbf{X}')$. x_t's gold label $y(\mathbf{X}, t)$ is set to $y(\mathbf{X}')$. Let $f(\mathbf{X})$ be the class assigned to the hybrid document by a task method, and let ϕ be an explanation method as defined above. Let $\text{rmax}(\mathbf{X}, \phi)$ denote the position of the maximally relevant word in $\mathbf{X}$ for the predicted class $f(\mathbf{X})$. If this maximally relevant word comes from a document with the correct gold label, the explanation method is awarded a hit:

$$\text{hit}(\phi, \mathbf{X}) = \mathbb{I}[y\big(\mathbf{X}, \text{rmax}(\mathbf{X}, \phi)\big) = f(\mathbf{X})] \quad (1)$$

where $\mathbb{I}[P]$ is 1 if P is true and 0 otherwise. In Fig 1 (bottom), the explanation method $\text{grad}^{\text{L2}}_{1p}$ places rmax outside the correct (underlined) fragment. Therefore, it does not get a hit point, while $\text{limsse}^{\text{ms}}_{s}$ does.

The pointing game accuracy of an explanation method is calculated as its total number of hit points divided by the number of possible hit points. This is a form of the pointing game paradigm from computer vision (Zhang et al., 2016).

2.2 Large context: Morphosyntactic agreement paradigm

Many natural languages display morphosyntactic agreement between words v and w. A DNN that

[1] Consider deciding the number of *[verb]* in "the children in the green house said that the big telescope *[verb]*" vs. "the children in the green house who broke the big telescope *[verb]*". The local contexts of "children" or "*[verb]*" do not suffice to solve this problem, instead, the large context of the entire sentence has to be considered.

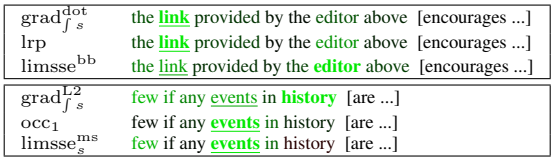

Figure 2: **Top:** verb context classified singular. Green: evidence for singular. Task method: GRU. **Bottom:** verb context classified plural. Green: evidence for plural. Task method: LSTM. Underlined: subject. Bold: rmax position.

predicts the agreeing feature in w should pay attention to v. For example, in the sentence "the children with the telescope are home", the number of the verb (plural for "are") can be predicted from the subject ("children") without looking at the verb. If the language allows for v and w to be far apart (Fig 3, top), successful task methods have to be able to handle large contexts.

Linzen et al. (2016) show that English verb number can be predicted by a unidirectional LSTM with accuracy $> 99\%$, based on left context alone. When a task method predicts the correct number, we expect successful explanation methods to place maximal relevance on the subject:

$$\text{hit}_{\text{target}}(\phi, \mathbf{X}) = \mathbb{I}[\text{rmax}(\mathbf{X}, \phi) = \text{target}(\mathbf{X})]$$

where $\text{target}(\mathbf{X})$ is the location of the subject, and rmax is calculated as above. Regardless of whether the prediction is correct, we expect rmax to fall onto a noun that has the predicted number:

$$\text{hit}_{\text{feat}}(\phi, \mathbf{X}) = \mathbb{I}[\text{feat}\big(\mathbf{X}, \text{rmax}(\mathbf{X}, \phi)\big) = f(\mathbf{X})]$$

where $\text{feat}(\mathbf{X}, t)$ is the morphological feature (here: number) of x_t. In Fig 2, rmax on "link" gives a $\text{hit}_{\text{target}}$ point (and a hit_{feat} point), rmax on "editor" gives a hit_{feat} point. $\text{grad}^{\text{L2}}_{\int s}$ does not get any points as "history" is not a plural noun.

Labels for this task can be automatically generated using part-of-speech taggers and parsers, which are available for many languages.

3 Explanation methods

In this section, we define the explanation methods that will be evaluated. For our purpose, explanation methods produce word relevance scores $\phi(t, k, \mathbf{X})$, which are specific to a given class k and a given input $\mathbf{X}$. $\phi(t, k, \mathbf{X}) > \phi(t', k, \mathbf{X})$ means that x_t contributed more than $x_{t'}$ to the task method's (potential) decision to classify $\mathbf{X}$ as k.

3.1 Gradient-based explanation methods

Gradient-based explanation methods approximate the contribution of some DNN input i to some output o with o's gradient with respect to i (Simonyan et al., 2014). In the following, we consider two output functions $o(k, \mathbf{X})$, the unnormalized class score $s(k, \mathbf{X})$ and the class probability $p(k|\mathbf{X})$:

$$s(k, \mathbf{X}) = \vec{w}_k \cdot \vec{h}(\mathbf{X}) + b_k \tag{2}$$

$$p(k|\mathbf{X}) = \frac{\exp\big(s(k, \mathbf{X})\big)}{\sum_{k'=1}^{K} \exp\big(s(k', \mathbf{X})\big)} \tag{3}$$

where k is the target class, $\vec{h}(\mathbf{X})$ the document representation (e.g., an RNN's final hidden layer), $\vec{w}_k$ (resp. b_k) k's weight vector (resp. bias).

The simple gradient of $o(k, \mathbf{X})$ w.r.t. i is:

$$\text{grad}_1(i, k, \mathbf{X}) = \frac{\partial o(k, \mathbf{X})}{\partial i} \tag{4}$$

grad_1 underestimates the importance of inputs that saturate a nonlinearity (Shrikumar et al., 2017). To address this, Sundararajan et al. (2017) integrate over all gradients on a linear interpolation $\alpha \in [0, 1]$ between a baseline input $\bar{\mathbf{X}}$ (here: all-zero embeddings) and $\mathbf{X}$:

$$\begin{aligned}\text{grad}_{\int}(i, k, \mathbf{X}) &= \int_{\alpha=0}^{1} \frac{\partial o(k, \bar{\mathbf{X}} + \alpha(\mathbf{X} - \bar{\mathbf{X}}))}{\partial i} \partial\alpha \\ &\approx \frac{1}{M} \sum_{m=1}^{M} \frac{\partial o(k, \bar{\mathbf{X}} + \frac{m}{M}(\mathbf{X} - \bar{\mathbf{X}}))}{\partial i}\end{aligned} \tag{5}$$

where M is a big enough constant (here: 50).

In NLP, symbolic inputs (e.g., words) are often represented as one-hot vectors $\vec{x}_t \in \{1, 0\}^{|V|}$ and embedded via a real-valued matrix: $\vec{e}_t = \mathbf{M}\vec{x}_t$. Gradients are computed with respect to individual entries of $\mathbf{E} = [\vec{e}_1 \ldots \vec{e}_{|\mathbf{X}|}]$. Bansal et al. (2016) and Hechtlinger (2016) use the L2 norm to reduce vectors of gradients to single values:

$$\phi_{\text{grad}^{\text{L2}}}(t, k, \mathbf{X}) = ||\text{grad}(\vec{e}_t, k, \mathbf{E})|| \tag{6}$$

where $\text{grad}(\vec{e}_t, k, \mathbf{E})$ is a vector of elementwise gradients w.r.t. $\vec{e}_t$. Denil et al. (2015) use the dot product of the gradient vector and the embedding[2], i.e., the gradient of the "hot" entry in $\vec{x}_t$:

$$\phi_{\text{grad}^{\text{dot}}}(t, k, \mathbf{X}) = \vec{e}_t \cdot \text{grad}(\vec{e}_t, k, \mathbf{E}) \tag{7}$$

We use "grad_1" for Eq 4, "$\text{grad}_{\int}$" for Eq 5, "$_p$" for Eq 3, "$_s$" for Eq 2, "L2" for Eq 6 and "dot" for Eq 7. This gives us eight explanation methods: $\text{grad}^{\text{L2}}_{1s}$, $\text{grad}^{\text{L2}}_{1p}$, $\text{grad}^{\text{dot}}_{1s}$, $\text{grad}^{\text{dot}}_{1p}$, $\text{grad}^{\text{L2}}_{\int s}$, $\text{grad}^{\text{L2}}_{\int p}$, $\text{grad}^{\text{dot}}_{\int s}$, $\text{grad}^{\text{dot}}_{\int p}$.

[2]For $\text{grad}^{\text{dot}}_{\int}$, replace $\vec{e}_t$ with $\vec{e}_t - \bar{\vec{e}}_t$. Since our baseline embeddings are all-zeros, this is equivalent.

3.2 Layer-wise relevance propagation

Layer-wise relevance propagation (LRP) is a backpropagation-based explanation method developed for fully connected neural networks and CNNs (Bach et al., 2015) and later extended to LSTMs (Arras et al., 2017b). In this paper, we use Epsilon LRP (Eq 58, Bach et al. (2015)). Remember that the activation of neuron j, a_j, is the sum of weighted upstream activations, $\sum_i a_i w_{i,j}$, plus bias b_j, squeezed through some nonlinearity. We denote the pre-nonlinearity activation of j as a'_j. The relevance of j, $R(j)$, is distributed to upstream neurons i proportionally to the contribution that i makes to a'_j in the forward pass:

$$R(i) = \sum_j R(j) \frac{a_i w_{i,j}}{a'_j + \text{esign}(a'_j)} \tag{8}$$

This ensures that relevance is conserved between layers, with the exception of relevance attributed to b_j. To prevent numerical instabilities, $\text{esign}(a')$ returns $-\epsilon$ if $a' < 0$ and ϵ otherwise. We set $\epsilon = .001$. The full algorithm is:

$$R(L_{k'}) = s(k, \mathbf{X}) \mathbb{I}[k' = k]$$

... recursive application of Eq 8 ...

$$\phi_{\text{lrp}}(t, k, \mathbf{X}) = \sum_{j=1}^{\dim(\vec{e}_t)} R(e_{t,j})$$

where L is the final layer, k the target class and $R(e_{t,j})$ the relevance of dimension j in the t'th embedding vector. For $\epsilon \to 0$ and provided that all nonlinearities up to the unnormalized class score are relu, Epsilon LRP is equivalent to the product of input and raw score gradient (here: $\text{grad}^{\text{dot}}_{1s}$) (Kindermans et al., 2016). In our experiments, the second requirement holds only for CNNs.

Experiments by Ancona et al. (2017) (see §6) suggest that LRP does not work well for LSTMs if all neurons – including gates – participate in backpropagation. We therefore use Arras et al. (2017b)'s modification and treat sigmoid-activated gates as time step-specific weights rather than neurons. For instance, the relevance of LSTM candidate vector $\vec{g}_t$ is calculated from memory vector $\vec{c}_t$ and input gate vector $\vec{i}_t$ as

$$R(g_{t,d}) = R(c_{t,d}) \frac{g_{t,d} \cdot i_{t,d}}{c_{t,d} + \text{esign}(c_{t,d})}$$

This is equivalent to applying Eq 8 while treating $\vec{i}_t$ as a diagonal weight matrix. The gate neurons in $\vec{i}_t$ do not receive any relevance themselves. See supplementary material for formal definitions of Epsilon LRP for different architectures.

3.3 DeepLIFT

DeepLIFT (Shrikumar et al., 2017) is another backpropagation-based explanation method. Unlike LRP, it does not explain $s(k, \mathbf{X})$, but $s(k, \mathbf{X}) - s(k, \bar{\mathbf{X}})$, where $\bar{\mathbf{X}}$ is some baseline input (here: all-zero embeddings). Following Ancona et al. (2018) (Eq 4), we use this backpropagation rule:

$$R(i) = \sum_j R(j) \frac{a_i w_{i,j} - \bar{a}_i w_{i,j}}{a'_j - \bar{a}'_j + \text{esign}(a'_j - \bar{a}'_j)}$$

where $\bar{a}$ refers to the forward pass of the baseline. Note that the original method has a different mechanism for avoiding small denominators; we use esign for compatibility with LRP. The DeepLIFT algorithm is started with $R(L_{k'}) = \big(s(k, \mathbf{X}) - s(k, \bar{\mathbf{X}})\big) \mathbb{I}[k' = k]$. On gated (Q)RNNs, we proceed analogous to LRP and treat gates as weights.

3.4 Cell decomposition for gated RNNs

The cell decomposition explanation method for LSTMs (Murdoch and Szlam, 2017) decomposes the unnormalized class score $s(k, \mathbf{X})$ (Eq 2) into additive contributions. For every time step t, we compute how much of $\vec{c}_t$ "survives" until the final step T and contributes to $s(k, \mathbf{X})$. This is achieved by applying all future forget gates $\vec{f}$, the final $\tanh$ nonlinearity, the final output gate $\vec{o}_T$, as well as the class weights of k to $\vec{c}_t$. We call this quantity "net load of t for class k":

$$\text{nl}(t, k, \mathbf{X}) = \vec{w}_k \cdot \Big(\vec{o}_T \odot \tanh\big((\prod_{j=t+1}^{T} \vec{f}_j) \odot \vec{c}_t \big) \Big)$$

where $\odot$ and $\prod$ are applied elementwise. The relevance of t is its gain in net load relative to $t-1$: $\phi_{\text{decomp}}(t, k, \mathbf{X}) = \text{nl}(t, k, \mathbf{X}) - \text{nl}(t-1, k, \mathbf{X})$. For GRU, we change the definition of net load:

$$\text{nl}(t, k, \mathbf{X}) = \vec{w}_k \cdot \big((\prod_{j=t+1}^{T} \vec{z}_j) \odot \vec{h}_t \big)$$

where $\vec{z}$ are GRU update gates.

3.5 Input perturbation methods

Input perturbation methods assume that the removal or masking of relevant inputs changes the

output (Zeiler and Fergus, 2014). Omission-based methods remove inputs completely (Kádár et al., 2017), while occlusion-based methods replace them with a baseline (Li et al., 2016b). In computer vision, perturbations are usually applied to patches, as neighboring pixels tend to correlate (Zintgraf et al., 2017). To calculate the omit_N (resp. occ_N) relevance of word x_t, we delete (resp. occlude), one at a time, all N-grams that contain x_t, and average the change in the unnormalized class score from Eq 2:

$$\phi_{[\text{omit}|\text{occ}]_N}(t,k,\mathbf{X}) = \sum_{j=1}^{N} \big[s(k,[\vec{e}_1 \dots \vec{e}_{|\mathbf{X}|}]) - s(k,[\vec{e}_1 \dots \vec{e}_{t-N-1+j}] \| \bar{\mathbf{E}} \| [\vec{e}_{t+j} \dots \vec{e}_{|\mathbf{X}|}])\big] \frac{1}{N}$$

where $\vec{e}_t$ are embedding vectors, $\|$ denotes concatenation and $\bar{\mathbf{E}}$ is either a sequence of length zero (ϕ_{omit}) or a sequence of N baseline (here: all-zero) embedding vectors (ϕ_{occ}).

3.6 LIMSSE: LIME for NLP

Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016) is a framework for explaining predictions of complex classifiers. LIME approximates the behavior of classifier f in the neighborhood of input $\mathbf{X}$ with an interpretable (here: linear) model. The interpretable model is trained on samples $\mathbf{Z}_1 \dots \mathbf{Z}_N$ (here: $N = 3000$), which are randomly drawn from $\mathbf{X}$, with "gold labels" $f(\mathbf{Z}_1) \dots f(\mathbf{Z}_N)$.

Since RNNs and CNNs respect word order, we cannot use the bag of words sampling method from the original description of LIME. Instead, we introduce Local Interpretable Model-agnostic Substring-based Explanations (LIMSSE). LIMSSE uniformly samples a length l_n (here: $1 \leq l_n \leq 6$) and a starting point s_n, which define the substring $\mathbf{Z}_n = [\vec{x}_{s_n} \dots \vec{x}_{s_n+l_n-1}]$. To the linear model, $\mathbf{Z}_n$ is represented by a binary vector $\vec{z}_n \in \{0,1\}^{|\mathbf{X}|}$, where $z_{n,t} = \mathbb{I}[s_n \leq t < s_n + l_n]$.

We learn a linear weight vector $\hat{\vec{v}}_k \in \mathbb{R}^{|\mathbf{X}|}$, whose entries are word relevances for k, i.e., $\phi_{\text{limsse}}(t,k,\mathbf{X}) = \hat{v}_{k,t}$. To optimize it, we experiment with three loss functions. The first, which we will refer to as $\text{limsse}^{\text{bb}}$, assumes that our DNN is a total black box that delivers only a classification:

$$\hat{\vec{v}}_k = \underset{\vec{v}_k}{\text{argmin}} \sum_n -\big[\log\big(\sigma(\vec{z}_n \cdot \vec{v}_k)\big)\mathbb{I}[f(\mathbf{Z}_n) = k] + \log\big(1 - \sigma(\vec{z}_n \cdot \vec{v}_k)\big)\mathbb{I}[f(\mathbf{Z}_n) \neq k]\big]$$

where $f(\mathbf{Z}_n) = \text{argmax}_{k'}\big(p(k'|\mathbf{Z}_n)\big)$. The black box approach is maximally general, but insensitive to the magnitude of evidence found in $\mathbf{Z}_n$. Hence, we also test magnitude-sensitive loss functions:

$$\hat{\vec{v}}_k = \underset{\vec{v}_k}{\text{argmin}} \sum_n \big(\vec{z}_n \cdot \vec{v}_k - o(k,\mathbf{Z}_n)\big)^2$$

where $o(k,\mathbf{Z}_n)$ is one of $s(k,\mathbf{Z}_n)$ or $p(k|\mathbf{Z}_n)$. We refer to these as $\text{limsse}_s^{\text{ms}}$ and $\text{limsse}_p^{\text{ms}}$.

4 Experiments

4.1 Hybrid document experiment

For the hybrid document experiment, we use the 20 newsgroups corpus (topic classification) (Lang, 1995) and reviews from the 10th yelp dataset challenge (binary sentiment analysis)[3]. We train five DNNs per corpus: a bidirectional GRU (Cho et al., 2014), a bidirectional LSTM (Hochreiter and Schmidhuber, 1997), a 1D CNN with global max pooling (Collobert et al., 2011), a bidirectional Quasi-GRU (QGRU), and a bidirectional Quasi-LSTM (QLSTM). The Quasi-RNNs are 1D CNNs with a feature-wise gated recursive pooling layer (Bradbury et al., 2017). Word embeddings are $\mathbb{R}^{300}$ and initialized with pre-trained GloVe embeddings (Pennington et al., 2014)[4]. The main layer has a hidden size of 150 (bidirectional architectures: 75 dimensions per direction). For the QRNNs and CNN, we use a kernel width of 5. In all five architectures, the resulting document representation is projected to 20 (resp. two) dimensions using a fully connected layer, followed by a softmax. See supplementary material for details on training and regularization.

After training, we sentence-tokenize the test sets, shuffle the sentences, concatenate ten sentences at a time and classify the resulting hybrid documents. Documents that are assigned a class that is not the gold label of at least one constituent word are discarded (yelp: $< 0.1\%$; 20 newsgroups: 14% - 20%). On the remaining documents, we use the explanation methods from §3 to find the maximally relevant word for each prediction. The random baseline samples the maximally relevant word from a uniform distribution.

For reference, we also evaluate on a **human judgment** benchmark (Mohseni and Ragan (2018), Table 2, C11-C15). It contains

[3] www.yelp.com/dataset_challenge

[4] http://nlp.stanford.edu/data/glove.840B.300d.zip

column	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23	C24	C25	C26	C27
	hybrid document experiment										man. groundtruth					morphosyntactic agreement experiment											
	yelp					20 newsgroups					20 newsgroups					$\mathrm{hit}_{\mathrm{target}}$ $f(\mathbf{X}) = y(\mathbf{X})$								$\mathrm{hit}_{\mathrm{feat}}$ $f(\mathbf{X}) \neq y(\mathbf{X})$			
ϕ	GRU	QGRU	LSTM	QLSTM	CNN	GRU	QGRU	LSTM	QLSTM	CNN	GRU	QGRU	LSTM	QLSTM	CNN	GRU	QGRU	LSTM	QLSTM	GRU	QGRU	LSTM	QLSTM	GRU	QGRU	LSTM	QLSTM
$\mathrm{grad}^{\mathrm{L2}}_{1s}$	.61	.68	.67	.70	.68	.45	.47	.25	.33	.79	.26	.31	.07	.18	.74	.48	.23	.63	.19	.52	.27	.73	.22	.09	.11	.19	.19
$\mathrm{grad}^{\mathrm{L2}}_{1p}$	.57	.67	.67	.70	.74	.40	.43	.26	.34	.70	.18	.35	.07	.13	.66	.48	.22	.63	.18	.53	.26	.73	.21	.09	.09	.18	.11
$\mathrm{grad}^{\mathrm{L2}}_{\int s}$	.71	.66	.69	.71	.70	.58	.32	.26	.21	.82	.23	.15	.11	.08	.76	.69	.67	.68	.51	.73	.70	.75	.55	.19	.22	.20	.20
$\mathrm{grad}^{\mathrm{L2}}_{\int p}$	.71	.70	.72	.71	.77	.56	.34	.30	.23	.81	.13	.08	.14	.01	.78	.68	.77	.50	.70	.74	.82	.54	.78	.19	.21	.19	.30
$\mathrm{grad}^{\mathrm{dot}}_{1s}$	.88	.85	.81	.77	.86	.79	.76	.59	.72	.89	.80	.70	.14	.47	.79	.81	.62	.73	.56	.85	.66	.81	.59	.42	.34	.46	.36
$\mathrm{grad}^{\mathrm{dot}}_{1p}$	.92	.88	.84	.79	.95	.78	.72	.59	.72	.81	.71	.59	.20	.44	.69	.79	.58	.74	.54	.83	.61	.81	.56	.41	.33	.46	.35
$\mathrm{grad}^{\mathrm{dot}}_{\int s}$	.84	.90	.85	.87	.87	.81	.68	.60	.68	.89	.82	.64	.21	.26	.80	.90	.87	.78	.84	.94	.92	.83	.89	.54	.51	.46	.52
$\mathrm{grad}^{\mathrm{dot}}_{\int p}$	.86	.89	.84	.89	**.96**	.80	.69	.62	.73	.89	.80	.53	.40	.54	.78	.87	.85	.68	.84	.93	.92	.74	.93	.53	.48	.42	.51
omit_1	.79	.82	.85	.87	.61	.78	.75	.54	.76	.82	.80	.48	.33	.48	.65	.81	.81	.79	.80	.86	.87	.86	.84	.43	.45	.44	.45
omit_3	.89	.80	.89	.88	.59	.79	.71	.72	.81	.76	.77	.37	.36	.49	.61	.74	.77	.73	.73	.82	.84	.82	.79	.41	.45	.42	.46
omit_7	.92	.88	.91	.91	.70	.79	.77	.77	.84	.84	.77	.49	.44	.55	.65	.76	.80	.66	.74	.85	.88	.78	.80	.40	.48	.43	.47
occ_1	.80	.71	.74	.84	.61	.78	.73	.60	.77	.82	.77	.49	.19	.10	.65	**.91**	.85	**.86**	.86	.94	.88	.89	.88	.50	.44	.46	.47
occ_3	.92	.61	.93	.85	.59	.78	.63	.74	.74	.76	.74	.37	.32	.35	.61	.74	.73	.71	.72	.78	.76	.76	.76	.43	.37	.41	.43
occ_7	.92	.77	.93	.90	.70	.78	.62	.74	.77	.84	.74	.35	.43	.39	.65	.64	.65	.63	.65	.73	.73	.72	.73	.36	.35	.39	.43
decomp	.79	.88	.92	.88	-	.75	.79	.77	.80	-	.54	.36	.72	.51	-	.84	.87	**.86**	.90	.90	.93	**.92**	**.96**	.52	.58	**.57**	**.63**
lrp	.92	.87	.91	.84	.86	.82	.83	.79	.85	.89	**.85**	.72	.74	.81	.79	.90	**.90**	**.86**	**.91**	**.95**	**.95**	.91	.95	.58	**.60**	.52	**.63**
deeplift	.91	.89	**.94**	.85	.87	.82	.83	.78	.84	.89	.84	.72	.70	.81	.80	**.91**	**.90**	.85	**.91**	**.95**	**.95**	.90	.95	**.59**	.59	.52	**.63**
$\mathrm{limsse}^{\mathrm{bb}}$	.81	.82	.83	.84	.78	.78	.81	.78	.80	.84	.52	.53	.53	.54	.57	.43	.41	.44	.42	.54	.51	.56	.52	.39	.43	.42	.41
$\mathrm{limsse}^{\mathrm{ms}}_{s}$	**.94**	**.94**	.93	**.93**	.91	**.85**	**.87**	**.83**	**.86**	.89	**.85**	**.84**	**.76**	**.84**	**.82**	.62	.62	.67	.63	.75	.74	.82	.75	.52	.53	.55	.53
$\mathrm{limsse}^{\mathrm{ms}}_{p}$	.87	.88	.85	.86	.94	**.85**	.86	**.83**	**.86**	**.90**	.81	.80	.74	.76	.76	.62	.62	.67	.63	.75	.74	.82	.75	.51	.53	.55	.53
random	.69	.67	.70	.69	.66	.20	.19	.22	.22	.21	.09	.09	.06	.06	.08	.27	.27	.27	.27	.33	.33	.33	.33	.12	.13	.12	.12
last	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	.66	.67	.66	.67	.76	.77	.76	.77	.21	.27	.25	.26
N	$7551 \leq N \leq 7554$					$3022 \leq N \leq 3230$					$137 \leq N \leq 150$					$N \approx 1400000$								$N \approx 20000$			

Table 2: Pointing game accuracies in hybrid document experiment (left), on manually annotated benchmark (middle) and in morphosyntactic agreement experiment (right). $\mathrm{hit}_{\mathrm{target}}$ (resp. $\mathrm{hit}_{\mathrm{feat}}$): maximal relevance on subject (resp. on noun with the predicted number feature). Bold: top explanation method. Underlined: within 5 points of top explanation method.

188 documents from the 20 newsgroups test set (classes sci.med and sci.electronics), with one manually created list of relevant words per document. We discard documents that are incorrectly classified (20% - 27%) and define: $\mathrm{hit}(\phi, \mathbf{X}) = \mathbb{I}[\mathrm{rmax}(\mathbf{X}, \phi) \in \mathrm{gt}(\mathbf{X})]$, where $\mathrm{gt}(\mathbf{X})$ is the manual ground truth.

4.2 Morphosyntactic agreement experiment

For the morphosyntactic agreement experiment, we use automatically annotated English Wikipedia sentences by Linzen et al. (2016)[5]. For our purpose, a sample consists of: all words preceding the verb: $\mathbf{X} = [x_1 \cdots x_T]$; part-of-speech (POS) tags: $\mathrm{pos}(\mathbf{X}, t) \in \{\text{VBZ, VBP, NN, NNS, } \ldots\}$; and the position of the subject: $\mathrm{target}(\mathbf{X}) \in [1, T]$. The number feature is derived from the POS:

$$\mathrm{feat}(\mathbf{X}, t) = \begin{cases} \text{Sg} & \text{if } \mathrm{pos}(\mathbf{X}, t) \in \{\text{VBZ, NN}\} \\ \text{Pl} & \text{if } \mathrm{pos}(\mathbf{X}, t) \in \{\text{VBP, NNS}\} \\ \text{n/a} & \text{otherwise} \end{cases}$$

The gold label of a sentence is the number of its verb, i.e., $y(\mathbf{X}) = \mathrm{feat}(\mathbf{X}, T + 1)$.

[5] www.tallinzen.net/media/rnn_agreement/agr_50_mostcommon_10K.tsv.gz

As task methods, we replicate Linzen et al. (2016)'s unidirectional LSTM ($\mathbb{R}^{50}$ randomly initialized word embeddings, hidden size 50). We also train unidirectional GRU, QGRU and QLSTM architectures with the same dimensionality. We use the explanation methods from §3 to find the most relevant word for predictions on the test set. As described in §2.2, explanation methods are awarded a $\mathrm{hit}_{\mathrm{target}}$ (resp. $\mathrm{hit}_{\mathrm{feat}}$) point if this word is the subject (resp. a noun with the predicted number feature). For reference, we use a random baseline as well as a baseline that assumes that the most relevant word directly precedes the verb.

5 Discussion

5.1 Explanation methods

Our experiments suggest that explanation methods for neural NLP differ in quality.

As in previous work (see §6), **gradient L2 norm** ($\mathrm{grad}^{\mathrm{L2}}$) performs poorly, especially on RNNs. We assume that this is due to its inability to distinguish relevances for and against k.

Gradient embedding dot product ($\mathrm{grad}^{\mathrm{dot}}$) is competitive on CNN (Table 2, $\mathrm{grad}^{\mathrm{dot}}_{1p}$ C05, $\mathrm{grad}^{\mathrm{dot}}_{1s}$ C10, C15), presumably because relu is linear on positive inputs, so gradients are exact in-

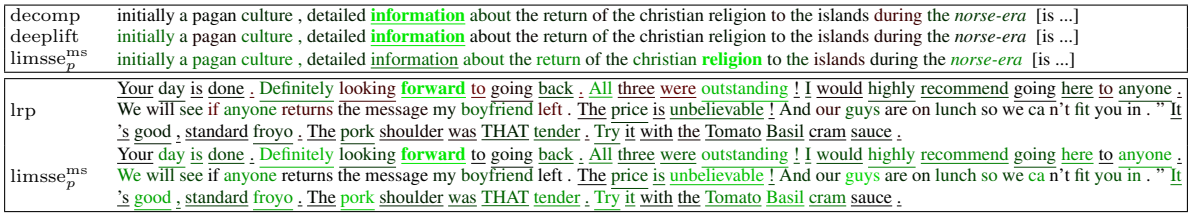

Figure 3: **Top:** verb context classified singular. Task method: LSTM. **Bottom:** hybrid yelp review, classified positive. Task method: QLSTM.

stead of approximate. grad^{dot} also has decent performance for GRU ($\text{grad}_{1p}^{\text{dot}}$ C01, $\text{grad}_{\int s}^{\text{dot}}$ C{06, 11, 16, 20, 24}), perhaps because GRU hidden activations are always in [-1,1], where $\tanh$ and σ are approximately linear.

Integrated gradient ($\text{grad}_{\int}$) mostly outperforms simple gradient (grad_1), though not consistently (C01, C07). Contrary to expectation, integration did not help much with the failure of the gradient method on LSTM on 20 newsgroups ($\text{grad}_1^{\text{dot}}$ vs. $\text{grad}_{\int}^{\text{dot}}$ in C08, C13), which we had assumed to be due to saturation of $\tanh$ on large absolute activations in $\vec{c}$. Smaller intervals may be needed to approximate the integration, however, this means additional computational cost.

The gradient of $s(k, \mathbf{X})$ performs better or similar to the gradient of $p(k|\mathbf{X})$. The main exception is yelp ($\text{grad}_{1s}^{\text{dot}}$ vs. $\text{grad}_{1p}^{\text{dot}}$, C01-C05). This is probably due to conflation by $p(k|\mathbf{X})$ of evidence for k (numerator in Eq 3) and against competitor classes (denominator). In a two-class scenario, there is little incentive to keep classes separate, leading to information flow through the denominator. In future work, we will replace the two-way softmax with a one-way sigmoid such that $\phi(t, 0, \mathbf{X}) := -\phi(t, 1, \mathbf{X})$.

LRP and **DeepLIFT** are the most consistent explanation methods across evaluation paradigms and task methods. (The comparatively low pointing game accuracies on the yelp QRNNs and CNN (C02, C04, C05) are probably due to the fact that they explain $s(k, .)$ in a two-way softmax, see above.) On CNN (C05, C10, C15), LRP and $\text{grad}_{1s}^{\text{dot}}$ perform almost identically, suggesting that they are indeed quasi-equivalent on this architecture (see §3.2). On (Q)RNNs, modified LRP and DeepLIFT appear to be superior to the gradient method (lrp vs. $\text{grad}_{1s}^{\text{dot}}$, deeplift vs. $\text{grad}_{\int s}^{\text{dot}}$, C01-C04, C06-C09, C11-C14, C16-C27).

Decomposition performs well on LSTM, especially in the morphosyntactic agreement experiment, but it is inconsistent on other architectures. Gated RNNs have a long-term additive and a multiplicative pathway, and the decomposition method only detects information traveling via the additive one. Miao et al. (2016) show qualitatively that GRUs often reorganize long-term memory abruptly, which might explain the difference between LSTM and GRU. QRNNs only have additive recurrent connections; however, given that $\vec{c}_t$ (resp. $\vec{h}_t$) is calculated by convolution over several time steps, decomposition relevance can be incorrectly attributed inside that window. This likely is the reason for the stark difference between the performance of decomposition on QRNNs in the hybrid document experiment and on the manually labeled data (C07, C09 vs. C12, C14). Overall, we do not recommend the decomposition method, because it fails to take into account all routes by which information can be propagated.

Omission and occlusion produce inconsistent results in the hybrid document experiment. Shrikumar et al. (2017) show that perturbation methods can lack sensitivity when there are more relevant inputs than the "perturbation window" covers. In the morphosyntactic agreement experiment, omission is not competitive; we assume that this is because it interferes too much with syntactic structure. occ_1 does better (esp. C16-C19), possibly because an all-zero "placeholder" is less disruptive than word removal. But despite some high scores, it is less consistent than other explanation methods.

Magnitude-sensitive **LIMSSE** ($\text{limsse}^{\text{ms}}$) consistently outperforms black-box LIMSSE ($\text{limsse}^{\text{bb}}$), which suggests that numerical outputs should be used for approximation where possible. In the hybrid document experiment, magnitude-sensitive LIMSSE outperforms the other explanation methods (exceptions: C03, C05). However, it fails in the morphosyntactic agreement experiment (C16-C27). In fact, we expect LIMSSE to be unsuited for *large context*

problems, as it cannot discover dependencies whose range is bigger than a given text sample. In Fig 3 (top), $\text{limsse}_p^{\text{ms}}$ highlights *any* singular noun without taking into account how that noun fits into the overall syntactic structure.

5.2 Evaluation paradigms

The assumptions made by our automatic evaluation paradigms have exceptions: (i) the correlation between fragment of origin and relevance does not always hold (e.g., a positive review may contain negative fragments, and will almost certainly contain neutral fragments); (ii) in morphological prediction, we cannot always expect the subject to be the only predictor for number. In Fig 2 (bottom) for example, "few" is a reasonable clue for plural despite not being a noun. This imperfect ground truth means that absolute pointing game accuracies should be taken with a grain of salt; but we argue that this does not invalidate them for comparisons.

We also point out that there are characteristics of explanations that may be desirable but are not reflected by the pointing game. Consider Fig 3 (bottom). Both explanations get hit points, but the lrp explanation appears "cleaner" than $\text{limsse}_p^{\text{ms}}$, with relevance concentrated on fewer tokens.

6 Related work

6.1 Explanation methods

Explanation methods can be divided into local and global methods (Doshi-Velez and Kim, 2017). Global methods infer general statements about what a DNN has learned, e.g., by clustering documents (Aubakirova and Bansal, 2016) or n-grams (Kádár et al., 2017) according to the neurons that they activate. Li et al. (2016a) compare embeddings of specific words with reference points to measure how drastically they were changed during training. In computer vision, Simonyan et al. (2014) optimize the input space to maximize the activation of a specific neuron. Global explanation methods are of limited value for explaining a specific prediction as they represent average behavior. Therefore, we focus on local methods.

Local explanation methods explain a decision taken for one specific input at a time. We have attempted to include all important local methods for NLP in our experiments (see §3). We do not address self-explanatory models (e.g., attention (Bahdanau et al., 2015) or rationale models (Lei et al., 2016)), as these are very specific architectures that may not be not applicable to all tasks.

6.2 Explanation evaluation

According to Doshi-Velez and Kim (2017)'s taxonomy of explanation evaluation paradigms, *application-grounded* paradigms test how well an explanation method helps real users solve real tasks (e.g., doctors judge automatic diagnoses); *human-grounded* paradigms rely on proxy tasks (e.g., humans rank task methods based on explanations); *functionally-grounded* paradigms work without human input, like our approach.

Arras et al. (2016) (cf. Samek et al. (2016)) propose a functionally-grounded explanation evaluation paradigm for NLP where words in a correctly (resp. incorrectly) classified document are deleted in descending (resp. ascending) order of relevance. They assume that the fewer words must be deleted to reduce (resp. increase) accuracy, the better the explanations. According to this metric, LRP (§3.2) outperforms grad^{L2} on CNNs (Arras et al., 2016) and LSTMs (Arras et al., 2017b) on 20 newsgroups. Ancona et al. (2017) perform the same experiment with a binary sentiment analysis LSTM. Their graph shows occ_1, $\text{grad}_1^{\text{dot}}$ and $\text{grad}_{\int}^{\text{dot}}$ tied in first place, while LRP, DeepLIFT and the gradient L1 norm lag behind. Note that their treatment of LSTM gates in LRP / DeepLIFT differs from our implementation.

An issue with the word deletion paradigm is that it uses syntactically broken inputs, which may introduce artefacts (Sundararajan et al., 2017). In our hybrid document paradigm, inputs are syntactically intact (though semantically incoherent at the document level); the morphosyntactic agreement paradigm uses unmodified inputs.

Another class of functionally-grounded evaluation paradigms interprets the performance of a secondary task method, on inputs that are derived from (or altered by) an explanation method, as a proxy for the quality of that explanation method. Murdoch and Szlam (2017) build a rule-based classifier from the most relevant phrases in a corpus (task method: LSTM). The classifier based on decomp (§3.4) outperforms the gradient-based classifier, which is in line with our results. Arras et al. (2017a) build document representations by summing over word embeddings weighted by relevance scores (task method: CNN). They show that K-nearest neighbor performs better on doc-

ument representations derived with LRP than on those derived with grad^{L2}, which also matches our results. Denil et al. (2015) condense documents by extracting top-K relevant sentences, and let the original task method (CNN) classify them. The accuracy loss, relative to uncondensed documents, is smaller for grad^{dot} than for heuristic baselines.

In the domain of human-based evaluation paradigms, Ribeiro et al. (2016) compare different variants of LIME (§3.6) by how well they help non-experts clean a corpus from words that lead to overfitting. Selvaraju et al. (2017) assess how well explanation methods help non-experts identify the more accurate out of two object recognition CNNs. These experiments come closer to real use cases than functionally-grounded paradigms; however, they are less scalable.

7 Summary

We conducted the first comprehensive evaluation of explanation methods for NLP, an important undertaking because there is a need for understanding the behavior of DNNs.

To conduct this study, we introduced evaluation paradigms for explanation methods for two classes of NLP tasks, small context tasks (e.g., topic classification) and large context tasks (e.g., morphological prediction). Neither paradigm requires manual annotations. We also introduced LIMSSE, a substring-based explanation method inspired by LIME and designed for NLP.

Based on our experimental results, we recommend LRP, DeepLIFT and LIMSSE for small context tasks and LRP and DeepLIFT for large context tasks, on all five DNN architectures that we tested. On CNNs and possibly GRUs, the (integrated) gradient embedding dot product is a good alternative to DeepLIFT and LRP.

8 Code

Our implementation of LIMSSE, the gradient, perturbation and decomposition methods can be found in our branch of the `keras` package: `www.github.com/NPoe/keras`. To re-run our experiments, see scripts in `www.github.com/NPoe/neural-nlp-explanation-experiment`. Our LRP implementation (same repository) is adapted from Arras et al. (2017b)[6].

[6] `https://github.com/ArrasL/LRP_for_LSTM`

9 Acknowledgement

We gratefully acknowledge funding for this work by the European Research Council (ERC #740516).

References

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2017. A unified view of gradient-based attribution methods for deep neural networks. In *Conference on Neural Information Processing System*, Long Beach, USA.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2018. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*, Vancouver, Canada.

Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2016. Explaining predictions of non-linear classifiers in NLP. In *First Workshop on Representation Learning for NLP*, pages 1–7, Berlin, Germany.

Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017a. What is relevant in a text document?: An interpretable machine learning approach. *PloS one*, 12(8):e0181142.

Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017b. Explaining recurrent neural network predictions in sentiment analysis. In *Eighth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 159–168, Copenhagen, Denmark.

Malika Aubakirova and Mohit Bansal. 2016. Interpreting neural networks to improve politeness comprehension. In *Empirical Methods in Natural Language Processing*, page 2035–2041, Austin, USA.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, San Diego, USA.

Trapit Bansal, David Belanger, and Andrew McCallum. 2016. Ask the GRU: Multi-task learning for deep text recommendations. In *ACM Conference on Recommender Systems*, pages 107–114, Boston, USA.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media.

James Bradbury, Stephen Merity, Caiming Xiong, and Richard Socher. 2017. Quasi-recurrent neural networks. In *International Conference on Learning Representations*, Toulon, France.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Misha Denil, Alban Demiraj, and Nando de Freitas. 2015. Extraction of salient sentences from labelled documents. In *International Conference on Learning Representations*, San Diego, USA.

Finale Doshi-Velez and Been Kim. 2017. A roadmap for a rigorous science of interpretability. *CoRR*, abs/1702.08608.

Bryce Goodman and Seth Flaxman. 2016. European union regulations on algorithmic decision-making and a "right to explanation". In *ICML Workshop on Human Interpretability in Machine Learning*, pages 26–30, New York, USA.

Yotam Hechtlinger. 2016. Interpretation of prediction models using the input gradient. In *Conference on Neural Information Processing Systems*, Barcelona, Spain.

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

Akos Kádár, Grzegorz Chrupała, and Afra Alishahi. 2017. Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*, 43(4):761–780.

Pieter-Jan Kindermans, Kristof Schütt, Klaus-Robert Müller, and Sven Dähne. 2016. Investigating the influence of noise and distractors on the interpretation of neural networks. In *Conference on Neural Information Processing Systems*, Barcelona, Spain.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, San Diego, USA.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *International Conference on Machine Learning*, pages 331–339, Tahoe City, USA.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Empirical Methods in Natural Language Processing*, pages 107–117, Austin, USA.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016a. Visualizing and understanding neural models in NLP. In *NAACL-HLT*, pages 681–691, San Diego, USA.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016b. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Yajie Miao, Jinyu Li, Yongqiang Wang, Shi-Xiong Zhang, and Yifan Gong. 2016. Simplifying long short-term memory acoustic models for fast training and decoding. In *International Conference on Acoustics, Speech and Signal Processing*, pages 2284–2288.

Sina Mohseni and Eric D Ragan. 2018. A human-grounded evaluation benchmark for local explanations of machine learning. *arXiv preprint arXiv:1801.05075*.

W James Murdoch and Arthur Szlam. 2017. Automatic rule extraction from long short term memory networks. In *International Conference on Learning Representations*, Toulon, France.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should I trust you?: Explaining the predictions of any classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, San Francisco, California.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. 2016. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 618–626, Honolulu, Hawaii.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153, Sydney, Australia.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations*, Banff, Canada.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, Sydney, Australia.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833, Zürich, Switzerland.

Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. 2016. Top-down neural attention by excitation backprop. In *European Conference on Computer Vision*, pages 543–559, Amsterdam, Netherlands.

Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. 2017. Visualizing deep neural network decisions: Prediction difference analysis. In *International Conference on Learning Representations*, Toulon, France.

10 Supplementary material

11 Corpora and data preprocessing

The 20 newsgroups corpus (Lang, 1995) was downloaded using the Python `sklearn` package (Pedregosa et al., 2011), removing all headers, footers and quotes. The corpus contains 18,846 posts and comes with a training and test set. We randomly split the latter into a heldout and a test set.

For sentiment analysis we use the Pennsylvania subset of the 10th yelp dataset challenge[7]. It contains 206,338 reviews with 1 to 5 star ratings. 1 or 2 stars are mapped to "negative", 4 or 5 stars to "positive", 3 star reviews are discarded. We randomly split the data into training, heldout and test sets (90%/5%/5%). On both corpora, we use NLTK (Bird et al., 2009) for word and sentence tokenization. Words with a frequency rank above 50000 are mapped to *oov*. To create hybrid documents, we sentence-tokenize the test sets, shuffle, and then concatenate ten sentences at a time.

The manually annotated 20 newsgroups documents were obtained from Mohseni and Ragan (2018)[8]. The relevance ground truth consists of one list of lowercased word types per document. There are a number of mismatches between the ground truth and the documents (e.g., one list contains *rays* but its document only contains *x-rays*). This made some reverse engineering necessary: Given $\mathbf{X}$ and its list, we add t to $\mathrm{gt}(\mathbf{X})$ if lower-cased x_t is a prefix or suffix of at least one word type in the list.

For the morphosyntactic agreement experiment, we use Linzen et al. (2016)'s corpus of 1,577,211 English Wikipedia sentences with automatic morphosyntactic annotation[9]. We replicate the original dataset sizes (9% train, 1% heldout, 90% test). Like in the original corpus, words with a frequency rank above 10,000 are replaced by their part-of-speech tag.

12 Neural networks

Every neural network used in our paper is made up of a word embedding matrix, followed by a core layer, followed by a fully-connected layer with softmax activation.

In the hybrid document experiment, the $|V| \times 300$ embedding matrix is initialized with GloVe embeddings (Pennington et al., 2014)[10], which are fine-tuned during training. The core layer is a bidirectional Gated Recurrent Unit (GRU, Cho et al. (2014)), bidirectional Long-Short Term Memory Network (LSTM, Hochreiter and Schmidhuber (1997)), bidirectional Quasi-GRU or Quasi-LSTM (Bradbury et al., 2017), or a 1D Convolutional Neural Network (CNN) with global max pooling (Collobert et al., 2011). In all cases, the core layer has a hidden size of 150 (bidirectional architectures: 75 per direction), for QRNNs and CNN, we use a kernel width of 5. For regularization, we use 50% dropout between layers and on hidden-to-hidden connections (GRU/LSTM only).

We minimize categorical crossentropy using Adam (Kingma and Ba, 2015), with learning rate 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and batch size 8. Heldout accuracy is monitored; after two stagnant epochs, the learning rate is halved, and after 5 (yelp), resp. 25 (20 newsgroups), stagnant epochs, training is stopped and the model from the best epoch is stored. Final test set accuracies are .964/.954/.965/.959/.957 on yelp and .727/.716/.730/.735/.705 on 20 newsgroups (GRU/QGRU/LSTM/QLSTM/CNN).

In the morphosyntactic agreement experiment, the $|V| \times 50$ embedding matrix is randomly initialized. All (Q)RNNs are unidirectional and have a hidden size of 50. QRNN kernel width is 5. The core layer is followed by a fully connected 50×2 layer with softmax activation. We minimize categorical crossentropy using Adam (see above), with early stopping after 20 epochs based on heldout accuracy, and a batch size of 16. Final test set accuracies are .991/.985/.990/.986 (GRU/QGRU/LSTM/QLSTM). Contrary to Linzen et al. (2016), we do not train an ensemble.

12.1 GRU

$$
\begin{aligned}
\vec{h}_0 &= 0 \\
\vec{z}_t &= \sigma(\mathbf{V_z}\vec{e}_t + \mathbf{U_z}\vec{h}_{t-1} + \vec{b}_z) \\
\vec{r}_t &= \sigma(\mathbf{V_r}\vec{e}_t + \mathbf{U_r}\vec{h}_{t-1} + \vec{b}_r) \\
\vec{g'}_t &= \mathbf{V}\vec{e}_t + \mathbf{U}(\vec{r}_t \odot \vec{h}_{t-1}) + \vec{b} \\
\vec{g}_t &= \tanh(\vec{g'}_t) \\
\vec{h}_t &= \vec{z}_t \odot \vec{h}_{t-1} + (\vec{1} - \vec{z}_t) \odot \vec{g}_t
\end{aligned}
$$

12.2 QGRU

$$
\begin{aligned}
\mathbf{Z} &= \sigma(\mathbf{V_z} \star [0 \ldots \vec{e}_1 \ldots \vec{e}_T] + \vec{b}_z) \\
\mathbf{G'} &= \mathbf{V} \star [0 \ldots \vec{e}_1 \ldots \vec{e}_T] + \vec{b} \\
\mathbf{G} &= \tanh(\mathbf{G'}) \\
\vec{h}_0 &= 0 \\
\vec{h}_t &= \vec{z}_t \odot \vec{h}_{t-1} + (1 - \vec{z}_t) \odot \vec{g}_t
\end{aligned}
$$

12.3 LSTM

$$
\begin{aligned}
\vec{c}_0 &= \vec{h}_0 = 0 \\
\vec{i}_t &= \sigma(\mathbf{V_i}\vec{e}_t + \mathbf{U_i}\vec{h}_{t-1} + \vec{b}_i) \\
\vec{f}_t &= \sigma(\mathbf{V_f}\vec{e}_t + \mathbf{U_f}\vec{h}_{t-1} + \vec{b}_f) \\
\vec{o}_t &= \sigma(\mathbf{V_o}\vec{e}_t + \mathbf{U_o}\vec{h}_{t-1} + \vec{b}_o) \\
\vec{g'}_t &= \mathbf{V}\vec{e}_t + \mathbf{U}\vec{h}_{t-1} + \vec{b} \\
\vec{g}_t &= \tanh(\vec{g'}_t) \\
\vec{c}_t &= \vec{f}_t \odot \vec{c}_{t-1} + \vec{i}_t \odot \vec{g}_t \\
\vec{h}_t &= \vec{o}_t \odot \tanh(\vec{c}_t)
\end{aligned}
$$

12.4 QLSTM

$$
\begin{aligned}
\mathbf{I} &= \sigma(\mathbf{V_i} \star [0 \ldots \vec{e}_1 \ldots \vec{e}_T] + \vec{b}_i) \\
\mathbf{F} &= \sigma(\mathbf{V_f} \star [0 \ldots \vec{e}_1 \ldots \vec{e}_T] + \vec{b}_f) \\
\mathbf{O} &= \sigma(\mathbf{V_o} \star [0 \ldots \vec{e}_1 \ldots \vec{e}_T] + \vec{b}_o) \\
\mathbf{G'} &= \mathbf{V} \star [0 \ldots \vec{e}_1 \ldots \vec{e}_T] + \vec{b} \\
\mathbf{G} &= \tanh(\mathbf{G'}) \\
\vec{h}_0 &= \vec{c}_0 = 0 \\
\vec{c}_t &= \vec{f}_t \odot \vec{c}_{t-1} + \vec{i}_t \odot \vec{g}_t \\
\vec{h}_t &= \vec{o}_t \odot \tanh(\vec{c}_t)
\end{aligned}
$$

[7] www.yelp.com/dataset_challenge

[8] http://github.com/SinaMohseni/ML-Interpretability-Evaluation-Benchmark

[9] www.tallinzen.net/media/rnn_agreement/agr_50_mostcommon_10K.tsv.gz

[10] http://nlp.stanford.edu/data/glove.840B.300d.zip

12.5 CNN

$$\begin{aligned}\mathbf{G}' &= \mathbf{V} \star [0 \dots \vec{e}_1 \dots \vec{e}_T \dots 0] + \vec{b} \\ \mathbf{G} &= \text{relu}(\mathbf{G}') \\ h_d &= \max_t(g_{t,d})\end{aligned}$$

13 RGB coding in examples

$$\begin{aligned}\phi'(t,k,\mathbf{X}) &= \frac{\phi(t,k,\mathbf{X})}{\max_{t'}(1.1|\phi(t',k,\mathbf{X})|)} \\ R(t,k,\mathbf{X}) &= \phi'(t,k,\mathbf{X})\mathbb{I}[\phi(t,k,\mathbf{X}) < 0] \\ G(t,k,\mathbf{X}) &= \phi'(t,k,\mathbf{X})\mathbb{I}[\phi(t,k,\mathbf{X}) > 0] \\ B(t,k,\mathbf{X}) &= 0\end{aligned}$$

14 Epsilon LRP and DeepLIFT

In the following, we assume that the hidden layer relevance vector $R(\vec{h})$ (resp. $R(\vec{h}_T)$) has been backpropagated by the upstream fully connected layer using equations from Sections 3.2 and 3.3 (main paper). DeepLIFT can be derived by replacing h, g, g', e, c with $h - \bar{h}$, $g - \bar{g}$, $g' - \bar{g}'$, $e - \bar{e}$, $c - \bar{c}$. F is CNN / QRNN kernel width.

14.1 GRU

$$\begin{aligned}R(g_{t,d}) &= R(h_{t,d})\frac{g_{t,d} \cdot (1 - z_{t,d})}{h_{t,d} + \text{esign}(h_{t,d})} \\ R(e_{t,d}) &= \sum_{j=1}^{\dim(\vec{g}_t)} R(g_{t,j})\frac{e_{t,d} \cdot v_{d,j}}{g'_{t,j} + \text{esign}(g'_{t,j})} \\ R(h_{t-1,d}) &= R(h_{t,d})\frac{h_{t-1,d} \cdot z_{t,d}}{h_{t,d} + \text{esign}(h_{t,d})} \\ &\quad + \sum_{j=1}^{\dim(\vec{g}_t)} R(g_{t,j})\frac{h_{t-1,d} \cdot r_{t,d} \cdot u_{d,j}}{g'_{t,j} + \text{esign}(g'_{t,j})}\end{aligned}$$

14.2 QGRU

$$\begin{aligned}R(g_{t,d}) &= R(h_{t,d})\frac{g_{t,d} \cdot (1 - z_{t,d})}{h_{t,d} + \text{esign}(h_{t,d})} \\ R(h_{t-1,d}) &= R(h_{t,d})\frac{h_{t-1,d} \cdot z_{t,d}}{h_{t,d} + \text{esign}(h_{t,d})} \\ R(e_{t,d}) &= \sum_{j=1}^{\dim(\vec{g}_t)} \sum_{k=0}^{F-1} R(g_{t+k,j})\frac{e_{t,d} \cdot v_{k,d,j}}{g'_{t+k,j} + \text{esign}(g'_{t+k,j})}\end{aligned}$$

14.3 LSTM

$$\begin{aligned}R(c_{T+1,d}) &= 0 \\ R(c_{t,d}) &= R(h_{t,d})\frac{\tanh(c_{t,d}) \cdot o_{t,d}}{h_{t,d} + \text{esign}(h_{t,d})} \\ &\quad + R(c_{t+1,d})\frac{c_{t,d} \cdot f_{t+1,d}}{c_{t+1,d} + \text{esign}(c_{t+1,d})} \\ R(g_{t,d}) &= R(c_{t,d})\frac{g_{t,d} \cdot i_{t,d}}{c_{t,d} + \text{esign}(c_{t,d})} \\ R(e_{t,d}) &= \sum_{j=1}^{\dim(\vec{g}_t)} R(g_{t,j})\frac{e_{t,d} \cdot v_{d,j}}{g'_{t,j} + \text{esign}(g'_{t,j})} \\ R(h_{t-1,d}) &= \sum_{j=1}^{\dim(\vec{g}_t)} R(g_{t,j})\frac{h_{t-1,d} \cdot u_{d,j}}{g'_{t,j} + \text{esign}(g'_{t,j})}\end{aligned}$$

14.4 QLSTM

$$\begin{aligned}R(c_{T+1,d}) &= 0 \\ R(c_{t,d}) &= R(h_{t,d})\frac{\tanh(c_{t,d}) \cdot o_{t,d}}{h_{t,d} + \text{esign}(h_{t,d})} \\ &\quad + R(c_{t+1,d})\frac{c_{t,d} \cdot f_{t+1,d}}{c_{t+1,d} + \text{esign}(c_{t+1,d})} \\ R(g_{t,d}) &= R(c_{t,d})\frac{g_{t,d} \cdot i_{t,d}}{c_{t,d} + \text{esign}(c_{t,d})} \\ R(e_{t,d}) &= \sum_{j=1}^{\dim(\vec{g}_t)} \sum_{k=0}^{F-1} R(g_{t+k,j})\frac{e_{t,d} \cdot v_{k,d,j}}{g'_{t+k,j} + \text{esign}(g'_{t+k,j})}\end{aligned}$$

14.5 CNN

$$\begin{aligned}F' &= \frac{F-1}{2} \\ R(g_{t,d}) &= R(h_d) \cdot \mathbb{I}[\text{argmax}_{t'}(g_{t',d}) = t] \\ R(e_{t,d}) &= \sum_{j=1}^{\dim(\vec{g})} \sum_{k=-F'}^{F'} R(g_{t+k,j})\frac{e_{t,d} \cdot v_{k,d,j}}{g'_{t+k,j} + \text{esign}(g'_{t+k,j})}\end{aligned}$$

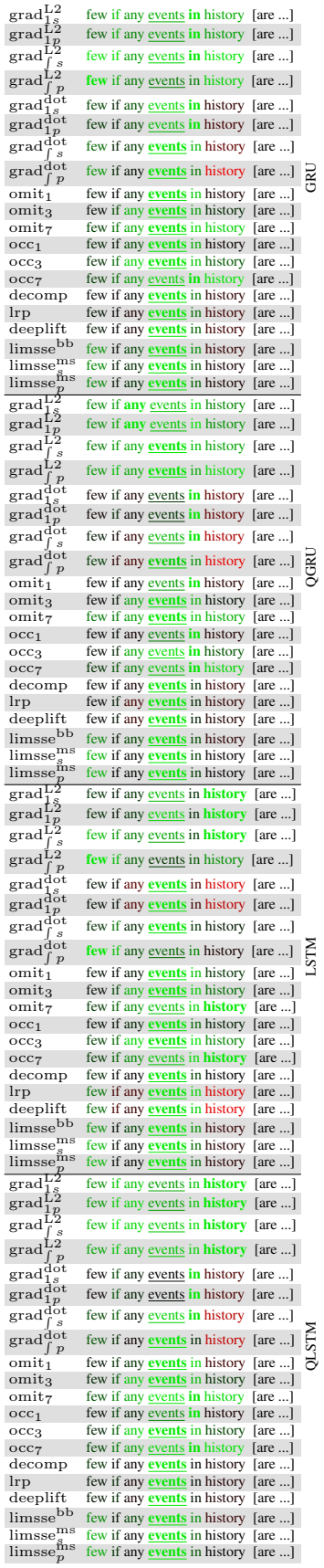

Figure 4: Verb context classified plural. Green (resp. red): evidence for (resp. against) the prediction. Underlined: subject. Bold: rmax position.

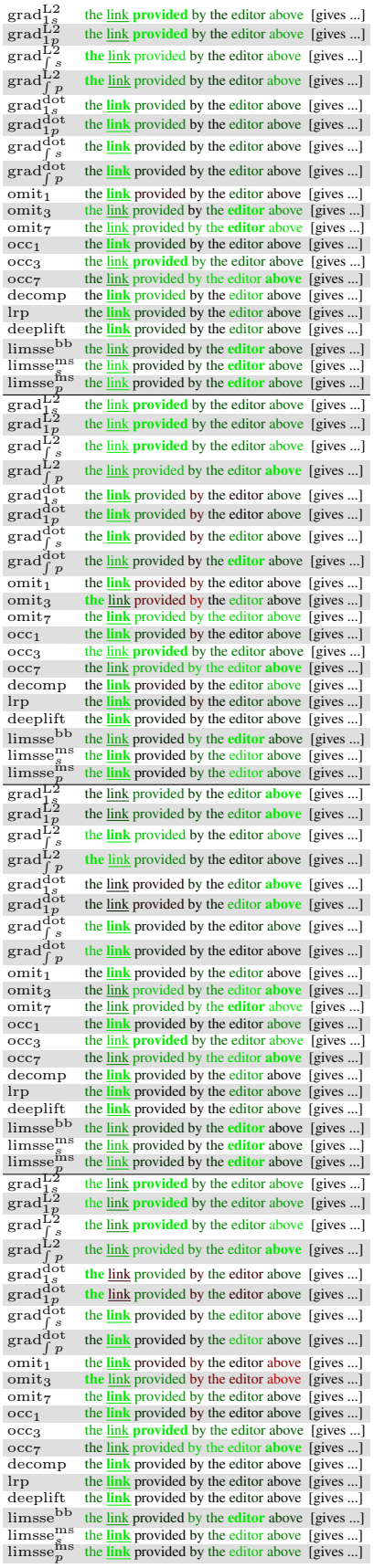

Figure 5: Verb context classified singular. Green (resp. red): evidence for (resp. against) the prediction. Underlined: subject. Bold: rmax position.

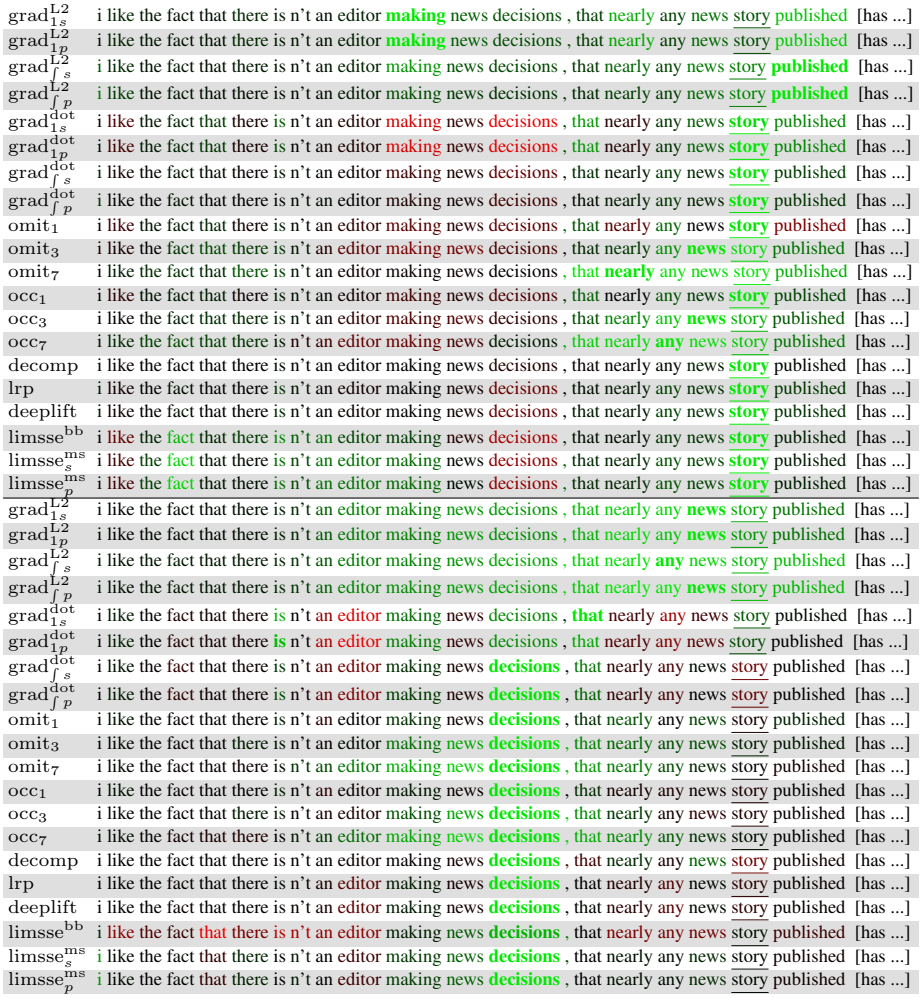

Figure 6: Verb context classified singular by GRU and plural by QLSTM. Green (resp. red): evidence for (resp. against) the prediction. Underlined: subject. Bold: rmax position.

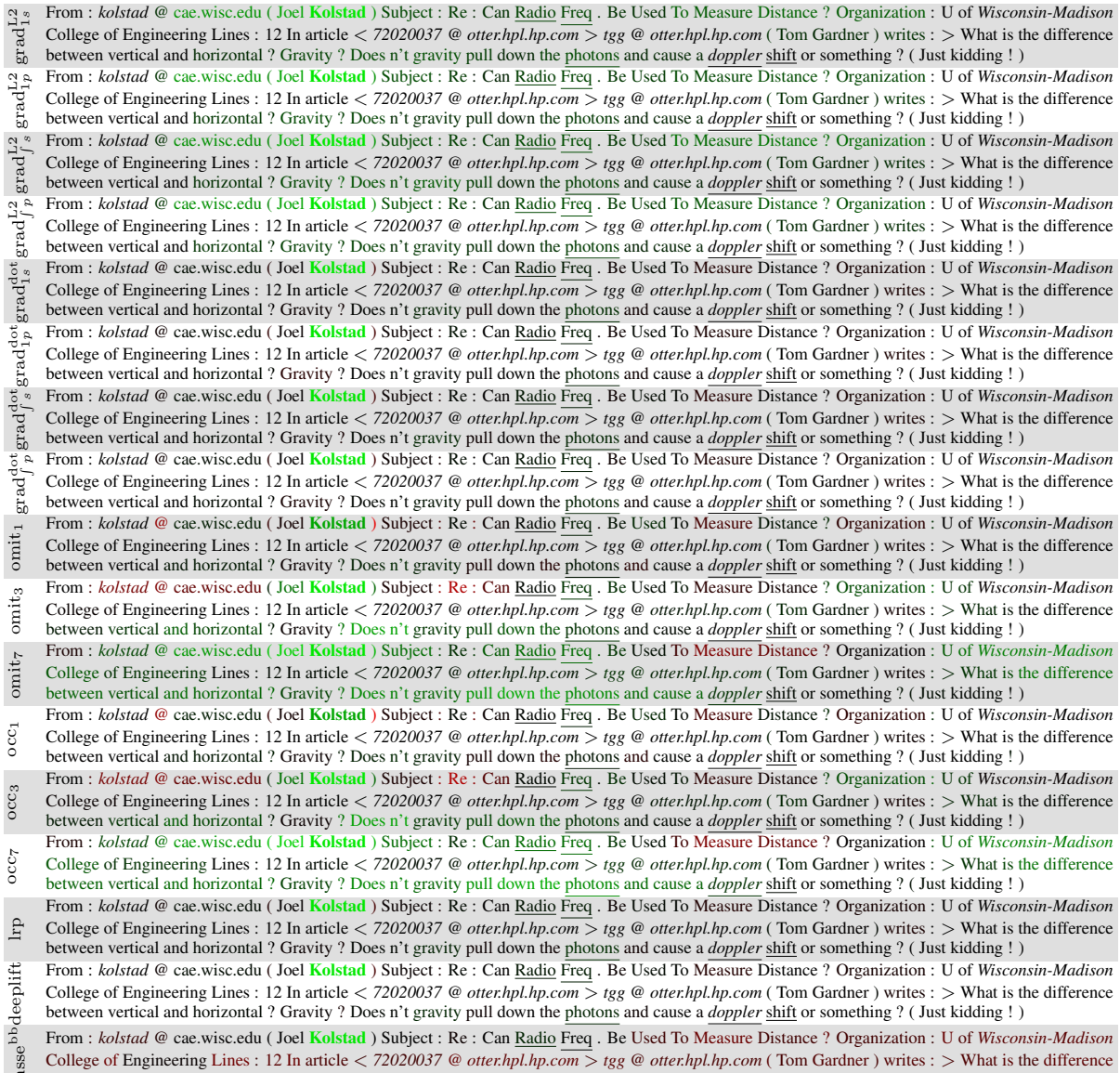

Figure 7: sci.electronics post (not hybrid). Underlined: Manual relevance ground truth. Green (resp. red): evidence for (resp. against) sci.electronics. Task method: CNN. Italics: OOV. Bold: rmax position.

$\text{grad}^{\text{L2}}_{1s}$	From : *chorley @ vms.ocom.okstate.edu* Subject : CS " gas " and **allergic** *response- Ques* . Lines : 6 Nntp-Posting-Host : *vms.ocom.okstate.edu* Organization : OSU College of Osteopathic Medicine This question derives from the Waco incident : Could CS (" gas ") particles create an allergic response which would result in *laryngospasm* and *asphyxiation* ? - especially in children . *DNC* in Ok. *OSU-COM* will *disavow* my opinion , and my existence , if necessary .
$\text{grad}^{\text{L2}}_{1p}$	From : *chorley @ vms.ocom.okstate.edu* Subject : CS " **gas** " and allergic *response- Ques* . Lines : 6 Nntp-Posting-Host : *vms.ocom.okstate.edu* Organization : OSU College of Osteopathic Medicine This question derives from the Waco incident : Could CS (" gas ") particles create an allergic response which would result in *laryngospasm* and *asphyxiation* ? - especially in children . *DNC* in Ok. *OSU-COM* will *disavow* my opinion , and my existence , if necessary .
$\text{grad}^{\text{L2}}_{\int s}$	From : *chorley @ vms.ocom.okstate.edu* Subject : CS " gas " and **allergic** *response- Ques* . Lines : 6 Nntp-Posting-Host : *vms.ocom.okstate.edu* Organization : OSU College of Osteopathic Medicine This question derives from the Waco incident : Could CS (" gas ") particles create an allergic response which would result in *laryngospasm* and *asphyxiation* ? - especially in children . *DNC* in Ok. *OSU-COM* will *disavow* my opinion , and my existence , if necessary .
$\text{grad}^{\text{L2}}_{\int p}$	From : *chorley @ vms.ocom.okstate.edu* Subject : CS " **gas** " and allergic *response- Ques* . Lines : 6 Nntp-Posting-Host : *vms.ocom.okstate.edu* Organization : OSU College of Osteopathic Medicine This question derives from the Waco incident : Could CS (" gas ") particles create an allergic response which would result in *laryngospasm* and *asphyxiation* ? - especially in children . *DNC* in Ok. *OSU-COM* will *disavow* my opinion , and my existence , if necessary .
$\text{grad}^{\text{dot}}_{1s}$	From : *chorley @ vms.ocom.okstate.edu* Subject : CS " gas " and **allergic** *response- Ques* . Lines : 6 Nntp-Posting-Host : *vms.ocom.okstate.edu* Organization : OSU College of Osteopathic Medicine This question derives from the Waco incident : Could CS (" gas ") particles create an allergic response which would result in *laryngospasm* and *asphyxiation* ? - especially in children . *DNC* in Ok. *OSU-COM* will *disavow* my opinion , and my existence , if necessary .
$\text{grad}^{\text{dot}}_{1p}$	From : *chorley @ vms.ocom.okstate.edu* Subject : CS " gas " and **allergic** *response- Ques* . Lines : 6 Nntp-Posting-Host : *vms.ocom.okstate.edu* Organization : OSU College of Osteopathic Medicine This question derives from the Waco incident : Could CS (" gas ") particles create an allergic response which would result in *laryngospasm* and *asphyxiation* ? - especially in children . *DNC* in Ok. *OSU-COM* will *disavow* my opinion , and my existence , if necessary .
$\text{grad}^{\text{dot}}_{\int s}$	From : *chorley @ vms.ocom.okstate.edu* Subject : CS " gas " and **allergic** *response- Ques* . Lines : 6 Nntp-Posting-Host : *vms.ocom.okstate.edu* Organization : OSU College of Osteopathic Medicine This question derives from the Waco incident : Could CS (" gas ") particles create an allergic response which would result in *laryngospasm* and *asphyxiation* ? - especially in children . *DNC* in Ok. *OSU-COM* will *disavow* my opinion , and my existence , if necessary .
$\text{grad}^{\text{dot}}_{\int p}$	From : *chorley @ vms.ocom.okstate.edu* Subject : CS " gas " and **allergic** *response- Ques* . Lines : 6 Nntp-Posting-Host : *vms.ocom.okstate.edu* Organization : OSU College of Osteopathic Medicine This question derives from the Waco incident : Could CS (" gas ") particles create an allergic response which would result in *laryngospasm* and *asphyxiation* ? - especially in children . *DNC* in Ok. *OSU-COM* will *disavow* my opinion , and my existence , if necessary .
omit_1	From : *chorley @ vms.ocom.okstate.edu* Subject : CS " gas " and **allergic** *response- Ques* . Lines : 6 Nntp-Posting-Host : *vms.ocom.okstate.edu* Organization : OSU College of Osteopathic Medicine This question derives from the Waco incident : Could CS (" gas ") particles create an allergic response which would result in *laryngospasm* and *asphyxiation* ? - especially in children . *DNC* in Ok. *OSU-COM* will *disavow* my opinion , and my existence , if necessary .
omit_3	From : *chorley @ vms.ocom.okstate.edu* Subject : CS " gas " and **allergic** *response- Ques* . Lines : 6 Nntp-Posting-Host : *vms.ocom.okstate.edu* Organization : OSU College of Osteopathic Medicine This question derives from the Waco incident : Could CS (" gas ") particles create an allergic response which would result in *laryngospasm* and *asphyxiation* ? - especially in children . *DNC* in Ok. *OSU-COM* will *disavow* my opinion , and my existence , if necessary .
omit_7	From : *chorley @ vms.ocom.okstate.edu* Subject : CS " gas " and allergic *response- Ques* . Lines : 6 Nntp-Posting-Host : *vms.ocom.okstate.edu* Organization : OSU College of Osteopathic **Medicine** This question derives from the Waco incident : Could CS (" gas ") particles create an allergic response which would result in *laryngospasm* and *asphyxiation* ? - especially in children . *DNC* in Ok. *OSU-COM* will *disavow* my opinion , and my existence , if necessary .
occ_1	From : *chorley @ vms.ocom.okstate.edu* Subject : CS " gas " and **allergic** *response- Ques* . Lines : 6 Nntp-Posting-Host : *vms.ocom.okstate.edu* Organization : OSU College of Osteopathic Medicine This question derives from the Waco incident : Could CS (" gas ") particles create an allergic response which would result in *laryngospasm* and *asphyxiation* ? - especially in children . *DNC* in Ok. *OSU-COM* will *disavow* my opinion , and my existence , if necessary .
occ_3	From : *chorley @ vms.ocom.okstate.edu* Subject : CS " gas " and **allergic** *response- Ques* . Lines : 6 Nntp-Posting-Host : *vms.ocom.okstate.edu* Organization : OSU College of Osteopathic Medicine This question derives from the Waco incident : Could CS (" gas ") particles create an allergic response which would result in *laryngospasm* and *asphyxiation* ? - especially in children . *DNC* in Ok. *OSU-COM* will *disavow* my opinion , and my existence , if necessary .
occ_7	From : *chorley @ vms.ocom.okstate.edu* Subject : CS " gas " and allergic *response- Ques* . Lines : 6 Nntp-Posting-Host : *vms.ocom.okstate.edu* Organization : OSU College of **Osteopathic** Medicine This question derives from the Waco incident : Could CS (" gas ") particles create an allergic response which would result in *laryngospasm* and *asphyxiation* ? - especially in children . *DNC* in Ok. *OSU-COM* will *disavow* my opinion , and my existence , if necessary .
decomp	From : *chorley @ vms.ocom.okstate.edu* Subject : CS " gas " and **allergic** *response- Ques* . Lines : 6 Nntp-Posting-Host : *vms.ocom.okstate.edu* Organization : OSU College of Osteopathic Medicine This question derives from the Waco incident : Could CS (" gas ") particles create an allergic response which would result in *laryngospasm* and *asphyxiation* ? - especially in children . *DNC* in Ok. *OSU-COM* will *disavow* my opinion , and my existence , if necessary .
lrp	From : *chorley @ vms.ocom.okstate.edu* Subject : CS " gas " and **allergic** *response- Ques* . Lines : 6 Nntp-Posting-Host : *vms.ocom.okstate.edu* Organization : OSU College of Osteopathic Medicine This question derives from the Waco incident : Could CS (" gas ") particles create an allergic response which would result in *laryngospasm* and *asphyxiation* ? - especially in children . *DNC* in Ok. *OSU-COM* will *disavow* my opinion , and my existence , if necessary .
deeplift	From : *chorley @ vms.ocom.okstate.edu* Subject : CS " gas " and **allergic** *response- Ques* . Lines : 6 Nntp-Posting-Host : *vms.ocom.okstate.edu* Organization : OSU College of Osteopathic Medicine This question derives from the Waco incident : Could CS (" gas ") particles create an allergic response which would result in *laryngospasm* and *asphyxiation* ? - especially in children . *DNC* in Ok. *OSU-COM* will *disavow* my opinion , and my existence , if necessary .
$\text{limsse}^{\text{bb}}$	From : *chorley @ vms.ocom.okstate.edu* Subject : CS " gas " and allergic *response- Ques* . Lines : 6 Nntp-Posting-Host : *vms.ocom.okstate.edu* Organization : OSU College of Osteopathic Medicine This question derives from the Waco incident : Could CS (" gas ") particles create an **allergic** response which would result in *laryngospasm* and *asphyxiation* ? - especially in children . *DNC* in Ok. *OSU-COM* will *disavow* my opinion , and my existence , if necessary .
$\text{limsse}^{\text{ns}}_{s}$	From : *chorley @ vms.ocom.okstate.edu* Subject : CS " gas " and allergic *response- Ques* . Lines : 6 Nntp-Posting-Host : *vms.ocom.okstate.edu* Organization : OSU College of Osteopathic Medicine This question derives from the Waco incident : Could CS (" gas ") particles create an **allergic** response which would result in *laryngospasm* and *asphyxiation* ? - especially in children . *DNC* in Ok. *OSU-COM* will *disavow* my opinion , and my existence , if necessary .
$\text{limsse}^{\text{ns}}_{p}$	From : *chorley @ vms.ocom.okstate.edu* Subject : CS " gas " and allergic *response- Ques* . Lines : 6 Nntp-Posting-Host : *vms.ocom.okstate.edu* Organization : OSU College of Osteopathic Medicine This question derives from the Waco incident : Could CS (" gas ") particles create an **allergic** response which would result in *laryngospasm* and *asphyxiation* ? - especially in children . *DNC* in Ok. *OSU-COM* will *disavow* my opinion , and my existence , if necessary .

Figure 8: sci.med post (not hybrid). Underlined: Manual relevance ground truth. Green (resp. red): evidence for (resp. against) sci.med. Task method: GRU. Italics: OOV. Bold: rmax position.

$\text{grad}^{\text{L2}}_{1s}$	If you find faith to be honest , show me how . David The whole **denominational** mindset only causes more problems , sadly . (See section 7 for details .) Thank you . 'The Armenians just shot and shot . Maybe *coz* they 're *'quality'* cars ; -) 200 *posts/day* . can you explain this or is it that they usually talk to stars more than regular players which explains the *hight* percentage of results after . It was produced in collaboration with the American College of *Surgeons* Commission on Cancer .
$\text{grad}^{\text{L2}}_{1p}$	If you find faith to be honest , show me how . David The whole **denominational** mindset only causes more problems , sadly . (See section 7 for details .) Thank you . 'The Armenians just shot and shot . Maybe *coz* they 're *'quality'* cars ; -) 200 *posts/day* . can you explain this or is it that they usually talk to stars more than regular players which explains the *hight* percentage of results after . It was produced in collaboration with the American College of *Surgeons* Commission on Cancer .
$\text{grad}^{\text{L2}}_{\int s}$	If you find faith to be honest , show me how . David The whole denominational mindset only causes more problems , sadly . (See section 7 for details .) Thank you . 'The Armenians just shot and shot . Maybe *coz* they 're *'quality'* cars ; -) 200 *posts/day* . can you explain this or is it that they usually talk to stars more than regular players which explains the *hight* percentage of results after . It was produced in collaboration with the American College of Surgeons Commission on Cancer .
$\text{grad}^{\text{L2}}_{\int p}$	If you find faith to be honest , show me how . David The whole denominational mindset only causes more problems , sadly . (See section 7 for details .) Thank you . 'The Armenians just shot and shot . Maybe *coz* they 're *'quality'* cars ; -) 200 *posts/day* . can you explain this or is it that they usually talk to stars more than regular players which explains the *hight* percentage of results after . It was produced in collaboration with the American College of *Surgeons* **Commission** on Cancer .
$\text{grad}^{\text{dot}}_{1s}$	If you find faith to be honest , show me how . David The whole denominational mindset only causes more problems , sadly . (See section 7 for details .) Thank you . 'The **Armenians** just shot and shot . Maybe *coz* they 're *'quality'* cars ; -) 200 *posts/day* . can you explain this or is it that they usually talk to stars more than regular players which explains the *hight* percentage of results after . It was produced in collaboration with the American College of *Surgeons* Commission on Cancer .
$\text{grad}^{\text{dot}}_{1p}$	If you find faith to be honest , show me how . David The whole denominational mindset only causes more problems , sadly . (See section 7 for details .) Thank you . 'The **Armenians** just shot and shot . Maybe *coz* they 're *'quality'* cars ; -) 200 *posts/day* . can you explain this or is it that they usually talk to stars more than regular players which explains the *hight* percentage of results after . It was produced in collaboration with the American College of *Surgeons* Commission on Cancer .
$\text{grad}^{\text{dot}}_{\int s}$	If you find faith to be honest , show me how . David The whole denominational mindset only causes more problems , sadly . (See section 7 for details .) Thank you . 'The **Armenians** just shot and shot . Maybe *coz* they 're *'quality'* cars ; -) 200 *posts/day* . can you explain this or is it that they usually talk to stars more than regular players which explains the *hight* percentage of results after . It was produced in collaboration with the American College of *Surgeons* Commission on Cancer .
$\text{grad}^{\text{dot}}_{\int p}$	If you find faith to be honest , show me how . David The whole denominational mindset only causes more problems , sadly . (See section 7 for details .) Thank you . 'The **Armenians** just shot and shot . Maybe *coz* they 're *'quality'* cars ; -) 200 *posts/day* . can you explain this or is it that they usually talk to stars more than regular players which explains the *hight* percentage of results after . It was produced in collaboration with the American College of *Surgeons* Commission on Cancer .
omit_1	If you find faith to be honest , show me how . David The whole denominational mindset only causes more problems , sadly . (See section 7 for details .) Thank you . 'The **Armenians** just shot and shot . Maybe *coz* they 're *'quality'* cars ; -) 200 *posts/day* . can you explain this or is it that they usually talk to stars more than regular players which explains the *hight* percentage of results after . It was produced in collaboration with the American College of *Surgeons* Commission on Cancer .
omit_3	If you find faith to be honest , show me how . David The whole denominational mindset only causes more problems , sadly . (See section 7 for details .) Thank you . 'The **Armenians** just shot and shot . Maybe *coz* they 're *'quality'* cars ; -) 200 *posts/day* . can you explain this or is it that they usually talk to stars more than regular players which explains the *hight* percentage of results after . It was produced in collaboration with the American College of *Surgeons* Commission on Cancer .
omit_7	If you find faith to be honest , show me how . David The whole denominational mindset only causes more problems , sadly . (See section 7 for details .) Thank you . 'The **Armenians** just shot and shot . Maybe *coz* they 're *'quality'* cars ; -) 200 *posts/day* . can you explain this or is it that they usually talk to stars more than regular players which explains the *hight* percentage of results after . It was produced in collaboration with the American College of *Surgeons* Commission on Cancer .
occ_1	If you find faith to be honest , show me how . David The whole denominational mindset only causes more problems , sadly . (See section 7 for details .) Thank you . 'The **Armenians** just shot and shot . Maybe *coz* they 're *'quality'* cars ; -) 200 *posts/day* . can you explain this or is it that they usually talk to stars more than regular players which explains the *hight* percentage of results after . It was produced in collaboration with the American College of *Surgeons* Commission on Cancer .
occ_3	If you find faith to be honest , show me how . David The whole denominational mindset only causes more problems , sadly . (See section 7 for details .) Thank you . 'The **Armenians** just shot and shot . Maybe *coz* they 're *'quality'* cars ; -) 200 *posts/day* . can you explain this or is it that they usually talk to stars more than regular players which explains the *hight* percentage of results after . It was produced in collaboration with the American College of *Surgeons* Commission on Cancer .
occ_7	If you find faith to be honest , show me how . David The whole denominational mindset only causes more problems , sadly . (See section 7 for details .) Thank you . 'The **Armenians** just shot and shot . Maybe *coz* they 're *'quality'* cars ; -) 200 *posts/day* . can you explain this or is it that they usually talk to stars more than regular players which explains the *hight* percentage of results after . It was produced in collaboration with the American College of *Surgeons* Commission on Cancer .
decomp	If you find faith to be honest , show me how . David The whole denominational mindset only causes more problems , sadly . (See section 7 for details .) Thank you . 'The **Armenians** just shot and shot . Maybe *coz* they 're *'quality'* cars ; -) 200 *posts/day* . can you explain this or is it that they usually talk to stars more than regular players which explains the *hight* percentage of results after . It was produced in collaboration with the American College of *Surgeons* Commission on Cancer .
lrp	If you find faith to be honest , show me how . David The whole denominational mindset only causes more problems , sadly . (See section 7 for details .) Thank you . 'The **Armenians** just shot and shot . Maybe *coz* they 're *'quality'* cars ; -) 200 *posts/day* . can you explain this or is it that they usually talk to stars more than regular players which explains the *hight* percentage of results after . It was produced in collaboration with the American College of *Surgeons* Commission on Cancer .
deeplift	If you find faith to be honest , show me how . David The whole denominational mindset only causes more problems , sadly . (See section 7 for details .) Thank you . 'The **Armenians** just shot and shot . Maybe *coz* they 're *'quality'* cars ; -) 200 *posts/day* . can you explain this or is it that they usually talk to stars more than regular players which explains the *hight* percentage of results after . It was produced in collaboration with the American College of *Surgeons* Commission on Cancer .
$\text{limsse}^{\text{bb}}$	If you find faith to be honest , show me how . David The whole denominational mindset only causes more problems , sadly . (See section 7 for details .) Thank you . 'The **Armenians** just shot and shot . Maybe *coz* they 're *'quality'* cars ; -) 200 *posts/day* . can you explain this or is it that they usually talk to stars more than regular players which explains the *hight* percentage of results after . It was produced in collaboration with the American College of *Surgeons* Commission on Cancer .
$\text{limsse}^{\text{ms}}_{s}$	If you find faith to be honest , show me how . David The whole denominational mindset only causes more problems , sadly . (See section 7 for details .) Thank you . 'The **Armenians** just shot and shot . Maybe *coz* they 're *'quality'* cars ; -) 200 *posts/day* . can you explain this or is it that they usually talk to stars more than regular players which explains the *hight* percentage of results after . It was produced in collaboration with the American College of *Surgeons* Commission on Cancer .
$\text{limsse}^{\text{ms}}_{p}$	If you find faith to be honest , show me how . David The whole denominational mindset only causes more problems , sadly . (See section 7 for details .) Thank you . 'The **Armenians** just shot and shot . Maybe *coz* they 're *'quality'* cars ; -) 200 *posts/day* . can you explain this or is it that they usually talk to stars more than regular players which explains the *hight* percentage of results after . It was produced in collaboration with the American College of *Surgeons* Commission on Cancer .

Figure 9: Hybrid newsgroup post, classified talk.politics.mideast. Green (resp. red): evidence for (resp. against) talk.politics.mideast. Underlined: talk.politics.mideast fragment. Italics: OOV. Task method: QGRU. Bold: rmax position.

$\text{grad}^{\text{L2}}_{1s}$	Fair enough . 2 . H. *Rahmi* , ed . There 's nothing punitive or unjust about it . Thanks (Deletion) It is not a question of grammar , it is a question of modelling . So far only two things seem to work : To kill it dead or to run into the house and close all doors and windows . FTP to ftp.uu.net : *graphics/jpeg/jpegsrc.v* ? .tar.Z Do n't forget to set binary mode when you FTP tar files . Interplanetary . : +49 231 755-4663 D-W4600 Dortmund 50 —— Fax : +49 231 **755-2386**
$\text{grad}^{\text{L2}}_{1p}$	Fair enough . 2 . H. *Rahmi* , ed . There 's nothing punitive or unjust about it . Thanks (Deletion) It is not a question of grammar , it is a question of modelling . So far only two things seem to work : To kill it dead or to run into the house and close all doors and windows . FTP to ftp.uu.net : *graphics/jpeg/jpegsrc.v* ? .tar.Z Do n't forget to set binary mode when you FTP tar files . Interplanetary . : +49 231 755-4663 D-W4600 Dortmund 50 —— Fax : +49 231 **755-2386**
$\text{grad}^{\text{L2}}_{\int s}$	Fair enough . 2 . H. *Rahmi* , ed . There 's nothing punitive or unjust about it . Thanks (Deletion) It is not a question of grammar , it is a question of modelling . So far only two things seem to work : To kill it dead or to run into the house and close all doors and windows . FTP to ftp.uu.net : *graphics/jpeg/jpegsrc.v* ? .tar.Z Do n't forget to set binary mode when you FTP tar files . Interplanetary . : +49 231 755-4663 D-W4600 Dortmund 50 —— Fax : +49 231 **755-2386**
$\text{grad}^{\text{L2}}_{\int p}$	Fair enough . 2 . H. *Rahmi* , ed . There 's nothing punitive or unjust about it . Thanks (Deletion) It is not a question of grammar , it is a question of modelling . So far only two things seem to work : To kill it dead or to run into the house and close all doors and windows . FTP to ftp.uu.net : *graphics/jpeg/jpegsrc.v* ? .tar.Z Do n't forget to set binary mode when you FTP tar files . Interplanetary . : +49 231 755-4663 D-W4600 Dortmund 50 —— Fax : +49 231 755-2386
$\text{grad}^{\text{dot}}_{1s}$	Fair enough . 2 . H. *Rahmi* , ed . There 's nothing punitive or unjust about it . Thanks (Deletion) It is not a question of grammar , it is a question of modelling . So far only two things seem to work : To kill it dead or to run into the house and close all doors and windows . FTP to ftp.uu.net : *graphics/jpeg/jpegsrc.v* ? .tar.Z Do n't forget to set binary mode when you FTP tar files . Interplanetary . : +49 231 755-4663 D-W4600 Dortmund 50 —— Fax : +49 231 **755-2386**
$\text{grad}^{\text{dot}}_{1p}$	Fair enough . 2 . H. *Rahmi* , ed . There 's nothing punitive or unjust about it . Thanks (Deletion) It is not a question of grammar , it is a question of modelling . So far only two things seem to work : To kill it dead or to run into the house and close all doors and windows . FTP to ftp.uu.net : *graphics/jpeg/jpegsrc.v* ? .tar.Z Do n't forget to set binary mode when you FTP tar files . Interplanetary . : +49 231 755-4663 D-W4600 Dortmund 50 —— Fax : +49 231 **755-2386**
$\text{grad}^{\text{dot}}_{\int s}$	Fair enough . 2 . H. *Rahmi* , ed . There 's nothing punitive or unjust about it . Thanks (Deletion) It is not a question of grammar , it is a question of modelling . So far only two things seem to work : To kill it dead or to run into the house and close all doors and windows . FTP to ftp.uu.net : *graphics/jpeg/jpegsrc.v* ? .tar.Z Do n't forget to set binary mode when you FTP tar files . Interplanetary . : +49 231 755-4663 D-W4600 Dortmund 50 —— Fax : +49 231 **755-2386**
$\text{grad}^{\text{dot}}_{\int p}$	Fair enough . 2 . H. *Rahmi* , ed . There 's nothing punitive or unjust about it . Thanks (Deletion) It is not a question of grammar , it is a question of modelling . So far only two things seem to work : To kill it dead or to run into the house and close all doors and windows . FTP to ftp.uu.net : *graphics/jpeg/jpegsrc.v* ? .tar.Z Do n't forget to set binary mode when you FTP tar files . Interplanetary . : +49 231 755-4663 D-W4600 Dortmund 50 —— Fax : +49 231 **755-2386**
omit_1	Fair enough . 2 . H. *Rahmi* , ed . There 's nothing punitive or unjust about it . Thanks (Deletion) It is not a question of grammar , it is a question of modelling . So far only two things seem to work : To kill it dead or to run into the house and close all doors and windows . FTP to ftp.uu.net : *graphics/jpeg/jpegsrc.v* ? .tar.Z Do n't forget to set binary mode when you FTP tar files . Interplanetary . : +49 231 755-4663 D-W4600 Dortmund 50 —— Fax : +49 231 **755-2386**
omit_3	Fair enough . 2 . H. *Rahmi* , ed . There 's nothing punitive or unjust about it . Thanks (Deletion) It is not a question of grammar , it is a question of modelling . So far only two things seem to work : To kill it dead or to run into the house and close all doors and windows . FTP to ftp.uu.net : *graphics/jpeg/jpegsrc.v* ? .tar.Z Do n't forget to set binary mode when you FTP tar files . Interplanetary . : +49 231 755-4663 D-W4600 Dortmund 50 —— Fax : **+49** 231 755-2386
omit_7	Fair enough . 2 . H. *Rahmi* , ed . There 's nothing punitive or unjust about it . Thanks (Deletion) It is not a question of grammar , it is a question of modelling . So far only two things seem to work : To kill it dead or to run into the house and close all doors and windows . FTP to ftp.uu.net : *graphics/jpeg/jpegsrc.v* ? .tar.Z Do n't forget to set binary mode when you FTP tar files . Interplanetary . : +49 231 755-4663 D-W4600 Dortmund **50** —— Fax : +49 231 755-2386
occ_1	Fair enough . 2 . H. *Rahmi* , ed . There 's nothing punitive or unjust about it . Thanks (Deletion) It is not a question of grammar , it is a question of modelling . So far only two things seem to work : To kill it dead or to run into the house and close all doors and windows . FTP to ftp.uu.net : *graphics/jpeg/jpegsrc.v* ? .tar.Z Do n't forget to set binary mode when you FTP tar files . Interplanetary . : +49 231 755-4663 D-W4600 Dortmund 50 —— Fax : +49 231 **755-2386**
occ_3	Fair enough . 2 . H. *Rahmi* , ed . There 's nothing punitive or unjust about it . Thanks (Deletion) It is not a question of grammar , it is a question of modelling . So far only two things seem to work : To kill it dead or to run into the house and close all doors and windows . FTP to ftp.uu.net : *graphics/jpeg/jpegsrc.v* ? .tar.Z Do n't forget to set binary mode when you FTP tar files . Interplanetary . : +49 231 755-4663 D-W4600 Dortmund 50 —— Fax : **+49** 231 755-2386
occ_7	Fair enough . 2 . H. *Rahmi* , ed . There 's nothing punitive or unjust about it . Thanks (Deletion) It is not a question of grammar , it is a question of modelling . So far only two things seem to work : To kill it dead or to run into the house and close all doors and windows . FTP to ftp.uu.net : *graphics/jpeg/jpegsrc.v* ? .tar.Z Do n't forget to set binary mode when you FTP tar files . Interplanetary . : +49 231 755-4663 D-W4600 Dortmund **50** —— Fax : +49 231 755-2386
decomp	Fair enough . 2 . H. *Rahmi* , ed . There 's nothing punitive or unjust about it . Thanks (Deletion) It is not a question of grammar , it is a question of modelling . So far only two things seem to work : To kill it dead or to run into the house and close all doors and windows . FTP to ftp.uu.net : *graphics/jpeg/jpegsrc.v* ? .tar.Z Do n't forget to set binary mode when you FTP tar files . Interplanetary . : +49 231 755-4663 D-W4600 Dortmund 50 —— Fax : +49 **231** 755-2386
lrp	Fair enough . 2 . H. *Rahmi* , ed . There 's nothing punitive or unjust about it . Thanks (Deletion) It is not a question of grammar , it is a question of modelling . So far only two things seem to work : To kill it dead or to run into the house and close all doors and windows . FTP to ftp.uu.net : *graphics/jpeg/jpegsrc.v* ? .tar.Z Do n't forget to set binary mode when you FTP tar files . Interplanetary . : +49 231 755-4663 D-W4600 Dortmund 50 —— Fax : +49 **231** 755-2386
deeplift	Fair enough . 2 . H. *Rahmi* , ed . There 's nothing punitive or unjust about it . Thanks (Deletion) It is not a question of grammar , it is a question of modelling . So far only two things seem to work : To kill it dead or to run into the house and close all doors and windows . FTP to ftp.uu.net : *graphics/jpeg/jpegsrc.v* ? .tar.Z Do n't forget to set binary mode when you FTP tar files . Interplanetary . : +49 231 755-4663 D-W4600 Dortmund 50 —— Fax : +49 **231** 755-2386
$\text{limsse}^{\text{bb}}$	Fair enough . 2 . H. *Rahmi* , ed . There 's nothing punitive or unjust about it . Thanks (Deletion) It is not a question of grammar , it is a question of modelling . So far only two things seem to work : To kill it dead or to run into the house and close all doors and windows . FTP to ftp.uu.net : *graphics/jpeg/jpegsrc.v* ? .tar.Z Do n't forget to set binary mode when you FTP tar files . Interplanetary . : +49 231 **755-4663** D-W4600 Dortmund 50 —— Fax : +49 231 755-2386
$\text{limsse}^{\text{ms}}_{s}$	Fair enough . 2 . H. *Rahmi* , ed . There 's nothing punitive or unjust about it . Thanks (Deletion) It is not a question of grammar , it is a question of modelling . So far only two things seem to work : To kill it dead or to run into the house and close all doors and windows . FTP to ftp.uu.net : *graphics/jpeg/jpegsrc.v* ? .tar.Z Do n't forget to set binary mode when you FTP tar files . Interplanetary . : +49 231 755-4663 D-W4600 Dortmund 50 —— Fax : +49 **231** 755-2386
$\text{limsse}^{\text{ms}}_{p}$	Fair enough . 2 . H. *Rahmi* , ed . There 's nothing punitive or unjust about it . Thanks (Deletion) It is not a question of grammar , it is a question of modelling . So far only two things seem to work : To kill it dead or to run into the house and close all doors and windows . FTP to ftp.uu.net : *graphics/jpeg/jpegsrc.v* ? .tar.Z Do n't forget to set binary mode when you FTP tar files . Interplanetary . : +49 **231** 755-4663 D-W4600 Dortmund 50 —— Fax : +49 231 755-2386

Figure 10: Hybrid newsgroup post, classified comp.windows.x. Green (resp. red): evidence for (resp. against) comp.windows.x. Underlined: comp.windows.x fragment. Italics: OOV. Task method: LSTM. Bold: rmax position. The telephone numbers in the last sentence appear in 3 comp.windows.x posts but nowhere else in the corpus.

$\text{grad}^{\text{L2}}_{1s}$	Sorry for any confusion I may have created . Jon Mandaville , Professor of History , Portland State University (Oregon) . Trying to mix up the lines is a dead end . matter of proving the track record of the # scientific method . – I know that I will have to pay tax when I go to register the car . It sounds like he has had another set back in his come back . Oh , and a prediction : Milt Cuyler . The **Xaw** support was missing from *OW2.0* but added in 3.0 . This slant *permeates* the text .
$\text{grad}^{\text{L2}}_{1p}$	Sorry for any confusion I may have created . Jon Mandaville , Professor of History , Portland State University (Oregon) . Trying to mix up the lines is a dead end . matter of proving the track record of the # scientific method . – I know that I will have to pay tax when I go to register the car . It sounds like he has had another set back in his come back . Oh , and a prediction : Milt **Cuyler** . The Xaw support was missing from *OW2.0* but added in 3.0 . This slant *permeates* the text .
$\text{grad}^{\text{L2}}_{\int s}$	**Sorry** for any confusion I may have created . Jon Mandaville , Professor of History , Portland State University (Oregon) . Trying to mix up the lines is a dead end . matter of proving the track record of the # scientific method . – I know that I will have to pay tax when I go to register the car . It sounds like he has had another set back in his come back . Oh , and a prediction : Milt Cuyler . The Xaw support was missing from *OW2.0* but added in 3.0 . This slant *permeates* the text .
$\text{grad}^{\text{L2}}_{\int p}$	**Sorry** for any confusion I may have created . Jon Mandaville , Professor of History , Portland State University (Oregon) . Trying to mix up the lines is a dead end . matter of proving the track record of the # scientific method . – I know that I will have to pay tax when I go to register the car . It sounds like he has had another set back in his come back . Oh , and a prediction : Milt Cuyler . The Xaw support was missing from *OW2.0* but added in 3.0 . This slant *permeates* the text .
$\text{grad}^{\text{dot}}_{1s}$	Sorry for any confusion I may have created . Jon Mandaville , Professor of History , Portland State University (Oregon) . Trying to mix up the lines is a dead end . matter of proving the track record of the # scientific method . – I know that I will have to pay tax when I go to register the car . It sounds like he has had another set back in his come back . Oh , and a prediction : Milt Cuyler . The **Xaw** support was missing from *OW2.0* but added in 3.0 . This slant *permeates* the text .
$\text{grad}^{\text{dot}}_{1p}$	Sorry for any confusion I may have created . Jon Mandaville , Professor of History , Portland State University (Oregon) . Trying to mix up the lines is a dead end . matter of proving the track record of the # scientific method . – I know that I will have to pay tax when I go to register the car . It sounds like he has had another set back in his come back . Oh , and a prediction : Milt Cuyler . The **Xaw** support was missing from *OW2.0* but added in 3.0 . This slant *permeates* the text .
$\text{grad}^{\text{dot}}_{\int s}$	Sorry for any confusion I may have created . Jon Mandaville , Professor of History , Portland State University (Oregon) . Trying to mix up the lines is a dead end . matter of proving the track record of the # scientific method . – I know that I will have to pay tax when I go to register the car . It sounds like he has had another set back in his come back . Oh , and a prediction : Milt Cuyler . The **Xaw** support was missing from *OW2.0* but added in 3.0 . This slant *permeates* the text .
$\text{grad}^{\text{dot}}_{\int p}$	Sorry for any confusion I may have created . Jon Mandaville , Professor of History , Portland State University (Oregon) . Trying to mix up the lines is a dead end . matter of proving the track record of the # scientific method . – I know that I will have to pay tax when I go to register the car . It sounds like he has had another set back in his come back . Oh , and a prediction : Milt Cuyler . The Xaw support was missing from *OW2.0* but added in 3.0 . This slant *permeates* the **text** .
omit_1	Sorry for any confusion I may have created . Jon Mandaville , Professor of History , Portland State University (Oregon) . Trying to mix up the lines is a dead end . matter of proving the track record of the # scientific method . – I know that I will have to pay tax when I go to register the car . It sounds like he has had another set back in his come back . Oh , and a prediction : Milt Cuyler . The **Xaw** support was missing from *OW2.0* but added in 3.0 . This slant *permeates* the text .
omit_3	Sorry for any confusion I may have created . Jon Mandaville , Professor of History , Portland State University (Oregon) . Trying to mix up the lines is a dead end . matter of proving the track record of the # scientific method . – I know that I will have to pay tax when I go to register the car . It sounds like he has had another set back in his come back . Oh , and a prediction : Milt Cuyler . The **Xaw** support was missing from *OW2.0* but added in 3.0 . This slant *permeates* the text .
omit_7	Sorry for any confusion I may have created . Jon Mandaville , Professor of History , Portland State University (Oregon) . Trying to mix up the lines is a dead end . matter of proving the track record of the # scientific method . – I know that I will have to pay tax when I go to register the car . It sounds like he has had another set back in his come back . Oh , and a prediction : Milt Cuyler . The Xaw **support** was missing from *OW2.0* but added in 3.0 . This slant *permeates* the text .
occ_1	Sorry for any confusion I may have created . Jon Mandaville , Professor of History , Portland State University (Oregon) . Trying to mix up the lines is a dead end . matter of proving the track record of the # scientific method . – I know that I will have to pay tax when I go to register the car . It sounds like he has had another set back in his come back . Oh , and a prediction : Milt Cuyler . The **Xaw** support was missing from *OW2.0* but added in 3.0 . This slant *permeates* the text .
occ_3	Sorry for any confusion I may have created . Jon Mandaville , Professor of History , Portland State University (Oregon) . Trying to mix up the lines is a dead end . matter of proving the track record of the # scientific method . – I know that I will have to pay tax when I go to register the car . It sounds like he has had another set back in his come back . Oh , and a prediction : Milt Cuyler . The **Xaw** support was missing from *OW2.0* but added in 3.0 . This slant *permeates* the text .
occ_7	Sorry for any confusion I may have created . Jon Mandaville , Professor of History , Portland State University (Oregon) . Trying to mix up the lines is a dead end . matter of proving the track record of the # scientific method . – I know that I will have to pay tax when I go to register the car . It sounds like he has had another set back in his come back . Oh , and a prediction : Milt Cuyler . The Xaw support was **missing** from *OW2.0* but added in 3.0 . This slant *permeates* the text .
decomp	Sorry for any confusion I may have created . Jon Mandaville , Professor of History , Portland State University (Oregon) . Trying to mix up the lines is a dead end . matter of proving the track record of the # scientific method . – I know that I will have to pay tax when I go to register the car . It sounds like he has had another set back in his come back . Oh , and a prediction : Milt Cuyler . The Xaw **support** was missing from *OW2.0* but added in 3.0 . This slant *permeates* the text .
lrp	Sorry for any confusion I may have created . Jon Mandaville , Professor of History , Portland State University (Oregon) . Trying to mix up the lines is a dead end . matter of proving the track record of the # scientific method . – I know that I will have to pay tax when I go to register the car . It sounds like he has had another set back in his come back . Oh , and a prediction : Milt Cuyler . The **Xaw** support was missing from *OW2.0* but added in 3.0 . This slant *permeates* the text .
deeplift	Sorry for any confusion I may have created . Jon Mandaville , Professor of History , Portland State University (Oregon) . Trying to mix up the lines is a dead end . matter of proving the track record of the # scientific method . – I know that I will have to pay tax when I go to register the car . It sounds like he has had another set back in his come back . Oh , and a prediction : Milt Cuyler . The **Xaw** support was missing from *OW2.0* but added in 3.0 . This slant *permeates* the text .
$\text{limsse}^{\text{bb}}$	Sorry for any confusion I may have created . Jon Mandaville , Professor of History , Portland State University (Oregon) . Trying to mix up the lines is a dead end . matter of proving the track record of the # scientific method . – I know that I will have to pay tax when I go to register the car . It sounds like he has had another set back in his come back . Oh , and a prediction : Milt Cuyler . The Xaw **support** was missing from *OW2.0* but added in 3.0 . This slant *permeates* the text .
$\text{limsse}^{\text{ns}}_{s}$	Sorry for any confusion I may have created . Jon Mandaville , Professor of History , Portland State University (Oregon) . Trying to mix up the lines is a dead end . matter of proving the track record of the # scientific method . – I know that I will have to pay tax when I go to register the car . It sounds like he has had another set back in his come back . Oh , and a prediction : Milt Cuyler . The Xaw **support** was missing from *OW2.0* but added in 3.0 . This slant *permeates* the text .
$\text{limsse}^{\text{ns}}_{p}$	Sorry for any confusion I may have created . Jon Mandaville , Professor of History , Portland State University (Oregon) . Trying to mix up the lines is a dead end . matter of proving the track record of the # scientific method . – I know that I will have to pay tax when I go to register the car . It sounds like he has had another set back in his come back . Oh , and a prediction : Milt Cuyler . The Xaw **support** was missing from *OW2.0* but added in 3.0 . This slant *permeates* the text .

Figure 11: Hybrid newsgroup post, classified comp.windows.x. Green (resp. red): evidence for (resp. against) comp.windows.x. Underlined: comp.windows.x fragment. Italics: OOV. Task method: QLSTM. Bold: rmax position.

Figure 12: Hybrid yelp review, classified positive. Green (resp. red): evidence for (resp. against) positive. Underlined: positive fragments. Italics: OOV. Task method: GRU. Bold: rmax position.

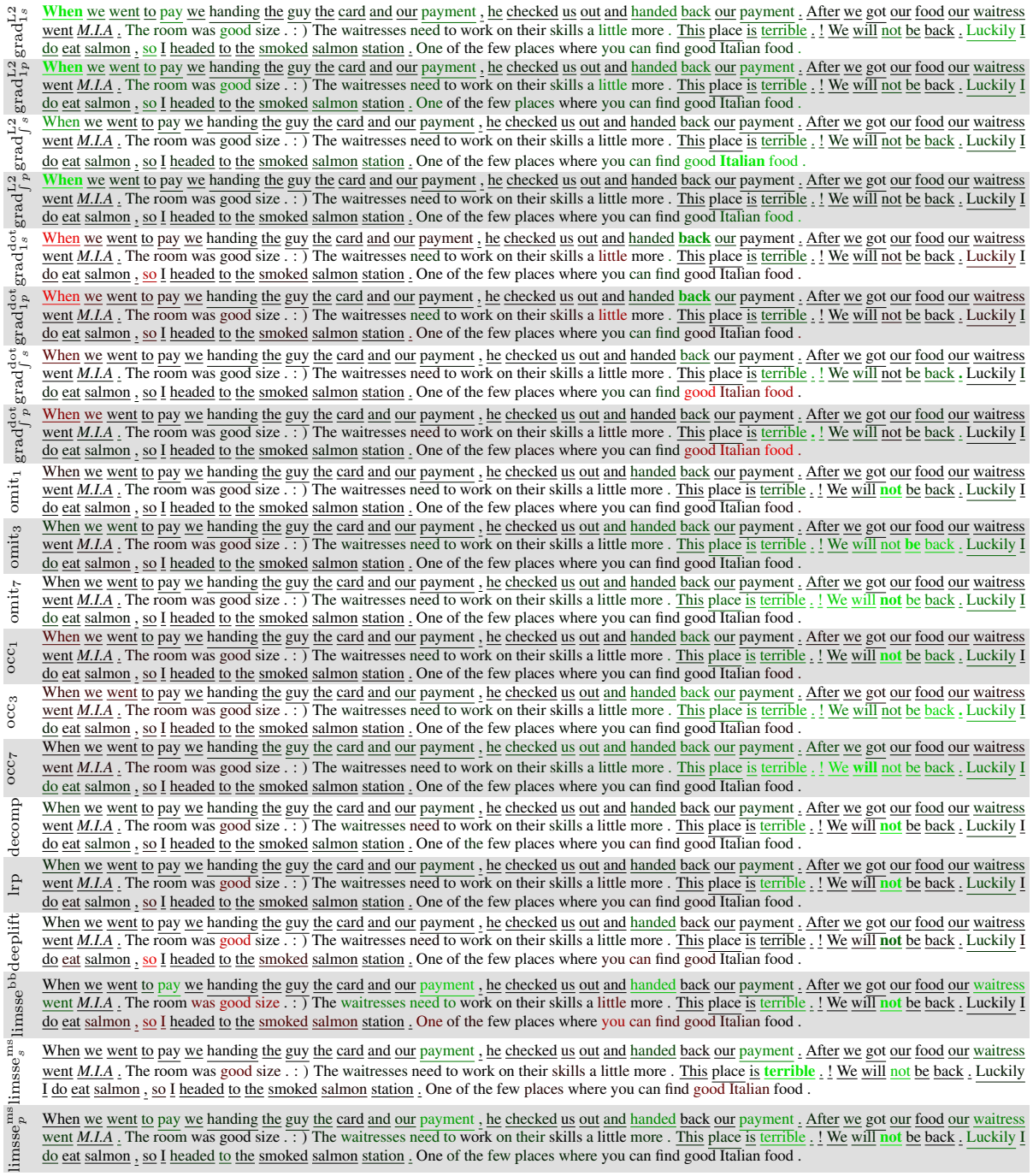

Figure 13: Hybrid yelp review, classified negative. Green (resp. red): evidence for (resp. against) negative. Underlined: negative fragments. Italics: OOV. Task method: LSTM. Bold: rmax position.