IDIAP RESEARCH REPORT

EVALUATING ATTENTION NETWORKS FOR ANAPHORA RESOLUTION

Jonathan Pilault Nikolaos Pappas
Lesly Miculicich Werlen Andrei Popescu-Belis

Idiap-RR-27-2017

SEPTEMBER 2017

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Evaluating Attention Networks for Anaphora Resolution

Jonathan Pilault[1], Nikolaos Pappas, Lesly Miculicich Werlen, Andrei Popescu-Belis

Idiap Research Institute, Martigny, Switzerland

September 25, 2017

Abstract

In this paper, we evaluate the results of using inter and intra attention mechanisms from two architectures, a Deep Attention Long Short-Term Memory-Network (LSTM-N) (Cheng et al., 2016) and a Decomposable Attention model (Parikh et al., 2016), for anaphora resolution, i.e. detecting coreference relations between a pronoun and a noun (its antecedent). The models are adapted from an entailment task, to address the pronominal coreference resolution task by comparing two pairs of sentences: one with the original sentences containing the antecedent and the pronoun, and another one with the pronoun replaced with a correct or an incorrect antecedent. The goal is thus to detect the correct replacements, assuming the original sentence pair entails the one with the correct replacement, but not one with an incorrect replacement. We use the CoNLL-2012 English dataset (Pradhan et al., 2012) to train the models and evaluate the ability to recognize correct and incorrect pronoun replacements in sentence pairs. We find that the Decomposable Attention Model performs better, while using a much simpler architecture. Furthermore, we focus on two previous studies that use intra- and inter-attention mechanisms, discuss how they relate to each other, and examine how these advances work to identify correct antecedent replacements.

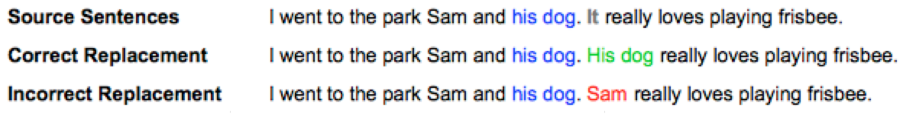
1 Introduction

Coreference resolution, in which the entities discussed in a given text are linked to all of the textual spans that refer to them, has been one of the key areas of NLP for several decades. Major modeling breakthroughs have been achieved, not surprisingly, following three successful shared tasks: MUC (Chinchor and Robinson, 1998), ACE (Doddington et al., 2004) and, most recently, CoNLL (Pradhan et al., 2011, 2012). Coreference resolution seeks to find and group the noun phrases in a text that refer to the same real-world entity. In this context, such noun phrases are called mentions, or anaphoric noun phrases. Mentions can be either nominal (including named entities) or pronominal. In this paper, we focus on the pronominal anaphora resolution task, which is an important and challenging part of the more general task of coreference resolution. Correct resolution of the antecedents of pronouns is important for a variety of other natural language processing tasks, including — but not limited to -– information retrieval, neural machine translation, and text understanding in dialog systems.

Recent research has greatly improved the state-of-the-art and allowed models to resolve coreference end-to-end and without hand-engineered features (Lee et al., 2017).

[1] Work done during an internship at the Idiap Research Institute, from March to August 2017. Contact email: Jonathan.Pilault@gmail.com.

While our approach is somewhat different from current trends in research, we build on some of the same ideas developed in the last decade in coreference resolution, i.e. a shift from rule-based, hand-crafted systems to statistical machine learning systems, see Mitkov (2002) for an overview. Classification has been a common approach to pronoun resolution, as seen for example in the work by Morton (2000) and Kehler (2004). Since there would be too many classes if we treated each antecedent candidate for each anaphor as a separate class, a binary classification approach can allow us to identify coreferent noun phrases from non-coreferent ones. With many candidates considered for each anaphor, the potential result of this approach is a set of candidates that are identified as coreferent with the anaphor. Our approach does assume that mentions have already been detected. Our models are fed during training hand-made pairs of sentence-groups (one or more consecutive sentences) that have either incorrect or correct pronoun replacements. Figure 1 is an example of such replacements.



Source Sentences	I went to the park Sam and his dog. It really loves playing frisbee.
Correct Replacement	I went to the park Sam and his dog. His dog really loves playing frisbee.
Incorrect Replacement	I went to the park Sam and his dog. Sam really loves playing frisbee.

Figure 1. Toy example of pronoun replacements. The mention *his dog* in blue should replace *it* in gray. *Sam* in red is an incorrect substitution and *his dog* in green is a correct substitution in the example shown. In the training data, the model is presented with a label, the original sentences and either the incorrect or the correct substitution in the target sentences.

We propose to train and test two models: a Deep Attention LSTM-N model (Cheng et al., 2016) and a Decomposable Attention model (Parikh et al., 2016) on the pronominal coreference resolution task. The first model extends the Long Short Term Memory architecture by performing shallow reasoning with memory and attention. The second one, which computes attention purely based on word embeddings, consists of feed-forward networks that operate largely independently of word order. The interest of the experiments is to determine whether or not we can correctly identify antecedents out of a list of mentions that are candidate antecedents.

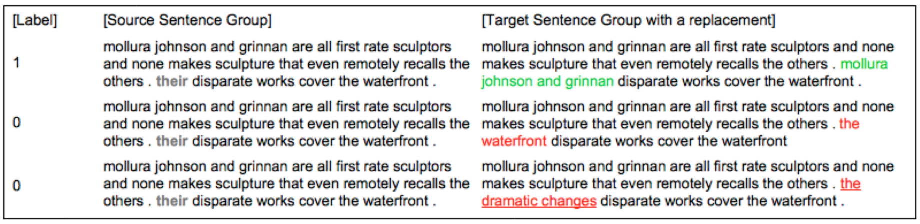
2 Data

The experiment used the CoNLL-2012 English dataset, which contains annotated co-reference relations. The dataset is substantially small, with 1.3 million words and 2802 documents in total. We do not use world knowledge, such as WordNet, animacy and gender lexicons, on which existing models have relied (Durrett and Klein, 2013). However, we have experimented Word2Vec (Mikolov et al., 2013) to initialize the model's word embeddings.[2]

The CoNLL-2012 shared task data has a relatively complex data format. The first and most time consuming step of this project has been to parse the data and reconstruct sentence-group pairs (source and target ones, as defined below). Available resources to extract coreference pairs are written in Java (Durrett and Klein, 2013; Lee et al., 2011; Björkelund and Farkas, 2012). We implemented a system to parse sentences, find suitable correct/incorrect replacements for pronouns. The parser would replace only one pronoun

[2] Word2Vec was pre-trained on the Google News corpus (3 billion words) to learn the word vector model (3 million 300-dimension English word vectors).

at a time. For example, if sentences contained two pronouns with antecedents, we would generate at least four examples: examples with incorrect and correct substitutions for the first pronoun, and examples with incorrect and correct substitutions for the second one. If multiple correct and unique candidate noun phrases existed in the dataset, we would create the same number of examples with correct substitution. For examples with incorrect replacements, we chose a random noun phrase from within and from outside of the scope of the sentences. Each line in our dataset contains a label, the original example and the example with a replacement, as shown in Figure 2.[3]



[Label]	[Source Sentence Group]	[Target Sentence Group with a replacement]
1	mollura johnson and grinnan are all first rate sculptors and none makes sculpture that even remotely recalls the others . their disparate works cover the waterfront .	mollura johnson and grinnan are all first rate sculptors and none makes sculpture that even remotely recalls the others . mollura johnson and grinnan disparate works cover the waterfront .
0	mollura johnson and grinnan are all first rate sculptors and none makes sculpture that even remotely recalls the others . their disparate works cover the waterfront .	mollura johnson and grinnan are all first rate sculptors and none makes sculpture that even remotely recalls the others . the waterfront disparate works cover the waterfront
0	mollura johnson and grinnan are all first rate sculptors and none makes sculpture that even remotely recalls the others . their disparate works cover the waterfront .	mollura johnson and grinnan are all first rate sculptors and none makes sculpture that even remotely recalls the others . the dramatic changes disparate works cover the waterfront .

Figure 2. Illustration of input data format. The format is similar to the one found for Natural Language Inference (entailment task), i.e. label | source sentences | target sentences (Bowman et al., 2015). Green noun-phrases are the correct antecedent replacement. Red noun-phrases are the incorrect antecedent replacement. The underlined replacement is in the document but not in the source sentence group.

The part-of-speech tags in the CoNLL-2012 dataset have allowed us to be mindful of correct sentence structures and grammatical form when generating new examples from sources sentences. We have used POS tags to catalog noun phrases, pronouns, articles during pre-processing. The steps to generate training and testing examples are the following ones:

1. For each article in the CoNLL-2012 dataset, noted $document_i$, we first loop through each m words in $document_i$ to store all possible noun-phrases $NP_i = \{np_j \mid j \in \mathbb{R} : 1 \leqslant j \leqslant m\}_i$, where j is the position of the noun phrase in a $document_i$.
2. For each article, we then loop through each word in the document, searching for pronouns POS tags with possible antecedents[4] $ANT_i = \{np_k \mid k \in \mathbb{R} : 1 \leqslant k \leqslant m\}_i$, where m is the number of words in the text.
3. When a pronoun and its position are found, we check if an antecedent exists in the sentence where the pronoun is positioned or the sentence immediately before. All words within the two sentences are said to be "within scope". All words outside of the two sentences are said to be "out of scope". We can define the start and end position of all words within scope as $start \in \mathbb{R} : 1 \leqslant start < end$ and $end \in \mathbb{R} : start < end \leqslant m$.
4. If the antecedent in step 3 is found, we create three new target sentences:
   (a) with the pronoun replaced by the correct antecedent $ant_x \in ANT_i$.
   (b) with the pronoun replaced by a random non-conreferent noun phrase np_x^{in} within scope such that $\{np_x^{in} \in NP_i \ \& \ np_x^{in} \notin ANT_i \mid x \in \mathbb{R} : \ start \leqslant x \leqslant end \ \& \ x \neq k\}$, where $start$ and end are defined in step 3.

[3] Additional examples are provided in the Appendix.
[4] Some pronouns may not have an antecedent; for example: "It is nice out".

(c) with the pronoun replaced by a random non-coreferent noun phrase np_x^{out} out of scope such that $\{np_x^{out} \in NP_i \ \& \ np_x^{out} \notin ANT_i \mid x \in \mathbb{R} : \ x < start \ or \ x > end \ \& \ x \neq k\}$, where *start* and *end* are defined in step 3.

In a preliminary experiment, we trained the models with a dataset that had only out-of-scope incorrect replacements, i.e. coming only from np_x^{out}. The resulting model performed very well, but actually was trained to recognize when replacements were out-of-scope and not if the replacement was correct or incorrect. This shows that replacements np_x^{in} should be included in the training data, since this makes the classification task focus on coreference. With a mix of data with both np_x^{in} and np_x^{out}, the model may be able learn to classify correct anaphoric mentions for in-scope and out-of scope replacements. Unfortunately, np_x^{in}, as defined in step 4c above, was not always present. To avoid throwing away examples with no np_x^{in} incorrect replacements, the dataset was populated with only one incorrect pronoun replacement np_x^{out} and one correct pronoun replacement ant_x. This technique has allowed us to augment our dataset with more non-coreferent examples. As a result, however, the resulting dataset is to some extent unbalanced. The number of examples per category type is detailed in Table 1.

Table 1. Number of examples per type in training/validation/test data.

Category	**Total examples**		
	train	validation	test
Correct replacement ant_x	19114	765	769
Incorrect replacement np_x^{in}	11468	470	463
Incorrect replacement np_x^{out}	19113	765	768
Total	49695	2000	2000

The determiners of the noun phrases (mostly definite or indefinite articles) were also integrated in our selection algorithm to make sure that candidate antecedents had the correct grammatical structure. Pronouns without antecedent were unchanged and no examples were created from such cases. Also, no examples with replacement were created from first person possessive pronouns since such pronouns generally have antecedents outside the scope of the text. For example, if a document is narrated in the first person, the identity of the narrator may not be explicitly found in the text. Also, for possessive pronouns in quotes or dialogue, we found that over 90% of antecedents were found outside a two-sentence scope. Finally, it must be mentioned that we had to control the total number of words in each example, to avoid using too much memory when training the network. Our sentence-groups were therefore limited to a maximum of two sentences. Our data contains 49,695 sentence-group pairs for training, 2,000 for validation and 2,000 for testing.

3 Architecture

3.1 Attention-based models under study

3.1.1 Long Short-Term Memory

The core of both models is a Long Short-Term Memory (LSTM) recurrent neural network (Hochreiter and Schmidhuber, 1997) which processes a variable-length sequence $x = (x_1, x_2, \ldots, x_n)$ by incrementally adding new content into a single memory slot, with gates controlling the extent to which new content should be memorized, old content should be erased, and current content should be exposed. At time step t, the memory c_t

and the hidden state h_t are updated with the following equations:

$$\begin{bmatrix} i_t \\ f_t \\ o_t \\ \hat{c}_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ tanh \end{bmatrix} W \cdot [h_{t-1}, x_t] \tag{1}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \hat{c}_t \tag{2}$$

$$h_t = o_t \odot tanh(c_t) \tag{3}$$

where i, f, and o are gate activations. Compared to the standard RNN, the LSTM uses additive memory updates and it separates the memory c from the hidden state h, which interacts with the environment when making predictions.

3.1.2 Deep Attention LSTM Network

We will first describe Cheng et al.'s Deep Attention LSTM Network (LSTMN) (Cheng et al., 2016). Cheng et al. changed the LSTM by adding a memory network in lieu of a memory cell, to extend the ability to memorize sequences under recursive compression. In a LSTM, the next state h_{t+1} is conditionally independent on states h_1, ..., h_{t-1} and tokens x_1, ..., x_t. In effect, h_{t+1} depends on h_t as it is assumed that the current state holds enough information about past states to predict the next state. Cheng et al. argue that the assumption may not hold when the sequence is long, and propose to address this potential limitation with an LSTMN, which may also model the structural relationship between tokens instead of the sequential token-by-token relationship seen in LSTMs. Similarly to Weston et al. (2014), the proposed Memory Network explicitly segregates memory storage from the neural network computation. The model is trained end-to-end with a memory addressing mechanism closely related to soft attention (Sukhbaatar et al., 2015).

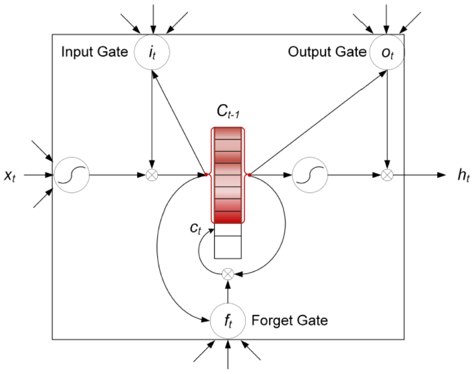


Figure 3. Long Short-Term Memory-Network. Color indicates degree of memory activation. Figure taken from: "Long Short-Term Memory-Networks for Machine Reading" (Cheng et al., 2016). The deep attention r gate is not shown in the figure.

The architecture of the LSTMN is shown in Figure 3. LSTMs maintain a hidden vector and a memory vector, while Memory Networks (Weston et al., 2014) have an array of key vectors to access a set of value vectors. The LSTMN uses a memory tape to store the contextual memory $C_{t-1} = (\tilde{c}_1, \dots, \tilde{c}_{t-1})$ and the previous hidden memory $H_{t-1} = (\tilde{h}_1, \dots, \tilde{h}_{t-1})$ derived from the attention mechanism. At time step t, the model computes the relation between x_t and $x_1, \dots, x_{t-1}$ through $h_1, \dots, h_{t-1}$ with an *intra-attention* layer:[5]

$$a_i^t = v^T tanh(W_h h_i + W_x x_t + W_{\tilde{h}} \tilde{h}_{t-1}) \tag{4}$$

$$s_i^t = softmax(a_i^t) \tag{5}$$

$\tilde{c}_t$ and $\tilde{h}_t$ are computed from the probability distribution s_i^t as:

$$\begin{bmatrix} \tilde{h}_t \\ \tilde{c}_t \end{bmatrix} = \sum_{i=1}^{t-1} s_i^t \cdot \begin{bmatrix} h_i \\ x_i \end{bmatrix} \tag{6}$$

Similarly to Equation 1, the input, forget, output and context gates of the LSTMN are then computed as follows:

$$\begin{bmatrix} i_t \\ f_t \\ o_t \\ \hat{c}_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ tanh \end{bmatrix} W \cdot [\tilde{h}_t, x_t] \tag{7}$$

$$c_t = f_t \odot \tilde{c}_t + i_t \odot \hat{c}_t \tag{8}$$

$$h_t = o_t \odot tanh(c_t) \tag{9}$$

Now that we have described mathematically the LSTMN, we will see how it can be used in an encoder-decoder architecture. Similarly to the architecture in Figure 5, the encoder's output will be one part of the decoder's input. Cheng et al. (2016) define two attention mechanisms. The first one, called "shallow attention fusion", is similar to the *inter-attention* introduced by Bahdanau et al. (2014). The other one, called "deep attention fusion", combines inter- and intra-attention, as shown in Figure 4.[6]

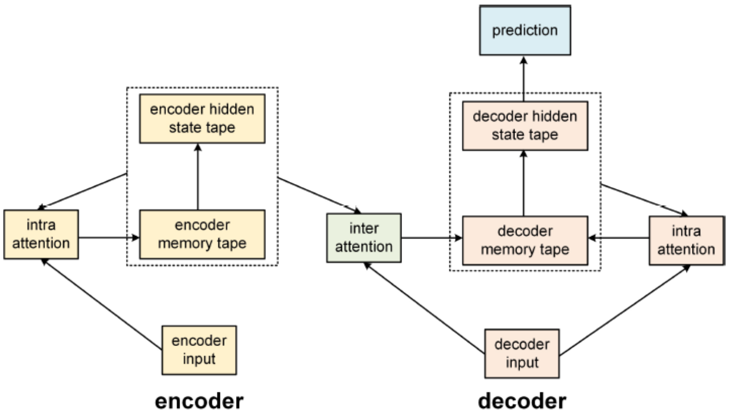


Figure 4. Encoder-Decoder LSTMN architecture with Deep Attention. From "Long Short-Term Memory-Networks for Machine Reading" (Cheng et al., 2016).

[5] "Intra-attention" stands for attention *within* the source or target sentence groups.
[6] "Inter-attention" stands for attention *between* source and target sentence-groups.

Given the encoder's context memory C_t and hidden memory H_t of the source sequence, we can compute the inter-attention between the source sequence and the target token y_t at time t in the decoder as follows:

$$b_i^t = u^T tanh(W_{\tilde{h}} \tilde{h}_i + W_y y_t + W_{\bar{h}} \bar{h}_{t-1}) \tag{10}$$

where y_t is the target token at time step t.

$$p_i^t = softmax(b_i^t) \tag{11}$$

The source's m token representation of the encoder's context memory and hidden memory becomes:

$$\begin{bmatrix} \bar{h}_t \\ \bar{c}_t \end{bmatrix} = \sum_{i=1}^{m} p_i^t \cdot \begin{bmatrix} \tilde{h}_i \\ \tilde{c}_i \end{bmatrix} \tag{12}$$

We then add another gate to the ones found in Equation 7:

$$r_t^{inter} = \sigma(W_r \cdot [\bar{h}_t, y_t]) \tag{13}$$

The decoder's context memory has an extra term from the inter-attention:

$$c_t^{dec} = r_t^{inter} \odot \bar{c}_t^{dec} + f_t^{dec} \odot \tilde{c}_t^{dec} + i_t^{dec} \odot \hat{c}_t^{dec} \tag{14}$$

$$h_t^{dec} = o_t^{dec} \odot tanh(c_t^{dec}) \tag{15}$$

We can then predict the label using a softmax (similarly to Equation 29 below):

$$label = softmax(h^{enc}, h^{dec}) \tag{16}$$

3.1.3 Decomposable Attention with feed-forward networks

We will now describe the second model that we have experimented with, Parikh et al.'s Decomposable Attention model (Parikh et al., 2016). The idea behind this model is similar to the previous model, since it relies on inter- and intra-attention mechanisms to make a prediction. The encoder-decoder networks are replaced entirely with feed-forward networks.

This model caught our interest for two reasons. Firstly, it uses significantly fewer parameters than the Deep Attention LSTMN model, while performing better on the Natural Language Inference task (according to the authors and to our experiments). It also operates largely independently of word order. Indeed, let us note $s1 = (s1_1, \ldots, s1_m)$ and $s2 = (s2_1, \ldots, s2_n)$ the source and target sentences with m and n tokens respectively, and $cl = (cl_1, \ldots, cl_k)$ are the labels at the encoding stage and k the number of output classes. We take F to be a feed-forward neural network with ReLU activations (Glorot et al., 2011). We compute the intra-attention as follows:

$$f_{ij} = F_{intra}(a_i)^T F_{intra}(a_j) \tag{17}$$

where F_{intra} is a feed-forward network. We then create a self-alignment matrix:

$$a'_i = \sum_{j=1}^{m} \frac{\exp(f_{ij})}{\sum_{l=1}^{m} \exp(f_{ik})} a_j \tag{18}$$

The input representation for subsequent steps is then defined as $\bar{a}_i = [a_i, a'_i]$ and analogously $\bar{b}_i := [b_i, b'_i]$. As for the inter-attention, unnormalized attention weights

e_{ij} are computed using via a soft-alignment matrix, a variant of the neural attention (Bahdanau et al., 2014):

$$e_{ij} = F^T_{inter}(\bar{a}_i)F_{inter}(\bar{b}_j) \tag{19}$$

where F_{inter} is a feed-forward network similar to F_{intra}. The normalized attention weights are as follows:

$$\alpha_j = \sum_{i=1}^{m} \frac{\exp(e_{ij})}{\sum\limits_{l=1}^{m} \exp(e_{ik})} \bar{a}_i \tag{20}$$

$$\beta_i = \sum_{j=1}^{n} \frac{\exp(e_{ij})}{\sum\limits_{l=1}^{n} \exp(e_{ik})} \bar{b}_j \tag{21}$$

Again using a feed-forward network G, we compute the alignment vectors for both source and target sentence-groups:

$$v_{1,i} = G([\bar{a}_i, \beta_i]) \quad \forall_i \in [1, \ldots, m] \tag{22}$$

$$v_{1,j} = G([\bar{b}_j, \alpha_j]) \quad \forall_j \in [1, \ldots, n] \tag{23}$$

Before feeding the results through a final feed-forward network to classify target sentences, we sum each alignment vector:

$$v_1 = \sum_{j=1}^{m} v_{1,i} \tag{24}$$

$$v_2 = \sum_{i=1}^{n} v_{1,j} \tag{25}$$

$$prediction = argmax(H([v_1, v_2])) \tag{26}$$

The only drawback from this approach compared to the LSTMN is that the classifier function H may need to be changed for other types of tasks, such as Language Modeling or Machine Translation. The Decomposable Attention model does not make use of Memory Networks as the Deep Attention LSTMN, but holds both source hidden state, target hidden state and the alignments between sentence-groups respectively in a $m \times m$, a $n \times n$ and three $m \times n$ weight matrices.

3.2 Baseline models

3.2.1 Encoder-decoder with stacked LSTMs

We used for comparison an encoder-decoder architecture with stacked bi-directional Long Short-Term Memory (LSTM) decoders on one side and stacked LSTMs on the decoder side. This baseline model is illustrated in Figure 5. While not shown in the figure, it is possible to connect the concatenated outputs of the bi-directional LSTM to another bi-directional LSTM. The output at time step m of the second bi-directional LSTM becomes the input of the second encoder layer.

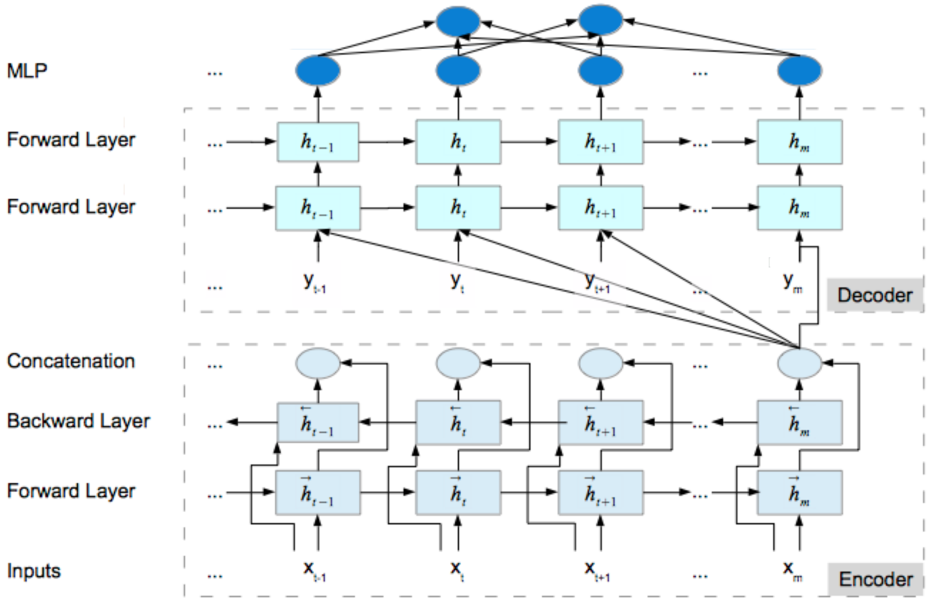


Figure 5. Stacked encoder-decoder bi-directional LSTM model used as a baseline (not all stacked layers are shown).

In comparison with the attention mechanisms of the Deep Attention LSTMN and the Decomposable Attention models, the encoder-decoder baseline model does not have any sort of attention. The encoder-decoder is defined as follows:

Encoder:

$$h_{x(t)} = biLSTM(\overleftarrow{h}_{x(t-1)}, \overrightarrow{h}_{x(t-1)}, x_{t-1}) \tag{27}$$

Decoder:

$$h_{\hat{y}(t)} = LSTM(h_{x(m)}, h_{\hat{y}(t-1)}, \hat{y}_{t-1}) \tag{28}$$

Prediction:

$$label = softmax(h_x, h_y) \tag{29}$$

where $h_{x(t)}$ is the hidden state of the encoder layers at token position t, $h_{y(t)}$ is the hidden state of the decoder layers at token position t, and $\hat{y}_t$ is the prediction of token y_t. The stacked layers allows us build a model with similar capacity to the Deep Attention LSTMN Network. Finally, we used word2vec to initialize the embeddings of both the source sentences and target sentences with pronoun replacement.

3.2.2 Language model

In addition to deep neural networks, we also experimented with a language model by using its perplexity to distinguish correct vs. incorrect replacements of a pronoun by an antecedent. We used the SRILM Language Modeling toolkit (Stolcke, 2002) with a trigram language model that was trained on the Europarl corpus (Koehn, 2002). Using the trained language model on our test data allowed us to extract perplexity of each sentence group with a replacement (i.e. target sentences).

To measure the probability of a sequence, we used the notion of perplexity the field of information theory (Shannon, 1948). Language can be considered to be a discrete information source which is generating a sequence of words $w_1, w_2, \ldots, w_m$ from a vocabulary set, $\mathbb{W}$. The probability of a symbol w_i is dependent upon the previous symbols $w_1, \ldots, w_{i-1}$. The information source's inherent per-word entropy H represents the amount of non-redundant information provided by each new word on average, defined in bits as (assuming ergodicity and a large enough m):

$$\hat{H} = -\frac{1}{m} \log_2 P(w_1, w_2, \ldots, w_m) \tag{30}$$

Perplexity is then defined as:

$$Perplexity = 2^{\hat{H}} \tag{31}$$

As suggested by Popescu (2009), the relationship between the perplexity classes and the prior coreference probability is straightforward. The lower the perplexity, the greater the coreference probability. As we had previously mentioned, each source sentence will have several corresponding target sentences with both correct and incorrect pronoun replacements. To classify the target sentences, we compared the perplexity acroos target sentences generated for each source sentences. Target sentences with the lowest perplexity were labeled as correct replacements. The other targets were classified as incorrect replacements.

Performing coreference resolution based on perplexity has three main limitations.

1. The first one is that perplexity will be higher when a pronoun is replaced by names since the names in question may be out of vocabulary words for the Language Model. When such cases arise, the model may assign a lower perplexity to a noun phrase that it is closer to an example seen during training to the expense of the correct pronoun replacement with the (proper) name of a person, animal or object.

2. The second limitation is that we need at least two generated examples to compare perplexity, since we are labeling correct substitutions as lowest perplexity target sentence. Source sentences with only one corresponding target sentences group were therefore excluded from our final results.

3. The third limitation is that we can have several correct replacements within one source sentence group. To make sure that we find all correct replacements, we assumed that sentence groups within ± 5% of the minimum perplexity should also be labeled as correct replacements.[7]

4 Experiments

In this section we present our experiments for evaluating the performance of a Deep Attention LSTMN (Section 3.1.2) and a Decomposable Attention model (Section 3.1.3) for coreference resolution, defined as distinguishing the replacement of a pronoun with its correct antecedent vs. with a wrong one, leveraging entailment models between the original and the "replaced" sentence groups. It is important to mention that we have focused on the models with deep attention fusion in Equations 10 and 11, since Cheng et al. (2016) reported higher performance on the Natural Language Inference task (i.e. entailment).

[7] We found, through trial and error on the validation set, that a variance of 5% gave an optimal F1 score.

4.1 Performance of the models: accuracy and kappa scores

Label classification accuracy along with model size and the number of epochs are shown in Table 2. Due to resource and time constraints, we were not able to run the Deep Attention LSTMN for more than 30 epochs (which took 2.5 days of training). We observed, though, that when we train it over 100 epochs, the best validation performance of the Decomposable Attention model occurs at epoch 72, with an accuracy of 72.2%, substantially higher than the value of 66.8% appearing in Table 2. It should be noted that randomly selecting between the binary classes would yield only 52.7% accuracy on the test data.

When limiting all systems to training over 30 epochs, the Deep Attention LSTMN performs best and achieves top validation accuracy at epoch 7. The Deep Attention LSTMN appears to learn the relationship between pronouns and antecedents much more quickly but also plateaus at 67.7% accuracy. To make sure that the model was not over fitting, we added a 20% dropout term to non-recurrent connections in the LSTMN, a technique which has proved to reduce over fitting in LSTMs in a variety of NLP tasks (Zaremba et al., 2014).

Using 20% dropout in the Deep Attention LSTMN, the best test accuracy only reaches 64.3% within 30 epochs. The attention mechanism in place helps converge towards higher accuracy. Without inter-attention and using LSTMs to encode and decode sentences, we arrive at 64.0% accuracy. When comparing with other high-capacity models such as the Bi-LSTM encoder-decoder, the test set accuracy decreases by 2.5 percentage points.

Table 2. Accuracy and model sizes. Classification accuracy results on the test set of the coreference task.

Models	**Total Epoch**	**Best Epoch**	**H**	$\|\Theta\|_M$	**Acc.**
SRILM	—	—	—	—	52.7%
Bi-LSTM encoder-decoder	30	19	450	3.2M	64.0%
Deep Attention LSTMN	30	7	450	3.4M	67.5%
Deep Attention LSTMN (20% dropout)	30	26	450	3.4M	64.3%
Decomposable Attention (20% dropout)	30	30	450	1.1M	66.8%

In Table 3 below, we present the performance on coreference resolution considered as a binary classification task (‘correc’ vs. ’incorrect’) in terms of F1 and kappa (κ) scores.[8] The Decomposable Attention model has the highest F1 score, topping the Deep Attention LSTMN by 0.1%.

Since the data is unbalanced, we also check if the results of the models surpass random classification, using Cohen’s kappa metric (Zaremba et al., 2014). The language model approach actually performs below chance, since the observed kappa is negative. On the other hand, the Deep Attention LSTMN and the Decomposable Attention model have similar observed kappa scores, close to 0.2, meaning above-chance classification performance, though far from a reliable human (often required to be above 0.67 or even 0.8). If we look at the True Negatives, we notice that the two models seem to perform better at classifying incorrect pronominal replacements. This is evidenced by the specificity of 68.4% and 68.3% vs. a precision of 31.4% and 32.2% for the Deep Attention LSTMN and the Decomposable Attention model respectively. If we compare models with and without inter-attention, the kappa scores are noticeably higher for models with inter-attention. The Bi-LSTM encoder-decoder architecture has an F1 score

[8] These scores were calculated using http://onlineconfusionmatrix.com/.

3.7 percentage points lower, and a kappa 0.06 points lower than the Deep Attention LSTMN.

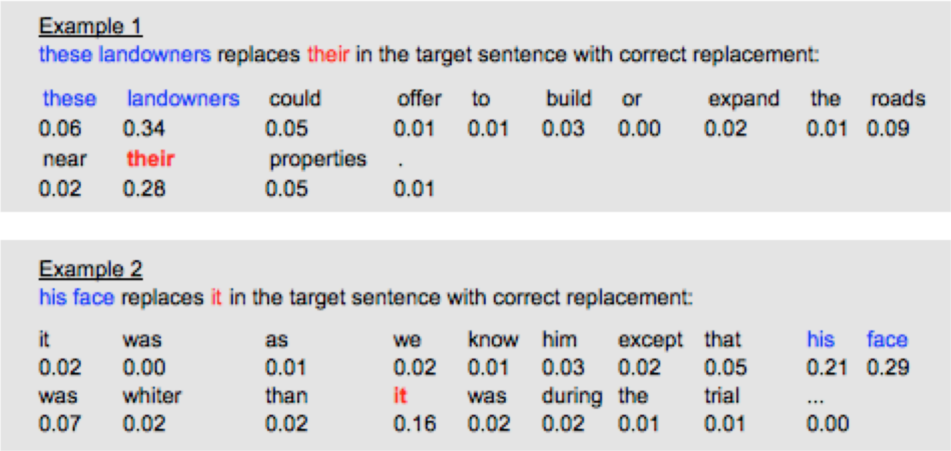
Table 3. Confusion matrix and kappa scores. Checking the statistical importance of our results with F1 scores and Cohen's Kappa.

Models	**TP**	**TN**	**FP**	**FN**	**F1 score**	κ
SRILM	209	844	407	540	30.6%	-0.06
Bi-LSTM encoder-decoder	223	1057	194	526	38.2%	0.16
Deep Attention LSTMN	235	1114	137	514	41.9%	0.22
Decomposable Attention (20% dropout)	241	1095	156	508	42.0%	0.22

4.2 Importance of attention weights

We now turn our analysis towards the importance of intra-attention and inter-attention in the models under study.

First though, we observed that *intra-attention* did not provide much benefits to model accuracy or training convergence speed. When we analyzed the intra-attention scores on test source and target sentences, we were not able to recognize a pattern. This observation was surprising, since the intuition was that the intra-attention would be able to make links between the pronouns and the correct replacements in the source sentence. During forward propagation, the attention layer a_i^t is computed based on the input x_t and hidden states h_i and $\tilde{h}_{t-1}$. During back propagation, the weights in Equation 4, namely W_x, W_h, and $W_{\tilde{h}}$, are updated with the gradient from the error signal from the decoder. However, the error signal does not seem to push the intra-attention layer with any useful updates in this coreference task. On the other hand, inter-attention between the replacement in the target sentences and the words of the source sentences seemed to be pointing towards the pronoun and the antecedent.



Example 1

these landowners replaces their in the target sentence with correct replacement:

these	landowners	could	offer	to	build	or	expand	the	roads
0.06	0.34	0.05	0.01	0.01	0.03	0.00	0.02	0.01	0.09
near	their	properties	.						
0.02	0.28	0.05	0.01						

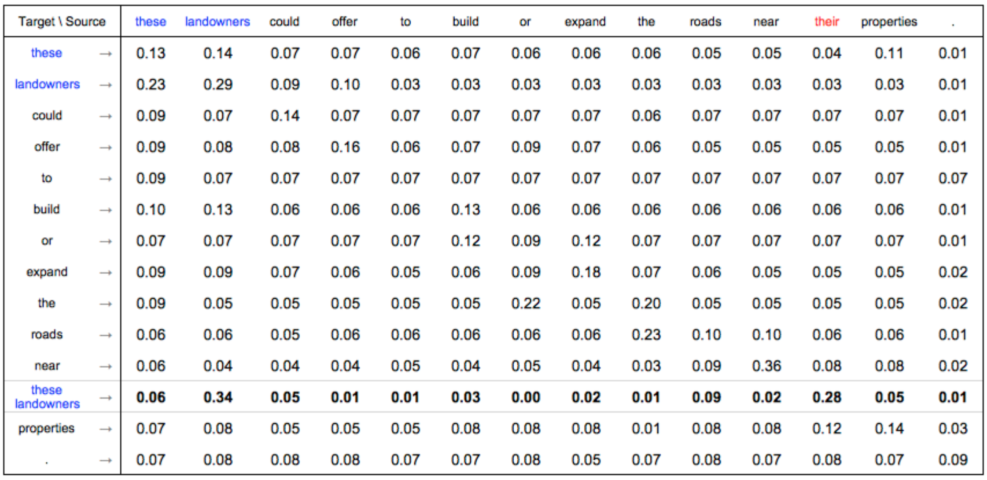
Example 2

his face replaces it in the target sentence with correct replacement:

it	was	as	we	know	him	except	that	his	face
0.02	0.00	0.01	0.02	0.01	0.03	0.02	0.05	0.21	0.29
was	whiter	than	it	was	during	the	trial	...	
0.07	0.02	0.02	0.16	0.02	0.02	0.01	0.01	0.00	

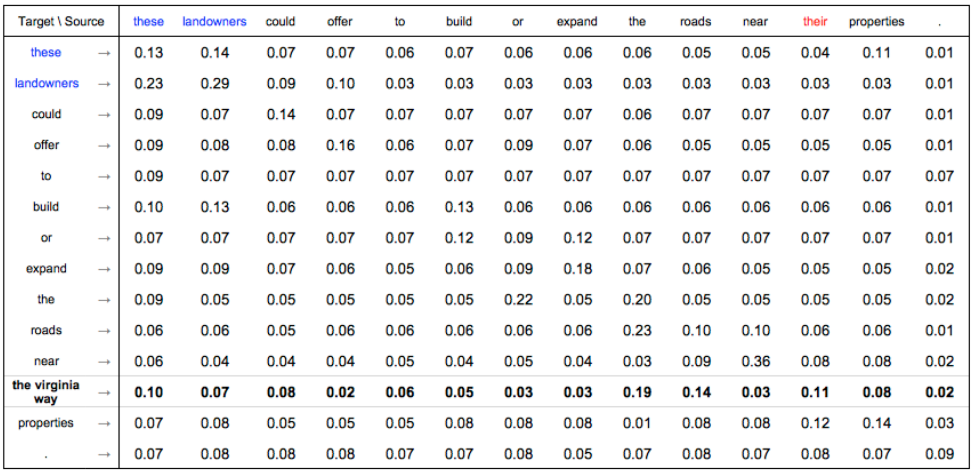
Figure 6. Examples of inter-attention on the source sentence. The correct antecedent is in blue and the pronoun to be replaced is in red. Inter-attention scores from the Deep Attention LSTM-N are found below each word.

From the two examples in Figure 6, we can see that the *inter-attention* score puts more weight on the pronoun in the source sentence. Interestingly, it seems that attention identifies a link between the pronoun and the correct antecedent. If we compare inter-attention scores p_i^t (see Equation 11) for other words in the target sentences in Figure 7a and with the incorrect replacement in Figure 7b, we can see that the inter-attention on the source sentence pronoun is higher for the correct anaphoric mention.



Target \ Source	these	landowners	could	offer	to	build	or	expand	the	roads	near	their	properties	.
these →	0.13	0.14	0.07	0.07	0.06	0.07	0.06	0.06	0.06	0.05	0.05	0.04	0.11	0.01
landowners →	0.23	0.29	0.09	0.10	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.01
could →	0.09	0.07	0.14	0.07	0.07	0.07	0.07	0.07	0.06	0.07	0.07	0.07	0.07	0.01
offer →	0.09	0.08	0.08	0.16	0.06	0.07	0.09	0.07	0.06	0.05	0.05	0.05	0.05	0.01
to →	0.09	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07
build →	0.10	0.13	0.06	0.06	0.06	0.13	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.01
or →	0.07	0.07	0.07	0.07	0.07	0.12	0.09	0.12	0.07	0.07	0.07	0.07	0.07	0.01
expand →	0.09	0.09	0.07	0.06	0.05	0.06	0.09	0.18	0.07	0.06	0.05	0.05	0.05	0.02
the →	0.09	0.05	0.05	0.05	0.05	0.05	0.22	0.05	0.20	0.05	0.05	0.05	0.05	0.02
roads →	0.06	0.06	0.05	0.06	0.06	0.06	0.06	0.06	0.23	0.10	0.10	0.06	0.06	0.01
near →	0.06	0.04	0.04	0.04	0.05	0.04	0.05	0.04	0.03	0.09	0.36	0.08	0.08	0.02
these landowners →	**0.06**	**0.34**	**0.05**	**0.01**	**0.01**	**0.03**	**0.00**	**0.02**	**0.01**	**0.09**	**0.02**	**0.28**	**0.05**	**0.01**
properties →	0.07	0.08	0.05	0.05	0.05	0.08	0.08	0.08	0.01	0.08	0.08	0.12	0.14	0.03
. →	0.07	0.08	0.08	0.08	0.07	0.07	0.08	0.05	0.07	0.08	0.07	0.08	0.07	0.09

(a) Inter-attention scores for a correct replacement.



Target \ Source	these	landowners	could	offer	to	build	or	expand	the	roads	near	their	properties	.
these →	0.13	0.14	0.07	0.07	0.06	0.07	0.06	0.06	0.06	0.05	0.05	0.04	0.11	0.01
landowners →	0.23	0.29	0.09	0.10	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.01
could →	0.09	0.07	0.14	0.07	0.07	0.07	0.07	0.07	0.06	0.07	0.07	0.07	0.07	0.01
offer →	0.09	0.08	0.08	0.16	0.06	0.07	0.09	0.07	0.06	0.05	0.05	0.05	0.05	0.01
to →	0.09	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07
build →	0.10	0.13	0.06	0.06	0.06	0.13	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.01
or →	0.07	0.07	0.07	0.07	0.07	0.12	0.09	0.12	0.07	0.07	0.07	0.07	0.07	0.01
expand →	0.09	0.09	0.07	0.06	0.05	0.06	0.09	0.18	0.07	0.06	0.05	0.05	0.05	0.02
the →	0.09	0.05	0.05	0.05	0.05	0.05	0.22	0.05	0.20	0.05	0.05	0.05	0.05	0.02
roads →	0.06	0.06	0.05	0.06	0.06	0.06	0.06	0.06	0.23	0.10	0.10	0.06	0.06	0.01
near →	0.06	0.04	0.04	0.04	0.05	0.04	0.05	0.04	0.03	0.09	0.36	0.08	0.08	0.02
the virginia way →	**0.10**	**0.07**	**0.08**	**0.02**	**0.06**	**0.05**	**0.03**	**0.03**	**0.19**	**0.14**	**0.03**	**0.11**	**0.08**	**0.02**
properties →	0.07	0.08	0.05	0.05	0.05	0.08	0.08	0.08	0.01	0.08	0.08	0.12	0.14	0.03
. →	0.07	0.08	0.08	0.08	0.07	0.07	0.08	0.05	0.07	0.08	0.07	0.08	0.07	0.09

(b) Inter-attention scores for an incorrect replacement.

Figure 7. Attention scores of correct (a) and incorrect (a) antecedent replacements. The correct antecedent in blue and the pronoun to be replaced in red. Inter-attention scores p_i^t in Equation 11 from the Deep Attention LSTM-N are found below each word. The incorrect replacement noun phrase is in bold and black. Scores should be read from left (target word) to right. The score of the words that replace the pronoun in the target sentence is an average.

Figures 7a and 7b illustrate on two examples that the inter-attention scores are higher on the source sentences pronoun for both correct and incorrect replacements in the target sentences. On average over the test set, we can expect a 69% higher inter-attention score p_i^t when a correct replacement is inserted vs. a incorrect replacement, as shown in Figure 8. Also, the correct replacement noun phrase displays less score variation with a coefficient of variation of 35%, which is 26% lower than the correct replacement noun-phrase's inter-attention score p_i^t. We have noticed that the attention scores of other words on the source sentence are close to three times lower than the score of the correct replacement. This observation suggests that the inter-attention mechanism has a preference for correct replacements and can contribute to identify such anaphoric mentions.

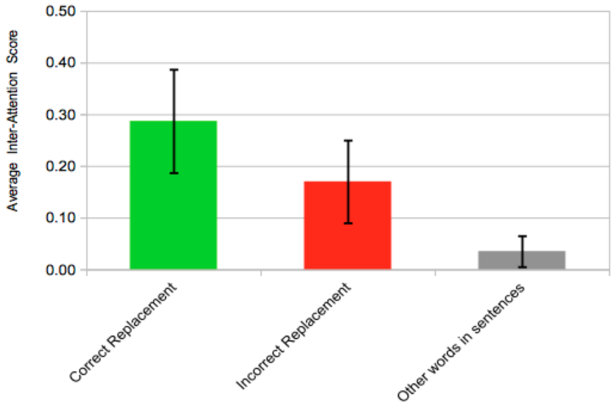


Figure 8. Mean inter-attention of the LSTMN for the anaphoric pronouns of the source sentences. Average values of the inter-attention scores at the position of the noun-phrase replacements and for other words in the target sentences. Scores were computed over the test set.

5 Conclusion

The models that were used in this study provided interesting insights into the links that an inter-attention mechanism can create between a source sentence and target sentence in the context of classifying correct antecedents in a pronominal coreference resolution task. The reference to pronouns in the source sentence allowed the model to classify incorrect and correct pronoun replacements in the target sentences. Higher capacity models in general have performed better on the coreference tasks that we have designed for this experiment. Clearly though, the dependencies learned go further than the SRILM and the bi-LSTM encoder-decoder model since the inter-attention scores allow the model to focus to a certain extent on the link between pronouns and their antecedents. The experiment can be extended by studying the performance of lower capacity models and rule-based models. Moreover, recalculating performance using the MUC, B^3 and $CEAF_{\phi 4}$ measures designed specifically for coreference would allow an explicit comparison with other coreference resolution systems. Applying these measures would however require a substantial refactoring of the task from the entailment format (adapted to models designed for Natural Language Inference) to the proper coreference task. Finally, and since high capacity models typically perform better with data training sets larger than 50k examples, adding more examples could potentially increase performance.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014. URL http://arxiv.org/abs/1409.0473.

Anders Björkelund and Richárd Farkas. Data-driven multilingual coreference resolution using resolver stacking. In *Joint Conference on EMNLP and CoNLL - Shared Task*, CoNLL '12, pages 49–55, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=2391181.2391185.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. *CoRR*, abs/1508.05326, 2015. URL http://arxiv.org/abs/1508.05326.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar F. Zaidan. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, pages 22–64, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-937284-12-1. URL http://dl.acm.org/citation.cfm?id=2132960.2132964.

Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. *CoRR*, abs/1601.06733, 2016. URL http://arxiv.org/abs/1601.06733.

N. Chinchor and P. Robinson. Appendix e: Muc-7 named entity task definition (version 3.5). In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*, 1998. URL http://www.aclweb.org/anthology/M98-1028.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Proceedings of LREC*, 2004.

Greg Durrett and Dan Klein. Easy victories and uphill battles in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL http://proceedings.mlr.press/v15/glorot11a.html.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. ISSN 0899-7667.

Andrew Kehler. The (non)utility of predicate-argument frequencies for pronoun interpretation. In *In Proceedings of 2004 North American chapter of the Association for Computational Linguistics annual meeting*, pages 289–296, 2004.

Philipp Koehn. Europarl: A multilingual corpus for evaluation of machine translation. Draft, 2002.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, CONLL Shared Task '11, pages 28–34, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 9781937284084. URL http://dl.acm.org/citation.cfm?id=2132936.2132938.

K. Lee, L. He, M. Lewis, and L. Zettlemoyer. End-to-end Neural Coreference Resolution. *ArXiv e-prints*, jul 2017.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013. URL http://arxiv.org/abs/1310.4546.

R Mitkov. Anaphora resolution (studies in language and linguistics), 2002.

Thomas S. Morton. Coreference for nlp applications. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, pages 173–180, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics. doi: 10.3115/1075218.1075241. URL http://dx.doi.org/10.3115/1075218.1075241.

Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. *CoRR*, abs/1606.01933, 2016. URL http://arxiv.org/abs/1606.01933.

Octavian Popescu. Name perplexity. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, NAACL-Short '09, pages 153–156, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1620853.1620896.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, CONLL Shared Task '11, pages 1–27, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 9781937284084. URL http://dl.acm.org/citation.cfm?id=2132936.2132937.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W12-4501.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948. ISSN 1538-7305. doi: 10.1002/j.1538-7305.1948.tb01338.x. URL http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x.

Andreas Stolcke. SRILM – an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 901–904, 2002.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. Weakly supervised memory networks. *CoRR*, abs/1503.08895, 2015. URL http://arxiv.org/abs/1503.08895.

Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *CoRR*, abs/1410.3916, 2014. URL `http://arxiv.org/abs/1410.3916`.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *CoRR*, abs/1409.2329, 2014. URL `http://arxiv.org/abs/1409.2329`.

Appendices

A Implementation details

The starting point for this project was the model that Cheng et al. (2016) made public, implemented using the Torch libraries. We chose to use Torch as well and created clones of the LSTM-N encoders and decoders. The cloning process is memory intensive and did not allow us to effectively increase the source and target sentences size. Higher sentences size would allow us to create a larger dataset, with mention pairs that span multiple sentences. To enable this, we have rebuilt the initial code base in PyTorch. The PyTorch implementation did not yield the exact same results on the Language Modeling and Natural Language Inference tasks found in the paper (Cheng et al., 2016). For that reason, we continued using Torch for this experiment.

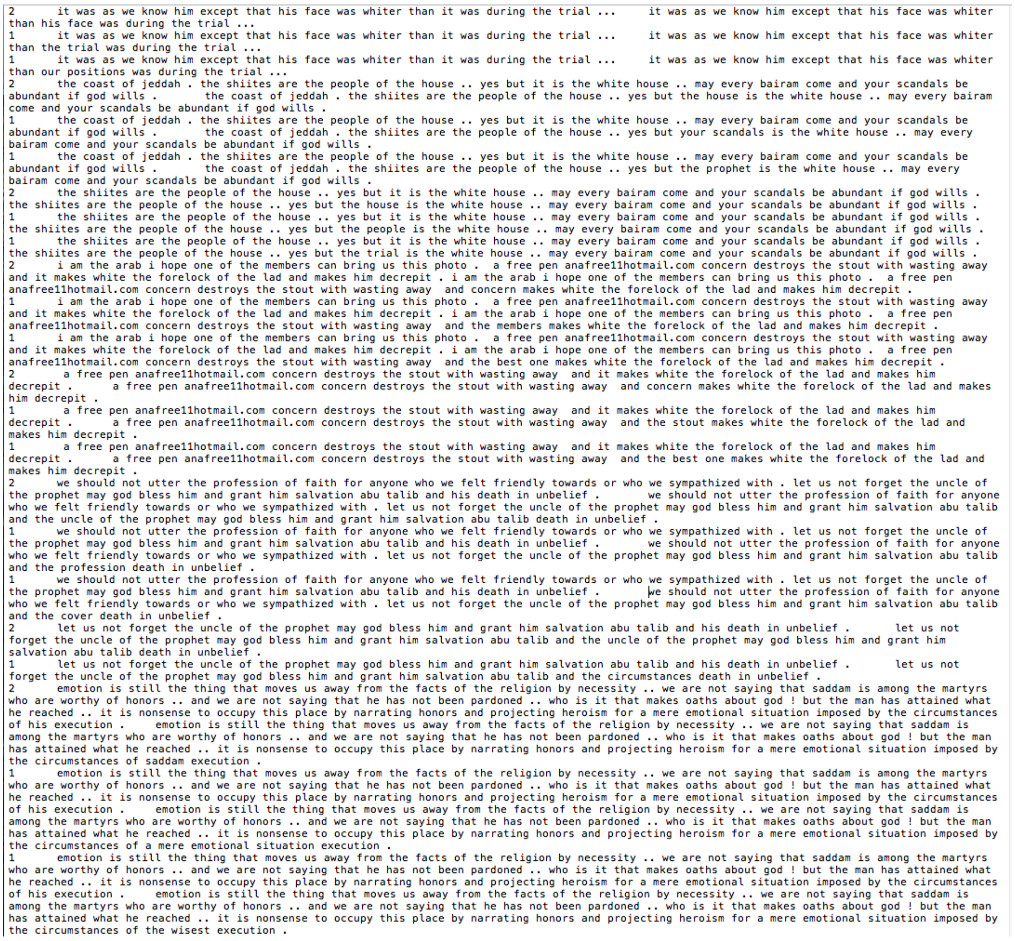
We also experimented with an application of the models to neural MT. For this experiment, OpenNMT helped set up baselines in translation. However, we had to develop our own implementation of beam search. Code will be available on Github at: https://github.com/jpilaul.

B Other possible applications

The Deep Attention LSTM-N architecture can be transformed for a variety of applications. For example, the encoder-decoders are natural fits for Neural Machine Translation (NMT) Tasks. We designed NMT models and trained them using the WMT 2013 Spanish-English dataset Callison-Burch et al. (2011). At the time of writing of this report, the models are still being trained. During the NMT experiments, we found that the softmax operations over large vocabulary were major bottlenecks in training speed. To increase training speed, we are currently integrating a hierarchical softmax operation.

C Supplemental data examples

The figure on the following page is an extract from the test data.



```
2       it was as we know him except that his face was whiter than it was during the trial ...     it was as we know him except that his face was whiter
than his face was during the trial ...
1       it was as we know him except that his face was whiter than it was during the trial ...     it was as we know him except that his face was whiter
than the trial was during the trial ...
1       it was as we know him except that his face was whiter than it was during the trial ...     it was as we know him except that his face was whiter
than our positions was during the trial ...
2       the coast of jeddah . the shiites are the people of the house .. yes but it is the white house .. may every bairam come and your scandals be
abundant if god wills .       the coast of jeddah . the shiites are the people of the house .. yes but the house is the white house .. may every bairam
come and your scandals be abundant if god wills .
1       the coast of jeddah . the shiites are the people of the house .. yes but it is the white house .. may every bairam come and your scandals be
abundant if god wills .       the coast of jeddah . the shiites are the people of the house .. yes but your scandals is the white house .. may every
bairam come and your scandals be abundant if god wills .
1       the coast of jeddah . the shiites are the people of the house .. yes but it is the white house .. may every bairam come and your scandals be
abundant if god wills .       the coast of jeddah . the shiites are the people of the house .. yes but the prophet is the white house .. may every
bairam come and your scandals be abundant if god wills .
2       the shiites are the people of the house .. yes but it is the white house .. may every bairam come and your scandals be abundant if god wills .
the shiites are the people of the house .. yes but the house is the white house .. may every bairam come and your scandals be abundant if god wills .
1       the shiites are the people of the house .. yes but it is the white house .. may every bairam come and your scandals be abundant if god wills .
the shiites are the people of the house .. yes but the people is the white house .. may every bairam come and your scandals be abundant if god wills .
1       the shiites are the people of the house .. yes but it is the white house .. may every bairam come and your scandals be abundant if god wills .
the shiites are the people of the house .. yes but the trial is the white house .. may every bairam come and your scandals be abundant if god wills .
2       i am the arab i hope one of the members can bring us this photo .  a free pen anafree11hotmail.com concern destroys the stout with wasting away
and it makes white the forelock of the lad and makes him decrepit . i am the arab i hope one of the members can bring us this photo .  a free pen
anafree11hotmail.com concern destroys the stout with wasting away  and concern makes white the forelock of the lad and makes him decrepit .
1       i am the arab i hope one of the members can bring us this photo .  a free pen anafree11hotmail.com concern destroys the stout with wasting away
and it makes white the forelock of the lad and makes him decrepit . i am the arab i hope one of the members can bring us this photo .  a free pen
anafree11hotmail.com concern destroys the stout with wasting away  and the members makes white the forelock of the lad and makes him decrepit .
1       i am the arab i hope one of the members can bring us this photo .  a free pen anafree11hotmail.com concern destroys the stout with wasting away
and it makes white the forelock of the lad and makes him decrepit . i am the arab i hope one of the members can bring us this photo .  a free pen
anafree11hotmail.com concern destroys the stout with wasting away  and the best one makes white the forelock of the lad and makes him decrepit .
2        a free pen anafree11hotmail.com concern destroys the stout with wasting away  and it makes white the forelock of the lad and makes him
decrepit .       a free pen anafree11hotmail.com concern destroys the stout with wasting away  and concern makes white the forelock of the lad and makes
him decrepit .
1        a free pen anafree11hotmail.com concern destroys the stout with wasting away  and it makes white the forelock of the lad and makes him
decrepit .       a free pen anafree11hotmail.com concern destroys the stout with wasting away  and the stout makes white the forelock of the lad and
makes him decrepit .
1        a free pen anafree11hotmail.com concern destroys the stout with wasting away  and it makes white the forelock of the lad and makes him
decrepit .       a free pen anafree11hotmail.com concern destroys the stout with wasting away  and the best one makes white the forelock of the lad and
makes him decrepit .
2       we should not utter the profession of faith for anyone who we felt friendly towards or who we sympathized with . let us not forget the uncle of
the prophet may god bless him and grant him salvation abu talib and his death in unbelief .       we should not utter the profession of faith for anyone
who we felt friendly towards or who we sympathized with . let us not forget the uncle of the prophet may god bless him and grant him salvation abu talib
and the uncle of the prophet may god bless him and grant him salvation abu talib death in unbelief .
1       we should not utter the profession of faith for anyone who we felt friendly towards or who we sympathized with . let us not forget the uncle of
the prophet may god bless him and grant him salvation abu talib and his death in unbelief .       we should not utter the profession of faith for anyone
who we felt friendly towards or who we sympathized with . let us not forget the uncle of the prophet may god bless him and grant him salvation abu talib
and the profession death in unbelief .
1       we should not utter the profession of faith for anyone who we felt friendly towards or who we sympathized with . let us not forget the uncle of
the prophet may god bless him and grant him salvation abu talib and his death in unbelief .       we should not utter the profession of faith for anyone
who we felt friendly towards or who we sympathized with . let us not forget the uncle of the prophet may god bless him and grant him salvation abu talib
and the cover death in unbelief .
2       let us not forget the uncle of the prophet may god bless him and grant him salvation abu talib and his death in unbelief .       let us not
forget the uncle of the prophet may god bless him and grant him salvation abu talib and the uncle of the prophet may god bless him and grant him
salvation abu talib death in unbelief .
1       let us not forget the uncle of the prophet may god bless him and grant him salvation abu talib and his death in unbelief .       let us not
forget the uncle of the prophet may god bless him and grant him salvation abu talib and the circumstances death in unbelief .
2       emotion is still the thing that moves us away from the facts of the religion by necessity .. we are not saying that saddam is among the martyrs
who are worthy of honors .. and we are not saying that he has not been pardoned .. who is it that makes oaths about god ! but the man has attained what
he reached .. it is nonsense to occupy this place by narrating honors and projecting heroism for a mere emotional situation imposed by the circumstances
of his execution .    emotion is still the thing that moves us away from the facts of the religion by necessity .. we are not saying that saddam is
among the martyrs who are worthy of honors .. and we are not saying that he has not been pardoned .. who is it that makes oaths about god ! but the man
has attained what he reached .. it is nonsense to occupy this place by narrating honors and projecting heroism for a mere emotional situation imposed by
the circumstances of saddam execution .
1       emotion is still the thing that moves us away from the facts of the religion by necessity .. we are not saying that saddam is among the martyrs
who are worthy of honors .. and we are not saying that he has not been pardoned .. who is it that makes oaths about god ! but the man has attained what
he reached .. it is nonsense to occupy this place by narrating honors and projecting heroism for a mere emotional situation imposed by the circumstances
of his execution .    emotion is still the thing that moves us away from the facts of the religion by necessity .. we are not saying that saddam is
among the martyrs who are worthy of honors .. and we are not saying that he has not been pardoned .. who is it that makes oaths about god ! but the man
has attained what he reached .. it is nonsense to occupy this place by narrating honors and projecting heroism for a mere emotional situation imposed by
the circumstances of a mere emotional situation execution .
1       emotion is still the thing that moves us away from the facts of the religion by necessity .. we are not saying that saddam is among the martyrs
who are worthy of honors .. and we are not saying that he has not been pardoned .. who is it that makes oaths about god ! but the man has attained what
he reached .. it is nonsense to occupy this place by narrating honors and projecting heroism for a mere emotional situation imposed by the circumstances
of his execution .    emotion is still the thing that moves us away from the facts of the religion by necessity .. we are not saying that saddam is
among the martyrs who are worthy of honors .. and we are not saying that he has not been pardoned .. who is it that makes oaths about god ! but the man
has attained what he reached .. it is nonsense to occupy this place by narrating honors and projecting heroism for a mere emotional situation imposed by
the circumstances of the wisest execution .
```

Figure 9. Examples extracted from the test dataset.