Learning Sentence Representation with Guidance of Human Attention

Shaonan Wang†, Jiajun Zhang‡, Chengqing Zong†‡
† National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
‡ CAS Center for Excellence in Brain Science and Intelligence Technology
*shaonan.wang, jjzhang, cqzong@nlpr.ia.ac.cn

Abstract

The most existing sentence representation models typically treat each word in sentences equally. However, extensive studies have proven that human read sentences by making a sequence of fixation and saccades [Rayner 1998], which is extremely efficient. In this paper, we propose two novel approaches, using significant predictors of human reading time, e.g., surprisal and word classes, implemented as attention models to improve representation capability of sentence embeddings. One approach utilizes surprisal directly as the attention weight over baseline models. The other one builds attention model with the help of POS tag and CCG super tag vectors which are trained together with word embeddings in the process of sentence representation learning. In experiments, we have evaluated our models on 24 textual semantic similarity datasets and the results demonstrate that the proposed models significantly outperform the state-of-the-art sentence representation models.

Introduction

Sentence representation aims to project sentence meaning into a uniform high dimensional space where sentences with similar meaning locate close to each other. Hence, sentence meaning become computable which is important for many natural language processing tasks, such as question answering, machine translation and so on. Recently, neural network based sentence representation models have drawn a lot of research attention and show advantages in representing sentence meaning [Le and Mikolov 2014][Kiros et al. 2015][Palangi et al. 2016]. However, the most existing sentence representation models typically treat each word in sentences equally. This is inconsistent with the way that human read and understand sentences, which is reading some words superficially and paying more attention to others. Human attention mechanism is extremely efficient and may lead to a more effective sentence representation model.

In the context of human information processing, attention is a mechanism that enhances some information and inhibits other information at a given moment [Smith and Kosslyn 2007]. Similarly in sentence processing, we would pay more attention to important words because of limited cognitive resources. Extensive studies have proven that reading time contains rich information about the way human process sentences. A range of computational models have been developed to account for human reading behaviors [Jurafsky 2002][Frank 2014][Hahn and Keller 2016]. Recently, there are also works trying to employ human reading time to improve the performance of natural language processing tasks, including sentence compression [Klerke, Goldberg, and Søgaard 2016] and part of speech tagging (POS) [Barrett et al. 2016]. However, to our best knowledge, there is no work of using human reading time to improve sentence representation. This paper is the first study toward this goal.

So far, extensive studies have proven that word surprisal and word attributes like POS tag, Combinatory Categorial Grammar (CCG) super tag, word length, and word frequency are all correlated with human reading time [Dember and Keller 2008][Frank et al. 2013][Barrett and Søgaard 2015]. This paper focuses on two kinds of predictors: surprisal as a continuous variable; POS tags and CCG supertags which are discretely variable. Surprisal, proposed by [Hale 2001] and [Levy 2008], measures the amount of information conveyed by a particular event. Generally, the higher surprisal value corresponds to the higher processing complexity and more reading time. Psycholinguistic experiments have shown that the different word classes like POS tags and CCG supertags are also correlated with human reading time. It is well-known that readers are more likely to fixate on words from open syntactic categories (verbs, nouns, adjectives) than on closed category items like prepositions and conjunctions [Rayner 1998][Nilsson and Nivre 2009]. Because reading time is correlated with human attention mechanism, these findings indicate that the predictors of human reading time mentioned above are crucial for simulating human attention mechanism to improve sentence representation.

In order to integrate surprisal, POS tag and CCG supertag into computational model to learn better sentence representation, we propose two novel approaches which are implemented as attention models, and called attention model with surprisal (ATT-SUR) and attention model with POS tag or CCG supertag (ATT-POS/ATT-CCG), respectively. Concerned on efficiency and complexity, we build our attention models over two baseline models, which stand for the state-of-the-art unsupervised/supervised methods. In our experiments, we have seen the significant improvements over the baselines in nearly all datasets.

To summarize, our main contributions include:
• We present two efficient approaches to use human atten-
tion mechanism to enhance performance of sentence representation, which are implemented as attention models over single sentences. Our approaches are easy to extend to other sentence representation models, and can also be seen as a general framework of integrating predictors of human reading time into sentence representation models.

- We show that the attention calculated by our proposed models is significantly correlated with human reading time, which proves the cognitive plausibility of the proposed attention models.
- We have evaluated our approaches on 24 SemEval datasets, which contain wide range of domains. The results show that our approaches significantly outperform the strong baselines, and some are even competitive with the best SemEval results, which are supervised methods with excessive hand-designed features.

**Related Works**

Conventionally, sentence meaning is represented by the bag-of-words model using abundant hand-crafted features. However, this method is not good enough to model complex interactions within words and sentences. The more recent works show that the neural network based models are particularly good at representing sentence meaning and achieve the state-of-the-art results in several tasks, such as semantic textual similarity task (Tai, Socher, and Manning 2015, He and Lin 2016), text classification (Lai et al. 2015, Chen et al. 2015) and so on. The existing sentence representation models can be summarized into two categories: supervised methods use training data to optimize the model parameters, which are more suitable for in-domain tasks; and unsupervised methods are trained on unlabeled datasets, which are more robust across different tasks (see more experimental support in Wieting et al. (2015)).

Another branch of related works is attention models. The attention mechanism was introduced in neural machine translation to utilize the contextual information of the source sentence in the decoding process (Bahdanau, Cho, and Bengio 2014). Afterwards, more attention models have been proposed in both computer vision and natural language processing. In computer vision area, the core concept of attention models is to focus on the important parts of the input image, instead of giving all pixels the same weight (Wang and Tax 2016). However, the most existing works in natural language processing follow the method of Bahdanau et al. (2014) and employ attention mechanism in tasks which involve sentence pairs modeling, such as machine reading (Hermann et al. 2015) and question answering (Xiong, Merity, and Socher 2016). These works ignore the attention mechanism in a single sentence which plays an important role in human reading behavior. Our work, however, explores attention models over single sentences with guidance of human attention.

The most similar work to our attention models are (Ling et al. 2015) and (Wang et al. 2016). Ling et al. (2015) explored the attention mechanism as weighted sum of context words to predict the target word in learning word representations. The weight corresponds to a set of parameters which determines the importance of each word in each relative position. Wang et al. (2016) employed the same method in sentence representation learning. However, this method gives different weight variable to different word in the different positions, which contains a large number of parameters and may cause the sparsity problem. In this paper, we propose two attention models with predictors of human reading time, which are effective and solve the sparsity problem. One approach assigns each word class (POS tag/CCG supertag) one vector and calculates the attention weight for each word class instead of each word. Another approach utilizes the surprisal information and directly uses surprisal value as the attention weight. Both approaches substantially improve the performance of sentence representation.

**Models**

**Baseline Models**

We build the sentence representation models by extending two baseline models with the proposed attention models. The both baselines use simple averaging model to represent sentences. Given a word sequence $x = <x_1, x_2, ..., x_n>$, the averaging model is described as:

$$g_{\text{sentence}}(x) = \frac{1}{n} \sum_{i=1}^{n} W_{x_i}$$  \hspace{1cm} (1)

Where $n$ is the length of the word sequence. The both baseline models learn the word embedding matrix $W_w$ with equation (1) and differ in training data and objectives. In what follows, we briefly describe the two baseline models.

- The Siamese CBOW (SCBOW) model (Kenter, Borisov, and de Rijke 2016) is an extension of CBOW model (Mikolov et al. 2013) toward sentence level, and predicts the target sentence by its surrounding sentences. This method needs successive sentences as the training corpus, e.g., sentences in the article. With a training objective of categorical cross-entropy same as the CBOW model, the SCBOW model outperforms the baseline additive model (simply average word embeddings in a sentence without updating) and more complex Skip-Thought model which utilizes the encoder-decoder neural network (Kiros et al. 2015).

- The Paragram-Phrase (PP) model (Wieting et al. 2015) trains sentence representation with a contrastive max-margin objective function. The training data consists of a set of phrase pairs $(x_1, x_2)$ from the Paraphrase Database (PPDB) (Ganitkevitch, Van Durme, and Callison-Burch 2013) and negative examples $(t_1, t_2)$ which are the most similar word pairs to $(x_1, x_2)$ generated in a mini-batch during optimization. Wieting et al. (2015) have proven that this method is extremely efficient and even competitive with systems tuned for the particular tasks.

**Our Models**

Inspired by human attention mechanism, we explore three significant predictors of human reading time, surprisal, POS tag and CCG supertag, which are implemented as attention
Attention Model with Surprisal. The concept of surprisal originates from the field of information theory. It is also known as self-information and measures the amount of information conveyed by the target. In language processing, the general form of surprisal is defined as:

$$Surprisal(x_t) = -\log(P(x_t|x_1,...,x_{t-1}))$$  \hspace{1cm} (2)

Where $x_1,...,x_{t-1}$ denote the already processed words and $x_t$ is the next word needs to be processed. Equation (2) shows that the surprisal $s$ at word $x_t$ corresponds to the negative logarithm of the conditional probability of $x_t$ given the sentential context $x_1...x_{t-1}$.

In practice, surprisal can be estimated with various language models ranging from the simplest statistical language models to the neural language models. Each word in a sentence corresponds to a surprisal value which indicates the log probability it should not occur in the context. Based on the assumption that the higher surprisal words convey more information and should gain more attention, we directly use surprisal as the attention weight in sentence representation. The proposed attention model with surprisal is computed as:

$$attention(x_t) = \frac{e^{S(x_t)}}{\sum_{i\in[1,\ldots,t]}e^{S(x_i)}}$$  \hspace{1cm} (3)

ATT-POS (ATT-CCG): Attention Model with POS Tags (CCG Supertags). Human reading behaviour is strongly influenced by the current reading word class. In this paper, we hypothesize that POS tag and CCG supertag of words are helpful in building the attention model for sentence representation. For instance, given a sentence a DT man NN with IN a DT hard JJ hat NN is VBZ dancing VBG, the better representation of the sentence should pay more attention on words with NN, JJ, VB and VBG tags, and less attention on words with DT and IN tags.

Different from the existing models, we assign POS tag (CCG supertag) of each word a vector and learn their values in the training process. The attention weights in sentence representation model are computed by the POS tag (CCG supertag) vectors and the word embedding vectors. The attention model with word classes is described as:

$$attention(x_t) = \frac{e^{W_{\alpha}^t \cdot C_w^t}}{\sum_{i\in[1,\ldots,t]}e^{W_{\alpha}^t \cdot C_w^i}}$$  \hspace{1cm} (4)

Where $W_{\alpha}$ is the word embeddings matrix and $C_w$ is the word classes matrix and both are updated during sentence representation training.

To build sentence representation model over the SCBOW model and the PP model, we use weighted summation of word embeddings instead of baseline averaging model, and the weight is calculated by equation (3) or equation (4). The sentence representation with attention is computed as:

$$g_{sentence}(x) = \frac{1}{n} \sum_{i}attention(x_i)W_{\alpha}^i$$  \hspace{1cm} (5)

Experiments and Results

Datasets

In our experiments, two baseline models use two different datasets. The SCBOW model uses the Toronto Book Corpus which contains 7,087 books collected from the web. The PP model is first trained with all of the PPDB XL datasets for 10 epochs and then trained 10 epochs on the entire PPDB XXL datasets which contains much larger paraphrases.

In evaluation, we use 24 SemEval datasets from semantic textual similarity (STS) task (year 2012-2016). The datasets covered wide range of domains like news, image and video descriptions, glosses, twitter, machine translation evaluation and so on. The STS task seeks to measure the degree of semantic equivalence between two sentences. Similarity score is expressed from complete equivalence to totally unrelated. For instance, two examples from the test dataset are shown below:

Some men are fighting. || Two men are fighting. || 4.250
A man is smoking. || A man is skating. || 0.500

Experimental Settings

In this paper, we use the Stanford POS tagger to assign POS tags and use C&C tools to assign CCG supertags in the training and testing datasets. The dimension of all the word embeddings is 300. The POS tag and CCG supertag vectors in the attention models are initialized randomly with size of 300, by drawing from a normal distribution with $\mu = 0.0$ and $\sigma = 0.01$.

In SCBOW model, we initialize the embedding layer with pre-trained word embeddings. We set the batch size to 100 and the number of negative examples to 2. The model is trained with AdaDelta for one epoch with the initial learning rate of 0.001. For the PP model, we directly use the trained model. For the Additive model in Table 1, we use the same pre-trained word embeddings as SCBOW model and use averaged word embeddings as the sentence representation.

The surprisal is calculated by a n-gram language model with SRILM tools. We use the 5 order n-gram model with modified Kneser-Ney smoothing method. We set surprisal value $x$ as $\min(\max(0, x), 10)$. The language model is trained on the Xinhua news dataset from Linguistic Data Consortium (LDC). We also train language model on the Toronto Book Corpus, but the performance is slightly worse.

1The corpus can be downloaded from [http://www.cs.toronto.edu/˜mbweb/](http://www.cs.toronto.edu/˜mbweb/) and we preprocess the corpus with scripts provided in Kenter et al. (2016).

2We use the 5 order n-gram model with pre-trained word embeddings. We set the batch size to 100 and the number of negative examples to 2. The model is trained with AdaDelta for one epoch with the initial learning rate of 0.001. For the PP model, we directly use the trained model. For the Additive model in Table 1, we use the same pre-trained word embeddings as SCBOW model and use averaged word embeddings as the sentence representation.

3The word embeddings is available at [https://github.com/mmihaltz/word2vec-GoogleNews-vectors](https://github.com/mmihaltz/word2vec-GoogleNews-vectors)

4The trained model is available at [http://ttic.uchicago.edu/˜wieting/](http://ttic.uchicago.edu/˜wieting/).
Table 1: Pearson rank correlation on SemEval textual similarity datasets. The bold scores in each row are the best result in the PP model column and the SCBOW model column, respectively. **Base** denotes baseline model.

|                  | Additive | Semeval-Best | PP       | SCBOW    |
|------------------|----------|--------------|----------|----------|
|                  |          |              | Base     | ATT-SUR  | ATT-SUR  | ATT-POS  | ATT-CCG  |
| MSRpar           | 0.423    | 0.734        | 0.476    | **0.484**| **0.429**| 0.425    | 0.414    | 0.419    |
| MSRvid           | 0.505    | 0.880        | 0.774    | **0.797**| 0.620    | 0.666    | 0.702    | **0.734**|
| OnWN             | 0.634    | 0.727        | 0.714    | **0.724**| 0.687    | 0.680    | 0.695    | **0.696**|
| SMTeurop         | 0.527    | 0.567        | 0.481    | **0.502**| 0.537    | 0.549    | 0.538    | **0.552**|
| SMNews           | 0.476    | 0.609        | 0.652    | **0.657**| 0.523    | 0.524    | 0.541    | **0.557**|
| **2012 Average** | 0.513    | 0.703        | 0.619    | **0.633**| 0.559    | 0.569    | 0.578    | **0.592**|
| FNWN             | 0.286    | 0.582        | 0.476    | **0.500**| 0.378    | 0.370    | 0.383    | **0.392**|
| OnWN             | 0.518    | 0.843        | 0.738    | **0.753**| 0.584    | **0.627**| 0.609    | 0.583    |
| headlines        | 0.649    | 0.784        | 0.733    | **0.746**| 0.693    | 0.705    | 0.704    | **0.711**|
| **2013 Average** | 0.484    | 0.736        | 0.649    | **0.666**| 0.552    | **0.567**| 0.565    | 0.562    |
| OnWN             | 0.612    | 0.875        | 0.812    | 0.819    | 0.686    | **0.713**| 0.705    | 0.693    |
| deft-forum       | 0.360    | 0.531        | 0.540    | **0.552**| 0.400    | 0.411    | 0.406    | **0.421**|
| deft-news        | 0.706    | 0.785        | 0.739    | **0.743**| 0.724    | **0.735**| 0.726    | 0.733    |
| headlines        | 0.647    | 0.784        | 0.707    | **0.719**| 0.652    | 0.655    | 0.666    | **0.668**|
| images           | 0.583    | 0.837        | 0.805    | **0.806**| 0.660    | 0.667    | 0.733    | **0.761**|
| tweets           | 0.648    | 0.792        | 0.769    | **0.777**| 0.708    | 0.689    | **0.754**| 0.723    |
| **2014 Average** | 0.593    | 0.767        | 0.729    | **0.736**| 0.638    | 0.645    | 0.665    | **0.666**|
| ans-forums       | 0.356    | 0.739        | **0.691**| 0.691    | 0.476    | **0.514**| 0.491    | 0.512    |
| ans-students     | 0.677    | 0.788        | 0.781    | **0.784**| 0.723    | 0.722    | 0.715    | **0.732**|
| belief           | 0.570    | 0.772        | 0.773    | **0.782**| 0.584    | 0.591    | 0.600    | **0.603**|
| headlines        | 0.672    | 0.842        | 0.764    | **0.771**| 0.713    | 0.723    | 0.720    | **0.727**|
| images           | 0.672    | 0.871        | 0.837    | **0.839**| 0.740    | 0.746    | 0.773    | **0.787**|
| **2015 Average** | 0.487    | 0.802        | 0.769    | **0.773**| 0.647    | 0.659    | 0.660    | **0.672**|
| answer           | 0.339    | 0.692        | **0.670**| 0.670    | **0.398**| 0.396    | 0.379    | 0.391    |
| deadlines        | 0.642    | 0.827        | 0.699    | **0.711**| 0.683    | 0.700    | **0.705**| 0.701    |
| plagiarism       | 0.615    | 0.841        | 0.802    | **0.817**| 0.708    | 0.711    | **0.736**| **0.736**|
| postediting      | 0.747    | 0.867        | 0.828    | **0.831**| **0.708**| 0.703    | 0.700    | 0.706    |
| question         | 0.487    | 0.747        | 0.535    | **0.546**| 0.580    | 0.627    | 0.623    | **0.652**|
| **2016 Average** | 0.566    | 0.795        | 0.707    | **0.715**| 0.615    | 0.627    | 0.629    | **0.637**|

**Textual Similarity**

Table 1 displays the results of the two baseline models and our models on all 24 textual similarity datasets. We also include two more baselines: the simple Additive model and the best SemEval model. Moreover, we implement the attention model introduced by Ling et al. (2015) over the SCBOW model. We assign POS tag (instead of context position) of each word a variable. But no improvements are found over the baseline SCBOW model. We check the results and find that most values in the word-POS matrix remain unchanged. Hence, the sparsity problem leads to the failure of this method.

From Table 1, we can see that the PP model outperforms the SCBOW model by a large margin. The PP model is trained with PPDB, which is constructed carefully from the big bilingual parallel corpus. Clearly, the dataset in PPDB consists rich semantic information which is extremely helpful for representing sentence meaning. However, SCBOW model uses textual context as training data, which is more easier to get and achieves an average 5 points improvement over the baseline Additive model. As Table 1 shows, the ATT-CCG model is better than ATT-SUR and ATT-POS, which indicates that the word classes with more fine-grained categories like CCG supertag are more suitable for building the attention model to improve sentence representation.

Table 1 also shows that our attention models are quite effective on most datasets. However, there are three exceptions: MSRpar-2012, answer-2016 and postediting-2016. Go back to the testing datasets, we find that sentences in MSRpar-2012 and postediting-2016 are really long, with an average length of 20.63 and 19.36. The answer-2016 dataset consists of large content of daily conversation, where function words are in majority. These characteristics may be the reason why our attention model fails.

Since attention model with surprisal and POS tag (CCG supertag) possesses quite different mechanism, we want to know if they are complementary or not. Therefore, we carried out extra experiments on using surprisal with ATT-POS...
(ATT-CCG) model and show results in Table 2. As it can be seen, the attention models with both surprisal and POS tag (CCG supertag) achieve the better results, which indicate that surprisal and POS tag (CCG supertag) contain different attention information and together they can construct a better attention model.

Table 2: The average pearson rank correlation on 24 SemEval textual similarity datasets. The baseline is SCBOW model and other models are built based on it.

| Baseline  | ATT-SUR  | ATT-POS |
|-----------|----------|---------|
| 0.608     | 0.619    | 0.626   |

| ATT-POS-SUR | ATT-CCG | ATT-CCG-SUR |
|-------------|---------|-------------|
| 0.629       | 0.633   | 0.638       |

Effects of the Attention Models

To delve deeper into how our models work, we compare the results of different attention models. Moreover, we examine the learned POS tag and CCG supertag vectors, and hope to investigate what these dense features capture.

Do ATT-SUR, ATT-POS and ATT-CCG Differ? The ATT-SUR model use surprisal directly as the word attention weight. While the ATT-POS and ATT-CCG model train the POS tag vectors and CCG supertag vectors, respectively, and calculate the attention by dot product with corresponding word embeddings. Due to different mechanism, we expect some difference in their attention value. To compare these attention models, we calculate the words with lowest (highest) attention values by averaging each word’s attention value in the testing data. The results are shown in Table 3.

Table 3: The five words with the lowest (highest) attention values by ATT-SUR, ATT-POS and ATT-CCG model in all test datasets.

| ATT-SUR    | ATT-POS    | ATT-CCG    |
|------------|------------|------------|
| low low    | low low    | low low    |
| pong to    | to skillet  | and skillet |
| donaldson  | dpx and    | uterus to   |
| palace     | dji below  | colander a  |
| basra      | liza       | suction nor |
| misconduct | relents    | foal the   |
|            |            | porch      |

In Table 3, as expected, we can see quite different patterns within different models. For the ATT-SUR, the words with lowest or highest attention do not show difference in word syntactic category. The five most lowest attention words are those often co-occur with a fix previous word. For example, pong often appears as ping pong and donaldson is a surname which is frequently used as william donaldson. The five highest attention words are infrequent words or words occur in the context which are not included in the training corpus. As for the ATT-POS model, five lowest attention words are function words like prepositions and conjunctions, while the highest attention words are content words. Similar results can be found in ATT-CCG model.

To show the attention variation in different word classes more clearly, we also calculate the attention value of each POS tag and CCG supertag by averaging words’ attention within the same tag. Figure 1 and Figure 2 present the relative attention value of each tag. From Figure 1, we can see that prepositions, conjunctions and adverbs have the lowest attention. The pronouns and content words like nouns, foreign words achieve the highest average attention. It is reasonable that pronouns gain high attention because the training corpus contains amount of dialogues which make the pronouns particularly important. Figure 2 shows similar trends on CCG supertags. The content words like nouns gain more attention, while the prepositions and conjunctions gain less attention.

Figure 1: The attention value percentage of the bottom ten POS tags (left) and the top ten POS tags (right)

Figure 2: The attention value percentage of the bottom ten CCG supertags (left) and the top ten CCG supertags (right)

What do POS Tag Vectors and CCG Supertag Vectors Capture? The POS tag vectors and CCG supertag vectors are trained together with word embeddings, and we wonder whether these vectors could carry some semantic information like the word embeddings.

We calculate the five most similar tags to NN which results NNS, PRP$, PSR, NNP and FW. The five most similar tags to IN are RB, TO, CC, UH, JJ. As for the CCG supertags, the five most similar tags to N is N/N, NP, NP[nb]/N, ((S\NP)\(S\NP))/S[b]\NP,
Figure 3: An example sentence from the Dundee corpus with the corresponding POS tag and CCG supertag, together with heatmaps of human first-pass reading time and attention calculated by our models with surprisal, POS and CCG, respectively.

\[ S[ng]\{NP\}/S[for]. \]

It shows that these vectors exhibit certain semantics within POS tags and CCG supertags. For instance, \textit{NN} is most similar to \textit{NNS} and the similar tags with \textit{IN} are all related to function words. The CCG supertag \textit{N} is most similar with \textit{N/N} and \textit{NP}, which are all related to nouns.

**Relation with Human Attention**

We have built attention models using significant predictors of human reading time to improve sentence representation, thus we wonder if the proposed attention models which are learned in large-scale corpus share similarity with human attention. Therefore, we calculate correlations between our proposed attention models (based on SCBOW model) and human reading time in the Dundee corpus (Kennedy et al. 2003). Results are shown in Table 4.

The Dundee corpus is widely used in research on human reading and consists of eye-tracking data for 10 subjects reading 20 newswire articles (about 51,000 words). It contains three measures of reading time:

- First-pass time (RTfpass) measures the total time spent reading a word before the first fixation on any other word. It is also known as gaze duration.
- Go-past time (RTgopast) measures the total time spent reading a word from the first fixation up to the first fixation on a word further to the right. This often includes the reading time on words to the left of the current word.
- Right-bounded time (RTrb) measures the total time spent reading a word before the first fixation on a word further to the right.

From Table 4, we can see that attention calculated by our models is highly correlated with human reading time, which proves the cognitive plausibility of the proposed attention models. Moreover, the ATT-CCG model is the most similar one with human reading time. In addition, Table 1 shows that ATT-CCG model gets the better results than ATT-SUR and ATT-POS. This indicates that predictors which are more correlated with human reading time can lead to a better sentence representation model.

**Conclusion and Future Work**

We have presented two novel approaches to enhance sentence representation models. Experimental evaluations show that our approaches lead to substantial gains in accuracy on nearly all 24 SemEval datasets. This is achieved by tackling predictors of human reading time as attention weights over baseline models.

The proposed approaches are effective and easy to extend to other sentence representation models. We have also shown the highly correlation between the attention calculated by our proposed models and human reading time, which proves the cognitive plausibility of the proposed attention models. Moreover, our models integrate human attention mechanism into computational model of sentence representation, which is a preliminary attempt toward the ultimate goal of simulating human language processing.

An interesting line of future work is to find other useful predictors of human attention. We are also interested in using prediction models of reading time to simulate human attention in representing sentences. Moreover, there is still room for improvement in our architecture, such as how to build the attention model, using other baseline models more than the SCBOW model and the PP model.

**References**

[Bahdanau, Cho, and Bengio 2014] Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by
[Zhu et al. 2015] Zhu, Y.; Kiros, R.; Zemel, R.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of the IEEE International Conference on Computer Vision, 19–27.