Modeling cloze probabilities and selectional preferences with neural networks

Kyröläinen Aki-Juhani, Luotolahti M. Juhani, Hakala Kai & Ginter Filip

akkyro@utu.fi, mjluot@utu.fi, kahaka@utu.fi, figint@utu.fi

University of Turku

There is ample evidence showing that during language processing we generate expectations about upcoming items and differences between expectations and upcoming tokens influence processing as measured by eye movements (Ehrlich & Rayner, 1981) and event-related brain potentials (Kutas & Hillyard, 1984). These studies closely connect to models on selectional preference, i.e. the semantic fit of a given item relative to its context, which have been proposed recently (Baroni & Lenci, 2010; Erk, Padó, & Padó, 2010; Lenci, 2011; Van de Cruys, 2014). Here, we continue this line of research and present an implementation of a neural network model trained on semantic vectors for Finnish. The results based on simplex clauses (subject, verb and object) show that the proposed model is able to successfully capture differences in expectations as measured by cloze probability and differences in selectional preference.

For the purpose of our study, an autoencoder-based neural network architecture (AE) was implemented (Bengio, 2009; Hinton & Zemel, 1994). Our model receives as its input the word vectors of the subject and the verb and attempts to predict the object word vector, which was not given as an input. In this respect, our system could be seen as a mapping in the vector space from subject and verb vectors into their most probable object vectors. The data come from the Finnish Internet Corpus containing approximately 3.7 billion tokens (Kanerva, Luotolahti, Laippala, & Ginter, 2014) and 1.4 million triplets were used to train the model. In our experiment the word vectors were trained with word2vec, 200 floats in length (Mikolov, Chen, Corrado, & Dean, 2013). These word vectors were concatenated to form a single vector, which was fed to a dense neural network layer with hyperbolic tangent activation and an output size of 200, the same size as a word vector. The output of this dense layer was fed to three different dense neural layers, each predicting a word vector (subject, verb, and object). These layers all used hyperbolic tangent as their activation.

The network was trained to minimize the mean squared error of these three vectors and it was implemented in Keras (Chollet, 2015).

To contrast the performance of the AE, we reimplemented a neural binary classifier (BiNN) based on the work of Van de Cruys (2014) in Keras. This model takes as its input the word vectors of a given triplet; subject, verb and the object. Given that this model is a binary classifier it requires both positive and negative examples as its training data. The positive examples were the same triplets which were used to train the AE and the negative examples were generated by randomly swapping the object of the same triplets mirroring our pseudo-disambiguation task setting (see below). The three word vectors were concatenated and then fed to a dense neural layer, with hyperbolic tangent activation function. This output, sized 200, was then fed to a dense layer with an output size of one and a sigmoid activation function limiting the output value between one and zero, suitable for binary prediction. The network was trained with binary cross-entropy as its loss function.

Expectations are traditionally estimated using a cloze task where participants are asked to fill in a blank. A cloze probability then reflects the probability of a particular item produced as a completion (Taylor, 1953). We used an on-line questionnaire to collect sentence completions for 150 transitive verbs and participants were instructed to provide three completions in order of preference. In total, 69 participants across Finland voluntarily took part in the experiment and provided 10,350 completions. Only transitive objects \( n = 9681 \) were included in the analysis and the cloze probability \( M = .19, SD = .08 \) was calculated for the verbs based on their most frequent completion. The results showed that the AE offered a
considerably better fit to the data based on Pearson correlation, $r(148) = .45, p < .001$, compared to the BiNN, $r(148) = .35, p < .001$.

Finally, to evaluate the performance of the neural network models in terms of selectional preference, a pseudo-disambiguation task (see Yarowsky, 1993) was implemented for triplets. This is a binary classification task where an observed triplet in a corpus is considered as a true instance and a false instance is constructed by randomly changing the object in a given triplet. For this task, two test conditions were created containing 2000 verbs. In the first condition, the pseudo-object was selected randomly and in the second condition, the frequency of the attested object and the pseudo-object was matched. Additionally, we implemented a simple fallback ngram model (Ngram) consisting of trigram (SVO), bigram (VO) and unigram frequencies (O). The task was carried out 1000 times in each condition and the results are presented in Table 1.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Avg. accuracy</th>
<th>Lower bound</th>
<th>Upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>AE Matched</td>
<td>0.828</td>
<td>0.816</td>
<td>0.842</td>
</tr>
<tr>
<td>AE Random</td>
<td>0.923</td>
<td>0.914</td>
<td>0.933</td>
</tr>
<tr>
<td>BiNN Matched</td>
<td>0.893</td>
<td>0.883</td>
<td>0.905</td>
</tr>
<tr>
<td>BiNN Random</td>
<td>0.905</td>
<td>0.894</td>
<td>0.916</td>
</tr>
<tr>
<td>Ngram Matched</td>
<td>0.781</td>
<td>0.768</td>
<td>0.794</td>
</tr>
<tr>
<td>Ngram Random</td>
<td>0.906</td>
<td>0.901</td>
<td>0.912</td>
</tr>
</tbody>
</table>

Table 1: Average accuracies of the models in the pseudo-disambiguation task across two conditions.

The AE achieved the highest average accuracy in the random condition and the difference, albeit small, between it and the BiNN was statistically significant, $t(1944.8) = 80.2, p < .001$. However, for the AE there was a drastic drop in average accuracy in the matched condition in contrast to the BiNN. Given that both neural network models were trained on the same semantic vectors, this difference is likely not simply related to frequency distributions but to how much information is available for a given model as the AE was only trained on the subject and verb vectors.

This study offers evidence that distributional models based on semantic vectors can capture, at least, certain aspects of contextual dependencies in production. The predictions of the implemented autoencoder were highly correlated with cloze probability and this model was able to extend to the prediction of selectional preference. At the same time, the results of the proposed model need to be tested in an on-line study such as a sentence reading task. Currently, we are collecting eye-tracking data in sentential context to test this. Conceptually, the implemented model offers an attractive perspective for modeling expectations in production and processing.

Bibliography


