Notation	Definition	Notation	Definition
$\mathbf{w}_i^{\mathrm{M}}$	Final word emb.	$\mathbf{z}_i$	Context emb.
W_f	Conv. weight	B_f	Conv. bias
$\mathbf{w}^{\mathrm{O}}$	Network output	W^{L}	Relation emb.
A^j	Input att.	A^{p}	Pooling att.
G	Correlation matrix		

Table 1: Overview of main notation.

in Figure 1. The input sentence is first encoded using word vector representations, exploiting the context and a positional encoding to better capture the word order. A primary attention mechanism, based on diagonal matrices is used to capture the relevance of words with respect to the target entities. To the resulting output matrix, one then applies a convolution operation in order to capture contextual information such as relevant n-grams, followed by max-pooling. A secondary attention pooling layer is used to determine the most useful convolved features for relation classification from the output based on an attention pooling matrix. The remainder of this section will provide further details about this architecture. Table 1 provides an overview of the notation we will use for this. The final output is given by a new objective function, described below.

3.1 Classification Objective

We begin with top-down design considerations for the relation classification architecture. For a given sentence S, our network will ultimately output some $\mathbf{w}^{\mathrm{O}}$. For every output relation $y \in \mathcal{Y}$, we assume there is a corresponding output embedding W_y^{L}, which will automatically be learnt by the network (dos Santos et al., 2015).

We propose a novel distance function $\delta_\theta(S)$ that measures the proximity of the predicted network output $\mathbf{w}^{\mathrm{O}}$ to a candidate relation y as follows.

$$\delta_\theta(S,y) = \left\| \frac{\mathbf{w}^{\mathrm{O}}}{|\mathbf{w}^{\mathrm{O}}|} - W_y^{\mathrm{L}} \right\| \tag{1}$$

using the L_2 norm (note that W_y^{L} are already normalized). Based on this distance function, we design a margin-based pairwise loss function $\mathcal{L}$ as

$$\begin{aligned} \mathcal{L} &= \big[\delta_\theta(S,y) + (1 - \delta_\theta(S,\hat{y}^-))\big] + \beta\|\theta\|^2 \\ &= \left[1 + \left\| \frac{\mathbf{w}^{\mathrm{O}}}{|\mathbf{w}^{\mathrm{O}}|} - W_y^{L} \right\| - \left\| \frac{\mathbf{w}^{\mathrm{O}}}{|\mathbf{w}^{\mathrm{O}}|} - W_{\hat{y}^-}^{\mathrm{L}} \right\|\right] \\ &\quad + \beta\|\theta\|^2, \end{aligned} \tag{2}$$

where 1 is the margin, β is a parameter, $\delta_\theta(S,y)$ is the distance between the predicted label embedding W^{L} and the ground truth label y and $\delta_\theta(S,\hat{y}^-)$ refers to the distance between $\mathbf{w}^{\mathrm{O}}$ and a selected incorrect relation label $\hat{y}^-$. The latter is chosen as the one with the highest score among all incorrect classes (Weston et al., 2011; dos Santos et al., 2015), i.e.

$$\hat{y}^- = \operatorname*{argmax}_{y' \in \mathcal{Y}, y' \neq y} \delta(S,y'). \tag{3}$$

This margin-based objective has the advantage of a strong interpretability and effectiveness compared with empirical loss functions such as the ranking loss function in the CR-CNN approach by dos Santos et al. (2015). Based on a distance function motived by word analogies (Mikolov et al., 2013b), we minimize the gap between predicted outputs and ground-truth labels, while maximizing the distance with the selected incorrect class. By minimizing this pairwise loss function iteratively (see Section 3.5), $\delta_\theta(S,y)$ are encouraged to decrease, while $\delta_\theta(S,\hat{y}^-)$ increase.

3.2 Input Representation

Given a sentence $S = (w_1, w_2, ..., w_n)$ with marked entity mentions $e_1(=w_p)$ and $e_2(=w_t)$, $(p, t \in [1, n]$, $p \neq t)$, we first transform every word into a real-valued vector to provide lexical-semantic features. Given a word embedding matrix W_V of dimensionality $d_{\mathrm{w}} \times |V|$, where V is the input vocabulary and d_{w} is the word vector dimensionality (a hyper-parameter), we map every w_i to a column vector $\mathbf{w}_i^{\mathrm{d}} \in \mathbb{R}^{d_{\mathrm{w}}}$.

To additionally capture information about the relationship to the target entities, we incorporate word position embeddings (WPE) to reflect the relative distances between the i-th word to the two marked entity mentions (Zeng et al., 2014; dos Santos et al., 2015). For the given sentence in Fig. 1, the relative distances of word "and" to entity e_1 "drinks" and e_2 "diabetes" are -1 and 6, respectively. Every relative distance is mapped to a randomly initialized position vector in $\mathbb{R}^{d_{\mathrm{p}}}$, where d_{p} is a hyper-parameter. For a given word i, we obtain two position vectors $\mathbf{w}_{i,1}^{\mathrm{p}}$ and $\mathbf{w}_{i,2}^{\mathrm{p}}$, with regard to entities e_1 and e_2, respectively. The overall word embedding for the i-th word is $\mathbf{w}_i^{\mathrm{M}} = [(\mathbf{w}_i^{\mathrm{d}})^\intercal, (\mathbf{w}_{i,1}^{\mathrm{p}})^\intercal, (\mathbf{w}_{i,2}^{\mathrm{p}})^\intercal]^\intercal$.

Using a sliding window of size k centered around the i-th word, we encode k successive