Dating Document using Graph Convolution Networks

Shikhar Vashishth, Shib S. Dasgupta, Swayambhu N. Ray, Partha Talukdar
Indian Institute of Science, Bangalore

Document Dating

Document Dating is the problem of automatically predicting the creation time of a document (called as Document Creation Time or DCT) based on its content.

Swiss adopted that form of taxation in **1995**. The concession was approved by the govt last September. **Four years after**, the IOC….

Document Creation Time (DCT)
[1999]

Motivation

Document creation time is at the core of many important tasks, such as, *information retrieval*, *temporal reasoning*, *text summarization*, *event detection*, and *analysis of historical text* etc. In all such tasks, the document date is assumed to be available and also accurate – a strong assumption, especially for arbitrary documents from the Web.

Contributions

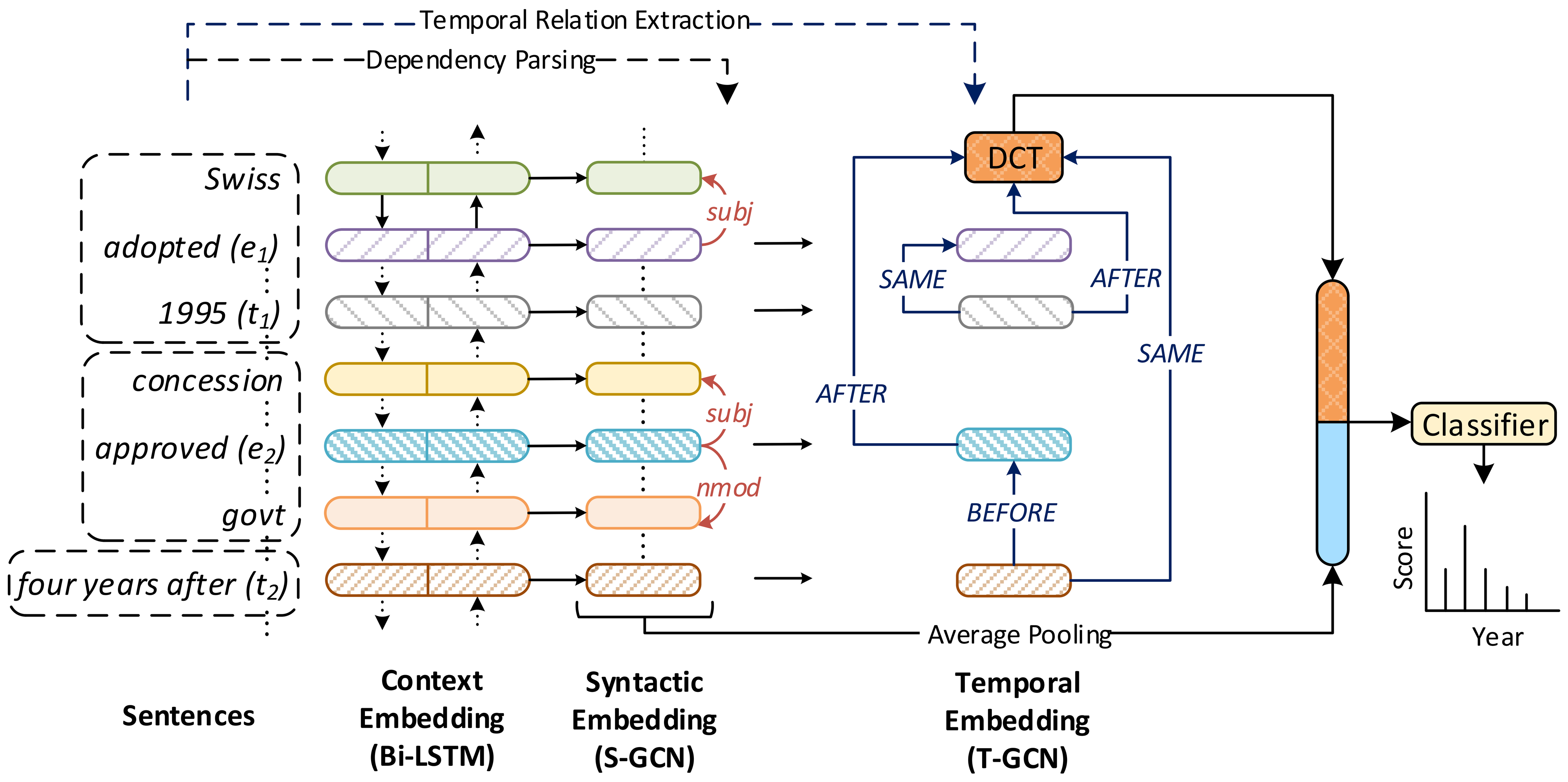
1. We propose NeuralDater, a Graph Convolution Network (GCN) based approach for document dating.
2. NeuralDater exploits syntactic as well temporal structure of the document, all within a principled joint model.
3. Through extensive experiments on multiple real-world datasets, we demonstrate NeuralDater's effectiveness over state-of-the-art baselines.

Source Code

The source code is available at:
`http://github.com/malllabiisc/neuraldater`

Contact: `shikhar@iisc.ac.in`

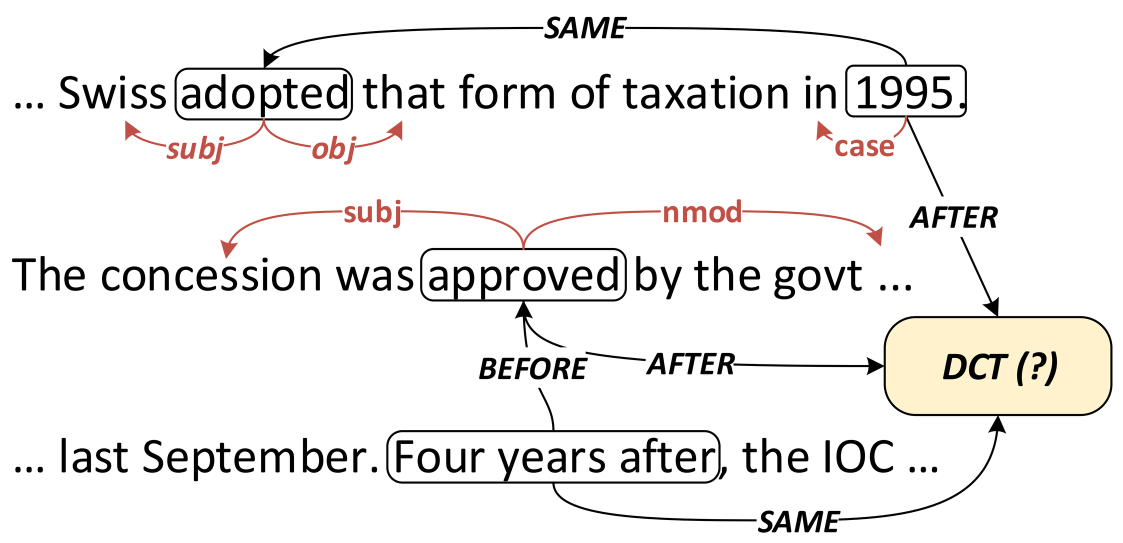
NeuralDater Overview

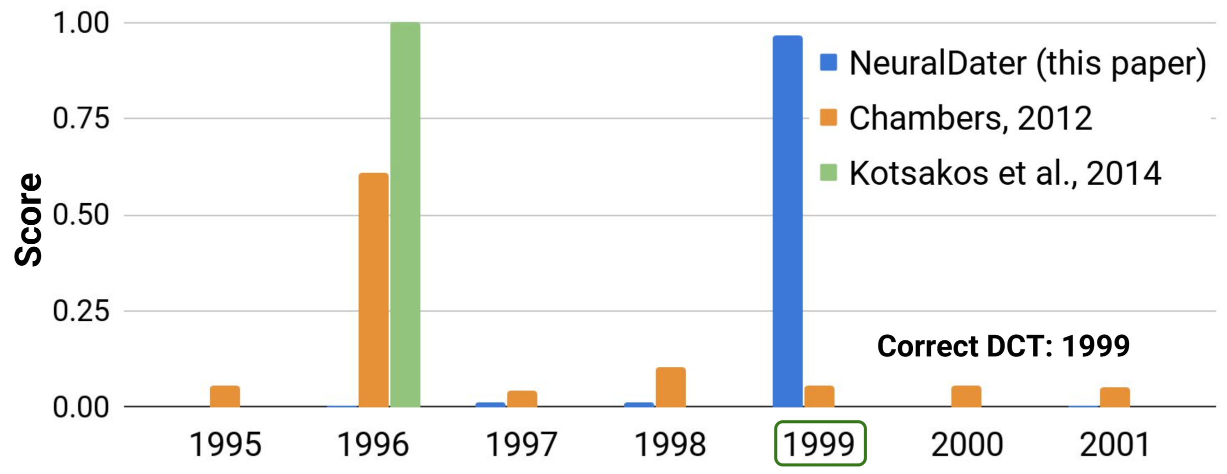


NeuralDater's architecture consists of four components:

1. **Context embedding:** Uses Bi-LSTM to encode context of each token in the document.
2. **Syntactic Embedding:** Employs GCN over dependency parse to encode syntactic information.
3. **Temporal Embedding:** Reasoning over temporal graph is performed using GCN.
4. **Classifier:** DCT embedding with averaged syntactic embedding are used for final prediction.

Method





Inference over temporal and syntactic graph structure of document allows NeuralDater to predict the correct document creation year 1999, whereas previous methods get misled by temporal expression 1995. Equation used for updating embeddings in GCN:

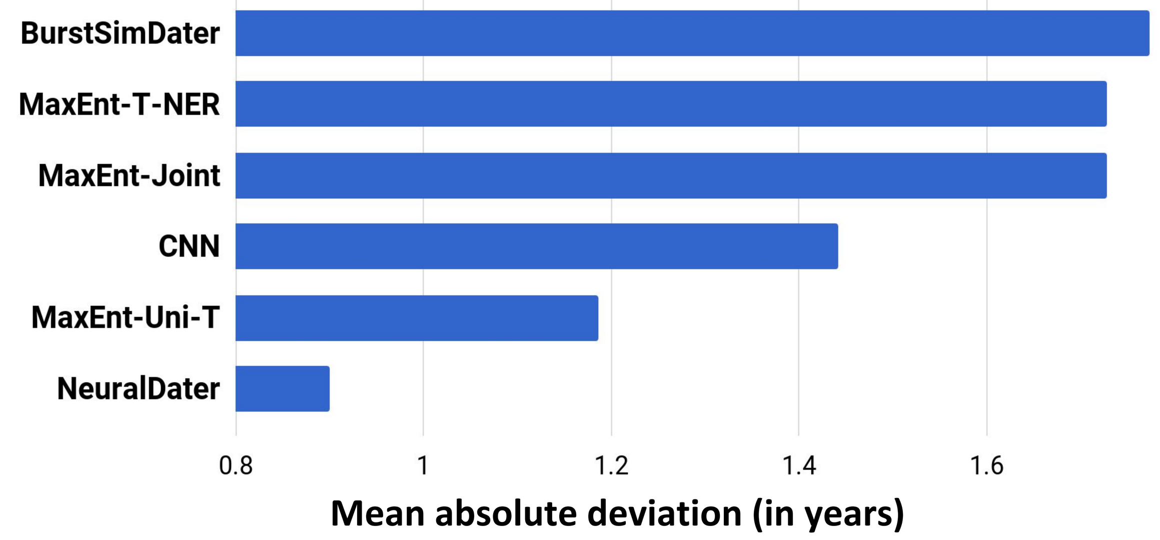
$$h_v^{k+1} = f\left(\sum_{u\in\mathcal{N}(v)} g_{l(u,v)}^k \times \left(W_{l(u,v)}^k h_u^k + b_{l(u,v)}^k\right)\right)$$

Here, $W_{l(u,v)}^k$ and $b_{l(u,v)}^k$ represent label dependent parameters of k^{th} GCN layer. $g_{l(u,v)}^k$ is the gating value and f is activation function used. $\mathcal{N}(v)$ represents the set of neighbors of node v and h_u^k is embedding of node u at k^{th} layer.

Results

Comparison of accuracy of different methods on APW and NYT datasets at year level granularity.

Method	APW	NYT
BurstySimDater	45.9	38.5
MaxEnt-Time+NER	52.5	42.3
MaxEnt-Joint	52.5	42.5
MaxEnt-Uni-Time	57.5	50.5
CNN	56.3	50.4
NeuralDater	**64.1**	**58.9**



NeuralDater significantly outperforms all other competitive baselines in terms of overall accuracy and mean absolute deviation.

Ablation Results

Accuracy of ablated version of NeuralDater for justifying importance of different components.

Method	Accuracy
T-GCN	57.3
S-GCN + T-GCN ($K=1$)	57.8
S-GCN + T-GCN ($K=2$)	58.8
S-GCN + T-GCN ($K=3$)	**59.1**
Bi-LSTM	58.6
Bi-LSTM + CNN	59.0
Bi-LSTM + T-GCN	60.5
Bi-LSTM + S-GCN + T-GCN (no gate)	62.7
Bi-LSTM + S-GCN + T-GCN ($K=1$)	**64.1**
Bi-LSTM + S-GCN + T-GCN ($K=2$)	63.8
Bi-LSTM + S-GCN + T-GCN ($K=3$)	63.3

References

[1] Nathanael Chambers. Labeling documents with timestamps: Learning from their time expressions. In *ACL '12*
[2] D. Kotsakos, T. Lappas, D. Kotzias, D. Gunopulos, N. Kanhabua, and K. NÃÿrvag. A burstinessaware approach for document dating. In *SemEval '15*