NLP 2020

Project presentation

Natural Language Processing A.Y 19/20

Andrea Bacciu

Professor: Roberto Navigli

Teaching assistants:

Edoardo Barba, Niccolò Campolungo, Simone Conia,

Caterina Lacerra, Luigi Procopio

Sapienza University of Rome

Named Entity Recognition - NER

Homework 1

Named Entity Recognition - NER

NER aims to Identify and classify the Named Entities

Classification into 4 classes:

PERson, **ORG**anization, **LOC**ation, **O**ther

John went to California to visit Google

John	went	to	California	to	visit	Google
PER	**O**	**O**	**LOC**	**O**	**O**	**ORG**

SAPIENZA

Task information

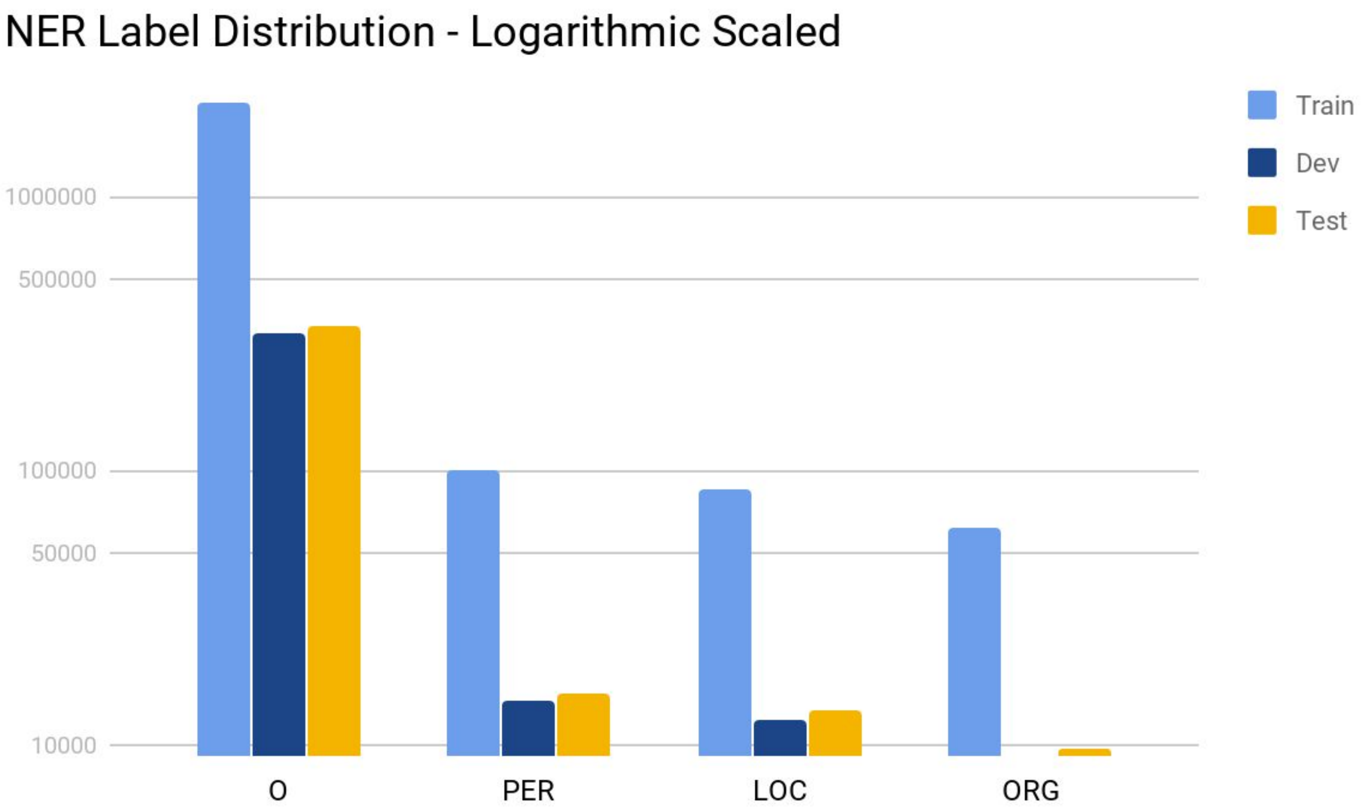
Dataset:
- English
- Public
- Metrics: Macro F1-Score
- Splitted in train, dev, test
- Already tokenized

Entity	Train	Dev	Test
O	2,177,423	315,809	335,567
PERSON	100,409	14,396	15,349
LOCATION	84,937	12,359	13,475
ORGANIZATION	61,988	9,043	9,645

Training:
- Optimizer: Adam
- Learning Rate = 0,001
- Loss function: Cross Entropy
- Epochs: 3
- Batch Size: 32
- DEV F1 early stopping

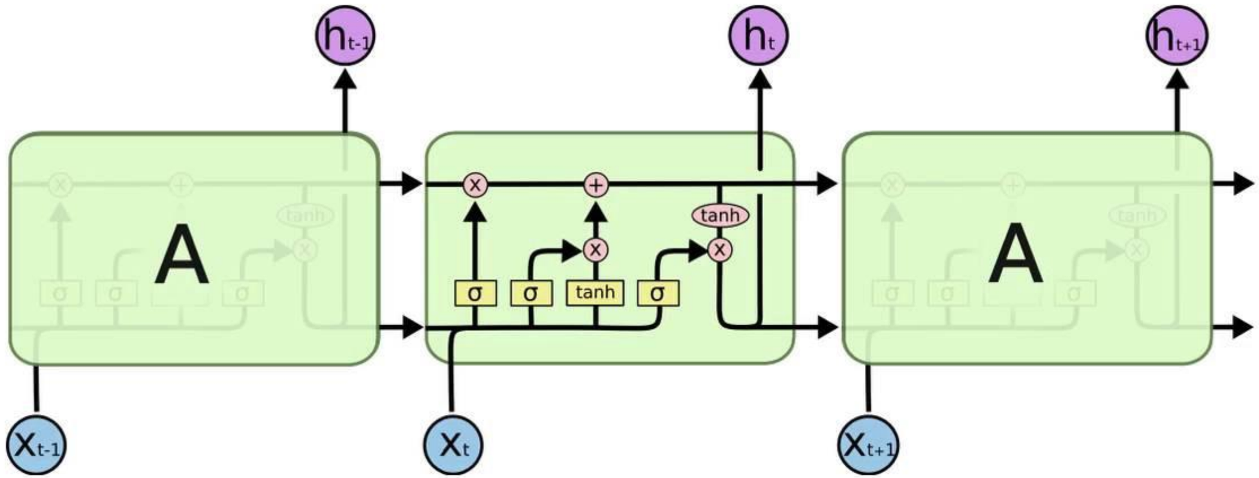
SAPIENZA

NER Label Distribution



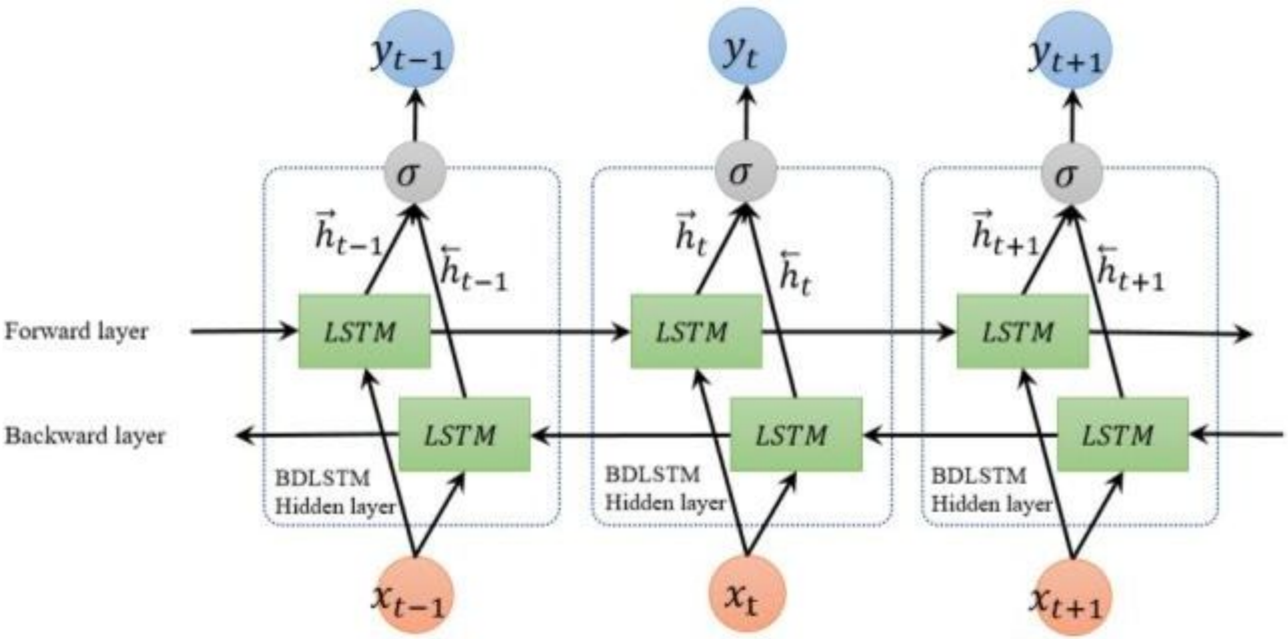
Word 2 Vec

LSTM - Long Short Term Memory



Able to capture long range dependency incorporating memory cell.

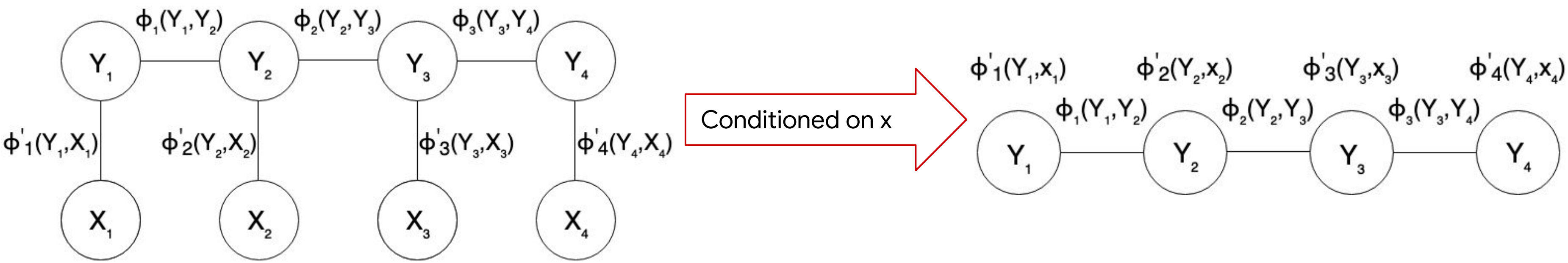
BiLSTM - Bidirectional LSTM



Able to combine the Forward and Backward pass to produce a context-aware output

Conditional Random Field - CRF

CRF structure



CRF helps improve the model's performance **when the current prediction is affected by the previous ones.**

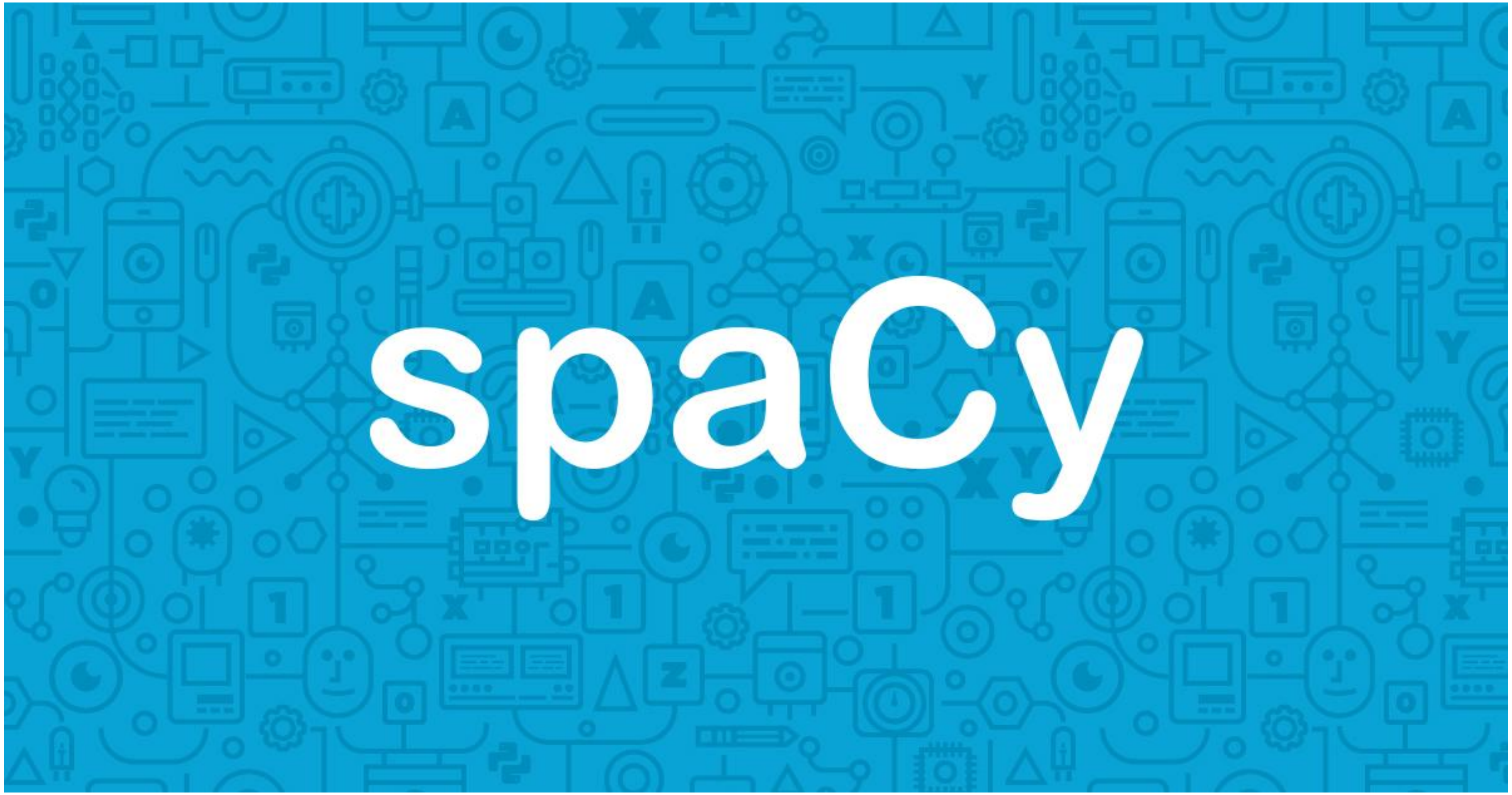
Because the CRF is a **discriminative** model, which are a supervised model class, capable of inferring knowledge from the observed data.

It models the conditional probability P(Y/X) i.e. X is always given or observed. Therefore the graph ultimately reduces to a simple chain.

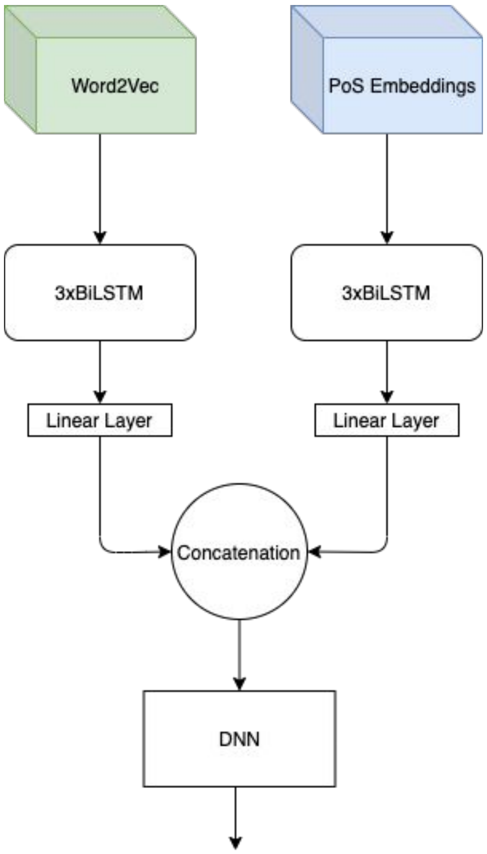
Part of Speech Embeddings

- PoS offer high quality information for the NER task.
- High correlation between labels and PoS tokens.
- I use the SpaCy model to extract the Part of Speech

PoS embeddings with SpaCy



NER Final Model

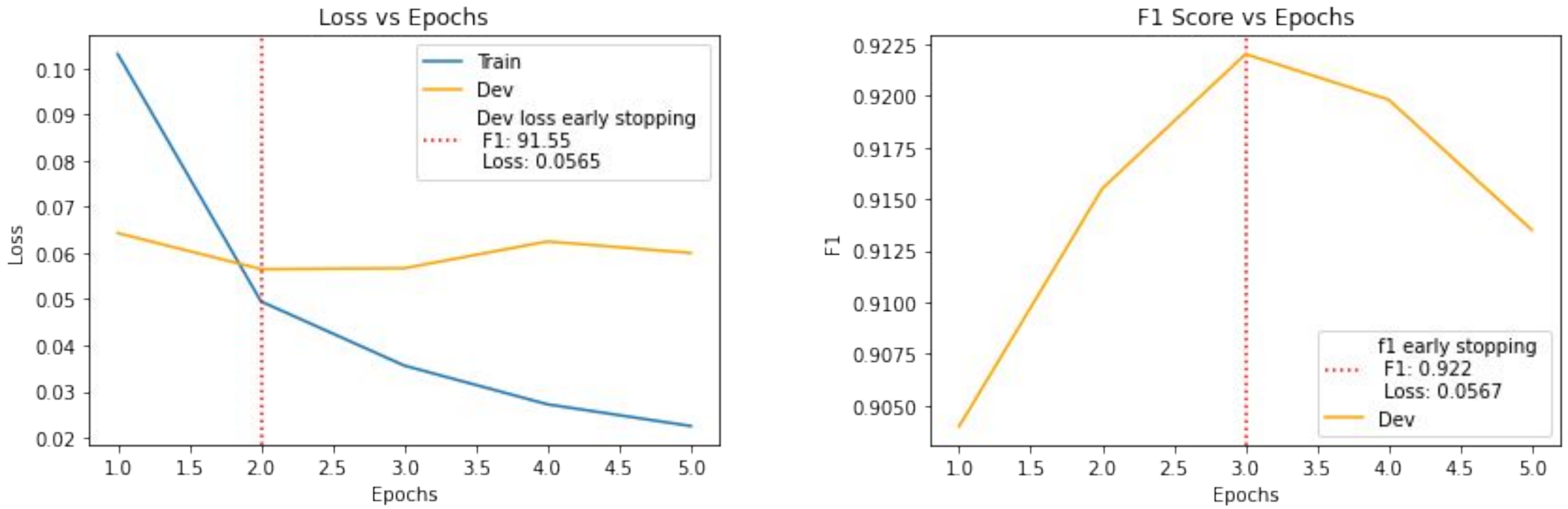


SAPIENZA

NER Hyperparameters

HParams	Value	Notes
Epochs	3	
Batch Size	32	
Optimizer	Adam	
Learning Rate	0.001	
Loss Function	Cross Entropy	
Dropout Embeddings	30%	
Word Emb dim	300	Word2Vec
Pos Emb dim	300	
Dropout BiLSTM	30%	
BiLSTM POS	300 out dim	x3 layer
BiLSTM Words	300 out dim	x3 layer
Linear Layer POS	300	
Linear Layer Words	300	
DNN Linear layer 1	150	Top
DNN Linear layer 2	75	Mid
DNN Linear layer 3	75	Bottom
DNN Linear layer 4	num class	output layer

Early stopping criteria



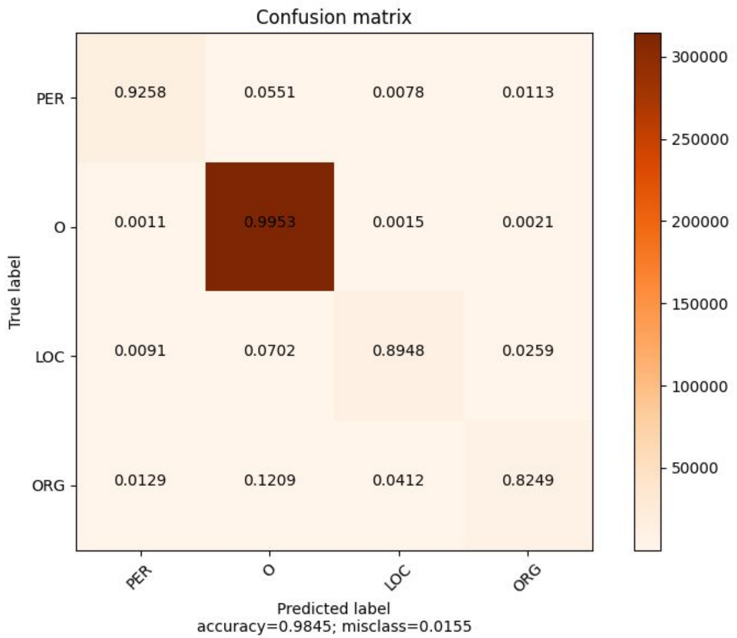
Results

Model	F1 - DEV	F1 - TEST
Baseline	24.88%	24.97%
Word Emb (100d) + BiLSTM	78.02%	78.73%
Word2vec + BiLSTM	80.25%	81.04%
Word2vec + BiLSTM + CRF	88.12%	89.76%
(POS emb + BiLSTM) + (W2V + BiLSTM) + CRF	92.11%	92.37%
(POS emb + BiLSTM) + (W2V + BiLSTM)	92.20%	92.66%

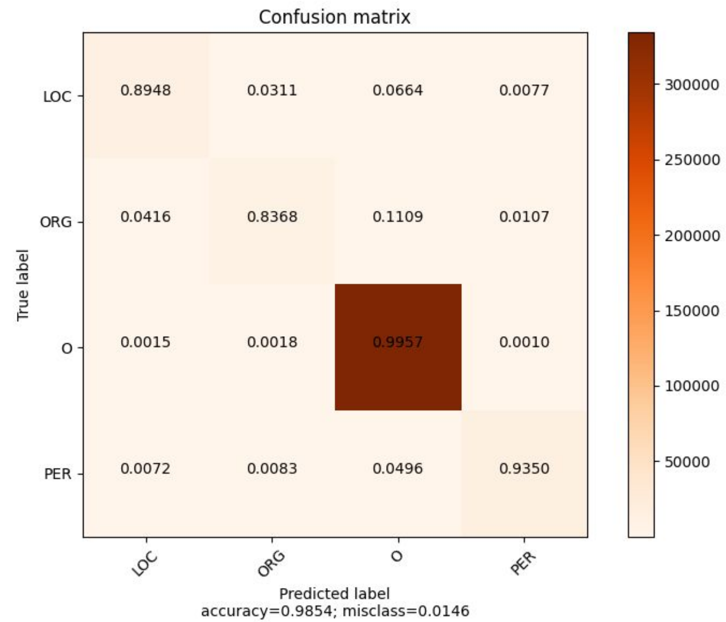
N.B: In the CRF experiments I used the Negative Log Likelihood Loss

Confusion Matrix: Dev vs Test

DEV



TEST



- The ORG class has worst predictions.
- The O class is better predicted.

Semantic Role Labeling

Homework 2

Semantic Role Labeling (SRL)

SRL is the task of addressing

"Who did What to Whom, How, Where and When?"

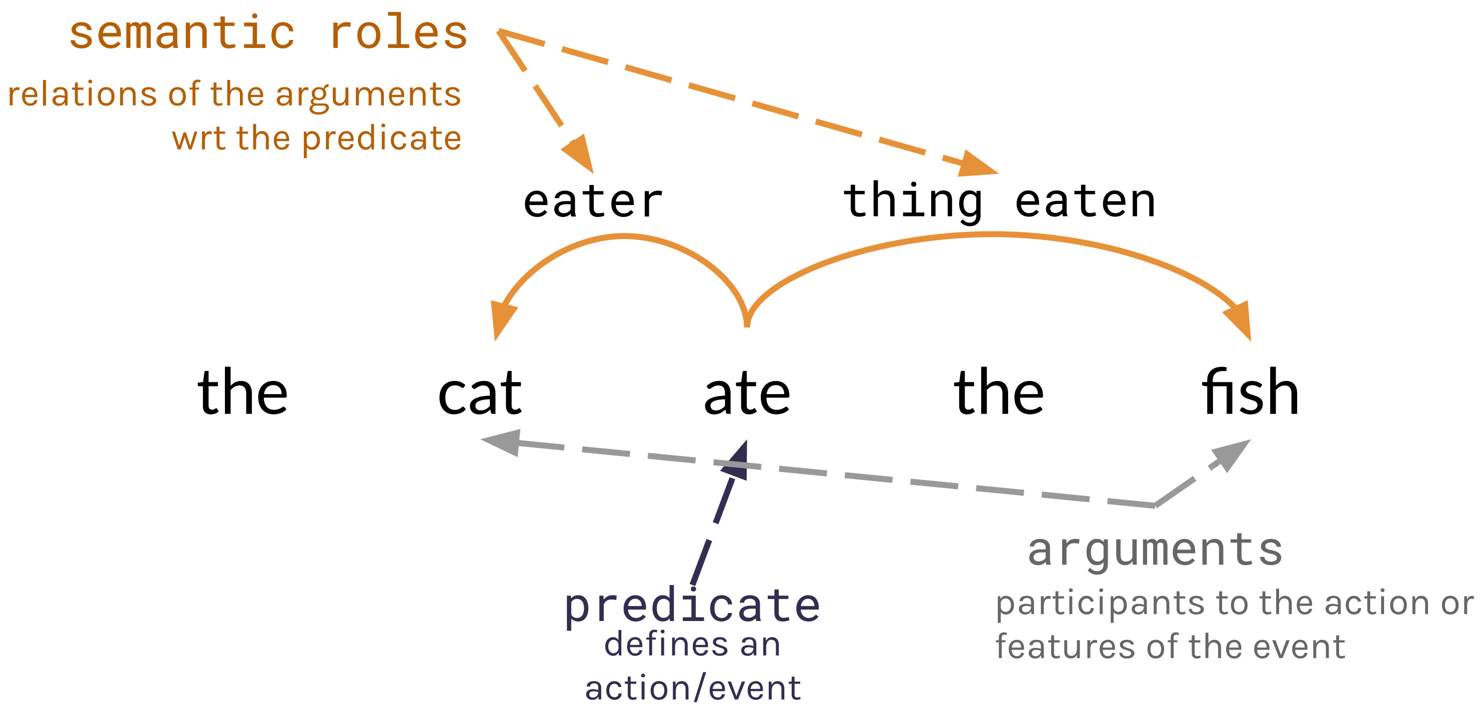
the cat ate the fish

- Who = the cat
- Did what = ate
- Whom = the fish

SRL: predicates, arguments and roles

SRL is the task of addressing

“Who did What to Whom, How, Where and When?”



The SRL pipeline

1 — PREDICATE IDENTIFICATION
Identify the predicate(s) in the target sentence

The cat **ate** the fish.

2 — PREDICATE DISAMBIGUATION
Associate the predicate with its sense

The cat **ate** the fish.

ate → **EAT/BITE**
→ CORRODE
→ CONSUME
→ ...

3 — ARGUMENT IDENTIFICATION
Identify the arguments of the predicate

The **cat** ate the **fish**.

4 — ARGUMENT CLASSIFICATION
Associate each argument with its class

The **cat** ate the **fish**.

cat → Agent
fish → Patient

The SRL pipeline

1. PREDICATE IDENTIFICATION
Identify the predicate(s) in the target sentence

The cat **ate** the fish.

2. PREDICATE DISAMBIGUATION
Associate the predicate with its sense

The cat **ate** the fish.

ate → **EAT/BITE**
→ CORRODE
→ CONSUME
→ ...

3. ARGUMENT IDENTIFICATION
Identify the arguments of the predicate

The **cat** ate the **fish**.

4. ARGUMENT CLASSIFICATION
Associate each argument with its class

The **cat** ate the **fish**.

cat → Agent
fish → Patient

Task information

Dataset:

- English
- Private from Sapienza NLP group
- Metrics: Macro F1-Score
- Splitted in train, dev, test
- Already tokenized
- Features: PoS, Lemma, Words, Dep. Relations, Dep Heads, Predicates

Training

- Optimizer: Adam
- Learning Rate = 0,001
- Loss function: Cross Entropy
- Epochs: 13
- Batch Size: 32
- Min Frequency = 2 on words and lemmas vocabulary

SAPIENZA

Sample

```
sentence_id: {
	“words”: [“The”, “cat”, “ate”, “the”, “fish”, “and”, “drank”, “the”, “milk”, “.”],
	“lemmas”: [“the”, “cat”, “eat”, “the”, “fish”, “and”, “drink”, “the”, “milk”, “.”],
	“pos_tags”: [“DET”, ..., “PUNCT”],
	“dependency_relations”: [“NMOD”, ..., “ROOT”, ..., “P”],
	“dependency_heads”: [1, 2, 0, ...],
	“predicates”: [“_”, “_”, “EAT_BITE”, “_”, “_”, “_”, “DRINK”, “_”, “_”, “_”],
	“roles”: {“2”: [“_”, “Agent”, “_”, “_”, “Patient”, “_”, “_”, “_”, “_”, “_”],
	          “6”: [“_”, “Agent”, “_”, “_”, “_”, “_”, “_”, “_”, “Patient”, “_”]}
}
```

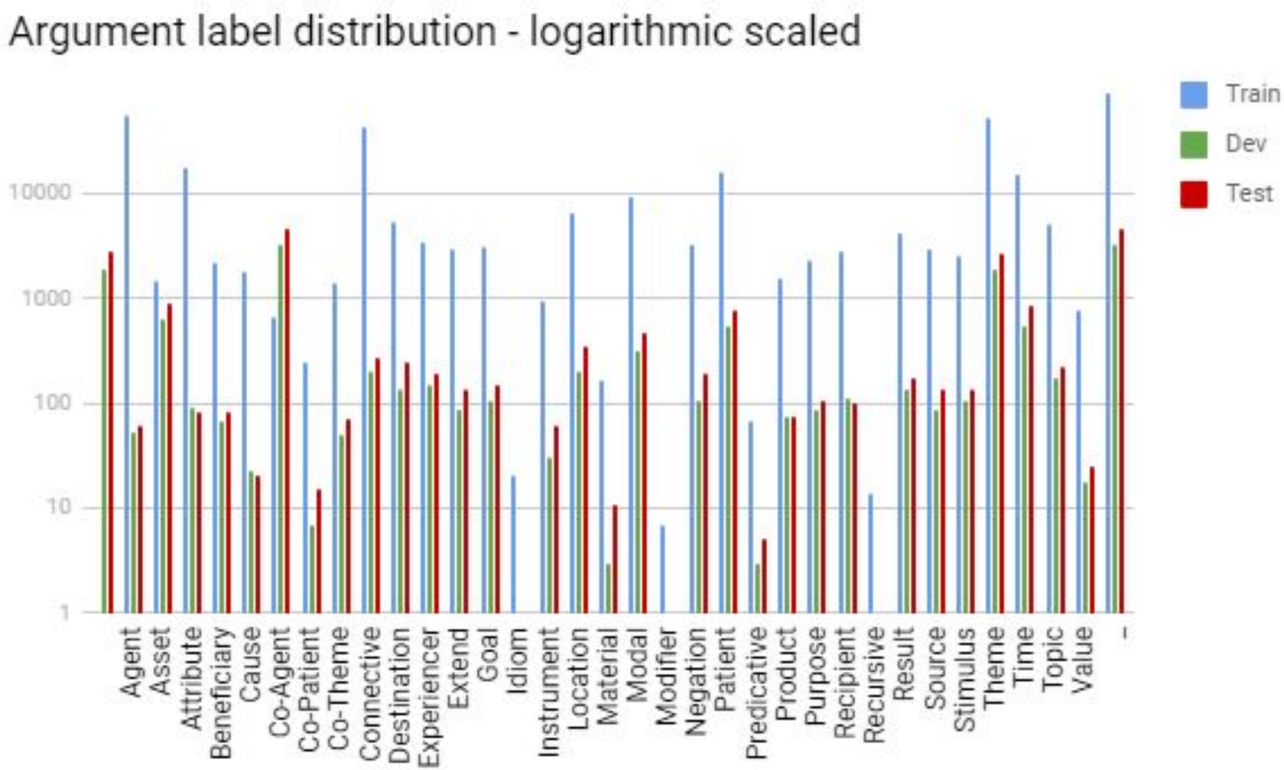
[_ , EAT_BITE, _, DRINK, _]

[_ , EAT_BITE, _, _, _]

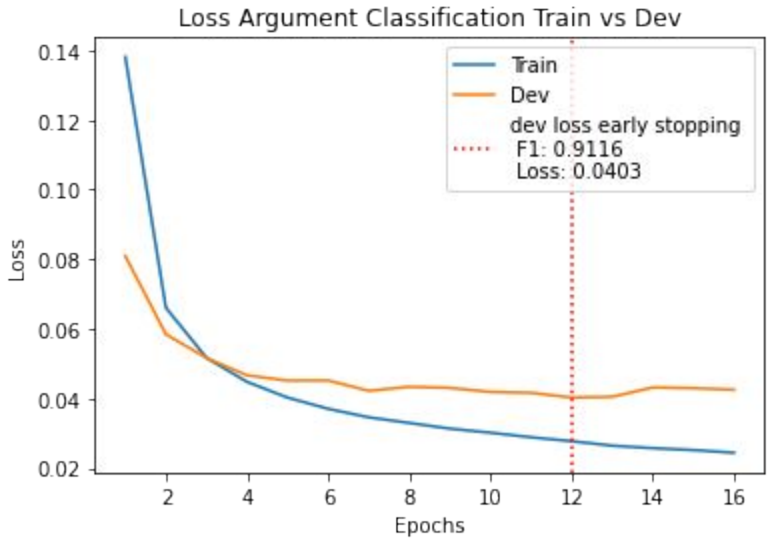
[_ , _, _, DRINK, _]

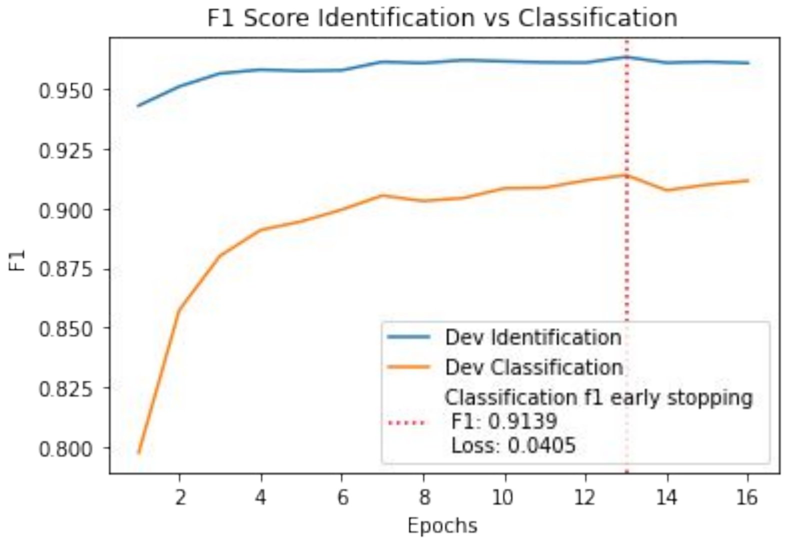
Task information

	DEV	TEST
Idiom	X	X
Modifier	X	X
Recursive	X	✓



Early stopping criteria

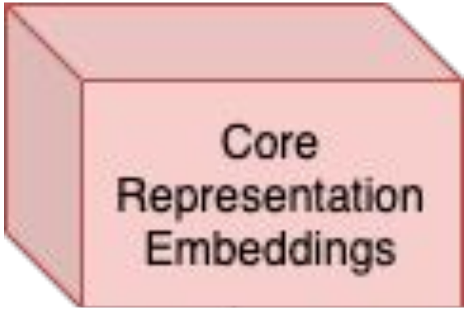




Word Representation

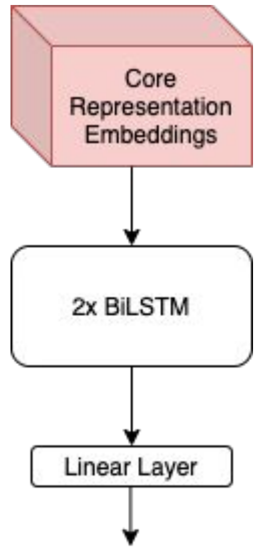
Base WR: Lemma emb ⊕ PoS emb ⊕ GloVe emb ⊕ Predicate emb

Core WR: Base WR ⊕ Dependency Relations Emb



Where ⊕ concatenation operator.
The formula are expressed w.r.t a single word in a sentence

The SRL pipeline



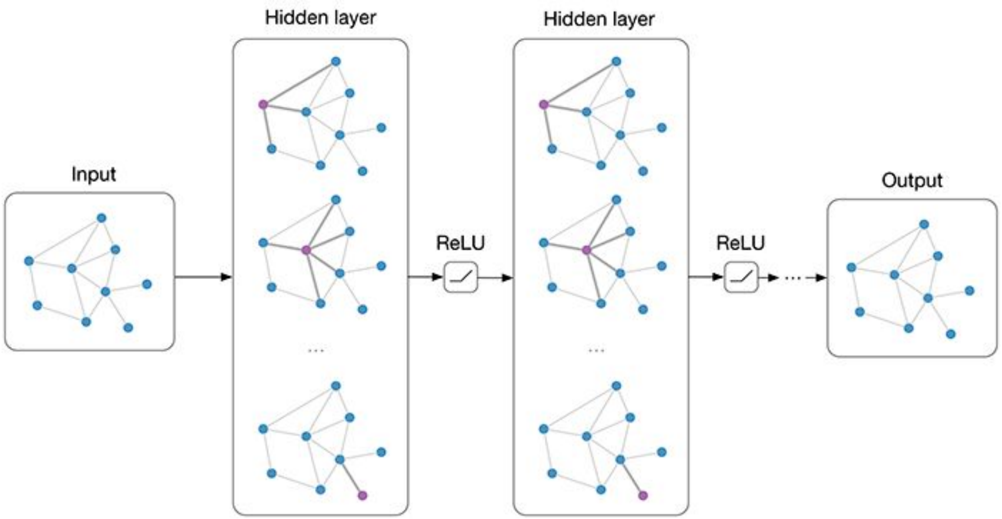
Syntactic Information - Marcheggiani et. al. 2017

Graph Convolutional Network - (Kipf and Welling)

- GCNs are neural networks that operate on graphs and induce features of nodes based on the properties of their neighbours

- Use Normalized Adjacency Matrix
  - A * D^(-1)

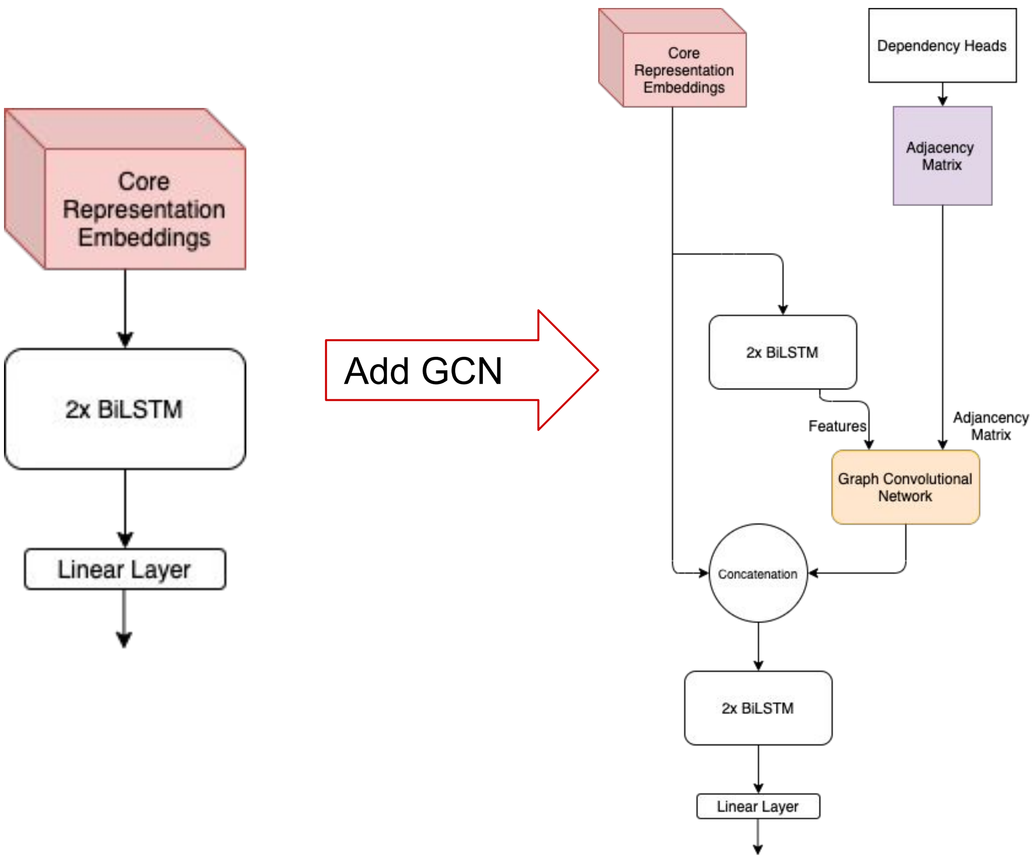
- GCN is not able to capture long dependencies between distant nodes in the graph.
  - Solve using context-aware input



“The semantic representations are closely related to syntactic ones”

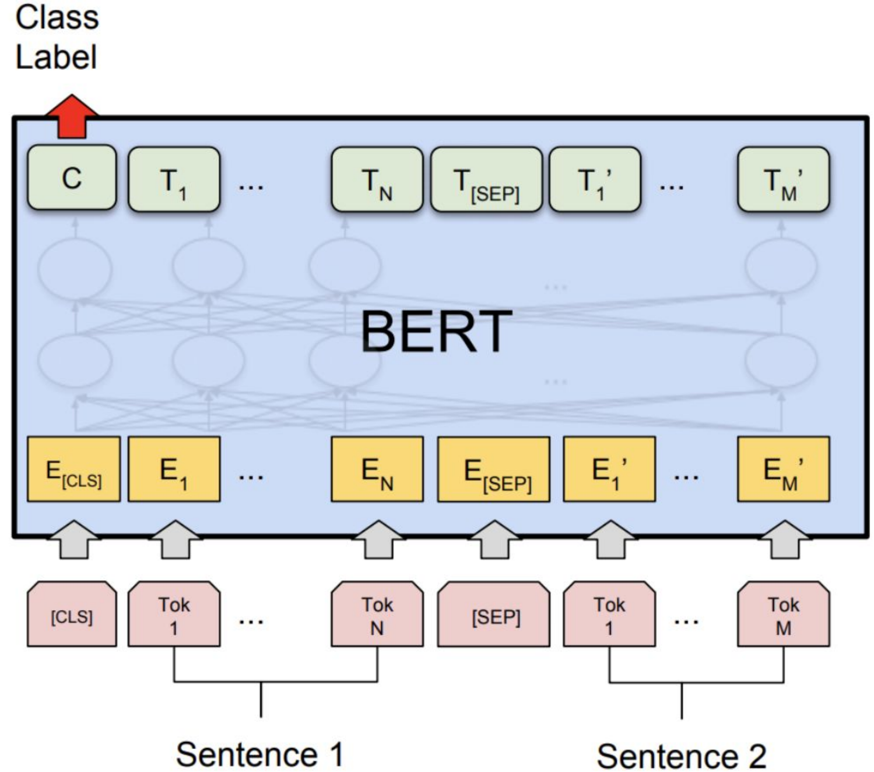
SAPIENZA

The SRL pipeline



SAPIENZA

BERT - Bidirectional Encoder Representation from Transfomer



BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova

(Bert Base uncased 768-dim)

SAPIENZA

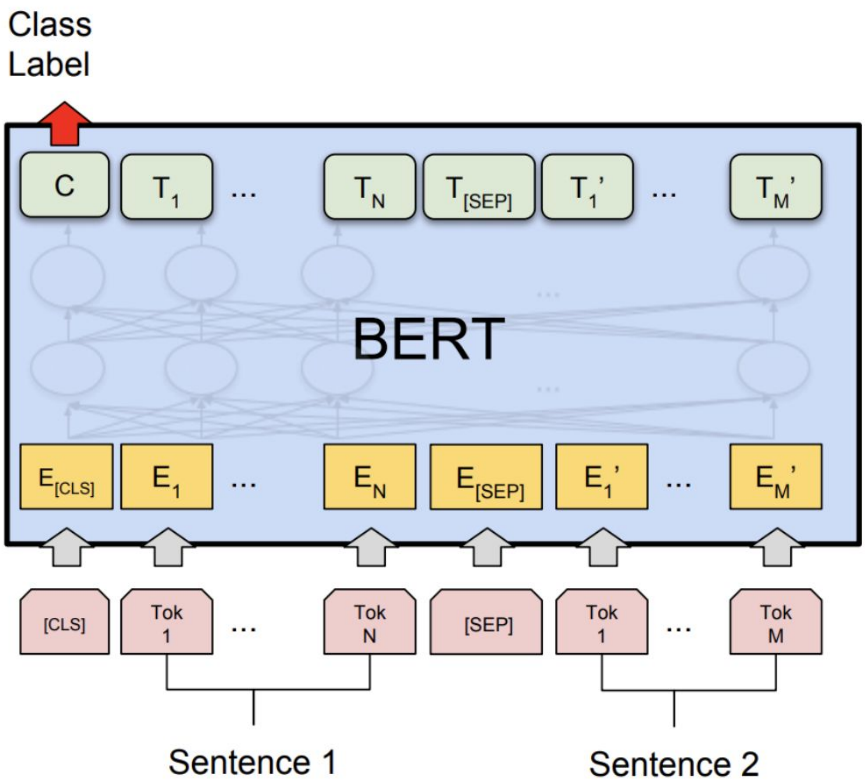
The Non-trainable Bert Embeddings

Vector representation based on the context

Use Word pieces vocabulary

- 30.000 word pieces in vocabulary
- Word segmentation based purely on frequency
- Least frequent words are divided into several word pieces
- *The infrequent 'Embeddings'* -> [*'em', '##bed', '##ding', '##s'*]
- Almost impossible encounter OOV word

I don't apply the fine-tuning to the bert model.



(Bert Base uncased 768-dim)

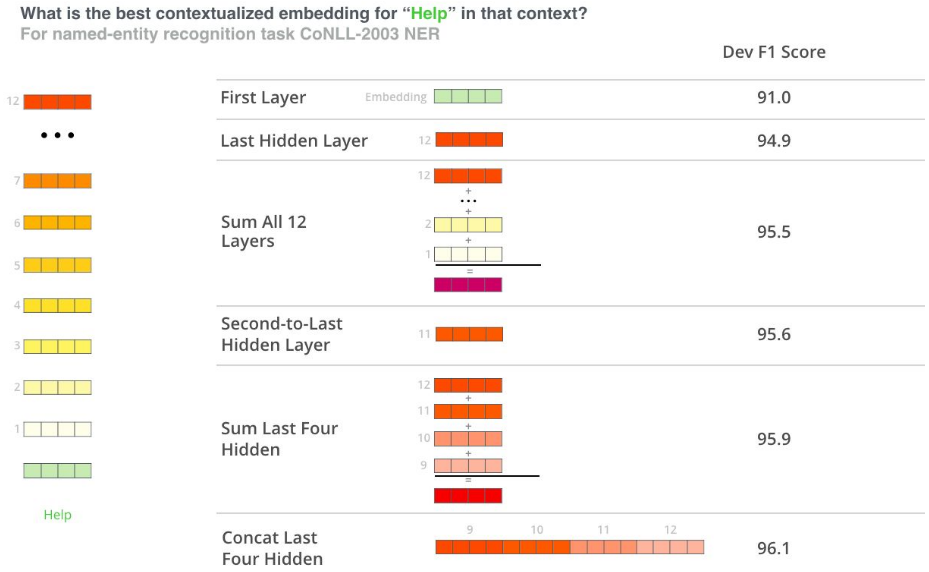
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova

The Non-trainable Bert Embeddings

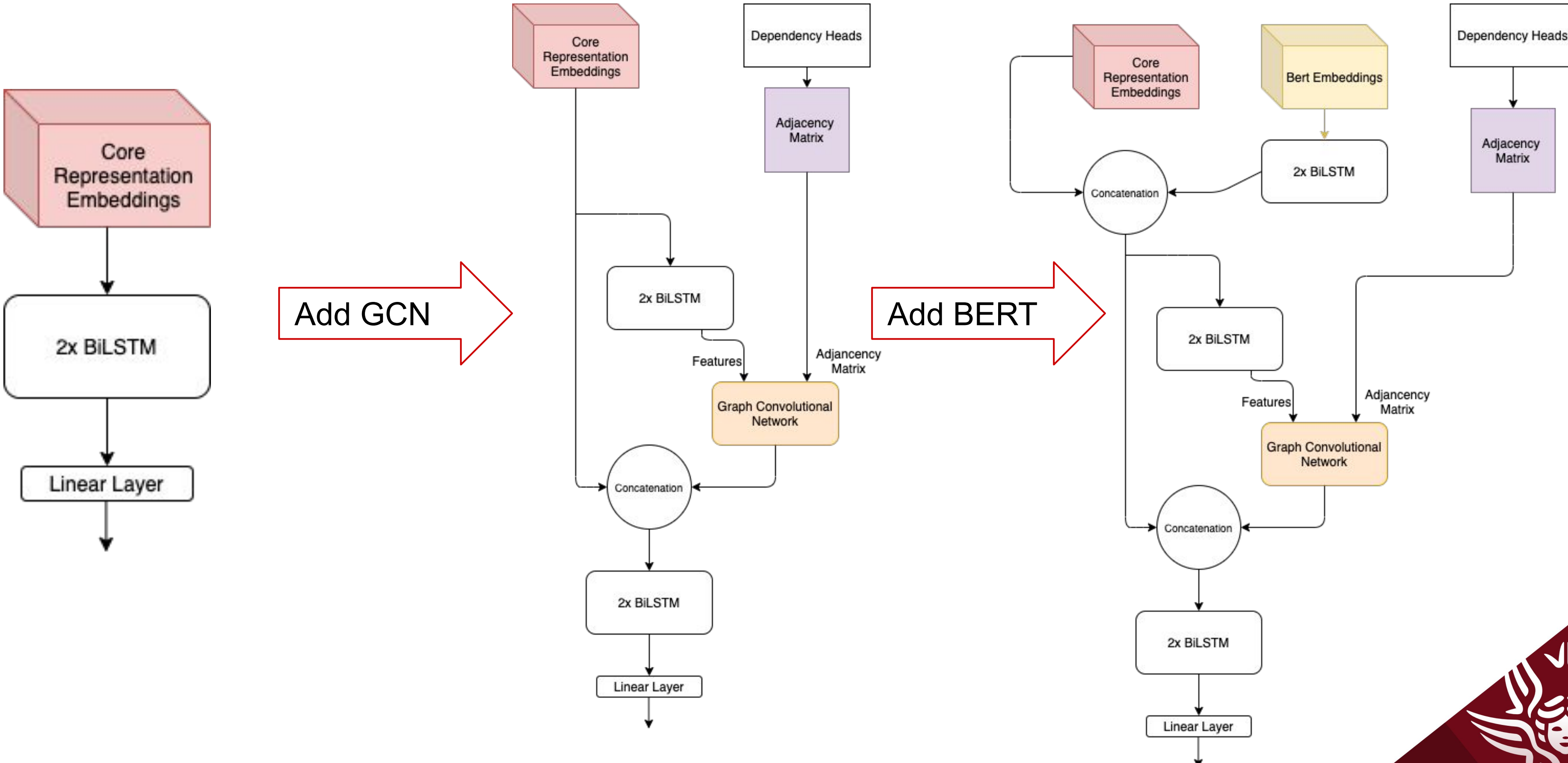
- To produce the embeddings: I summed the last 4 hidden layers
    - this pooling strategy is proven to be one of the most efficient with low memory consumption

- I removed the special tokens like [CLS] and [SEP]

- I merged the word pieces produced by a single words using the average of their embeddings.



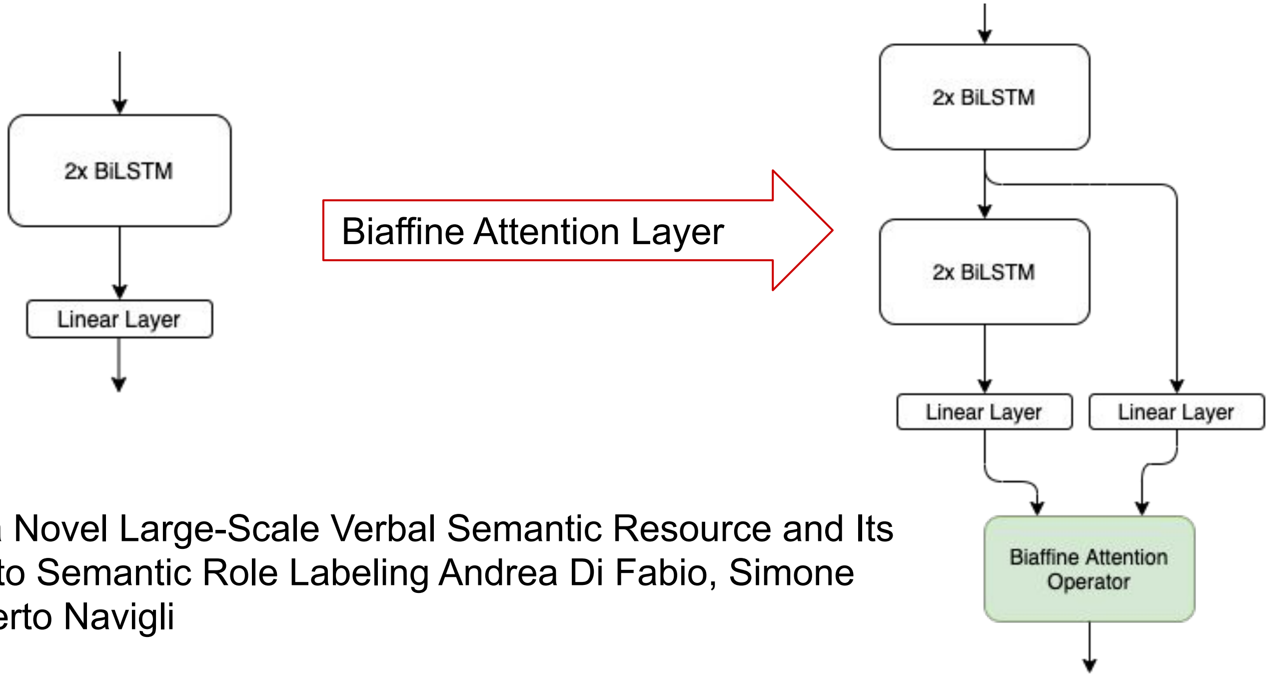
The SRL pipeline



SAPIENZA

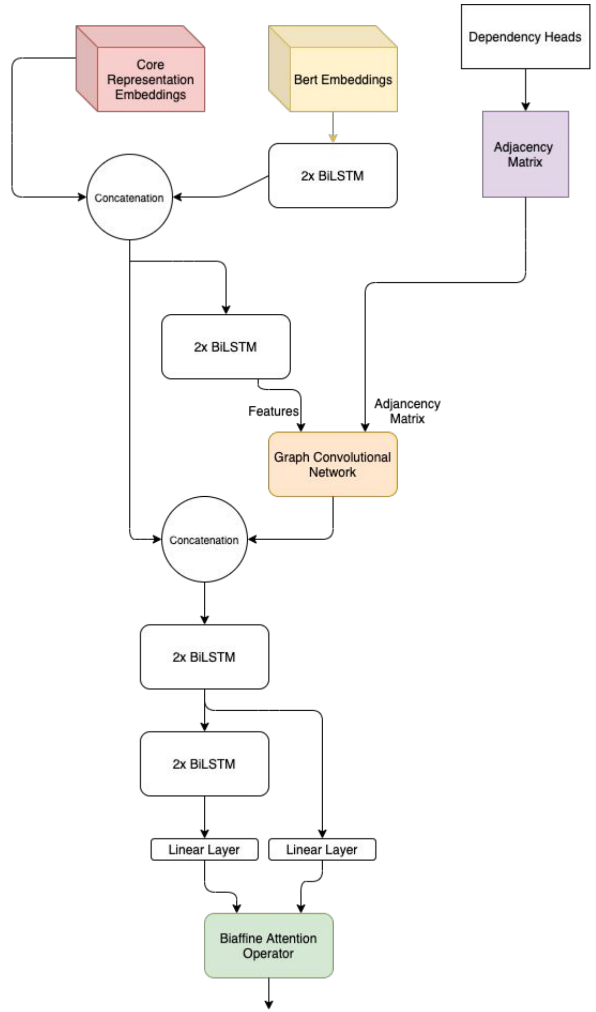
Biaffine Attention Layer

$$\text{Biaffine}(y_1, y_2) = \underbrace{\mathbf{y}_1^T \mathbf{U} \mathbf{y}_2}_{\text{Bilinear}} + \underbrace{\mathbf{W}(\mathbf{y}_1 \circ \mathbf{y}_2) + \mathbf{b}}_{\text{Linear}}$$



VerbAtlas: a Novel Large-Scale Verbal Semantic Resource and Its Application to Semantic Role Labeling Andrea Di Fabio, Simone Conia, Roberto Navigli

Final Model



Dependency Heads
Core Representation Embeddings
Bert Embeddings
Adjacency Matrix
2x BiLSTM
Concatenation
2x BiLSTM
Features
Adjacency Matrix
Graph Convolutional Network
Concatenation
2x BiLSTM
2x BiLSTM
Linear Layer
Linear Layer
Biaffine Attention Operator

SRL Hyperparameters

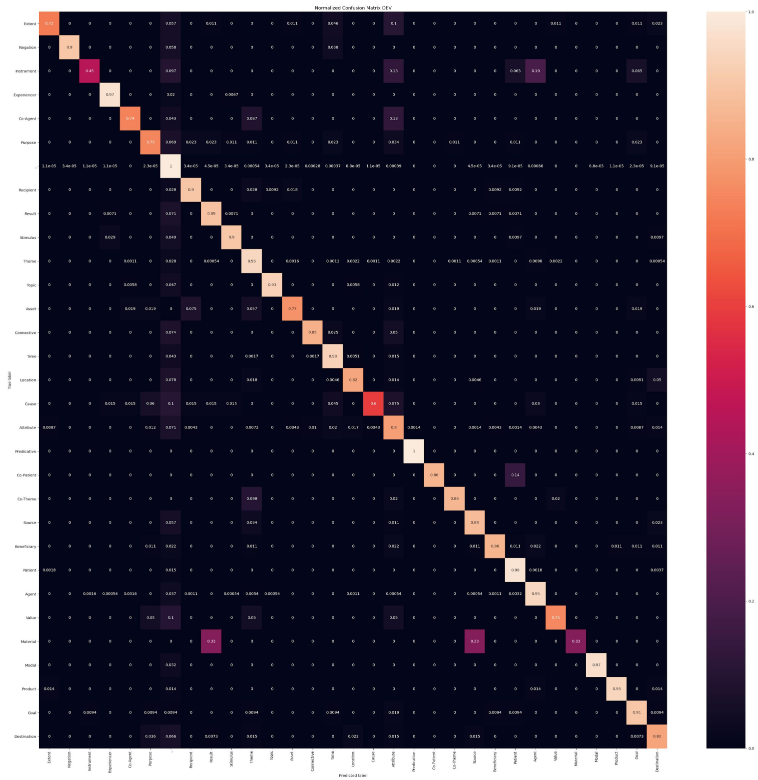
HParams	Value	Notes
Epochs	13	
Batch Size	32	
Optimizer	Adam	
Learning Rate	0.001	
Loss Function	Cross Entropy	
Word Vocab Min frequency	2	
Lemma Vocab Min frequency	2	
Dropout Embeddings	30%	
Word Emb dim	300	GloVe 6B
Bert Emb dim	768	Bert Base uncased
Pos Emb dim	300	
Lemma Emb dim	300	
Predicate Emb dim	400	
Dependency Relations Emb dim	300	
Num GCN-Convolutional Layer	2	
GCN-Convolutional Layer Hidden size	250	first layer
GCN-Convolutional Layer Hidden size	35	second layer
GCN Dropout	50%	
Dropout BiLSTM	30%	
BiLSTM Bert	300 out dim	x2 layer
BiLSTM GCN	300 out dim	x2 layer
BiLSTM Biaffine B1	300 out dim	x2 layer
BiLSTM Biaffine B2	300 out dim	x2 layer
Biaffine inputs	35, 35	y1, y2
Biaffine outputs	35	length of label vocab
Linear layer y1	35	
Linear layer y2	35	

Results

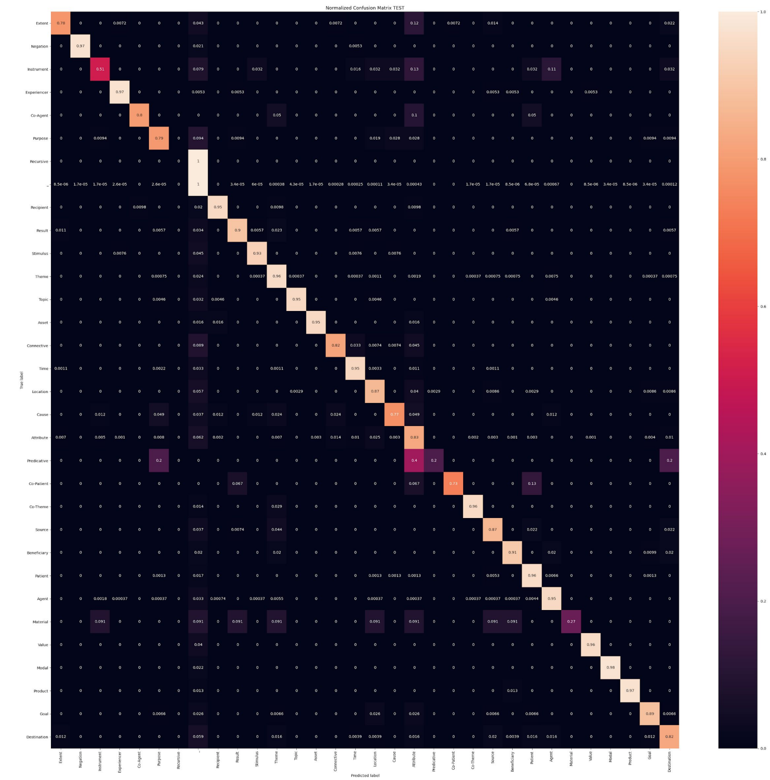
SRL Results				
Experiments	F1-Dev		F1-Test	
	Identification	Classification	Identification	Classification
Baseline + Base WR	90.84%	85.84%	91.80%	87.38%
Baseline + Core WR	94.05%	88.96%	94.52%	90.30%
Baseline + GCN + Core WR	95.68%	89.35%	96.62%	91.25%
Baseline + Bert + GCN + Core WR	95.80%	91.07%	95.90%	92.05%
Baseline + Bert + GCN + Biaffine + Core WR*	**96.32%**	**91.38%**	**96.87%**	**92.78%**
Baseline + Bert + Core WR	95.14%	90.71%	95.72%	91.98%
Baseline + GCN + Biaffine + Core WR	95.78%	89.70%	96.72%	92.28%
Baseline + Bert + Biaffine + Core WR	95.09%	90.80%	95.90%	92.21%

Confusion Matrix: Dev vs Test

DEV



TEST



- The _ class is better predicted on both dataset.
- The material class is poorly predicted in both dataset.
- Recursive on the test is poorly predicted, the train dataset has 14 Recursive examples.

SAPIENZA

Reference

- VerbAtlas: a Novel Large-Scale Verbal Semantic Resource and Its Application to Semantic Role Labeling Andrea Di Fabio, Simone Conia, Roberto NavigliVerbAtlas: a Novel Large-Scale Verbal Semantic Resource and Its Application to Semantic Role Labeling Andrea Di Fabio, Simone Conia, Roberto Navigli

- Encoding Sentences with Graph Convolutional Networks for Semantic Role Labeling Diego Marcheggiani, Ivan Titov

- Semi-supervised classification with graph convolutional networks Thomas N. Kipf, Max Welling

- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova

- Pytorch-crf package

SAPIENZA

Thank you for the attention

Andrea Bacciu