Multi-Way Classification of Relations Between Pairs of Entities

Sahitya Patel

Supervisor: Dr. Harish Karnick

June 2018

Department of Computer Science and Engineering

Indian Institute of Technology, Kanpur

Cite Us As

```
@MastersThesis{Sahitya:2018,
 author = { {Sahitya Patel} and Harish Karnick},
 title = {Multi-Way Classification of Relations Between Pairs of Entities},
 school = {Indian Institute of Technology Kanpur (IITK)},
 address = {India},
 year = 2018,
 month = 6
}
```

Outline

Introduction

- Relation Classification

Related Work

- Features for Relation Classification
- Previous Methods for Relation Classification

Proposed Models

- BGRU based models
- CNN-BGRU based models
- Combined Models

Results

- Dataset : SEMEVAL 2010 task 8
- Performance of Proposed Models
- Hyperparameters, Confusion Matrix
- Comparison with competing methods

Conclusion and Future Work

Introduction

Introduction

Information Extraction (IE)

IE = Identification of entities + Identification of relations between the entities

Relation Classification (RC)

- Relation classification is a task of assigning predefined relation labels to the entity pairs that occur in texts
- **[People]**$_{e1}$ have been moving back into **[downtown]**$_{e2}$
- Entity-Destination(e1,e2)
  - e1 = people
  - e2 = downtown
- H(S, e1, e2) = y
  - S = Sentence
  - e1, e2 = entities
  - y comes from a predefined set of relations

Introduction

Message-Topic(e1,e2) Examples

1. The final **[programme]$_{e1}$** detailed the **[history]$_{e2}$** of Russborough House
2. The **[letter]$_{e1}$** contains a description of the **[demolition]$_{e2}$** of the old synagogue
3. On 17 May 2005, the committee held a **[hearing]$_{e1}$** concerning specific **[allegations]$_{e2}$**
4. The **[newsletter]$_{e1}$** tells of practical **[projects]$_{e2}$** developed to help those affected by the pandemic

- Message-Topic(e1,e2) relation is conveyed in different ways in above examples
- An efficient system for Relation Classification needs to account for syntactic and semantic features of the overall sentence to correctly identify the relation
- SEMEVAL 2010 task 8 dataset
  - 9 directional relations + 1 other
  - Total 2*9+1 = 19 classes

Related Work

- Features for Relation Classification
  - Linguistic features
  - Positional features
- Previous Methods for Relation Classification
  - Att-BLSTM
  - Att-BGRU

Related Work : Features

Linguistic Features

- Word Embeddings
- Part of Speech tags (POS)
- WordNet tags
- Shortest Dependency Path (SDP)
- Grammar Relation tags (GR)

Word Positional Indicators

E1_START **People** **E1_END** have been moving back into **E2_START** **downtown** **E2_END**

Word Positional Embedding

People	have	been	moving	back	into	**downtown**
0	1	2	3	4	5	6
-6	-5	-4	-3	-2	-1	0

Related Work : Methods

Categories of Relation Classification Methods

1. Feature Based
2. Kernel Based
3. Neural Network Based

Common Framework for Neural Network based methods

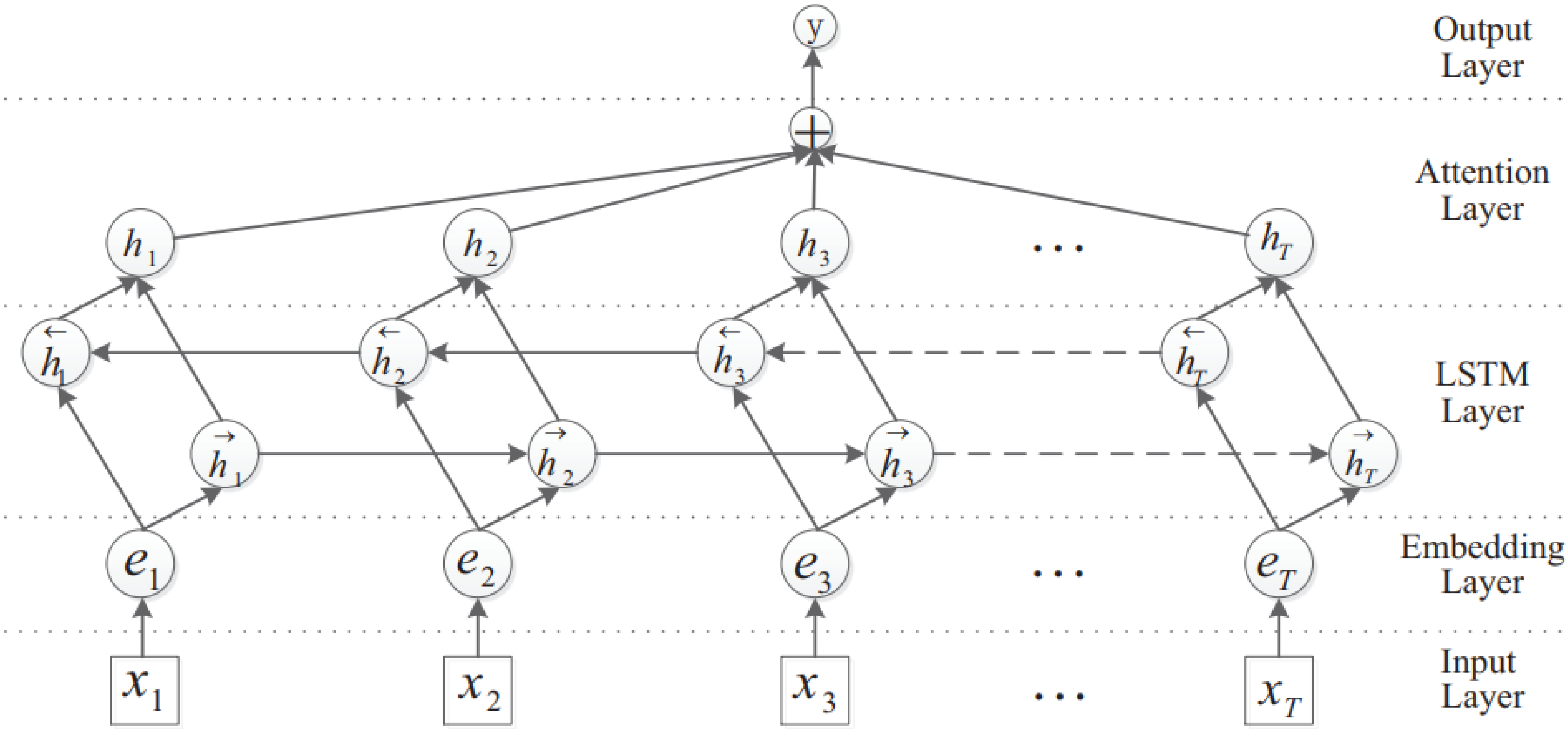
- Features: Input Sentence, Linguistic Features, Positional Features
- Neural Networks: CNN / RNN / LSTM / GRU
- Pooling over the sentence features
- Classification

Related Work : Methods

Relation Classification methods and their scores on SEMEVAL 2010 Task 8 Dataset

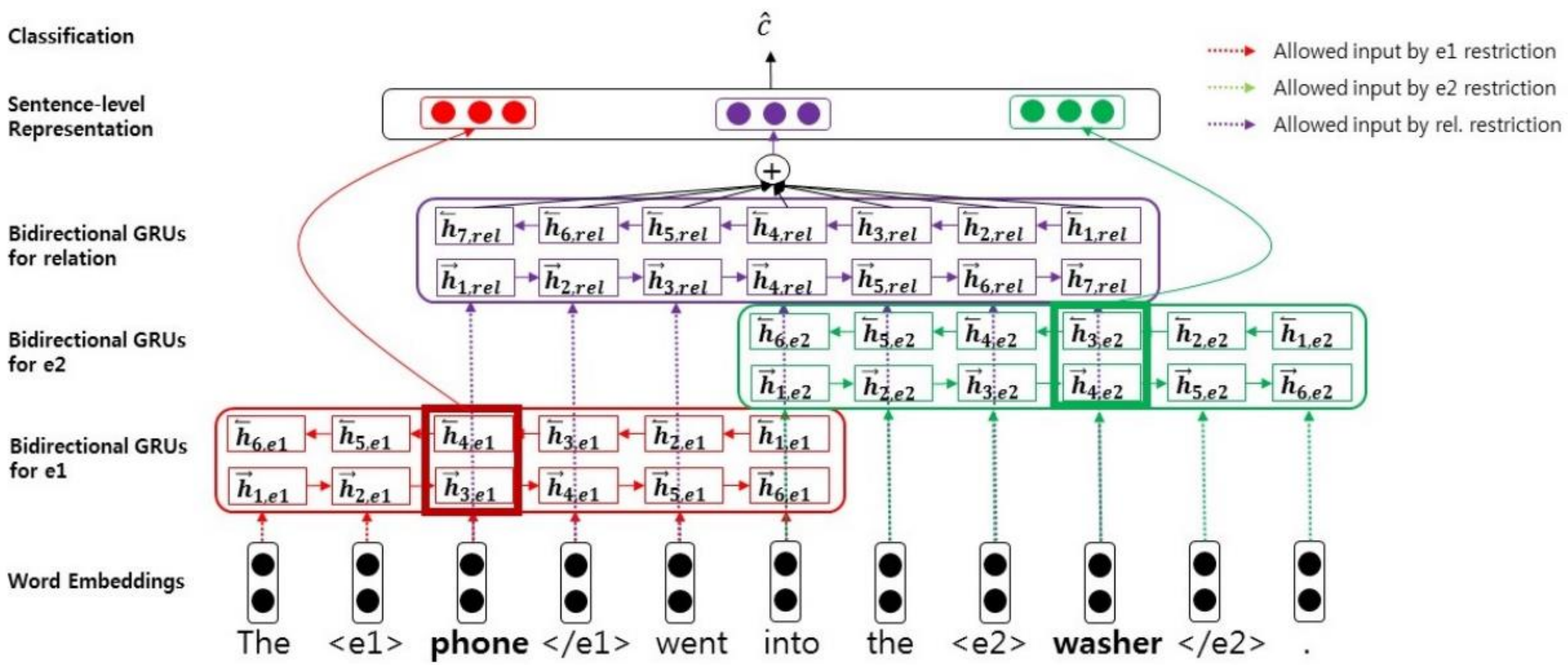
Classifier	Additional Information	F1
SVM [Rink and Harabagiu, 2010]	POS, WordNet, Pre xes and other morphological features, Dependency parse, Levin classed, PropBank, FanmeNet, NomLexPlus, Google n-gram, Paraphrases, TextRunner	**82.2**
MVRNN [Socher et al., 2012]	Word embeddings, POS, NER, WordNet	**82.4**
CNN [Zeng et al., 2014]	Word embeddings, Position embeddings, WordNet	**82.7**
FCM [Yu et al., 2014]	Word embeddings, Dependency parsing, NER	**83.0**
DepNN [Liu et al., 2015]	Word embeddings, SDP, NER	**83.6**
SDP-LSTM [Xu et al., 2015b]	Word embeddings, SDP, POS, GR, WordNet	**83.7**
Att-BLSTM [Zhou et al., 2016]	Word embeddings, Positional indicator	**84.0**
Ensemble Methods [Nguyen and Grishman, 2015]	Word embeddings, SDP, POS, NER, WordNet, CNNs+RNNs+Voting	**84.1**
CR-CNN [dos Santos et al., 2015]	Word embeddings, Position embeddings, Special ranking objective	**84.1**
att-BGRU [Kim and Lee, 2017]	Word embeddings, Range restricted	**84.1**
SPTree [Miwa and Bansal, 2016]	Word embeddings, SDP, POS	**84.4**
EAtt-BiGRU [Qin et al., 2017]	Word embeddings, Positional embeddings	**84.7**
ER-CNN+R-RNN [Vu et al., 2016]	Word embeddings, Extended middle context, Ensemble, Voting, Special ranking objective	**84.9**
depLCNN [Xu et al., 2015a]	Word embeddings, SDP, WordNet, Word around nominals, Negative sampling from NYT dataset	**85.6**
DRNN [Xu et al., 2016]	Word embeddings, SDP, POS, GR, WordNet, Data augmentation, (w/o data augmentation - 84.2)	**86.1**
BRCNN [Cai et al., 2016]	Word embeddings, SDP, POS, NER, WordNet, Data augmentation	**86.3**
Att-Input-Pooling-CNN [Wang et al., 2016]	Word embeddings, Position embeddings, Special ranking objective	**88.0**

Related Work : Att-BLSTM



Att-BLSTM - Bidirectional LSTM model with Attention [ZHOU et al., 2016]

Related Work : Att-BGRU



Att-BGRU - Multiple Range-Restricted Bidirectional GRUs with Attention [KEEM and LEE, 2017]

Related Work : Methods

Observations on existing methods

- Many of the methods use linguistic features
  - Costly
  - Error-Prone

- Mostly one-Layered architectures
  - One layer of CNN or RNN

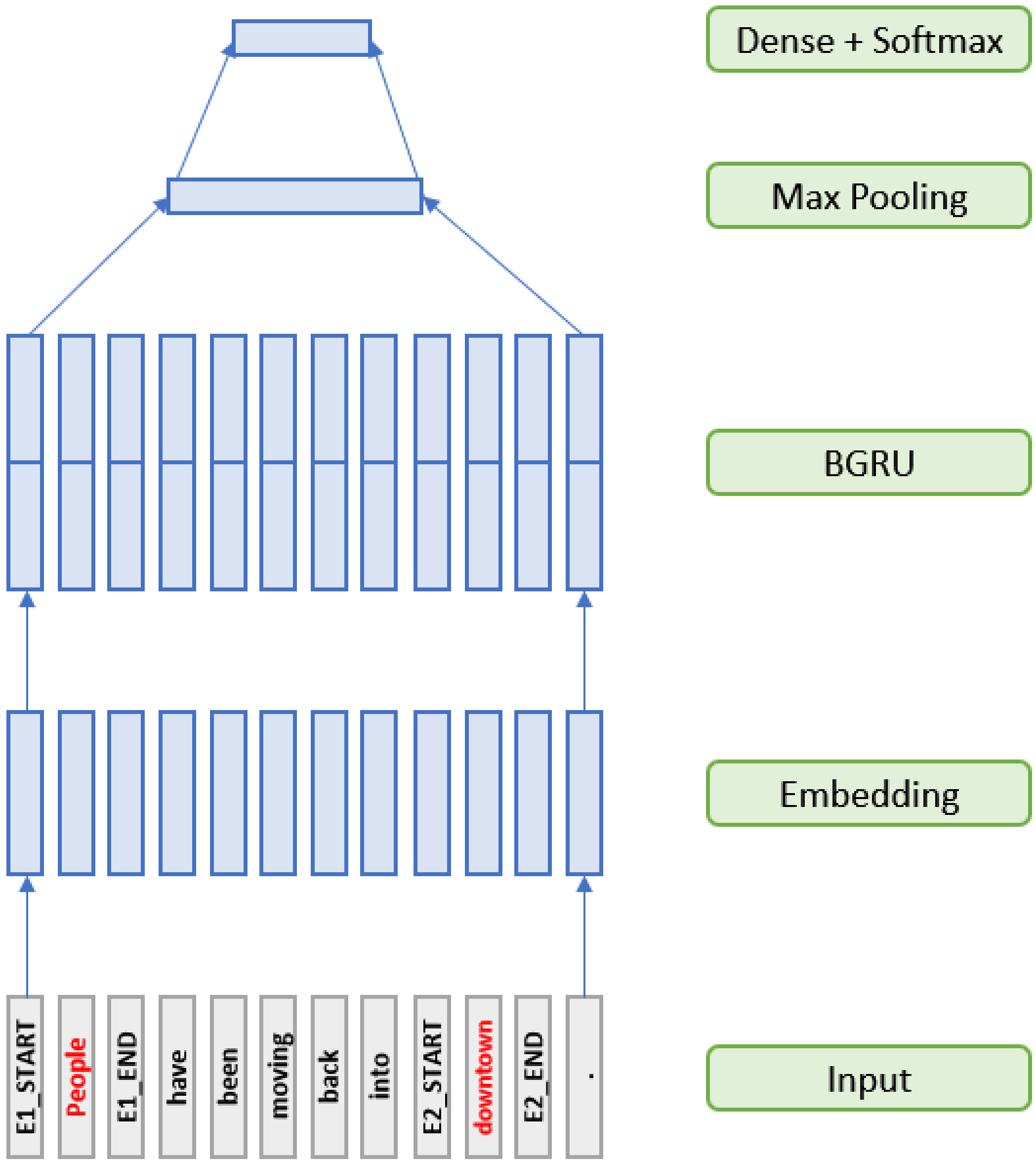
- Most of them use a single pooling – max pooling or attention pooling

Proposed Models

- BGRU based models
  - BGRU-M
  - BGRU-ME
- CNN-BGRU based models
  - CBGRU-ME
  - CBGRU-A
  - CBGRU-MEA

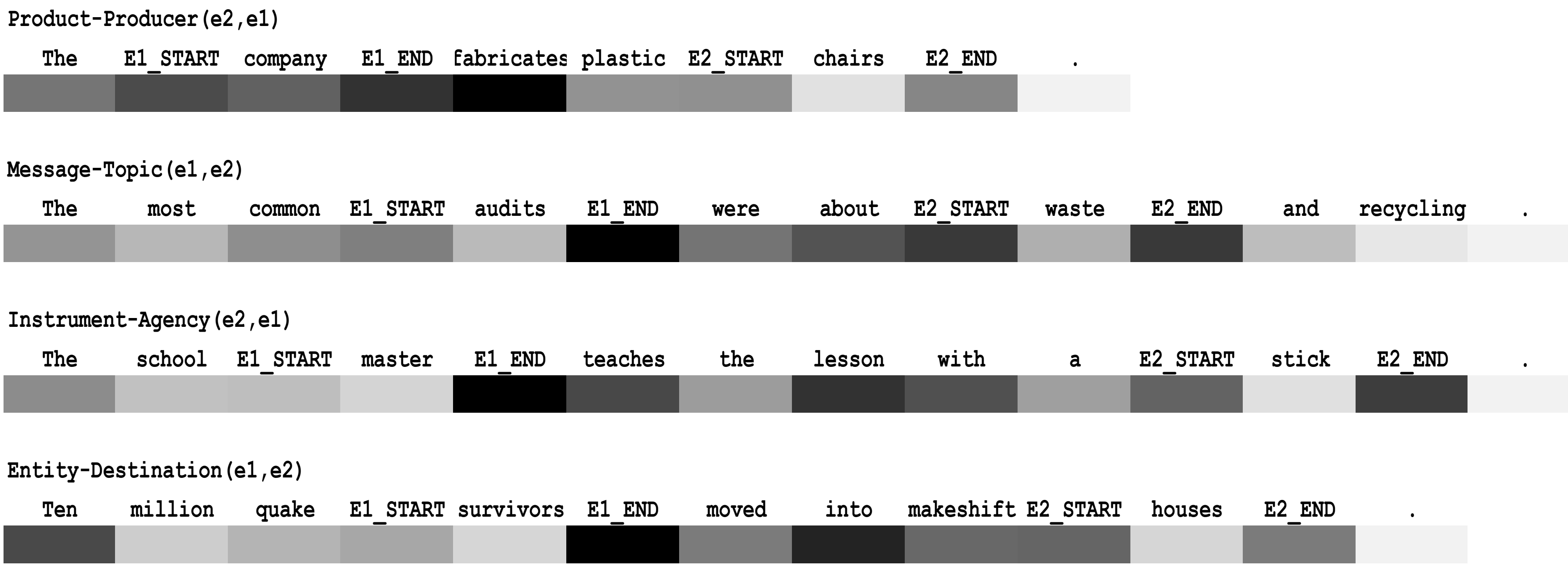
Proposed Model : BGRU-M

- Input
  - Positional Indicators
  - Tokenization
  - Padding
- Embedding
  - Google word2vec vectors
  - OOV
- Bidirectional GRU
  - Concatenation of forward and backward pass features
- Max pooling
  - Maximum over the length of the sentence
- Dense and Softmax
  - Cross-Entropy
  - Classification



Proposed Model : BGRU-M

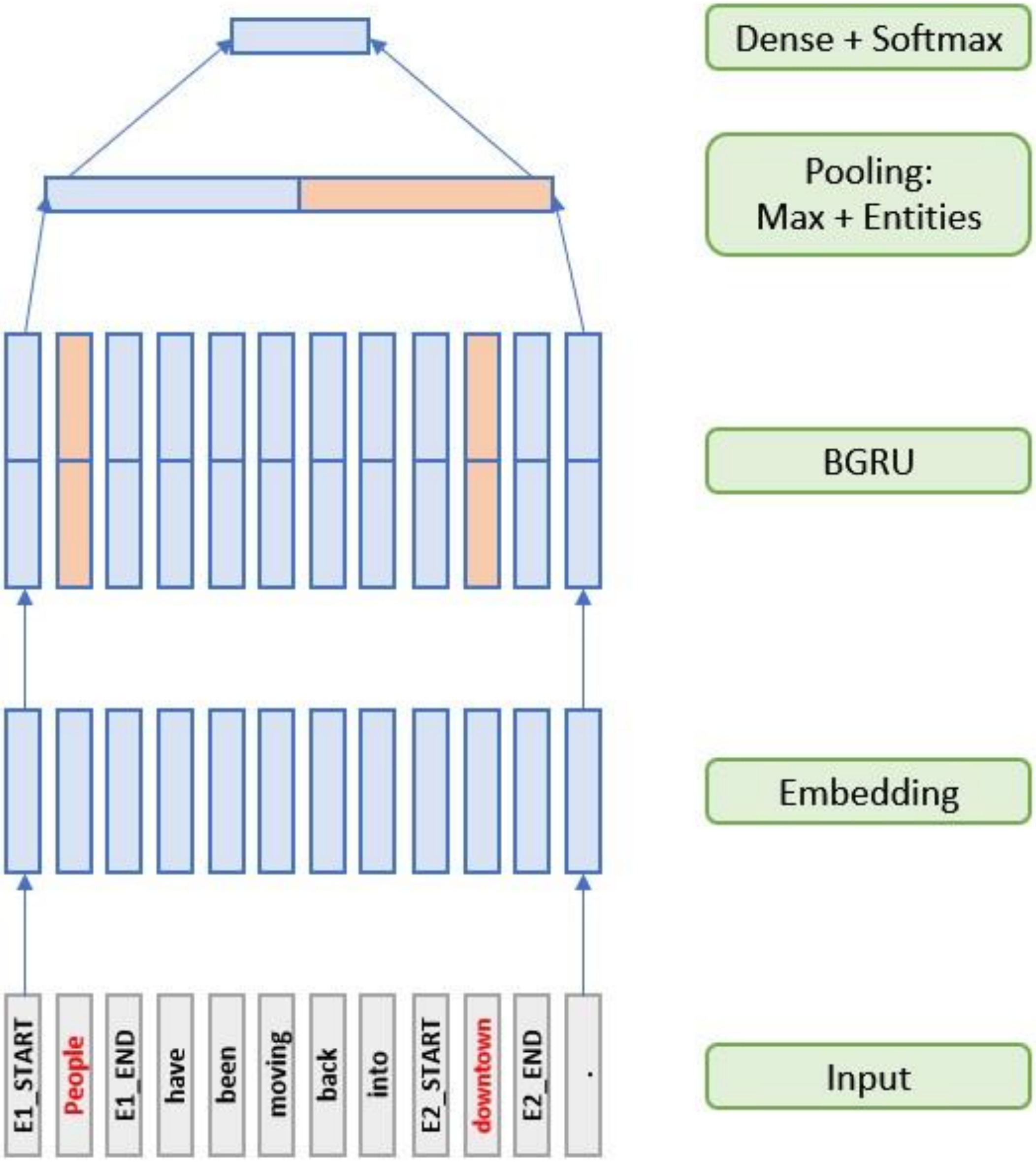
Max Pooling Visualization



- Features from verbs and several important words are pooled
    - *fabricates, were, teaches, moved, about, with, into*
- Features from irrelevant words are pooled
    - *The, Ten, quake, and*
- Features from entities are not pooled
    - *chairs, audits, master, houses*

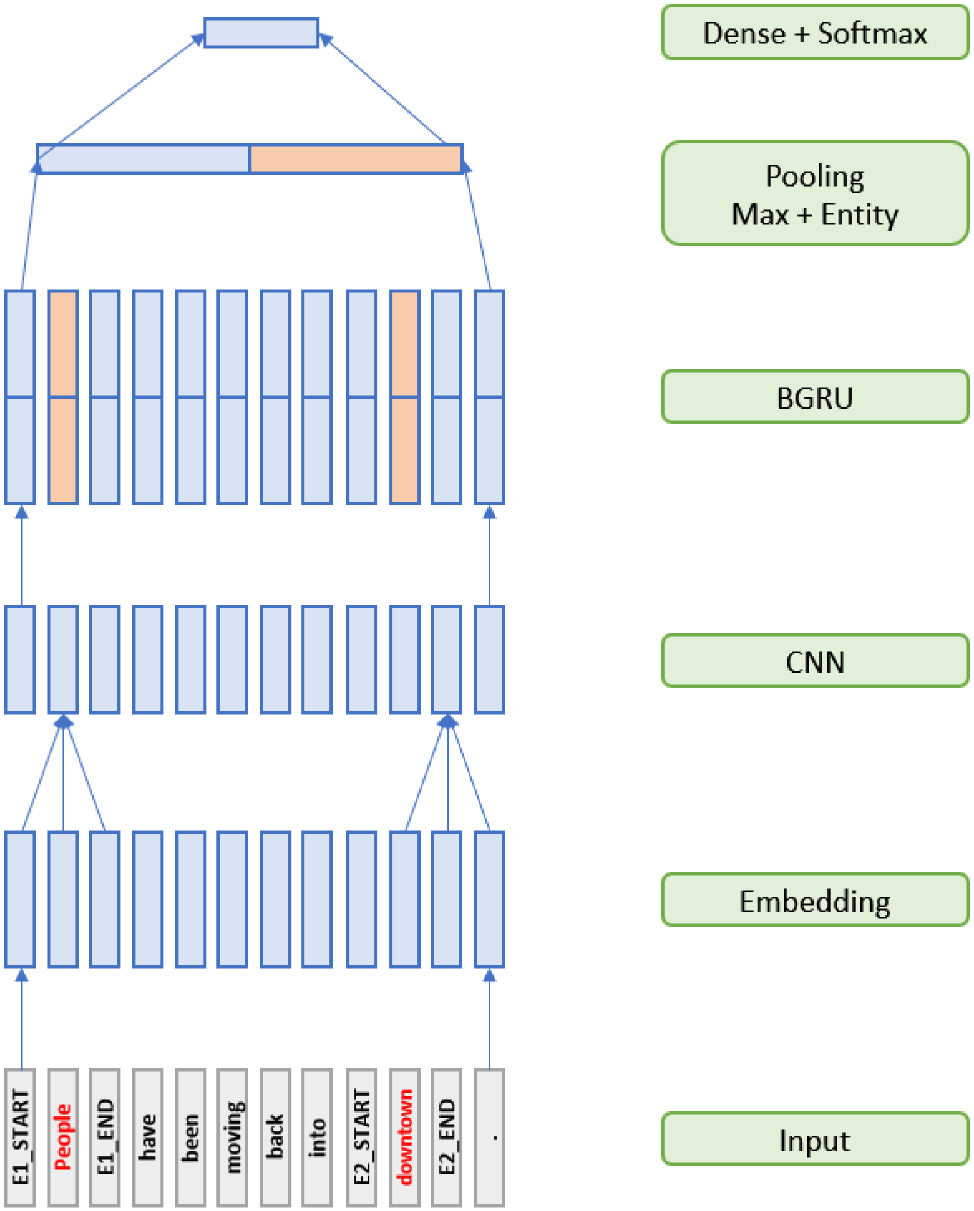
Proposed Model : BGRU-ME

- Entities Pooling
  - Entities are the most significant clues for Relation Classification
  - Maximum over the features of entities
- Concatenation of max and entities pooling features
- Better results than model BGRU-M

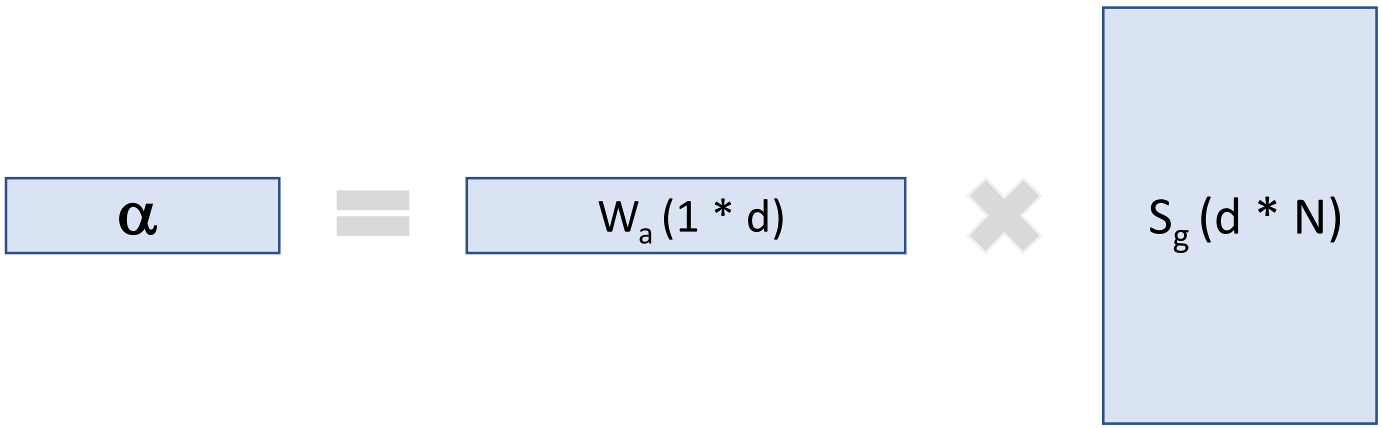


Proposed Model : CBGRU-ME

- CNN and BGRU – two layered architecture for RC
- CNN
  - N-gram features
  - Filter length : 3
  - Returns a sequence of local features
- BGRU
  - Returns a sequence of sentential features
  - Local as well as global features
- Better results than model BGRU-ME



Attention Pooling

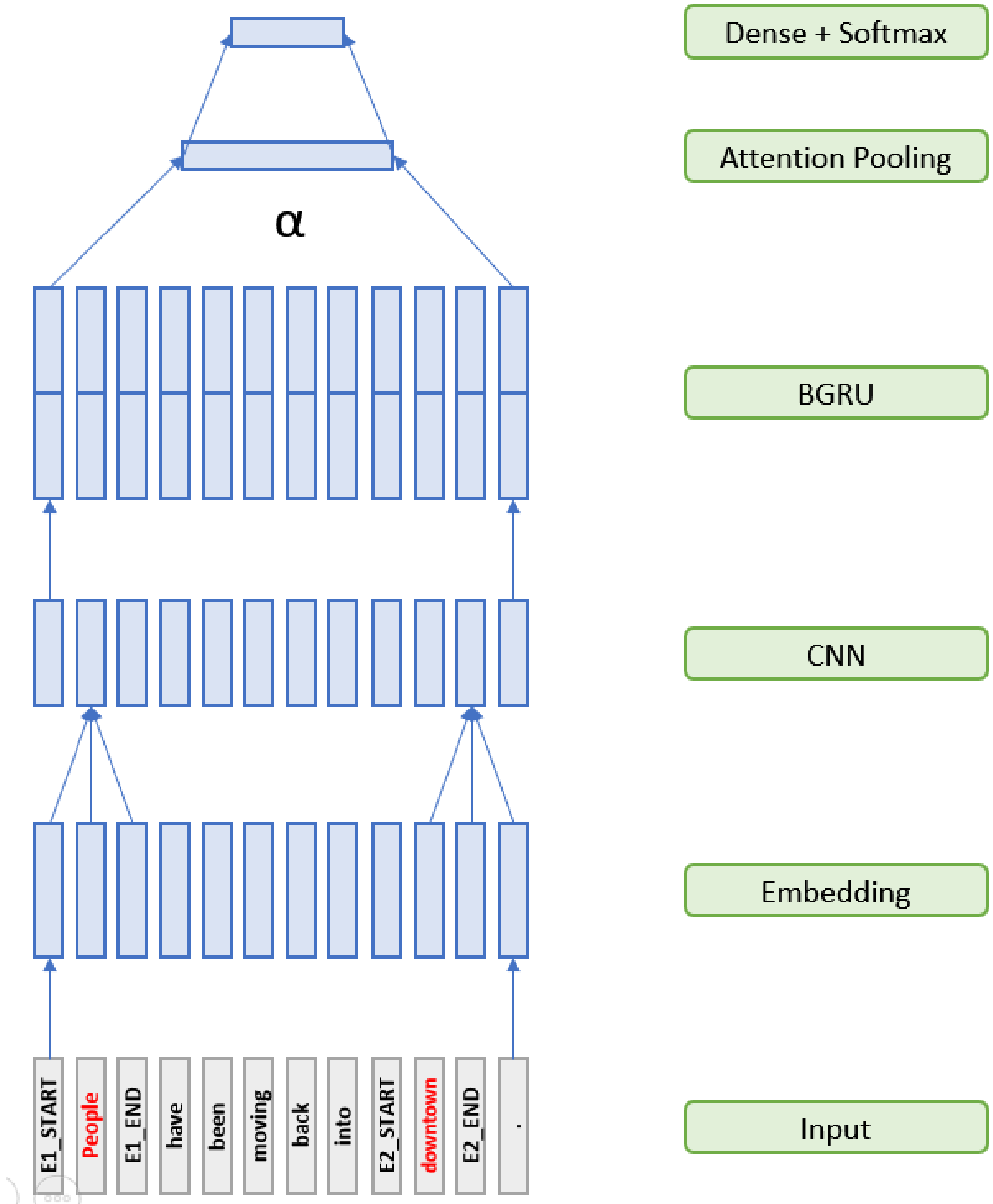


$$\alpha = \textit{softmax}\,(\,W_a * S_g + B_a\,)$$
$$S_{p\_attn} = S_g * \alpha^T$$

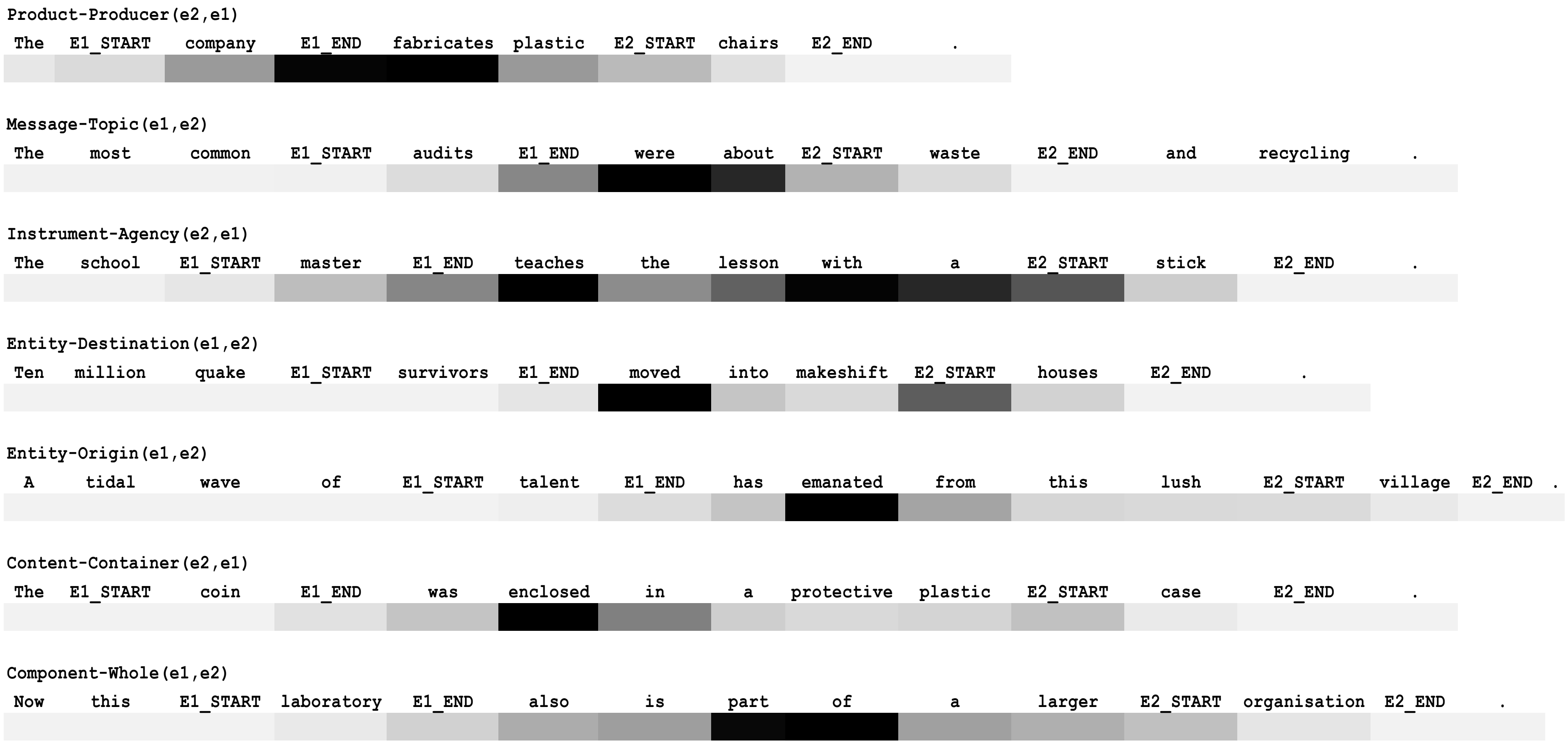
Sg	Features from BGRU
α	Attention Vector
S_{p_attn}	Sentence level featue vector

Proposed Model : CBGRU-A

- The attention mechanism produces a normalized vector α (whose size is the length of the sentence) using which we take a linear combination of BGRU features
- Importance of words in a sentence is learned



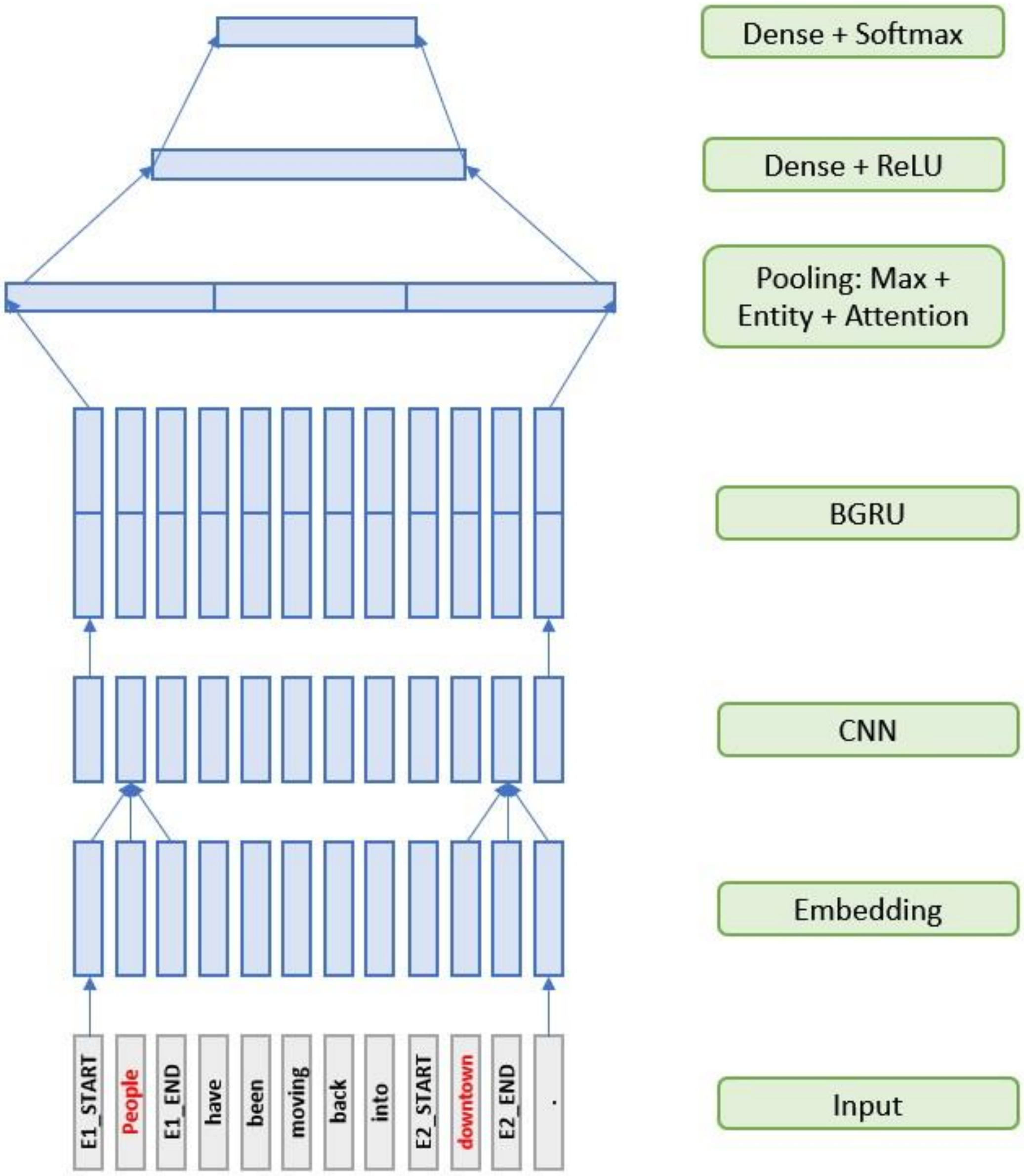
Proposed Model : CBGRU-A : Attention



- Features from verbs and other important words are pooled
  - *fabricates, teaches, moved, from, about, with, in, part of*
- Features from irrelevant words are not pooled
  - *The, most, Ten, quake, and*
- Features from some entities are not pooled
  - *chairs, audits, master, houses*

Proposed Model : CBGRU-MEA

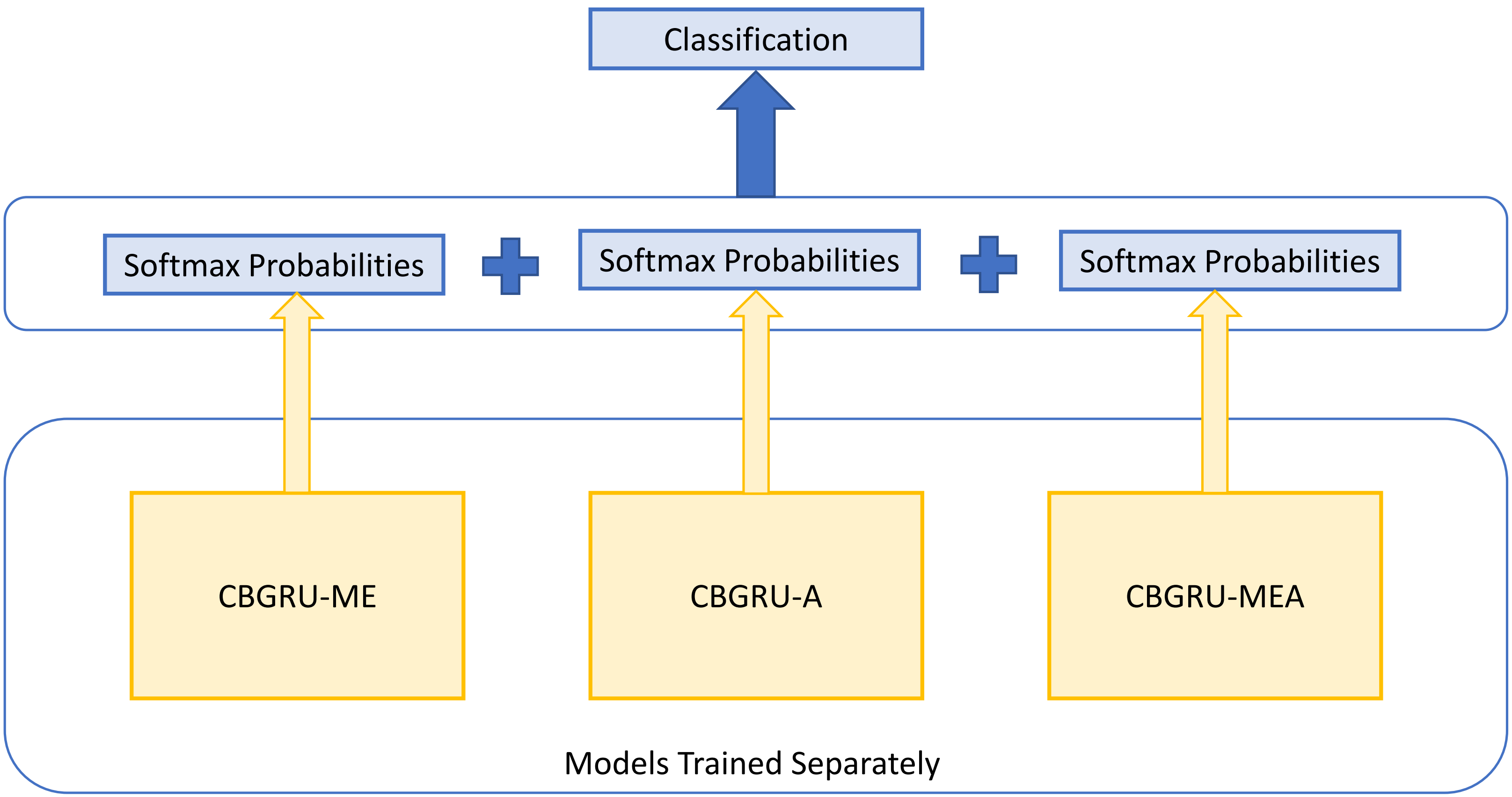
- Taking advantage of all kinds of pooling, namely max, entities, attention pooling
- Features vectors of all kinds of pooing are concatenated
- Extra fully-connected layer with ReLU activation
- Better results than all previously proposed models



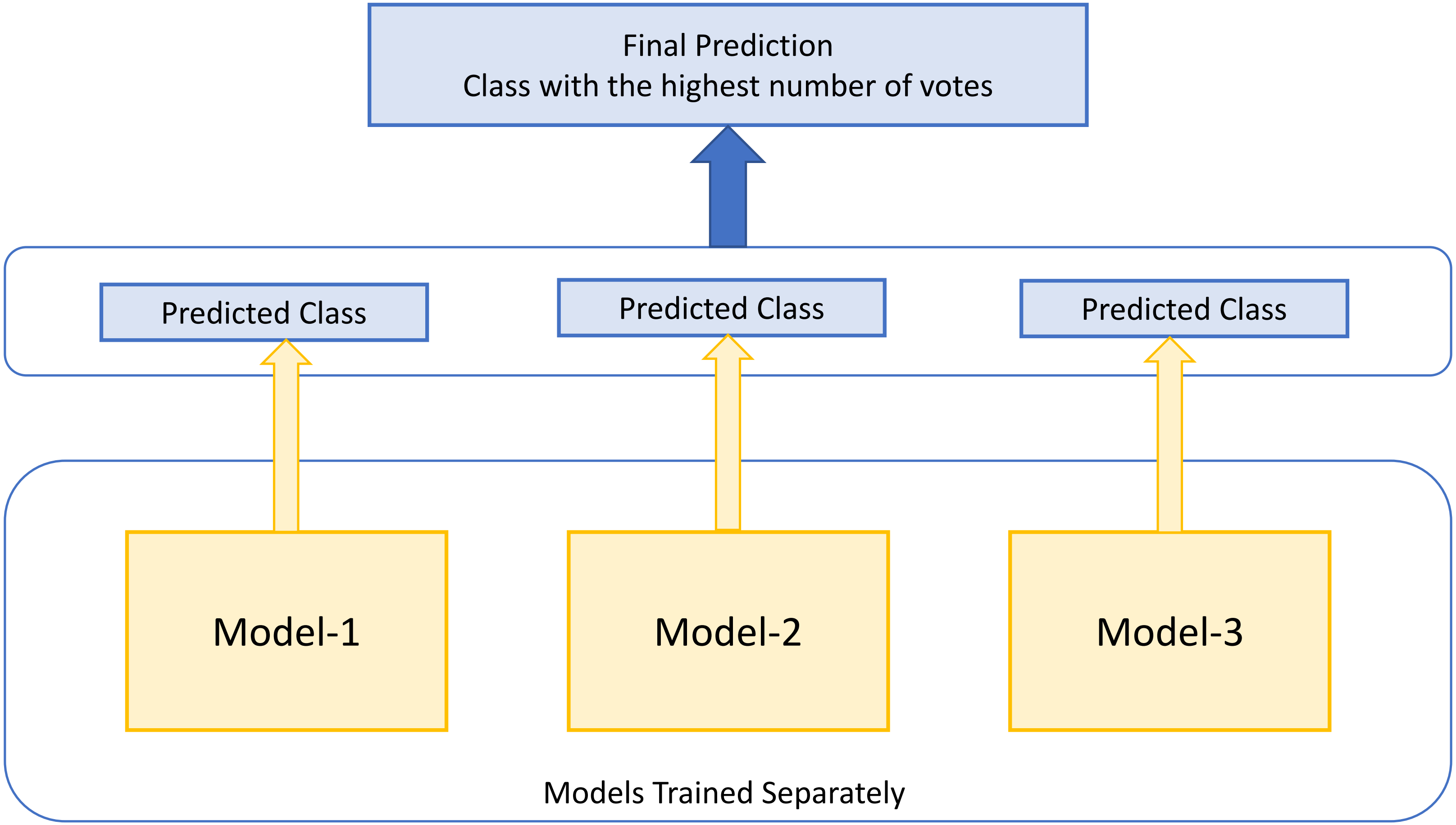
Combined Models

- Softmax Probabilities Sum
- Voting Scheme

Combined Models : Softmax Probabilities Sum



Combined Models : Voting Scheme



Results

- Dataset
- Results
  - Performance of proposed models
  - Hyperparameters
  - Confusion Matrix
- Comparison with Competing Methods
- Experiments

Dataset : SEMEVAL 2010 task 8

The <e1>company</e1> fabricates plastic <e2>chairs</e2>.
Product-Producer(e2,e1)

	Total
Training sentences	7208
Validation sentences	792
Test sentences	2717

Relation	Total	(e1,e2)	(e2,e1)
Cause-Effect	1331	478	853
Component-Whole	1253	632	621
Content-Container	732	527	205
Entity-Destination	1137	1135	2
Entity-Origin	974	779	195
Instrument-Agency	660	119	541
Member-Collection	923	110	813
Message-Topic	895	700	195
Product-Producer	948	431	517
Other	1864		

Relations	Description
Cause-Effect	An event or object leads to an effect. Example: Smoking causes cancer.
Instrument-Agency	An agent uses an instrument. Example: laser printer
Product-Producer	A producer causes a product to exist. Example: The farmer grows apples.
Content-Container	An object is physically stored in a delineated area of space, the container. Example: Earth is located in the Milky Way.
Entity-Origin	An entity is coming or is derived from an origin (e.g., position or material). Example: letters from foreign countries
Entity-Destination	An entity is moving towards a destination. Example: The boy went to bed.
Component-Whole	An object is a component of a larger whole. Example: My apartment has a large kitchen.
Member-Collection	A member forms a nonfunctional part of a collection. Example: There are many trees in the forest.
Message-Topic	An act of communication, whether written or spoken, is about a topic. Example: The lecture was about semantics.

[Hendrickx et al.,2009]

Dataset : SEMEVAL 2010 task 8

Evaluation Measure : Macro-Averaged F1-Score

- Nine proper classes are considered for calculating macro-averaged F1-score
- "Other" is excluded
- Official scorer script is provided with the dataset

$\mathcal{C}_p$ = { Cause-Effect, Instrument-Agency, Product-Producer, Content-Container, Entity-Origin, Entity-Destination, Component-Whole, Member-Collection, Message-Topic }

$|\mathcal{C}_p|$ = Number of classes in $\mathcal{C}_p$

$$Precision(\mathcal{P}_c) = \frac{TP_c}{TP_c + FP_c} \qquad Recall(\mathcal{R}_c) = \frac{TP_c}{TP_c + FN_c} \qquad F1\_score(\mathcal{F}_c) = \frac{2 * \mathcal{P}_c * \mathcal{R}_c}{\mathcal{P}_c + \mathcal{R}_c}$$

$$macro\_averaged\_F1\_score(\mathcal{F}) = \frac{\sum_{c \in \mathcal{C}_p} \mathcal{F}_c}{|\mathcal{C}_p|}$$

Results : Proposed Models

Performance of our Proposed Models for Relation Classification on SEMEVAL 2010 task 8 dataset

Model	F1-Score
BGRU-E	83.08
BGRU-A	83.30
BGRU-M	83.53
BGRU-ME	83.63
CBGRU-M	84.19
CBGRU-ME	84.35
CBGRU-A	84.69
CBGRU-MEA	**84.87**

BGRU	Bidirectional GRU
CBGRU	CNN-BGRU
M	Max Pooling
E	Entities Pooling
A	Attention Pooling

Results : Combined Models

Performance of our Combined Models for Relation Classification on SEMEVAL 2010 task 8 dataset

		Combined Model F1-Score	
Model	**Model F1-Score**	**Prob. Sum**	**Voting**
CBGRU-ME	84.35 84.33 84.27	84.56	84.65
CBGRU-A	84.69 84.65 84.63	85.10	84.69
CBGRU-MEA	84.87 84.83 84.81	**85.79**	**85.80**
CBGRU-ME CBGRU-A CBGRU-MEA	84.35 84.69 84.87	**85.50**	**85.62**
CBGRU-ME CBGRU-A	84.35 84.69	85.13	---
CBGRU-A CBGRU-MEA	84.69 84.87	85.34	---
CBGRU-ME CBGRU-MEA	84.35 84.87	85.43	---

BGRU	Bidirectional GRU
CBGRU	CNN-BGRU
M	Max Pooling
E	Entities Pooling
A	Attention Pooling

Results : Hyperparameters

	BGRU-{*}	CBGRU-{*}
Batch Size	100	128
Learning Rate	1	1
Optimizer	Adadelta	Adadelta
L2 Regularization Weight	0.00001	0.00001
Max Sentence Length	101	101
Word Embedding Vector Size	300	300
CNN Output Features (No of Filters)	-	256
CNN Filters Size (Window Length)	-	3
GRU Output Features	300	64
Dropout: Embedding	0.72	0.64
Dropout: BGRU	0.5	-
Dropout: CNN	-	0.32
Dropout: Pooling Layer	0.5	0.32
Early stopping epochs	20	20

Results : Confusion Matrix

Confusion Matrix for a combined model.
Three CBGRU-MEA models, with F1-scores – 84.87, 84.83, 84.81 – respectively, are combined by voting scheme.
Combined model F1 - 85.80

	Classified As												
	CE	**CW**	**CC**	**ED**	**EO**	**IA**	**MC**	**MT**	**PP**	**O**	**SUM**	**DIR_X**	**SUM**
CE	301	0	0	0	7	0	0	1	2	15	326	2	328
CW	1	250	5	2	1	9	8	6	1	22	305	7	312
CC	0	0	175	5	2	0	1	0	0	6	189	3	192
ED	0	2	7	270	0	0	0	1	0	11	291	1	292
EO	3	0	2	3	239	2	0	1	2	6	258	0	258
IA	0	3	0	2	3	122	0	0	5	20	155	1	156
MC	0	4	0	2	2	0	214	1	0	10	233	0	233
MT	0	0	0	1	3	0	1	248	0	7	260	1	261
PP	4	3	1	3	5	6	1	3	192	12	230	1	231
O	10	25	28	25	32	15	40	39	24	216	454	0	454
SUM	**319**	**287**	**218**	**313**	**294**	**154**	**265**	**300**	**226**	**325**	**2701**	**16**	**2717**

- All relations are the most confused with "Other" and vice-versa
- A larger number of wrong directions for **CW** relation
- Confused Group: **CW-MC**

CE	Cause-Effect	IA	Instrument-Agency
CW	Component-Whole	MC	Member-Collection
CC	Content-Container	MT	Message-Topic
ED	Entity-Destination	PP	Product-Producer
EO	Entity-Origin	O	Other
DIR X	Classified with wrong direction		

Comparison with Competing Methods - I

Comparison with Relation Classification methods which
do not use extra linguistic features and special ranking objective
Dataset - SEMEVAL 2010 task 8

Classifier	Additional Information	F1
Att-BLSTM [Zhou et al., 2016]	Word embeddings, Positional indicator	84.0
att-BGRU [Kim and Lee, 2017]	Word embeddings, Range restricted	84.1
EAtt-BiGRU [Qin et al., 2017]	Word embeddings, Positional embeddings	84.7
Our Model: CBGRU-MEA	**Word embeddings, Positional indicator**	**84.8**
Our Model: Combined CBGRU-MEA	**Word embeddings, Positional indicator**	**85.8**

Comparison with Competing Methods - ||

Comparison with Relation Classification methods which
do not use extra linguistic features and use special ranking objective
Dataset - SEMEVAL 2010 task 8

Classifier	Additional Information	F1
CR-CNN [dos Santos et al., 2015]	Word embeddings, Position embeddings, Special ranking objective	84.1
Our Model: CBGRU-MEA	**Word embeddings, Positional indicator**	**84.8**
ER-CNN+R-RNN [Vu et al., 2016]	Word embeddings, Extended middle context, Ensemble, Voting, Special ranking objective	84.9
Our Model: Combined CBGRU-MEA	**Word embeddings, Positional indicator**	**85.8**
Att-Input-Pooling-CNN [Wang et al., 2016]	Word embeddings, Position embeddings, Special ranking objective	88.0

Comparison with Competing Methods - |||

Comparison with Relation Classification methods which use extra linguistic features
Dataset - SEMEVAL 2010 task 8

Classifier	Additional Information	F1
SVM [Rink and Harabagiu, 2010]	POS, WordNet, Pre xes and other morphological features, Dependency parse, Levin classed, PropBank, FanmeNet, NomLex-Plus, Google n-gram, Paraphrases, TextRunner	82.2
MVRNN [Socher et al., 2012]	Word embeddings, POS, NER, WordNet	82.4
CNN [Zeng et al., 2014]	Word embeddings, Position embeddings, WordNet	82.7
FCM [Yu et al., 2014]	Word embeddings, Dependency parsing, NER	83.0
DepNN [Liu et al., 2015]	Word embeddings, SDP, NER	83.6
SDP-LSTM [Xu et al., 2015b]	Word embeddings, SDP, POS, GR, WordNet	83.7
Ensemble Methods [Nguyen and Grishman, 2015]	Word embeddings, SDP, POS, NER, WordNet, CNNs+RNNs+Voting	84.1
SPTree [Miwa and Bansal, 2016]	Word embeddings, SDP, POS	84.4
Our Model: CBGRU-MEA	**Word embeddings, Positional indicator**	**84.8**
depLCNN [Xu et al., 2015a]	Word embeddings, SDP, WordNet, Word around nominals, Negative sampling from NYT dataset	85.6
Our Model: Combined CBGRU-MEA	**Word embeddings, Positional indicator**	**85.8**
DRNN [Xu et al., 2016]	Word embeddings, SDP, POS, GR, WordNet, Data augmentation, (w/o data augmentation - 84.2)	86.1
BRCNN [Cai et al., 2016]	Word embeddings, SDP, POS, NER, WordNet, Data augmentation	86.3

Experiments

Performance of model CBGRU-MEA with POS and WordNet tags

Model	Model F1-Score	F1-Score	
		Probabilities Sum	Voting
CBGRU-MEA + WordNet	84.54	85.35	85.23
CBGRU-MEA + POS	84.64		
CBGRU-MEA + POS + WordNet	84.78		

- In the model CBGRU-MEA, POS embeddings and WordNet embeddings are appended to word embeddings
- Length of embeddings: 10
- Number of POS tags: 47
- Number of WordNet tags: 51

Sentence	The	E1_START	company	E1_END	fabricates	plastic	E2_START	chairs	E2_START	.
POS	DT	E1_START	NN	E1_END	VBZ	JJ	E2_START	NNS	E2_START	.
WordNet	OTHER	E1_START	noun.group	E1_END	verb.creation	noun.substance	E2_START	noun.artifact	E2_START	OTHER

Conclusion and Future Work

Conclusion

Proposed Models for Relation Classification

- BGRU based models : BGRU-M, BGRU-ME
- CNN-BGRU based models : CBGRU-ME, CBGRU-A, CBGRU-MEA
- Two-layered CNN-BGRU based models outperform the one-layered BGRU based models
- The model CBGRU-MEA reports the highest F1-score of **84.87** on SEMEVAL 2010 task 8 dataset, without using external NLP tools

Feature Pooling

- Novel entities pooling over BGRU features to pool entity specific features, which improves the results of relation classification
- Used three different kinds of pooling together, namely max, entities, and attention pooling
- Using all kinds of pooling together improves the results

Combined Models

- We experimented on combining architecturally same and different models by applying two methods, namely a voting scheme and summing softmax probabilities, which further improve the F1-score to **85.80** on SEMEVAL 2010 task 8 dataset

Future Work

- Deep architectures with data augmentation techniques
- Modelling relation classification task as a multi-task learning
- Incorporating AMR information for relation classification
- Alternate architectures to use linguistic features effectively

Thank You

Questions ?