Extracting Relations Between Chemicals and Genes in Biomedical Text

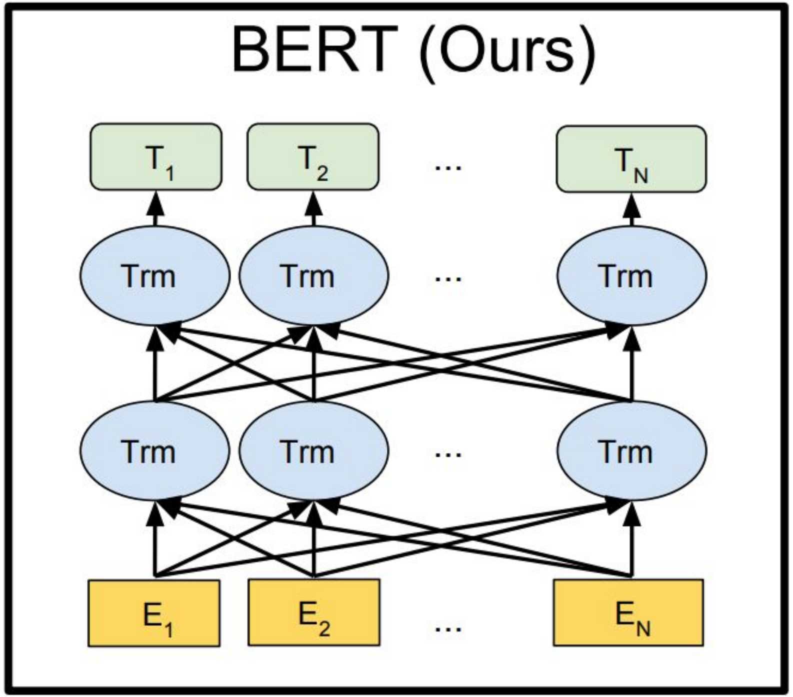
Chloë Smart

Motivation

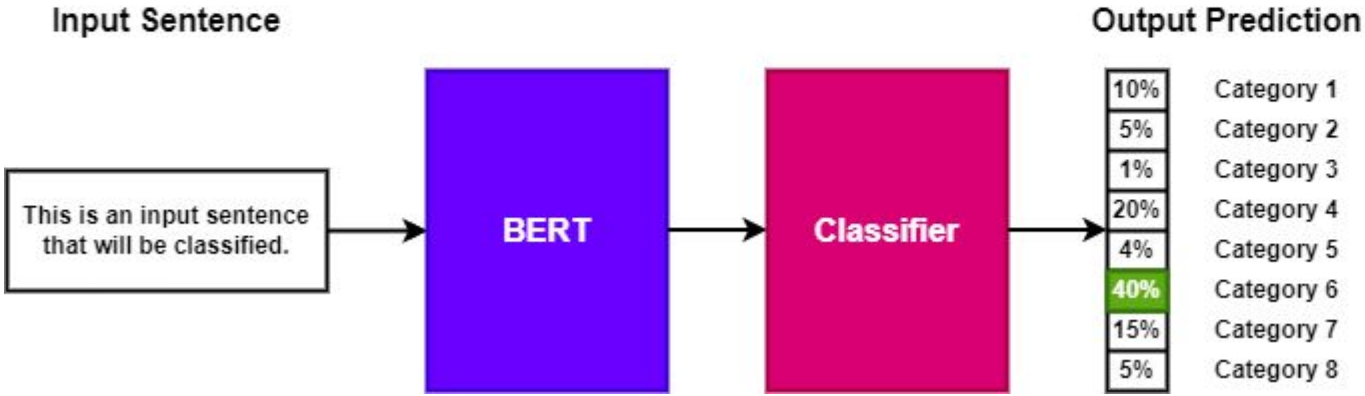
- Large amounts of research needed to develop a drug
- Causes bottlenecks within drug design and discovery
- Solution is automation with deep learning models

BERT

- BERT pre-trains bidirectional representations from text
- Can be used to perform biomedical relation extraction.
- PubMedBERT uses domain-specific pre-training

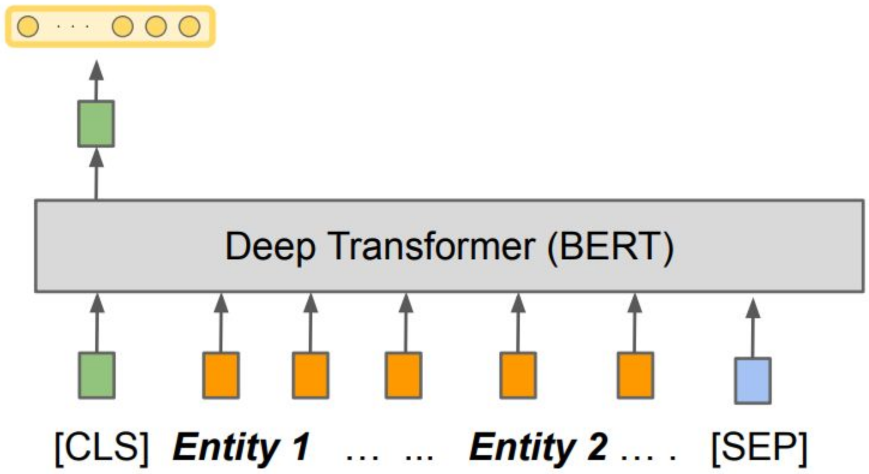


Text Classification



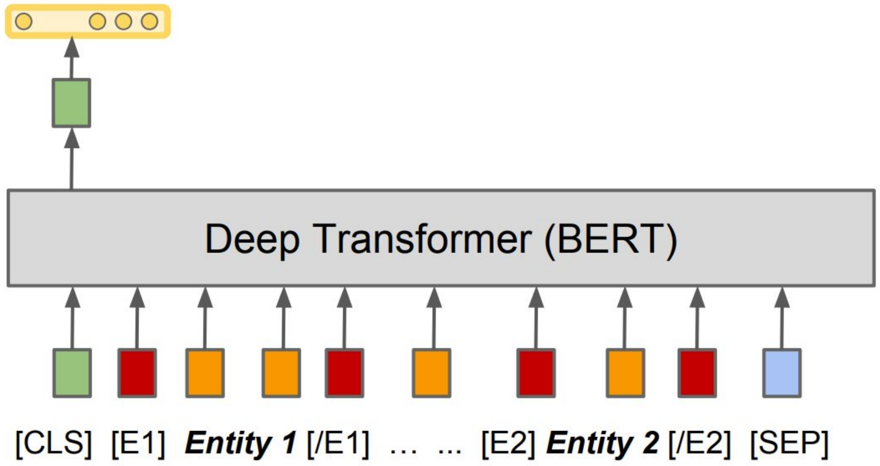
Sentence Representations

- [CLS] embeddings tend to represent the sentence as a whole.
- Can extract [CLS] embeddings and classify them



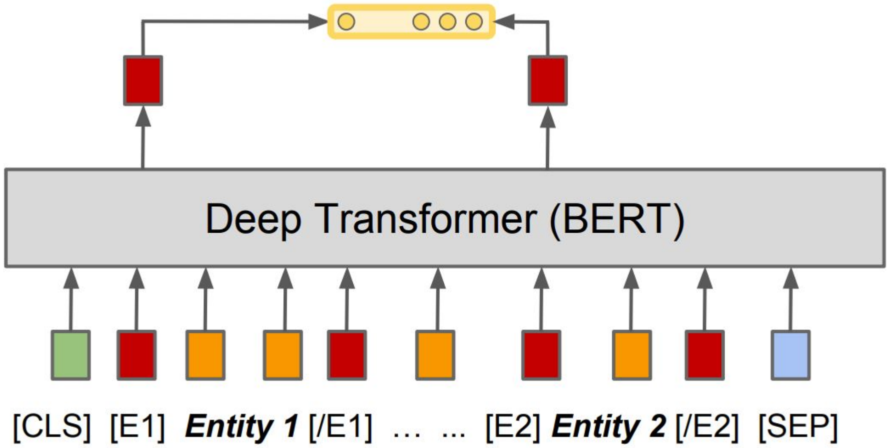
Entity Markers

- Can add tags around entities
- Gives BERT more information about which entities are of interest
- Could improve [CLS] representation



Entity Representations

- Could get more information from the entities themselves
- Extract beginning tags of each entity and concatenate them together
- **Will classifying entity representations be more effective than classifying [CLS] representations?**



Entity Information

- How much entity information does BERT require within the text input?

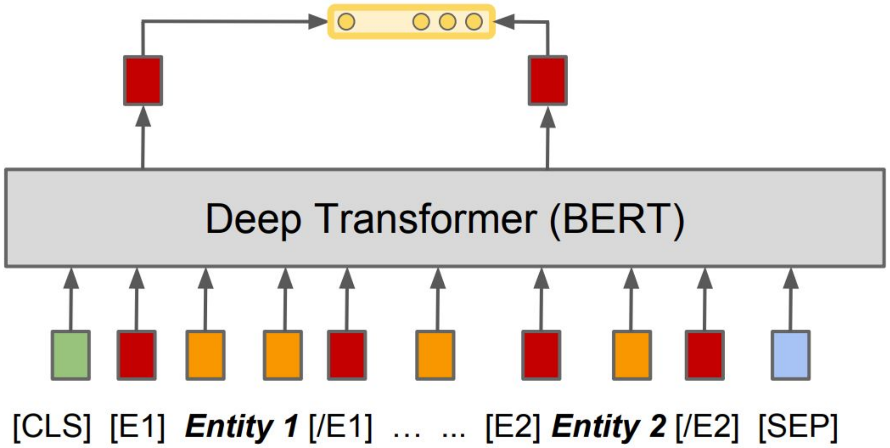
“**Paris** is the capital of **France**.”

“**City** is the capital of **Country**.”

“_ is the capital of _.”

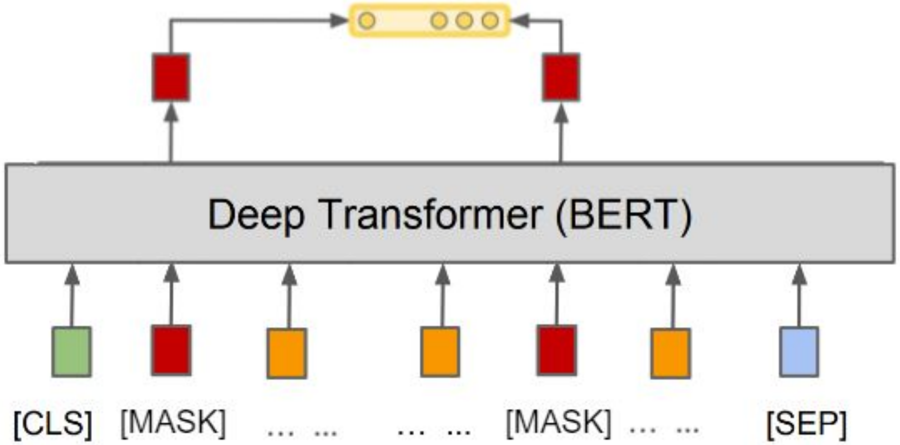
Context and Mention

- Contains both the entity mentions and surrounding context
- Equivalent structure to previous question



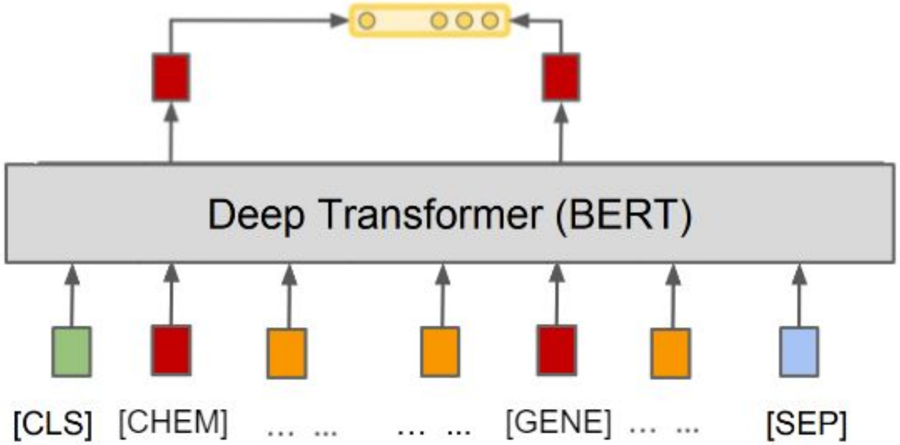
Context

- Can represent only context by using [MASK] tags
- Doesn't provide information about entities



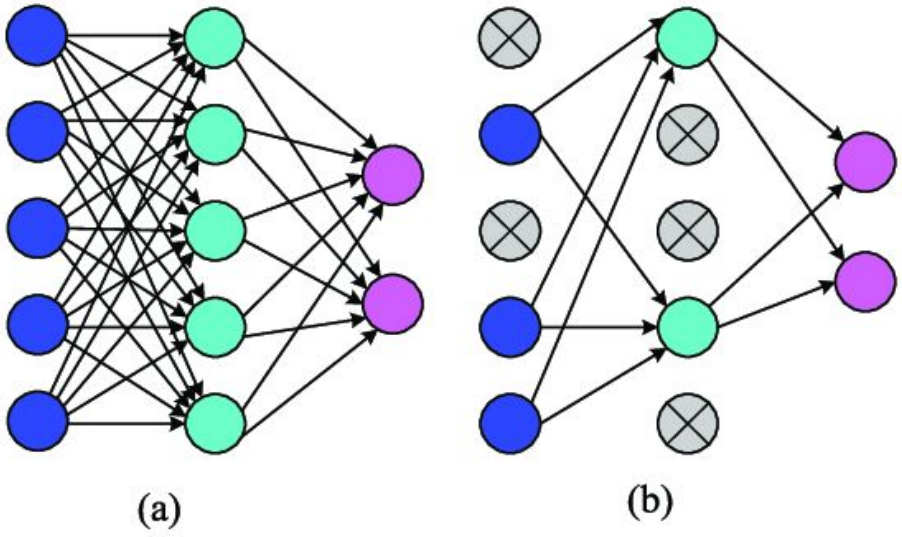
Context and Type

- [CHEM] for chemicals, [GENE] for genes
- Might give BERT more ideas about the surrounding context
- This is the ‘Special masking’ input

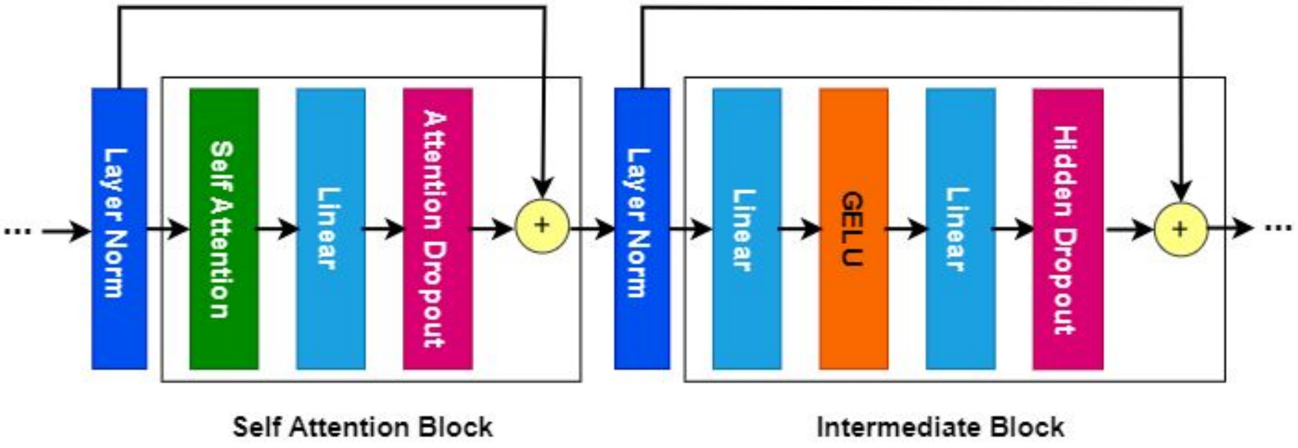


Dropout

- Can use dropout to regularise
- Promotes generalisation
- **What is the optimal dropout level for different sentence representations?**



Method

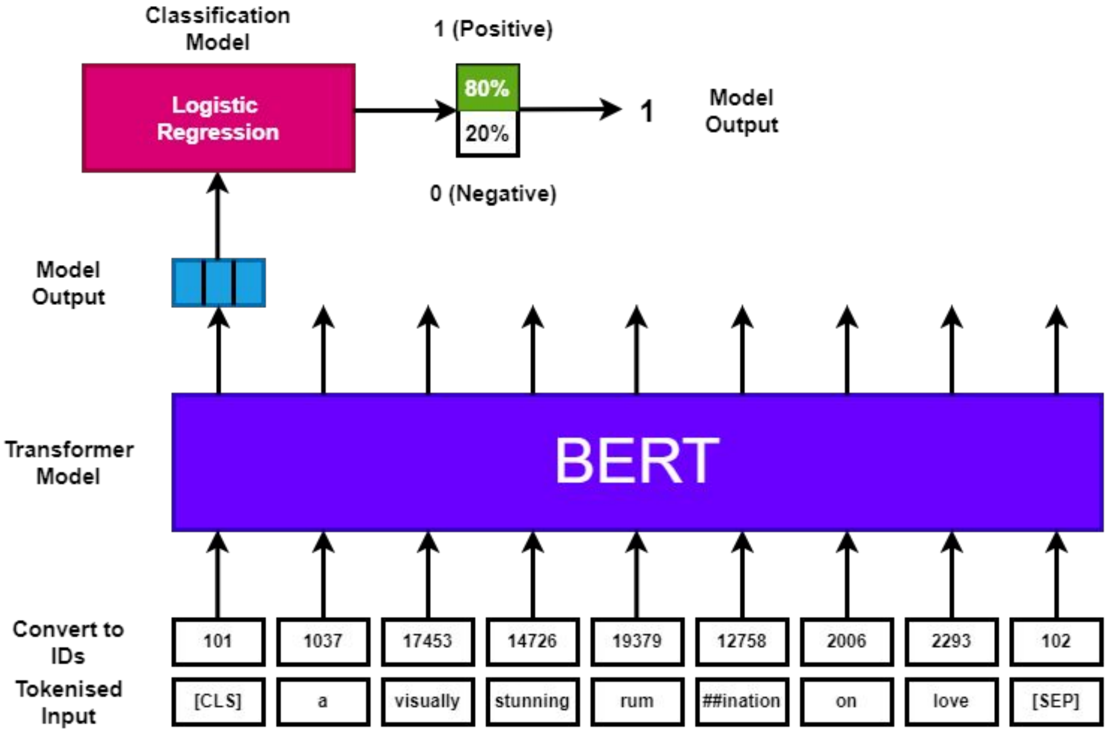


- Adjust **Attention_probs_dropout** and **Hidden_dropout_prob** parameters
- Values will range from 0.1-0.7, incrementing by 0.1 each time

DrugProt Dataset

- DrugProt is a dataset that contains information about the interactions between drugs and proteins.
- The dataset was created by collecting data from various sources, including public databases and literature.

Feature Extraction



Cartesian Product

Entity Pair	Tagged Sentence	Label
EACA - plasmin	[E1] EACA [/E1] inhibited the binding of [E2] plasmin [/E2] to gp330 slightly more than the binding of plasminogen to gp330.	INHIBITOR
EACA - gp330	[E1] EACA [/E1] inhibited the binding of plasmin to [E2] gp330 [/E2] slightly more than the binding of plasminogen to gp330.	NONE
EACA - plasminogen	[E1] EACA [/E1] inhibited the binding of plasmin to gp330 slightly more than the binding of [E2] plasminogen [/E2] to gp330.	INHIBITOR
EACA - gp330	[E1] EACA [/E1] inhibited the binding of plasmin to gp330 slightly more than the binding of plasminogen to [E2] gp330 [/E2].	NONE

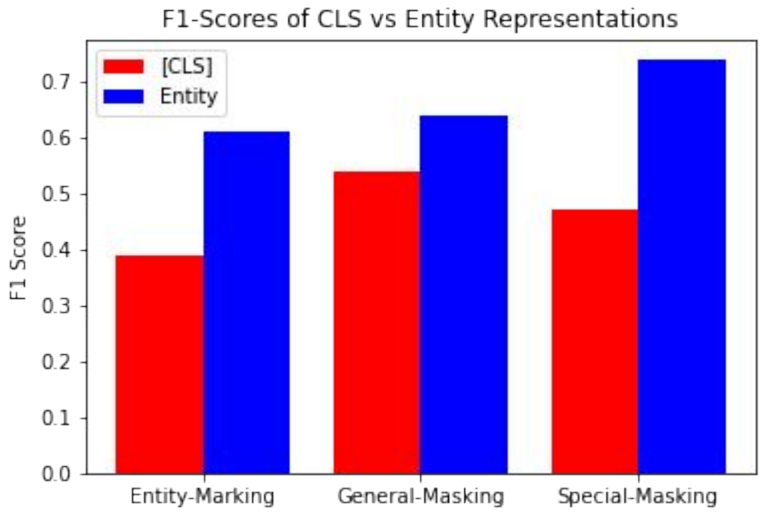
Results For Representations

Embeddings	F1-Score	Precision	Recall
[CLS]	0.39	0.74	0.35
[E1]/[E2]	**0.61**	**0.81**	**0.56**

- **Will classifying entity representations be more effective than classifying [CLS] representations? YES!**
- Could be potentially due to different vector sizes ([CLS] has size 768, [E1]+[E2] has size 1536).
- Increased dimension reduces confusion regarding class boundaries for logistic regression.

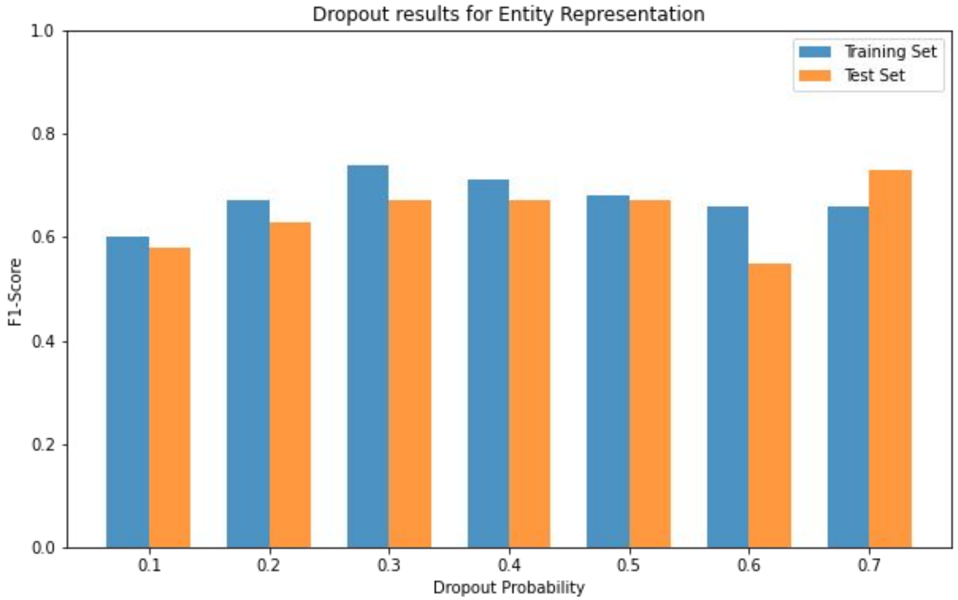
Results for Context

- **How much entity information does BERT require within the text input? CONTEXT + TYPE!**
- Using input with context and type provided the best results
- [CLS] representations were best with the input with just context



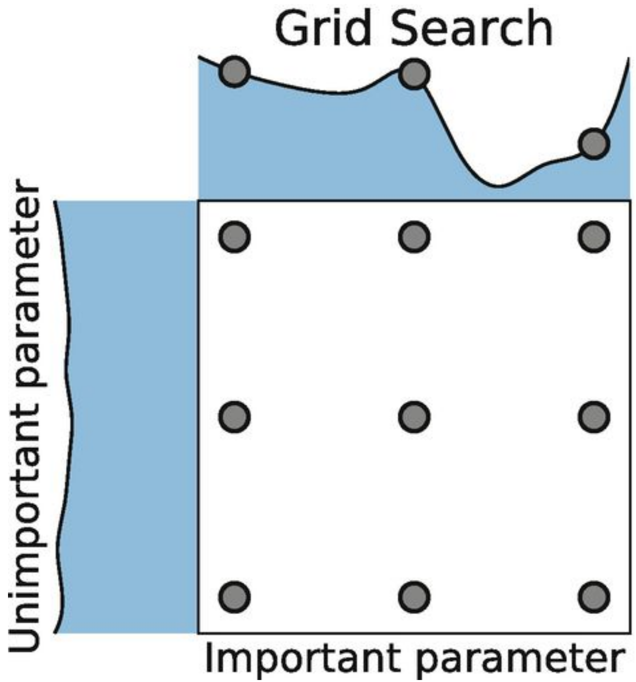
Results For Dropout

- **What is optimal dropout level for different sentence representations?**
- Best dropout = 0.5
- Smallest difference between training and test sets
- Excessive dropout causes degradation in performance



Discussion

- Can improve logistic regression by tuning its hyperparameters
- Grid search gave hyperparameters that improved F1 Scores by 10\%



Applying Results

- Used optimal logistic regression parameters
- Entity representations
- Context and type sentences
- 0.5 dropout
- Best relation F1 Score: 90% for PART-OF relation
- Overall model F1 score: 78%
- Shows promising results for biomedical relation extraction