Problem:

1. Overfitting
2. Matching based on Siamese at word level has many OOV due to OCR errors
3. Matching based on Siamese makes silly mistakes (Matching Name to Phone numbers)

Solution 1:

1. Use text classifier:

Train NER following: <https://towardsdatascience.com/named-entity-recognition-ner-meeting-industrys-requirement-by-applying-state-of-the-art-deep-698d2b3b4ede>

Based on char level ELMO to avoid OOV in case of

Tags:

<https://docs.google.com/spreadsheets/d/1FG0TLpaGOdaOsCXZ16_aVgf3v_BJk44nzetK0iG7-fI/edit#gid=0>

<https://www.kaggle.com/abhinavwalia95/entity-annotated-corpus>

Use SpaCy to train on new NERs:

https://spacy.io/api/annotation#named-entities

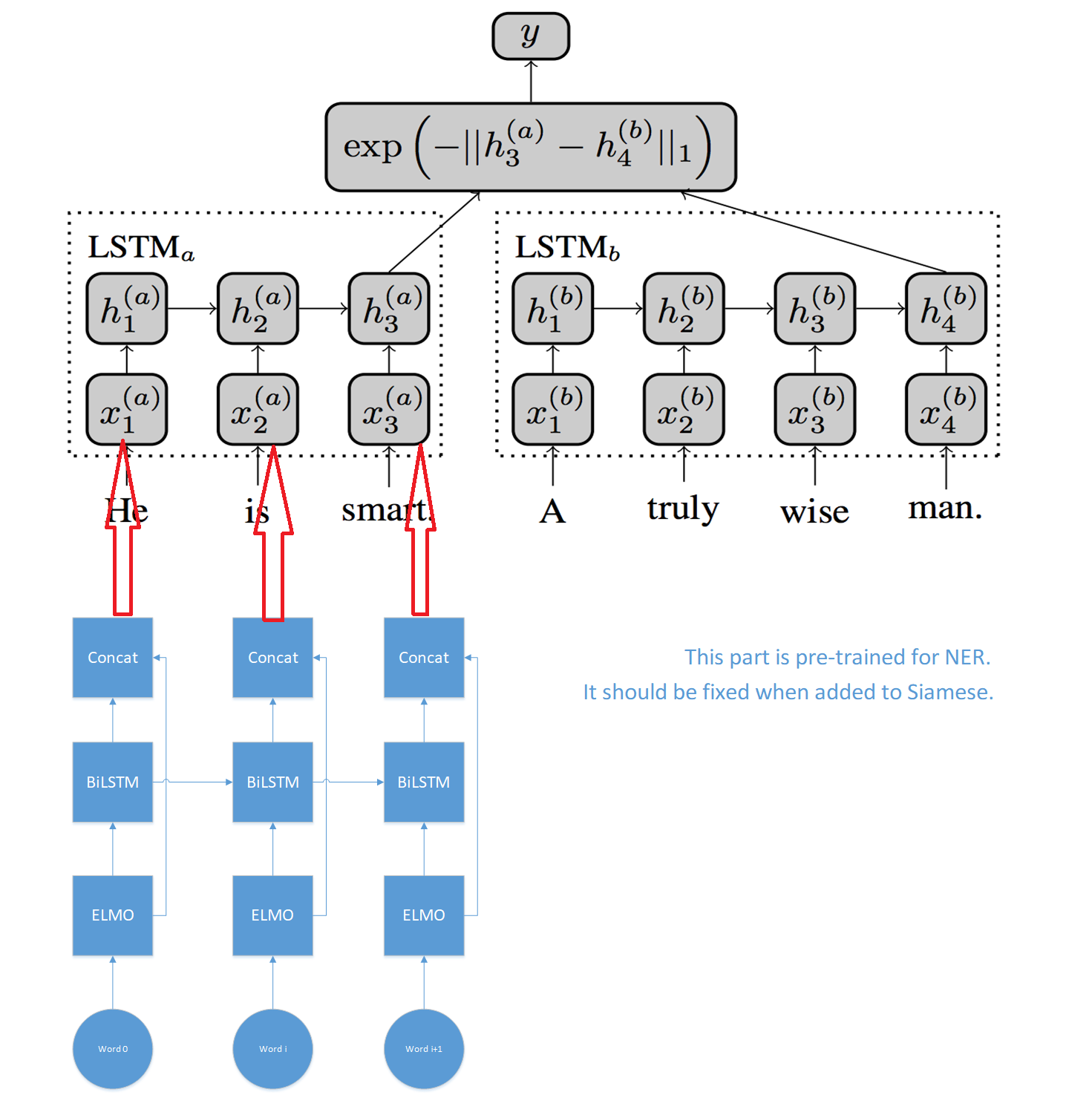
https://spacy.io/usage/examples

1. Use as input features to Siamese similarity network to improve matching

Solution 2:

Use

<https://spacy.io/usage/examples> https://github.com/explosion/spaCy/blob/master/examples/information\_extraction/entity\_relations.py





1. Use the text features with Siamese Network

