Primer AI's Systems for Acronym Identification and Disambiguation

Nicholas Egan, John Bohannon

contact: *negan at primer.ai*

Summary

- We built a transformer based model for acronym identification and an information retrieval system for acronym disambiguation
- Model performance was improved on both tasks by using distantly supervised auxiliary datasets we call AuxAI and AuxAD
- We identified duplication issues with the SciAD dataset, which we remedied in a dataset we call SciAD-dedupe

Our three datasets are available at github.com/PrimerAI/sdu-data

Acronym Identification (AI)

Extract short and long form acronyms within a sentence

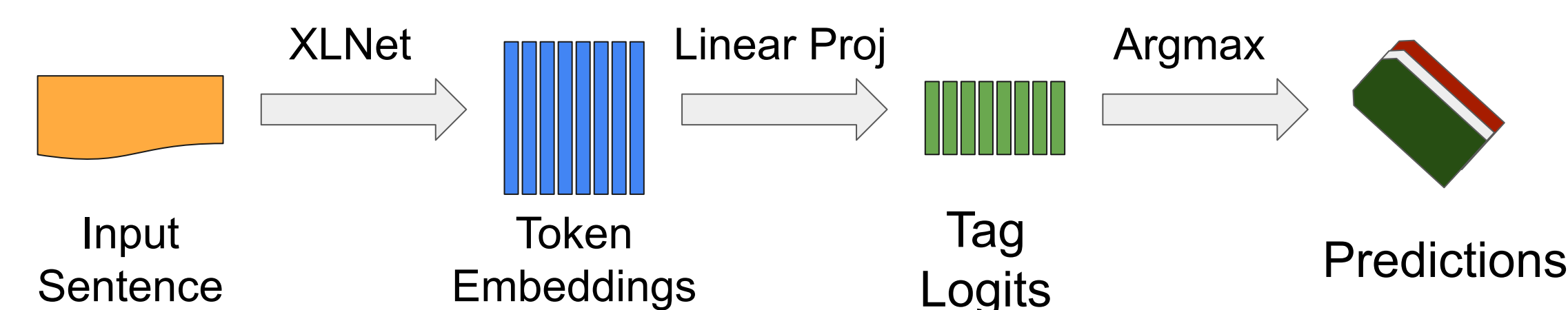a Convolutional Neural Network or CNN is used in ML

Acronym Disambiguation (AD)

Determine the expansion of a particular acronym given the sentence context

a Convolutional Neural Network or CNN is used in ML

Maximum Likelihood   Malware Landscape   Machine Learning

AI Model

XLNet based token-tagging approach:

Input Sentence → XLNet → Token Embeddings → Linear Proj → Tag Logits → Argmax → Predictions

We averaged logits across an ensemble of XLNet models trained with different hyperparameters on SciAI and AuxAI

AD System

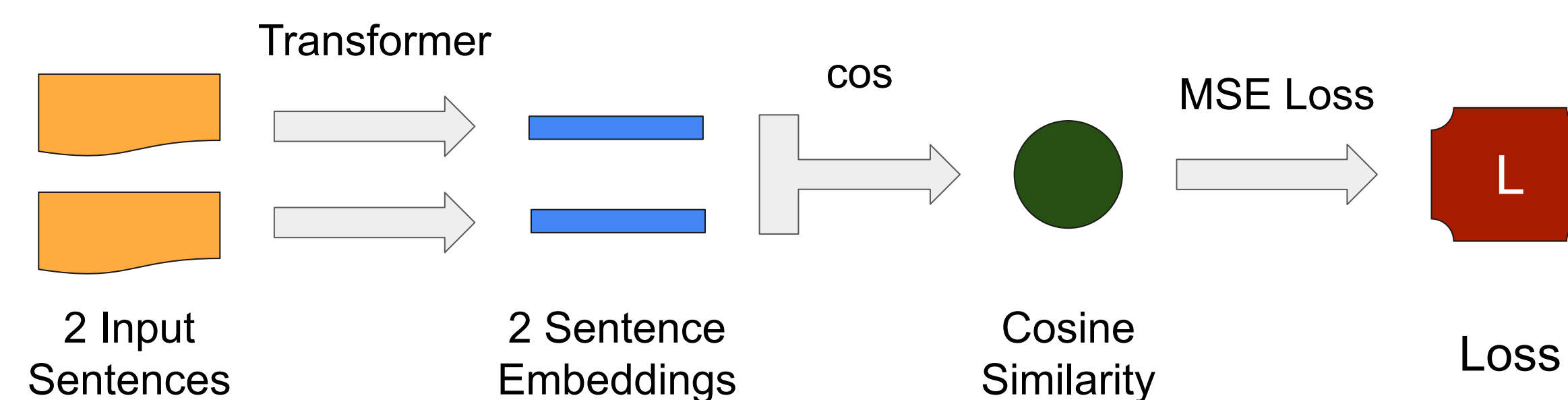**Training Time:** compute sentence embedding for every training example
Inference Time:
1. compute sentence embedding for test example
2. find closest training example by cosine similarity
3. use its label

Encoder Models for AD

Cosine similarity was computed across an ensemble of encoders:
- SIF (Arora, Liang, and Ma 2017)
- Sentence Transformers (Reimers and Gurevych 2019)
- Transformers trained as Twin Networks on SciAD and AuxAD:

2 Input Sentences → Transformer → 2 Sentence Embeddings → cos → Cosine Similarity → MSE Loss → Loss

AuxAI

314k additional AI training examples formed by:
1. Scraping acronyms with expansions from Abbreviations.com
2. Finding arXiv abstracts with those acronyms and expansions
3. Labeling an acronym in a sentence if:
   a. The acronym and expansion appear in the abstract
   b. OR the acronym is a common short form

AuxAD

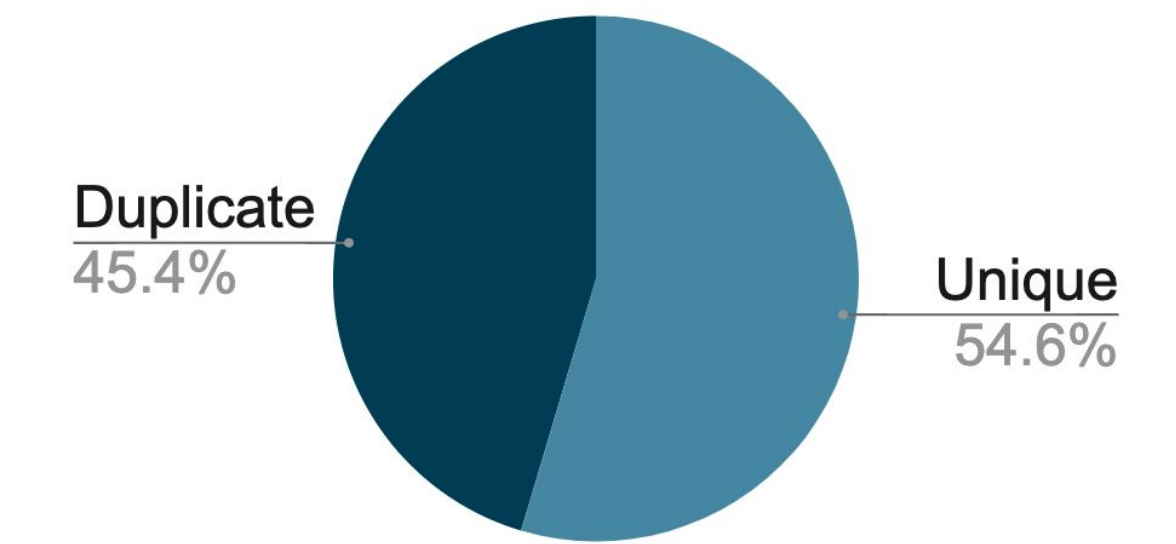113k additional AD training examples formed by:
1. Finding arXiv abstracts with acronyms/expansions in SciAD dictionary
2. Using the one-sense-per-discourse assumption to label other sentences in the abstract with the short form as corresponding to the long form

Duplicates in SciAD

The SciAD contains many duplicate examples with the same (sentence, acronym) pair.

The overlap of data between train and dev/test suggests that SciAD is a biased measurement of performance.

Duplicates in SciAD Dev

Duplicate 45.4%   Unique 54.6%

Interestingly, duplicate examples do not always have the same label:

- **10.5%** of duplicated examples in train/dev contain multiple labels
- **93.1%** of dev examples with a duplicate in train share a label with at least one of the duplicates
- **10.8%** of dev examples with a duplicate in train have a conflicting label with at least one of the duplicates

We release SciAD-dedupe as a deduplicated version of SciAD

AI Results

System	Dev	Test
Baseline	85.46	84.09
LSTM-CRF	-	86.55
XLNet, SciAI	92.17	-
XLNet, AuxAI	66.96	-
XLNet, AuxAI → SciAI	93.14	-
XLNet Ensemble	**93.63**	**92.60**

macro averaged F1 scores for SciAI dev and test

AD Results

System	Dev	Test	Ded.
Baseline	59.73	60.97	59.97
GAD	-	81.90	-
Our Ensemble	91.22	91.58	93.15

macro averaged F1 scores for SciAI dev, SciAI test, and SciAI-dedupe dev

Our AI system ranked 4th on the leaderboard, and our AD system ranked 7th