SQuAD: 100,000+ Questions for Machine Comprehension of Text

Rajpurkar et. al.

Presented By:
Manash
Co-founder, Bengali.AI
Research Assistant - Data Science SHOP UP

Lessons to learn from this paper

- How to build dataset for a unique problem
- Create baseline models
- Analyze a dataset based on **specific metrics (**create one when none is available**)**

Abstract

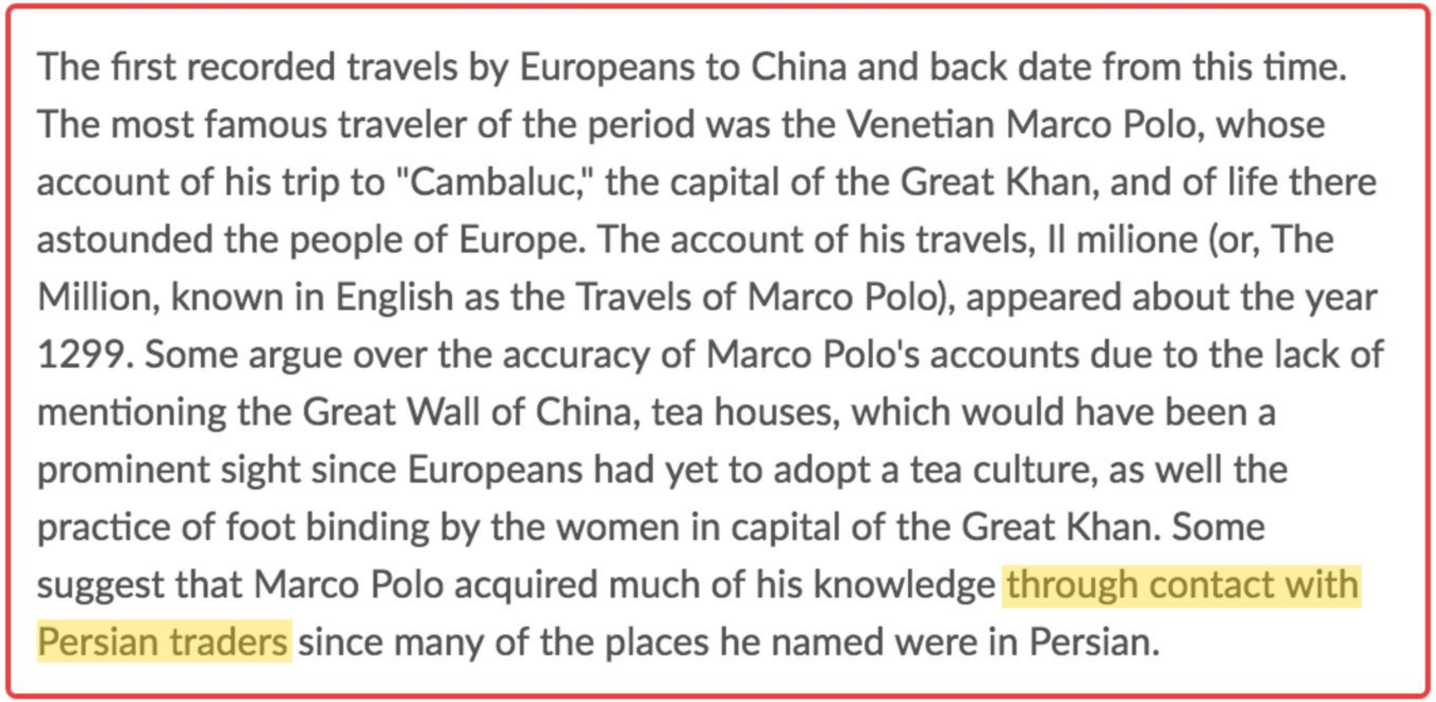
- **S**tanford **Qu**estion **A**nswering **D**ataset (SQuAD) is a **reading comprehension** dataset

- A strong **multinomial logistic regression model** was built which achieved an **F1 score** of **51.0%**

- Where human performance is 86.8%

- The main challenge is to build such dataset and see how machine performs on this abstract level problem.

Example



The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, Il milione (or, The Million, known in English as the Travels of Marco Polo), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge through contact with Persian traders since many of the places he named were in Persian.

How did some suspect that Polo learned about China instead of by actually visiting it?

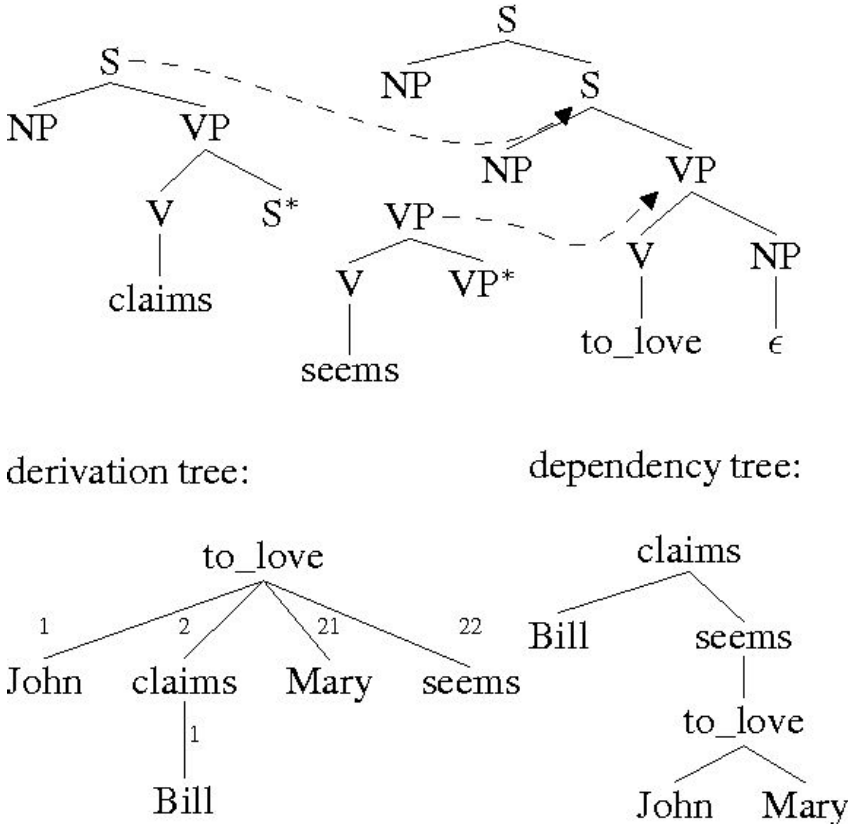
Answer: through contact with Persian traders

Properties of the dataset

- It does not provide a **list of answer choices for each question**.
    - System must select the answer from all possible spans in the passage.
- **Rich diversity** of answer types in SQuAD
- Automatic **techniques to quantify diversity**

Assessment of Difficulty : Which Method Worked

- A logistic regression model was built with a range of features.
  - Lexicalized and dependency tree path feature



Bengali.AI Community

Assessment of Difficulty : Which Method Didn’t Work

- Model performance **worsens** with increasing complexity of

1. Answer types

2. Syntactic divergence between the question and the sentence containing the answer; **but there is no degradation for humans**.

Existing Datasets

- Reading comprehension
- Open-domain Question Answering
- Cloze datasets
    - Answers are single word based

NLP Dataset Collection

Dataset Collection

Three stages of dataset collection:

1. Curating passages
2. Crowdsourcing question-answers on those passages
3. Obtaining additional answers

Dataset Analysis

1. **Diversity in Answers**

   a. Separating numerical and non numerical answers

   b. Non numericals are categorized using **POS** tags generated by Stanford CoreNLP

2. **Reasoning required to answer questions**

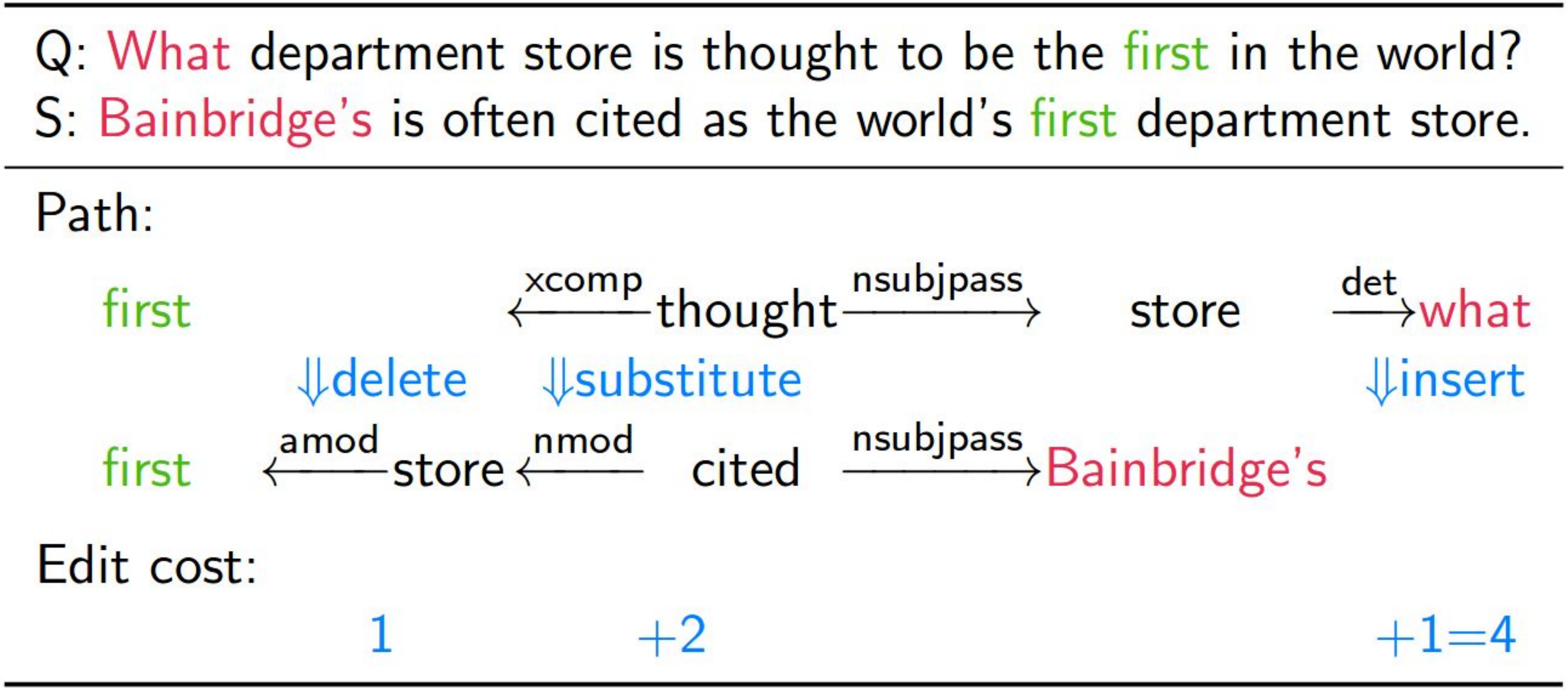
   a. Labeling total dataset into following categories

      i. **Lexical variation** (synonymy) - Question and answer sentences synonymous
      ii. **Lexical variation** (word knowledge) - Require word knowledge to resolve connection
      iii. **Syntactic variation** - Paraphrased form of a question doesn't match with the answer
      iv. **Multiple sentence reasoning -** Higher level fusion of multi sentence
      v. **Ambiguous**

Dataset Analysis (contd.)

3. **Stratification by syntactic divergence**

    - An automatic method to quantify the syntactic divergence



Q: What department store is thought to be the first in the world?

S: Bainbridge's is often cited as the world's first department store.

Path:

$$\text{first} \xleftarrow{\text{xcomp}} \text{thought} \xrightarrow{\text{nsubjpass}} \text{store} \xrightarrow{\text{det}} \text{what}$$

⇓delete ⇓substitute ⇓insert

$$\text{first} \xleftarrow{\text{amod}} \text{store} \xleftarrow{\text{nmod}} \text{cited} \xrightarrow{\text{nsubjpass}} \text{Bainbridge's}$$

Edit cost:

1 +2 +1=4

Methods to Solve

Candidate Answer Generation

- Assuming there are M tokens in a passage, there are O(M^2) possible answers to each question.

- Stanford CoreNLP was use to prune large possible answers.

Sliding Window Baseline

- For each candidate answer, all candidates that have maximum overlap were kept. Then answers were chosen using sliding window based method.

Methods to Solve (contd.) : Logistic Regression

Logistic Regression

- In this model, **several features** from each candidate answer were extracted.

Model Details

- **Loss:** Multiclass log-likelihood
- **Optimizer:** Adam (Learning Rate = 0.1)
- **Regularization:** L2
- **Regularization Strength:** 0.1 / number of batches

$$\begin{aligned}\text{minimize} \quad & -\frac{1}{M}\sum_{i=1}^{M}\log\frac{e^{W_{y_i}^T f(\mathbf{x}_i)+b_{y_i}}}{\sum_{j=1}^{C} e^{W_j^T f(\mathbf{x}_i)+b_j}}\\ \text{subject to} \quad & \|f(\mathbf{x}_i)\|_2=\alpha,\ \ \forall i=1,2,...M,\end{aligned}$$

Methods to Solve (contd.) : Features for LR Model

Matching Word Frequencies (Unigram & Bigram)

- Sum of TF-IDF of the words that occur in both the **question and the sentence containing the candidate answer.**

Root Match

- Matching the root of parsing dependencies from answer and question sentence.

Lengths

- Number of words to the left, right, inside span and in the whole sentence.

Methods to Solve (contd.) : Features for LR Model

Span Word Frequencies

- Sum of the TF-IDF of the words in the span of question, regardless of their position.

Constituent Label

- Parse tree label of the span, optionally combined with the wh-word in the question.

Span POS Tags

- Sequence of the POS tags in the span.

Methods to Solve (contd.) : Features for LR Model

Lexicalized

- Lemmas of question words combined with the lemmas of words within distance 2 to the span in the sentence based on the dependency parse trees.

Dependency Tree Paths

- For each word occurs in both the Q&A, the path in the dependency parse tree from the word in the sentence to the span.

Model Evaluation

1. **Exact Match**

2. **(Macro-averaged) F1 Score**

    - Loosely measures the **average overlap** between the prediction and ground truth answer.

    - Prediction and ground truth are considered as **bags of tokens**, then compute their F1.

    - Maximum F1 over all of the ground truth answers are taken for a given question, then average the score over all questions.

Human Performance

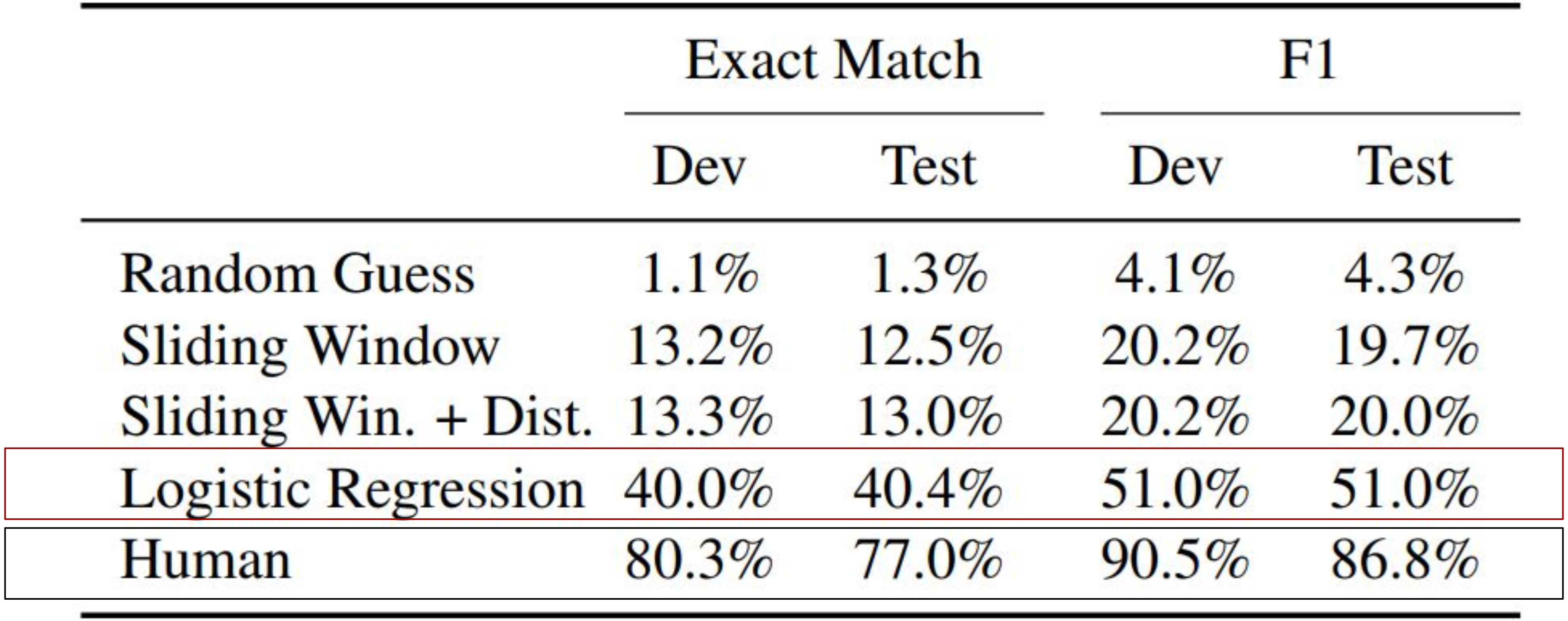
Exact Match Metric:

- 77% On Test Set

F1 Metric:

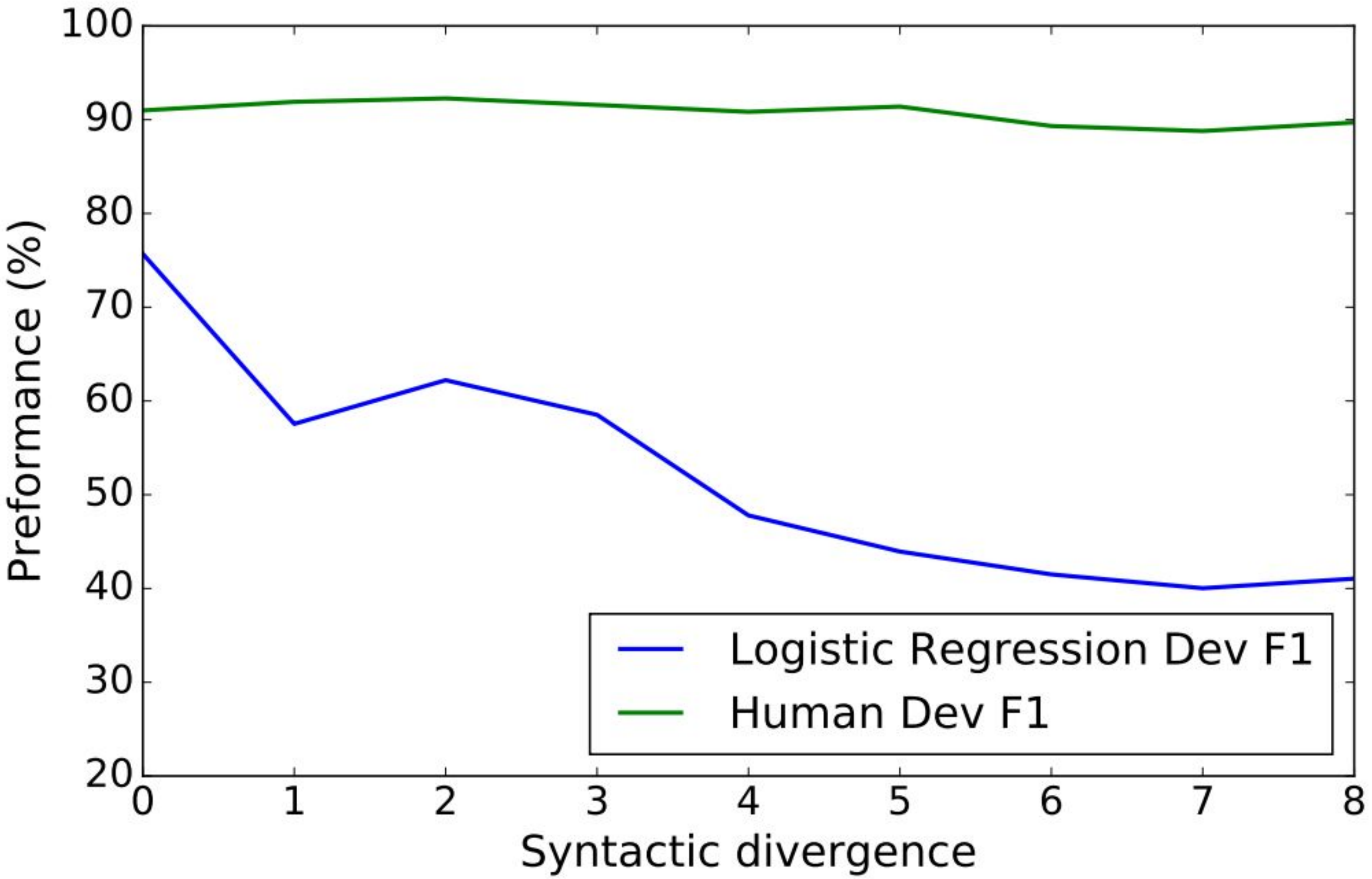
- 86.8% on Test Set

Performance Comparison



	Exact Match		F1	
	Dev	Test	Dev	Test
Random Guess	1.1%	1.3%	4.1%	4.3%
Sliding Window	13.2%	12.5%	20.2%	19.7%
Sliding Win. + Dist.	13.3%	13.0%	20.2%	20.0%
Logistic Regression	40.0%	40.4%	51.0%	51.0%
Human	80.3%	77.0%	90.5%	86.8%

Effect of Syntactic Divergence



Conclusion

- Main goal is to view datasets as a way to guide progress towards the end goal of NLU.

- Based on some unique features of this dataset it differs from the most of the traditional RC and Q&A datasets.

- A Logistic Regression model was proposed along with the dataset as a baseline model.

Thank you all