Towards Content Transfer through Grounded Text Generation

Shrimai Prabhumoye, Chris Quirk, Michel Galley

Carnegie Mellon University
Language Technologies Institute

Microsoft
Research

Motivation

- We constantly author text
- AI assistance deals with *form* (grammar, style, etc.)
- Our goal is to control for *content*

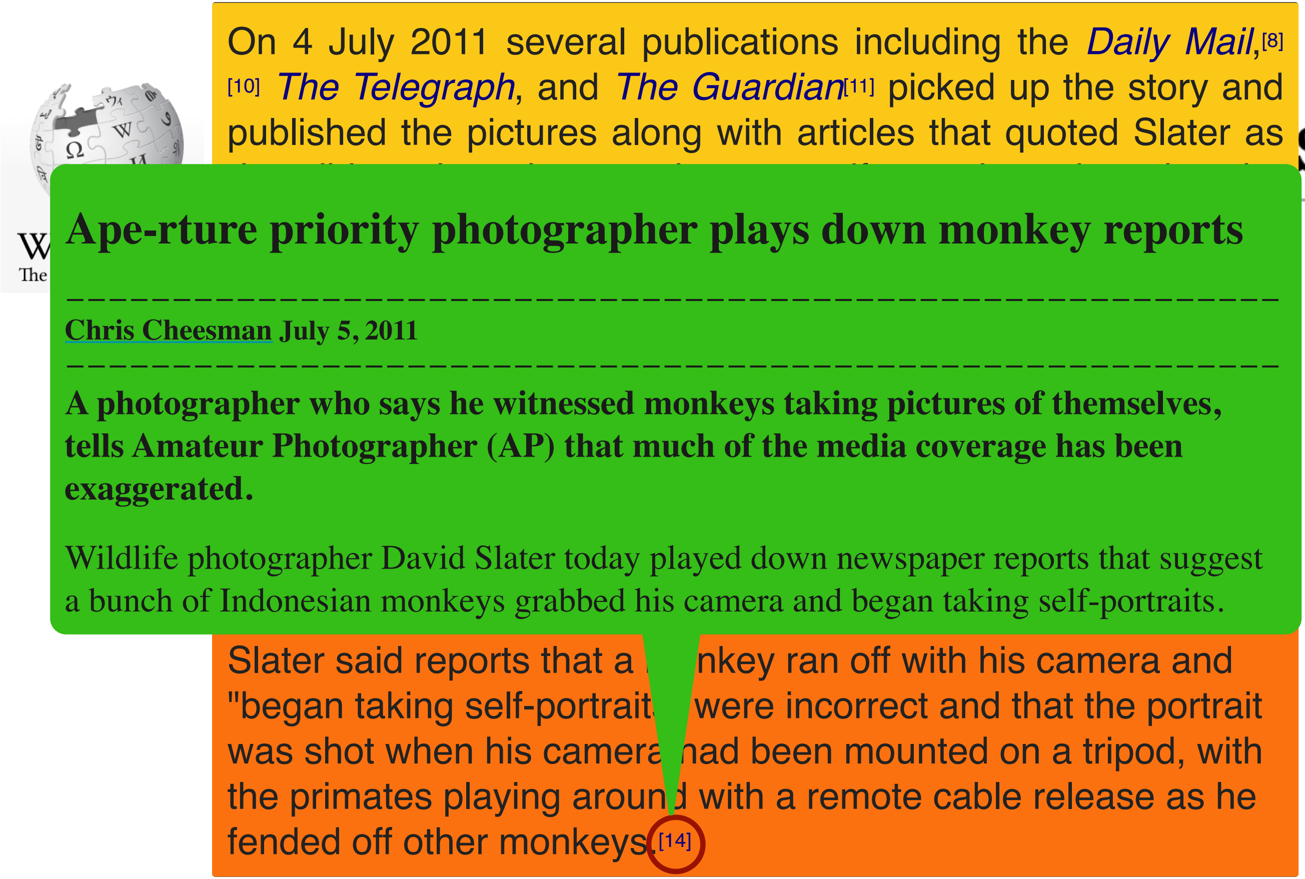
After graduate form Columbia University, Obama worked in Chicago.

After graduating from Columbia University, Obama worked in Chicago.

After graduating from Carnegie Mellon University, Obama worked in Chicago.

After graduating from Columbia University, Obama worked in Chicago.

What is our task?



On 4 July 2011 several publications including the *Daily Mail*,[8] [10] *The Telegraph*, and *The Guardian*[11] picked up the story and published the pictures along with articles that quoted Slater as

spute

Ape-rture priority photographer plays down monkey reports

Chris Cheesman July 5, 2011

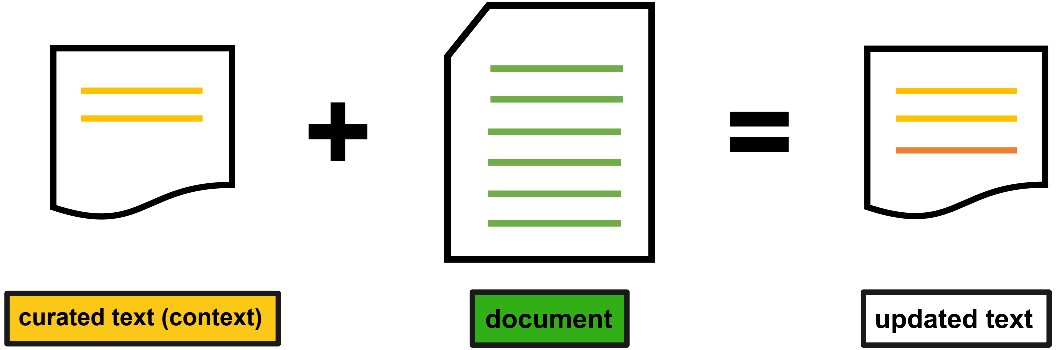
A photographer who says he witnessed monkeys taking pictures of themselves, tells Amateur Photographer (AP) that much of the media coverage has been exaggerated.

Wildlife photographer David Slater today played down newspaper reports that suggest a bunch of Indonesian monkeys grabbed his camera and began taking self-portraits.

Slater said reports that a nkey ran off with his camera and "began taking self-portrait were incorrect and that the portrait was shot when his camera had been mounted on a tripod, with the primates playing around with a remote cable release as he fended off other monkeys.[14]

Carnegie Mellon University
Language Technologies Institute

Primary Contribution



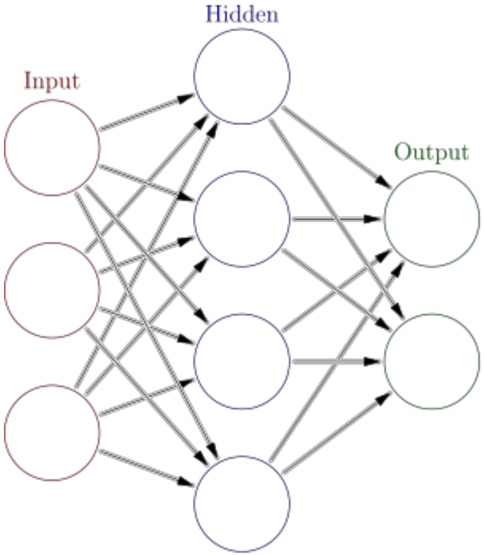
- design a task to perform content transfer from an unstructured source of information
- release dataset

Applications

- Maintain software documentation given incoming streams of text (email, software requirements etc)
- Legal precedent around a topic
- Inbox Summarization
- Updating Wikipedia articles

Overview

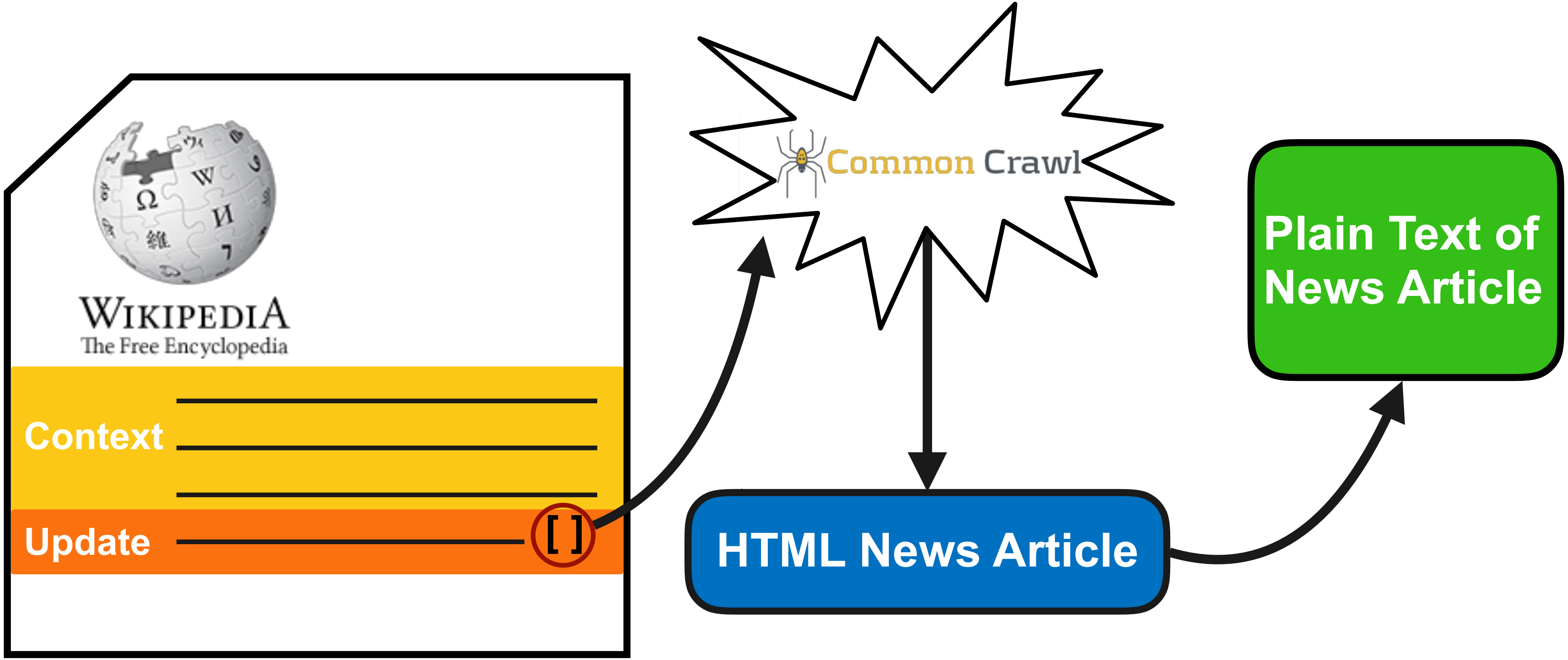
Dataset



Models

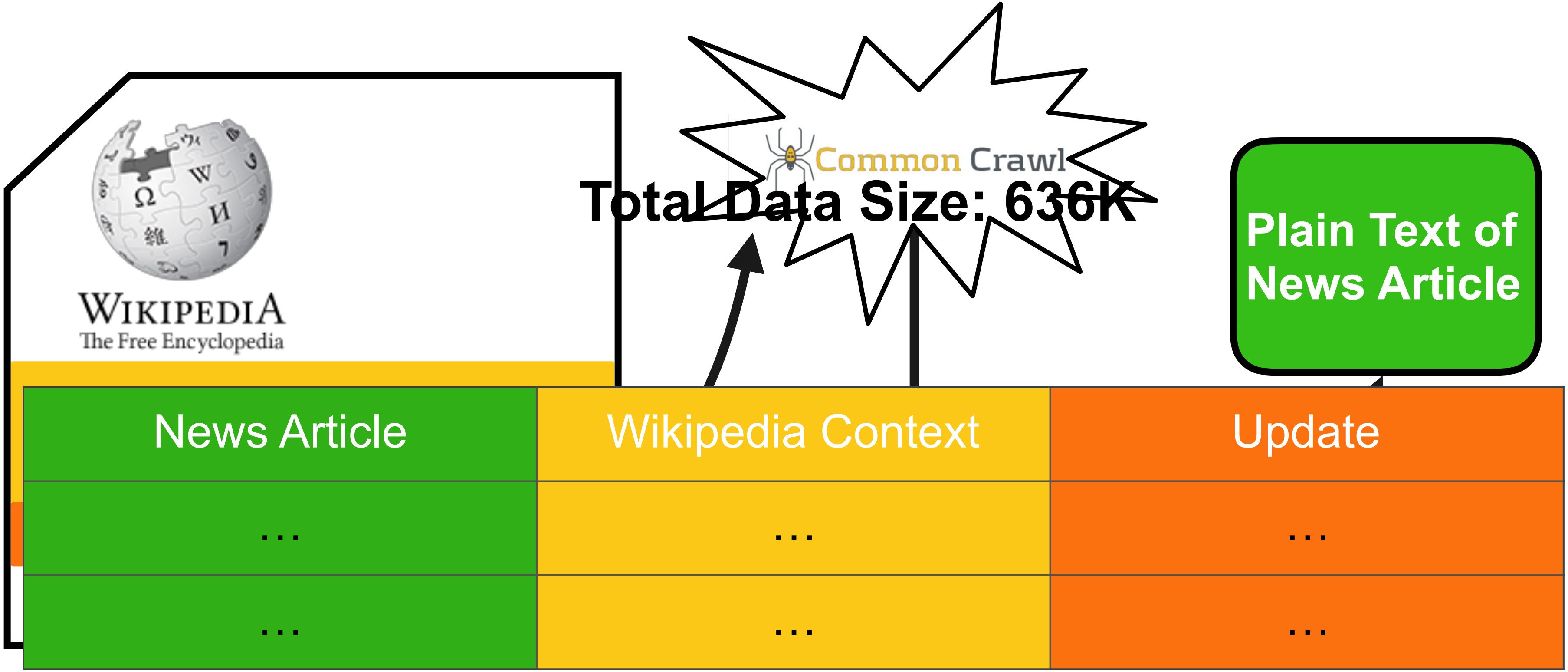
Evaluation

Data Creation Process



Carnegie Mellon University
Language Technologies Institute

Data Creation Process



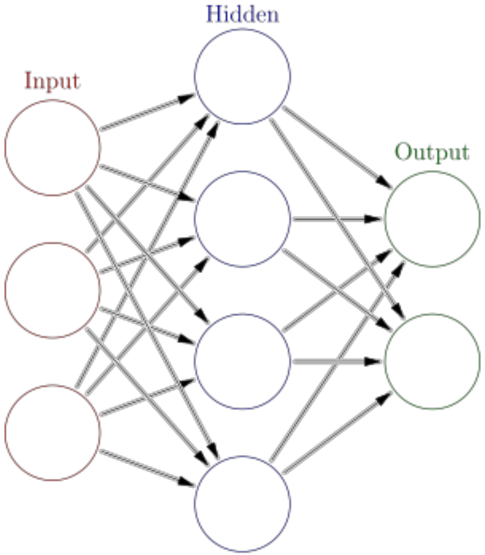
Total Data Size: 636K

Plain Text of News Article

News Article	Wikipedia Context	Update
…	…	…
…	…	…

Overview

Dataset



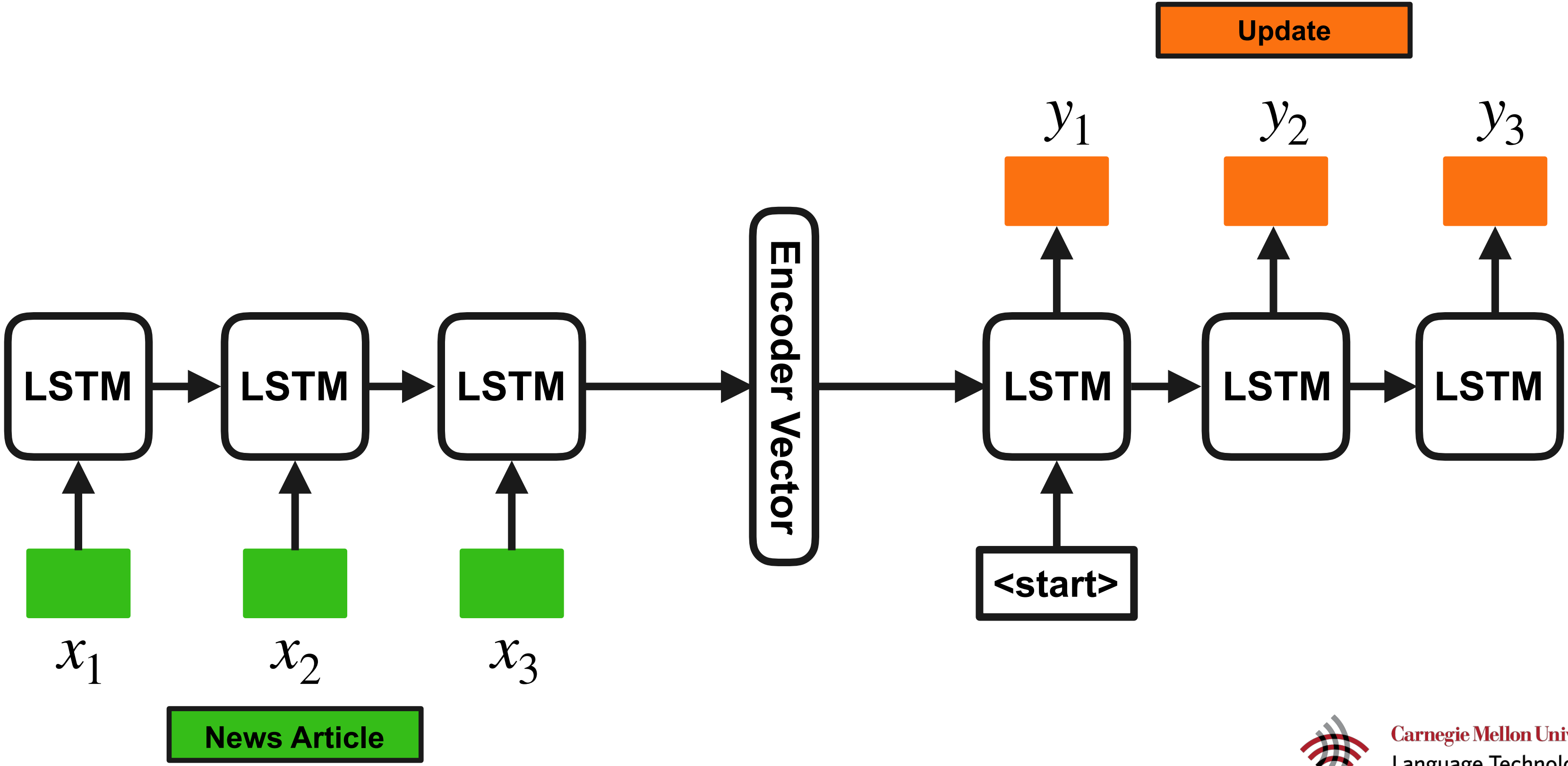
Models

Evaluation

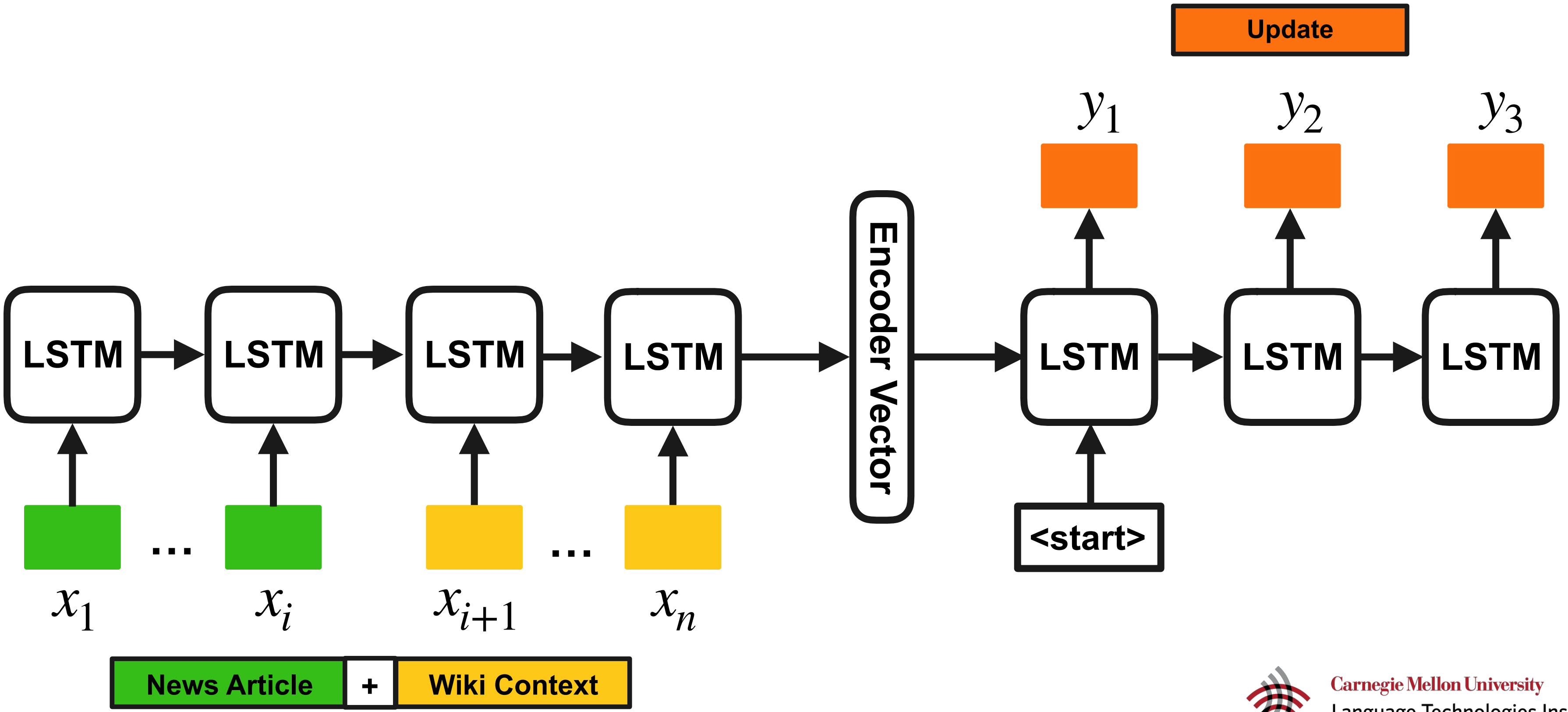
Models

- Generative Models
    - Context Agnostic Generative Model (CAG) — Baseline
    - Context Informed Generative Model (CIG)
    - Context Responsive Generative Model (CRG)
    - all models have global attention
- Extractive Models
    - SumBasic
    - Context Informed SumBasic
    - Oracle
- All models are simplistic to infer if context helps in generation

Context Agnostic Model (CAG) - Baseline

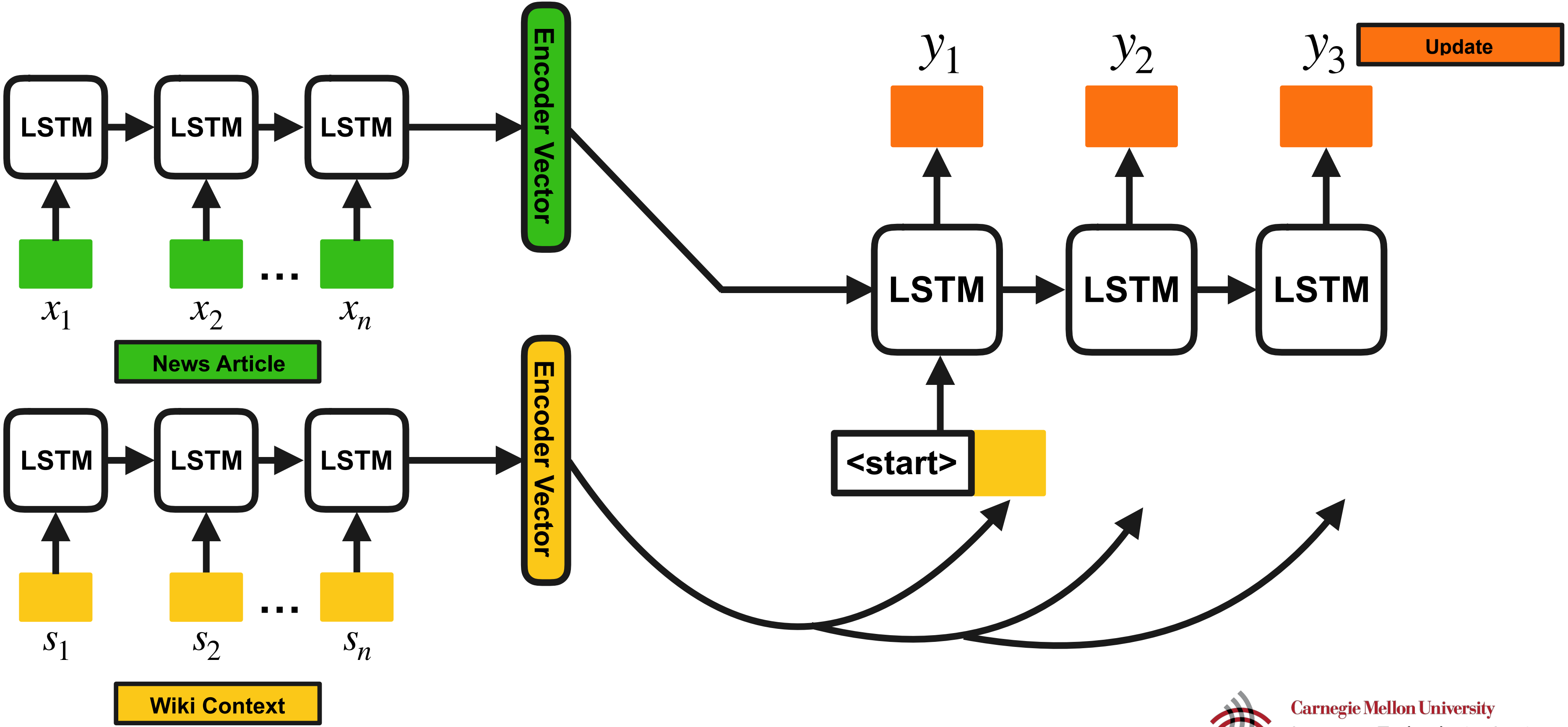


Context Informed Model (CIG)



Carnegie Mellon University
Language Technologies Institute

Context Responsive Model (CRG)

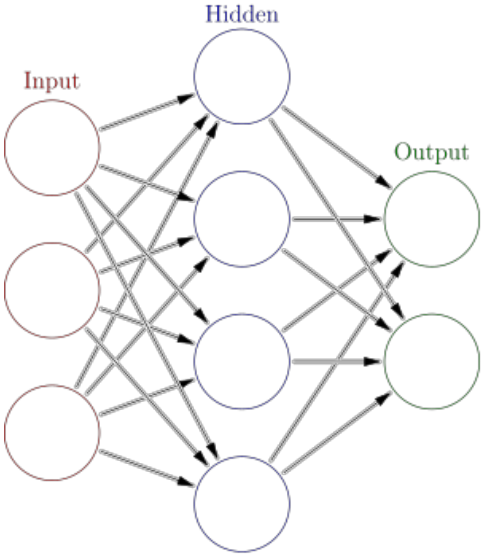


Extractive Models

- SumBasic: a model based on unigram probabilities
- Context Informed SumBasic: the unigram probabilities take into account the words in the Wikipedia context.
- Oracle: establish an upper limit attainable by extractive methods

Overview

Dataset



Models

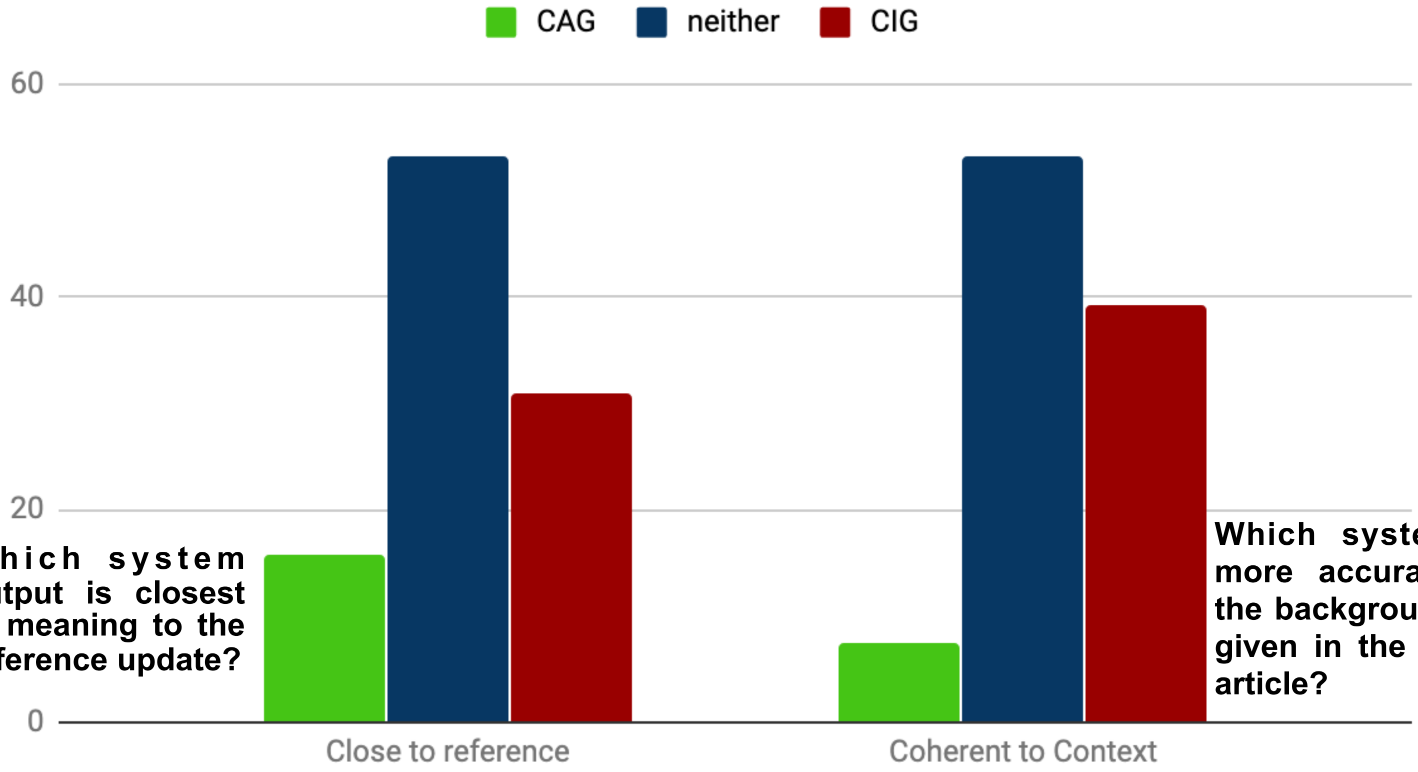
Evaluation

Automated Evaluation

Model	ROUGE-L
SumBasic	5.6 (5.6-5.7)
Context Informed SumBasic	7.0 (7.0-7.1)
Context Agnostic Generative Model	9.1 (9.0-9.2)
Context Informed Generative Model	**16.0 (15.9-16.1)**
Context Responsive Generative Model	14.7 (14.6-14.8)
Oracle	28.8 (28.7-29.0)

* METEOR and BLEU numbers are consistent with ROUGE-L

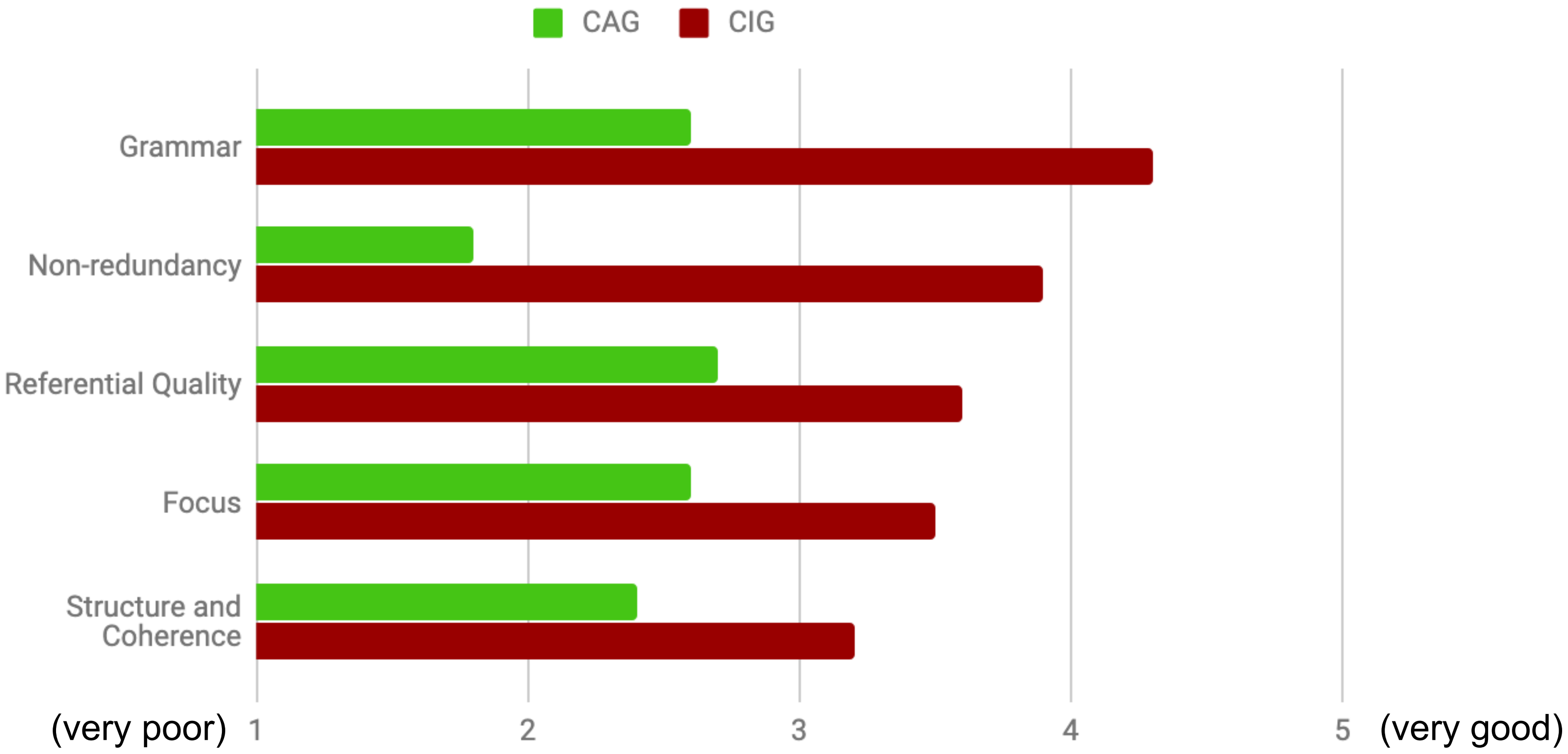
Relative Human Evaluation



Which system output is closest in meaning to the reference update?

Which system output is more accurate relative to the background information given in the snippet of the article?

Absolute Quality Evaluation



Carnegie Mellon University
Language Technologies Institute

Examples

Reference Update	CIG Generated Update
rob brydon, the comedian was born in baglan.	he was born in baglan
in may 2014 he was diagnosed with prostate cancer.	st. clair was diagnosed with prostate cancer.
on april 3, 2014, manning signed a one-year deal with the cincinnati bengals.	on march 9, 2014, manning signed a one-year contract with the cincinnati bengals.

Summary

- We have designed a new task for content transfer through grounded text generation
- Both automatic and human evaluation of our baseline models shows that context is important for the generation process — treating our task as summarization is not enough!
- The total size of the dataset is 636k
- Code and data can be found at https://github.com/shrimai/Towards-Content-Transfer-through-Grounded-Text-Generation
    - Code available for all the models
    - Raw data and the train data used in experiments

Thank You