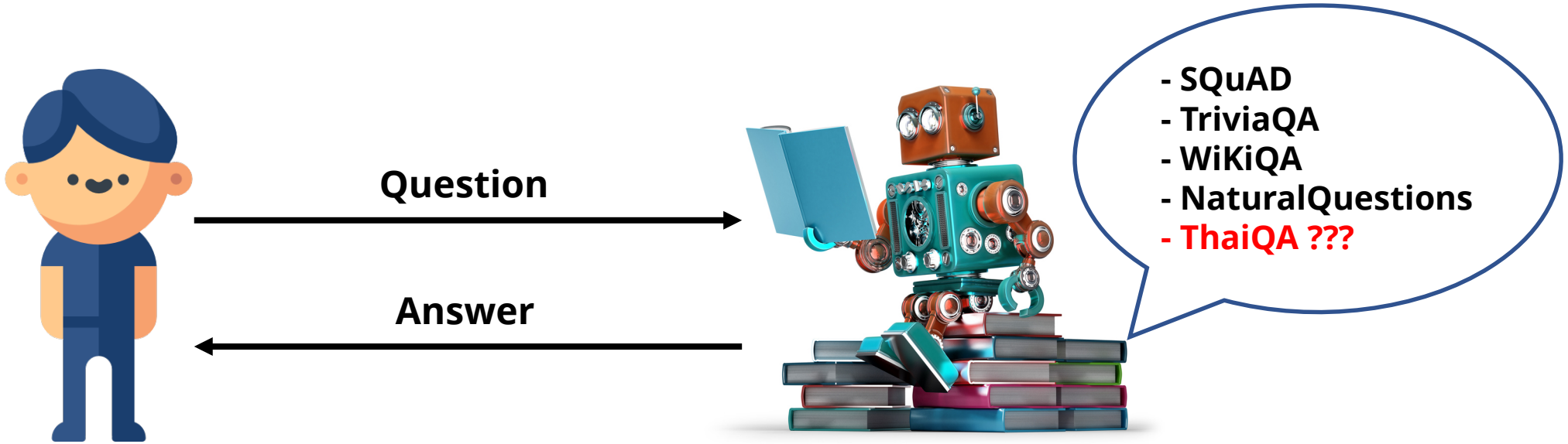
iSAI NLP 2019
The 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing
iSAI-NLP2019
Oct 30 - Nov 1, 2019 Chiang Mai, Thailand

The First Wikipedia Questions and Factoid Answers Corpus in the Thai Language

National Electronics and Computer Technology Center (NECTEC)

Motivation

- A question answering (QA) system is a challenging task in NLP.
- English QA datasets are available for research and benchmarking.
- A Thai QA dataset is a lack of benchmark datasets.



Thai QA Dataset

- A question-answer pairs dataset created by humans on a set of Thai Wikipedia articles
- An answer is a word, a segment of text, or a span appearing on a part of the corresponding reading passage.

Thai QA dataset

Types of Questions in QA Datasets

Simple (Factoid)

[what, where, who, when, which, how much/many]

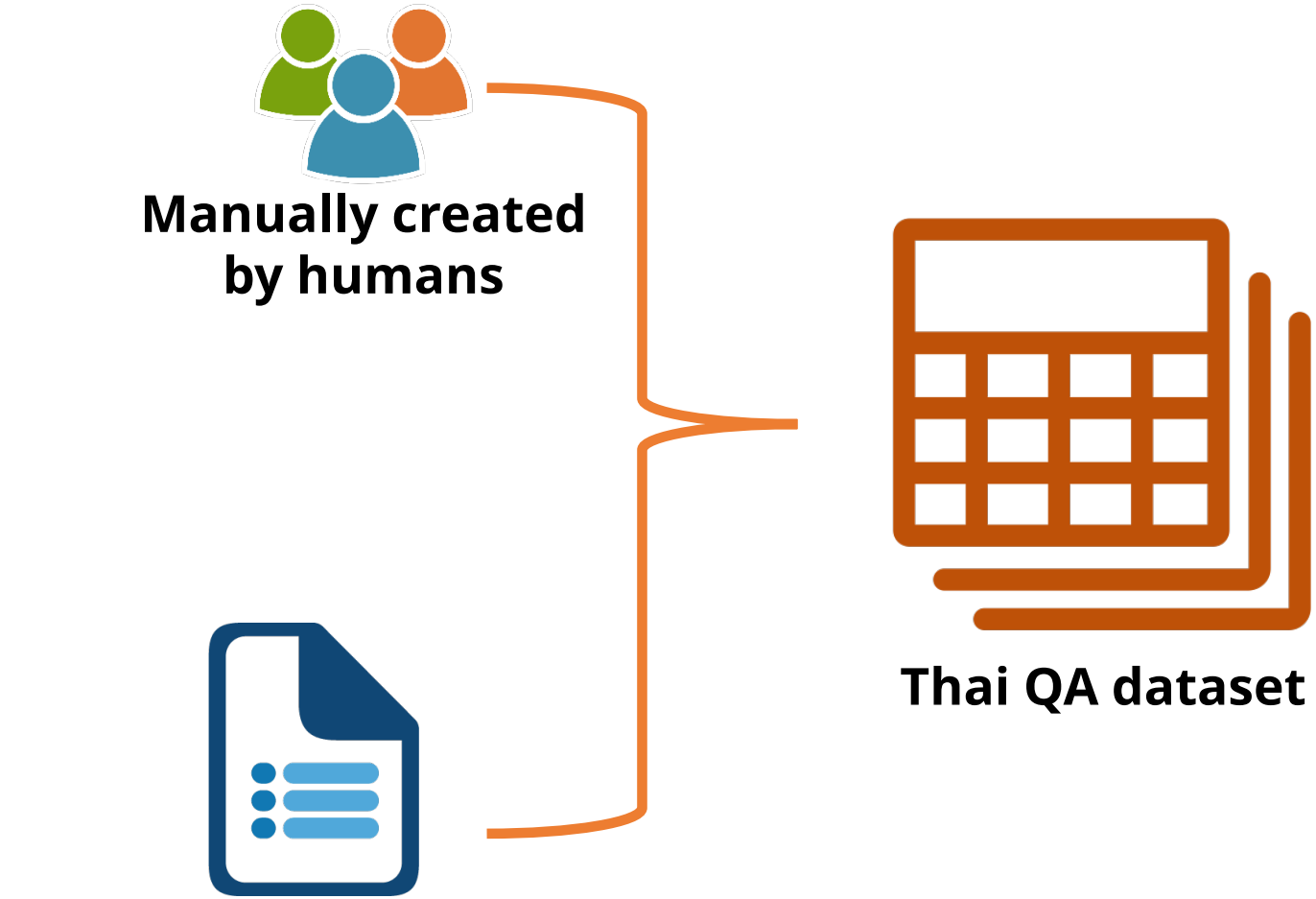
- A simple question is for a simple fact retrieved from a single document
  - Person name
  - Location
  - Date

Complex (Narrative)

[why, how]

- A question typically has long pieces as answers which may come from single or multiple documents
  - List questions
  - Hypothetical questions
  - Confirmation questions [yes/no]
  - Opinion questions

Thai QA Dataset Construction



Thai QA Annotation System

Creating both questions and answers

Annotators (Humans)

Question Answering Corpus (คลังถามตอบโดยใช้ข้อมูลจากวิกิพีเดียภาษาไทย)

ค้นหาชื่อเรื่อง (วิกิพีเดีย)

สุ่มเอกสาร

เอกสาร (ฝ่าปฏิบัติการสะท้านโลก 2)

ฝ่าปฏิบัติการสะท้านโลก 2

ฝ่าปฏิบัติการสะท้านโลก 2 () เป็นภาพยนตร์โลดโผน/สายลับลำดับที่ 2 ในชุด กำกับโดยจอห์น วู เขียนบทโดยโรเบิร์ต ทาวน์ นำแสดงโดยทอม ครูซ, ดูเกรย์ สก็อต, วิง เรมส์และแอนโทนี ฮ็อปกินส์ เข้าฉายเมื่อปี ค.ศ. 2000

เรื่องย่อ

เรื่องย่อ. อีธาน ฮันต์ถูกตามตัวหลังหน่วย IMF พบว่ามีคนใช้ตัวตนของเขาทำภารกิจปกป้องคร. เนโครวิดช์ ในการเดินทางมาสหรัฐอเมริกา แต่คร. เนโครวิดช์กลับเสียชีวิตในเหตุเครื่องบินตก คร. เนโครวิดช์เป็นนักชีวเคมีผู้สร้างไวรัสคิเมียราและยารักษาบิลเลโรฟอนให้บริษัทไบโอไซด์ ซึ่งไวรัสและยารักษาหายไป IMF เชื่อว่าคนที่เอาไปคือฌอน แอมโบรส อดีตสายลับ IMF ฮันต์จึงได้รับภารกิจให้ไปตามไวรัสและยารักษาคืนมาโดยเลือกสมาชิกได้เอง 2 คน แต่คนที่ 3 ต้องเป็นไนยาห์ นอร์ดอฟ-ฮอล นักโจรกรรมมืออาชีพที่เข้าถึงตัวแอมโบรสได้เพราะสองคนนี้เคยคบหากันมาก่อน หลังพบกับไนยาห์ ฮันต์เดินทางไปที่ซิดนีย์เพื่อพบกับลูกทีม 2 คนคือลูเทอร์ สติกเคลและบิลลี แบร์ด ฮันต์ สติกเคลและแบร์ดวางแผนลอบเข้าไปในบริษัทไบโอไซด์ ส่วนไนยาห์เข้าหาแอมโบรสเพื่อเก็บข้อมูลเกี่ยวกับไวรัส ที่งานแข่งม้า แอมโบรสนัดพบกับจอห์น ซี. แมคคอย ประธานบริหารบริษัทไบโอไซด์และใช้วิดีโอผู้ติดเชื้อคิเมียราเพื่อรีดทรัพย์ ไนยาห์แอบส่งวิดีโอให้ฮันต์ก่อนจะคืนให้แอมโบรส เมื่อฮันต์ตรวจสอบวิดีโอ เขาพบว่าไวรัสคิเมียราจะมีระยะฟักตัว

1 Annotator Mode

- Verifying question-answer pairs
- Managing question-answer pairs

Administrator (Linguist)

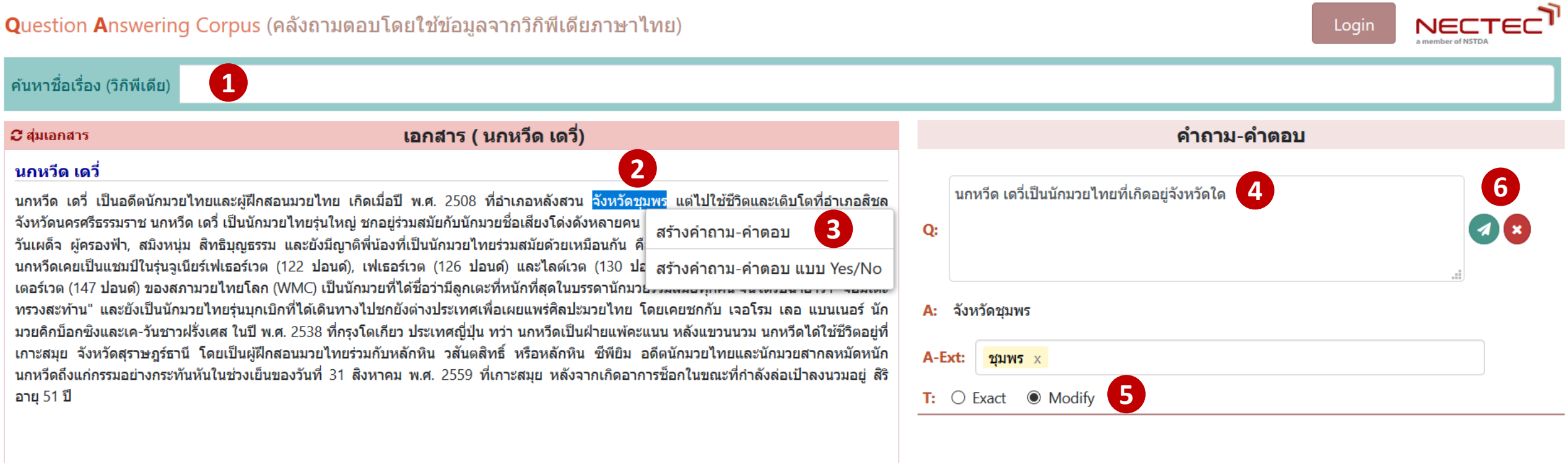
Question Answering Corpus (คลังถามตอบโดยใช้ข้อมูลจากวิกิพีเดียภาษาไทย)

สถิติโดยรวม | สถิติแต่ละผู้ใช้ | ค้นหาคำถาม-คำตอบ | กำกับ word focus

ชื่อผู้ใช้	ชุดข้อมูล	จำนวนเอกสาร	คำถาม (ทั้งหมด)
hawa_qa19	qa2019	1000	2000
may_qa19	qa2019	1050	1313
may_qa20	qa2019	1	2
narm_qa19	qa2019	2877	4217
narm_qa20	qa2019	0	0

2 Administrator Mode

UI of Annotation System



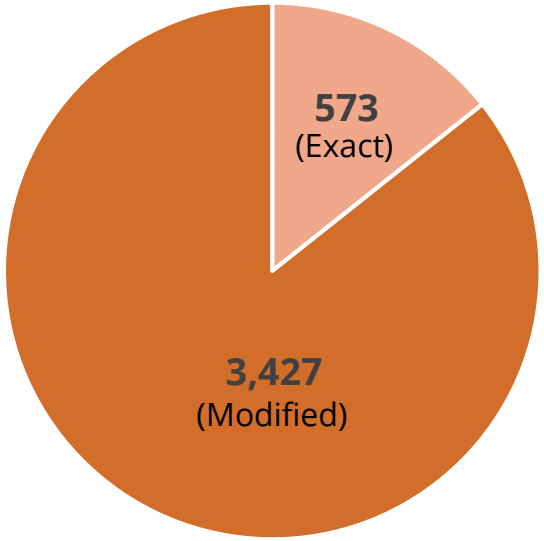
1. Searching an article
2. Highlighting text appearing an answer
3. Right clicking on the answer and selecting menu
4. Typing a question corresponding with the answer
5. Selecting types of the question
6. Clicking the save button

The Constraints of Creating a Question

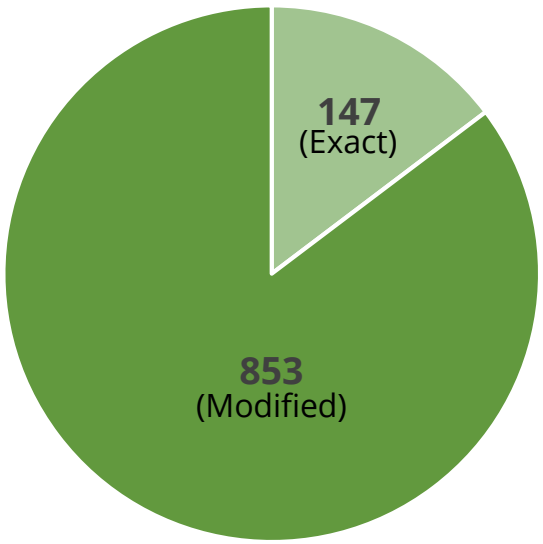
1. A question and an answer are created by a human in natural language on Thai Wikipedia article.
2. An answer is always a word, a segment of text, or a span appearing on a part of the context.
3. A question is a factoid question.
4. A question length must be less than 2 lines.
5. Typos from Thai Wikipedia should be avoided in a question and an answer.

Dataset Statistics

- The dataset contains 5,000 question-answer pairs created from 2,923 Thai Wikipedia articles.

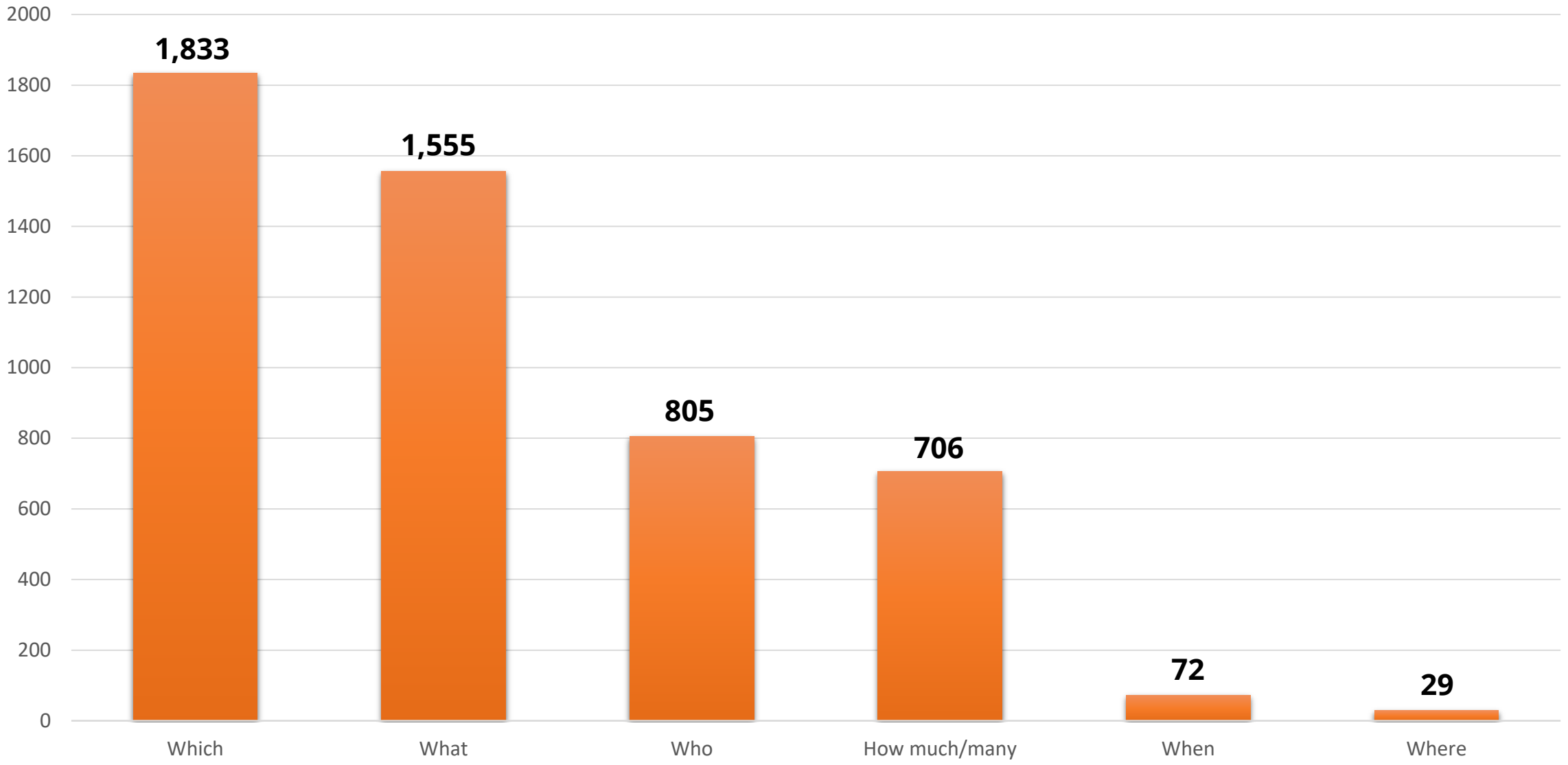


4,000 QA pairs
(Training dataset)



1,000 QA pairs
(Validation dataset)

The Number of QA Pairs by Interrogative Words



An Example of Question and Answer

In Thai

คำถาม

กีฬาประจำชาติแห่งแดนอาทิตย์อุทัยที่มีประวัติยาวนานคือกีฬาอะไร

คำตอบ

ซูโม่

Translation

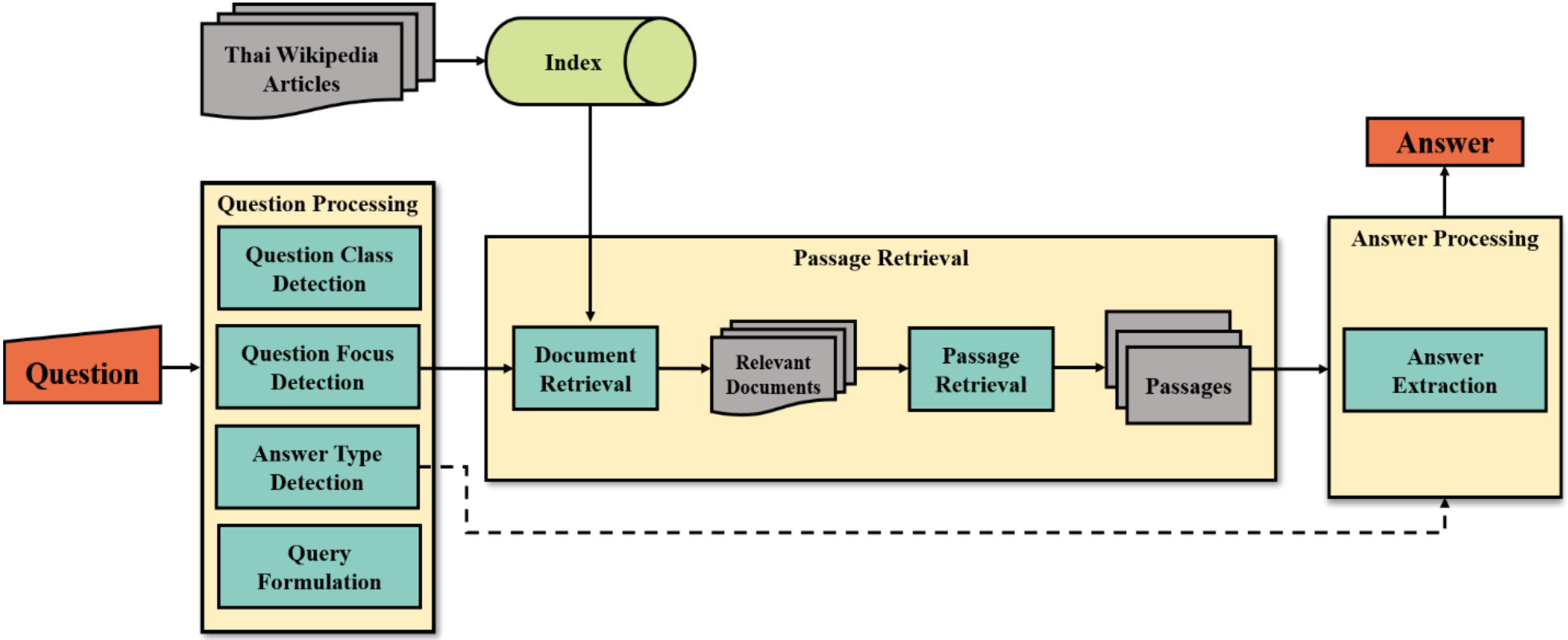
Question

What is the national sport of the land of the sun that has a long history?

Answer

Sumo

A Baseline of Thai Question Answering System

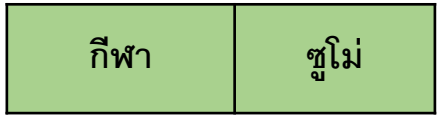


Evaluation

- We use exact match (EM) and F1 metrics, computed on common substring of word level between the predicted answer and the gold answer

Gold standard#1

Gold standard#2



Gold standard#3

Prediction#1

1) **EM** = **1**, Precision = 1/1 = 1, Recall = 1/1 = 1, **F1** = (2*1*1)/(1+1) = **1**

2) **EM** = **0**, Precision = 1/1 = 1, Recall = 1/2 = 0.5, **F1** = (2*1*0.5)/(1+0.5) = **0.66**

3) **EM** = **0**, Precision = 1/1 = 1, Recall = 1/3 = 0.33, **F1** = (2*1*0.33)/(1+0.33) = **0.49**

Prediction#2

1) **EM** = **0**, Precision = 1/2 = 0.5, Recall = 1/1 = 1, **F1** = (2*0.5*1)/(0.5+1) = **0.66**

2) **EM** = **0**, Precision = 1/2 = 0.5, Recall = 1/2 = 0.5, **F1** = (2*0.5*0.5)/(0.5+0.5) = **0.50**

3) **EM** = **0**, Precision = 1/2 = 0.5, Recall = 1/3 = 0.33, **F1** = (2*0.5*0.33)/(0.5+0.33) = **0.39**

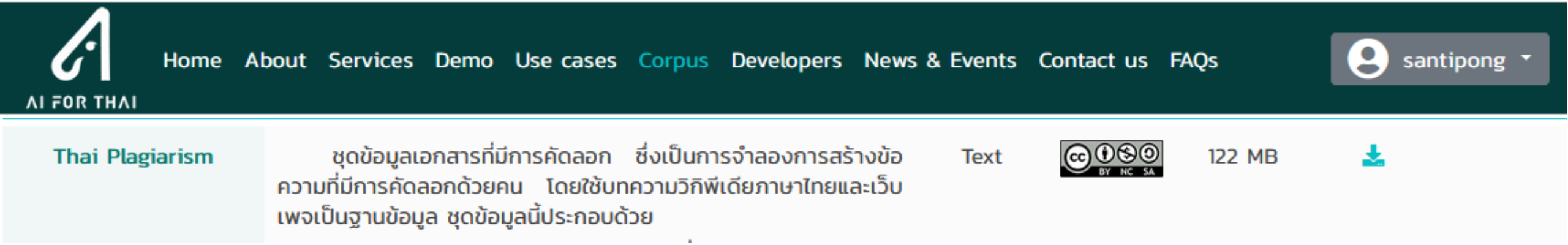
The Experimental Results

- The experimental setting was conducted for two tasks.

Method	Document Retriever (Accuracy)	Document Reader (Exact Match)
Baseline	71.24	25.92

How to Get the Dataset

- The corpus is free for research and development.
- You need to register to get the corpus.



https://aiforthai.in.th

Conclusions

- We designed and constructed a Thai QA dataset.
- The dataset consists of 5,000 question-answer pairs.
- The dataset was evaluated using the baseline system.
- The F1 score achieved 25.92% at 1-best accuracy.
- The dataset is free for research, available at http://aiforthai.in.th

Thank you

Santipong Thaiprayoon
Speech and Text Understanding Research Team (STU)
Artificial Intelligence Research Unit (AINRU)
National Electronics and Computer Technology Center (NECTEC)


Email: santipong.thaiprayoon@nectec.or.th