Thai Question Answering System and Dataset

Speech and Text Understanding Research Team (STU), NECTEC

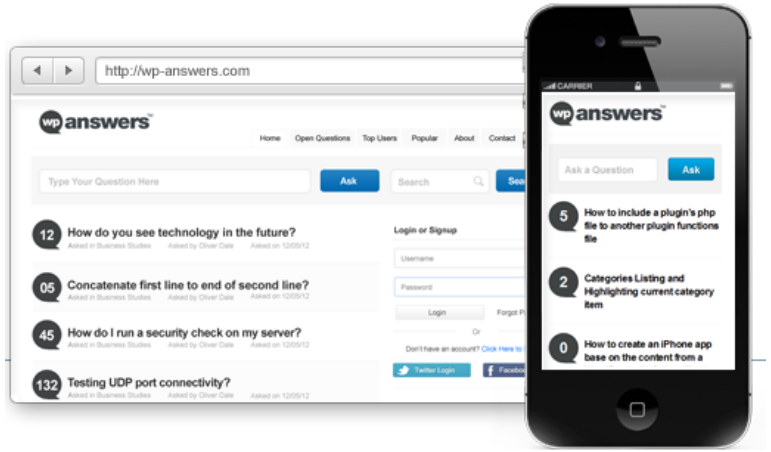
Outline

- What is a question answering system
  - QA vs. Chatbot
  - Types of QA systems
  - Types of questions
- Our proposed system
  - Question answering dataset
  - Question answering system
- Evaluation of QA system
  - Exact match (EM)
  - F1 score (Macro-average precision recall)

What is a question answering system

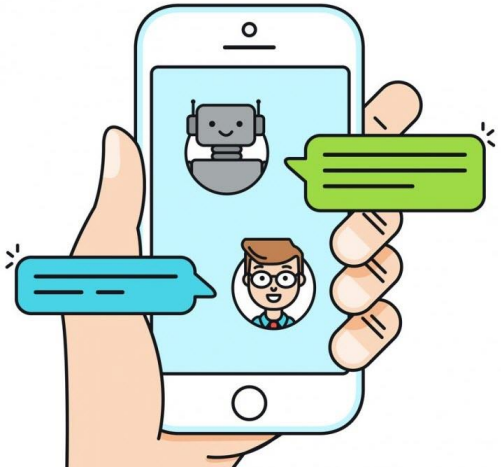
- Question Answering System (QAS) is an information retrieval system that automatically generates an concise answer of a question posed by human in natural language

How is the difference between QA and Chatbot



Question answering

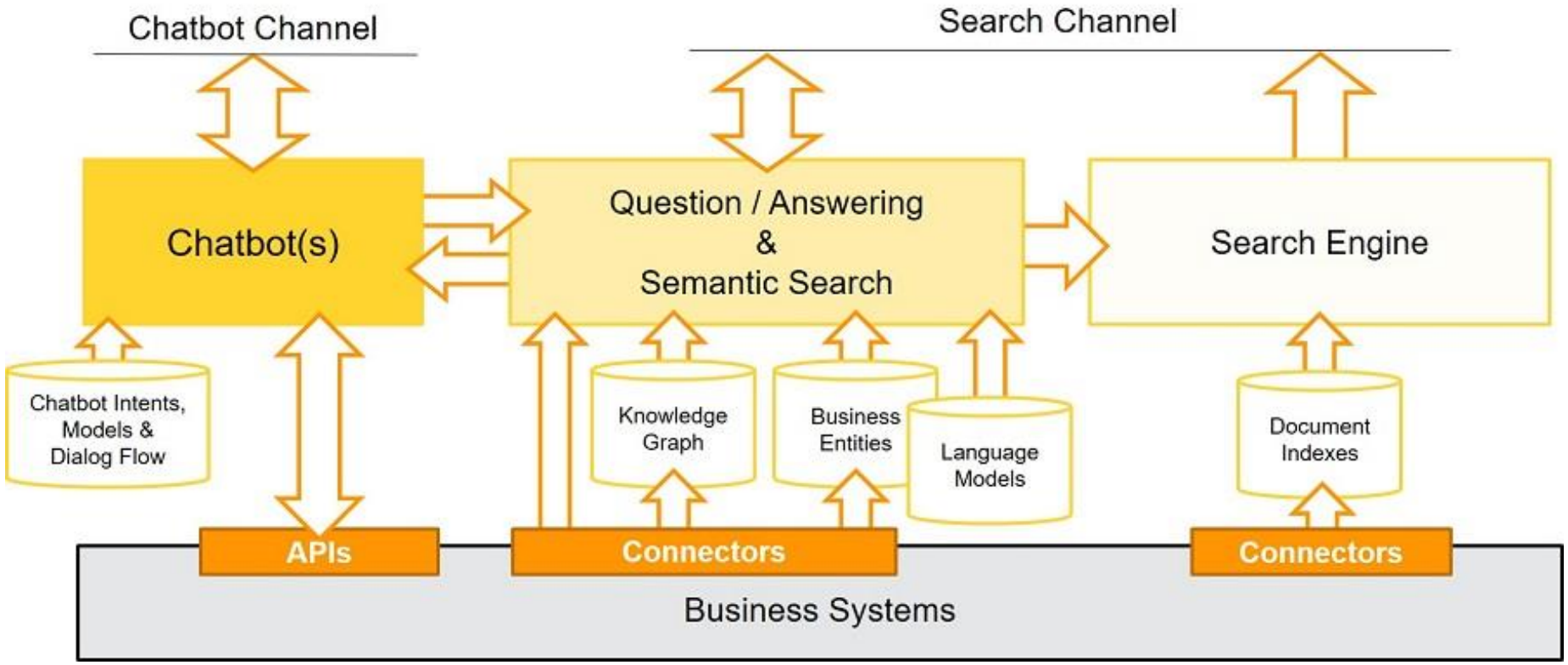
- QA system is a task of answering a question
- QA system handles broad domains of knowledge
- QA system is designed to answer a question straightforwardly



Chatbot

- Chatbot is a task of using conversation ability between users and a bot
- Chatbot handles deep dialogs and specific domains
- Chatbot is designed to promote, sell, or troubleshoot

How is the difference between QA and Chatbot (cont'd)



Both chatbot and QA system can be compatible task depending on users looking for what information are

Types of QA systems

- Open-domain deals with questions about nearly everything from several domains
  - + Covering wide range of queries
  - - Low accuracy
- Closed-domain deals with questions under specific domain
  - + High accuracy
  - - Limited coverage over the possible queries
  - - Needs domain expert

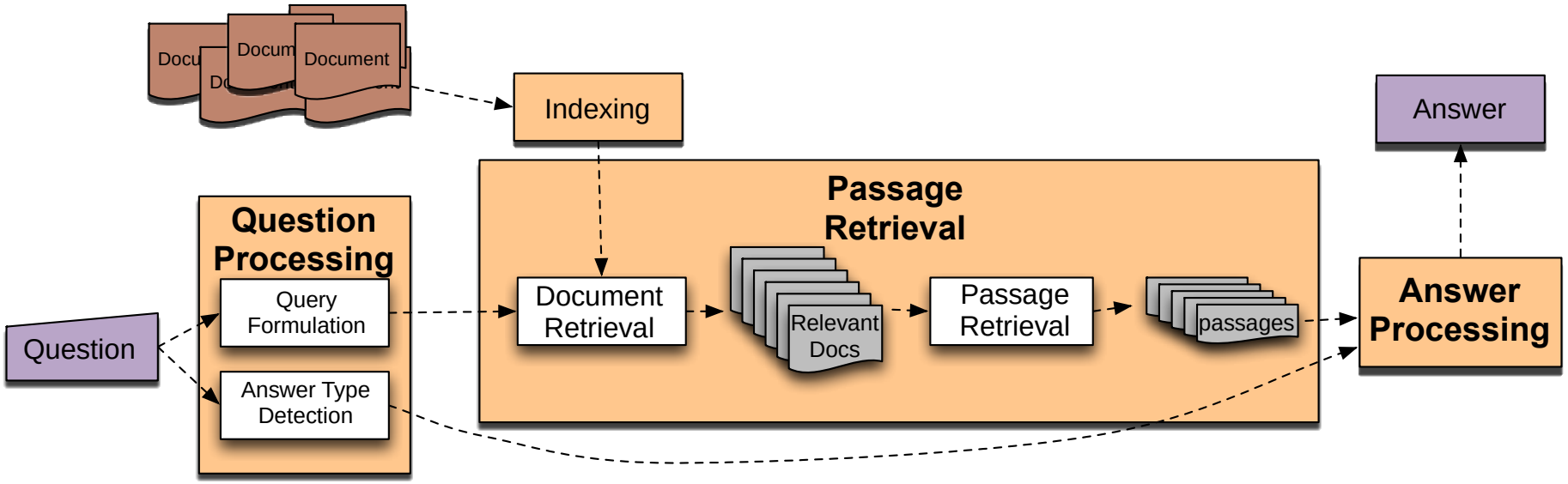
Paradigms of question answering system

- Information retrieval-based approaches
  - IR techniques first find relevant documents and passages
  - Using machine reading comprehension algorithms (MRC) to read these retrieved passages and draw an answer
  - TREC; Google
- Knowledge-based approaches
  - Build a semantic representation of the query
  - Map from this semantics to query structured data or resources
  - Ontologies (Wikipedia infoboxes, dbPedia, WordNet, Yago)
- Hybrid approaches
  - Find many candidate answers in both textual and knowledge sources based to answer questions
  - IBM Watson; Apple Siri; Wolfram Alpha

Types of questions

- Simple (factoid) questions [what,where,who,when,which,how many]: Questions are simple facts as answers, and these fact are retrieve from a single document
- Complex (narrative) questions [why,how]: questions typically have long pieces as answers which may come from single or multiple documents
  - List questions
  - Hypothetical questions
  - Confirmation questions [yes or no]
  - Causal questions [why or how]
  - Opinion questions

IR-based factoid question answering



- Question Processing
  - Detect question type, answer type, focus, relations
  - Formulate queries to send to a search engine
- Passage Retrieval
  - Retrieve ranked documents
  - Break into suitable passages and rerank
- Answer processing
  - Extract candidate answers
  - Rank candidates using evidence from text and external sources

Question processing

- Answer type detection
  - Decide the named entity type (person, place, time, etc.) of the answer
- Query formulation
  - Choose query keywords for the IR system
- Question type classification
  - Is this a definition question, a math question, a list question
- Question focus detection
  - Find string of words in the question that are likely to be replaced by the answer
- Relation extraction
  - Find relations between entities in the question

Question processing (cont'd)

- For example, *Which US state capital has the largest population?*
  - **Query:** US state capital has the largest population
  - **Answer type:** city
  - **Focus word:** state capital

Query formulation

- Query formulation is the task of creating a query which a list of tokens
  - To send to an IR to retrieve documents that may contain answer strings
  - Using query expansion methods to get documents related answer strings
- For example, *Which US state capital has the largest population?*
  - **Query:** ~~Which~~ US state capital has ~~the~~ largest population~~?~~

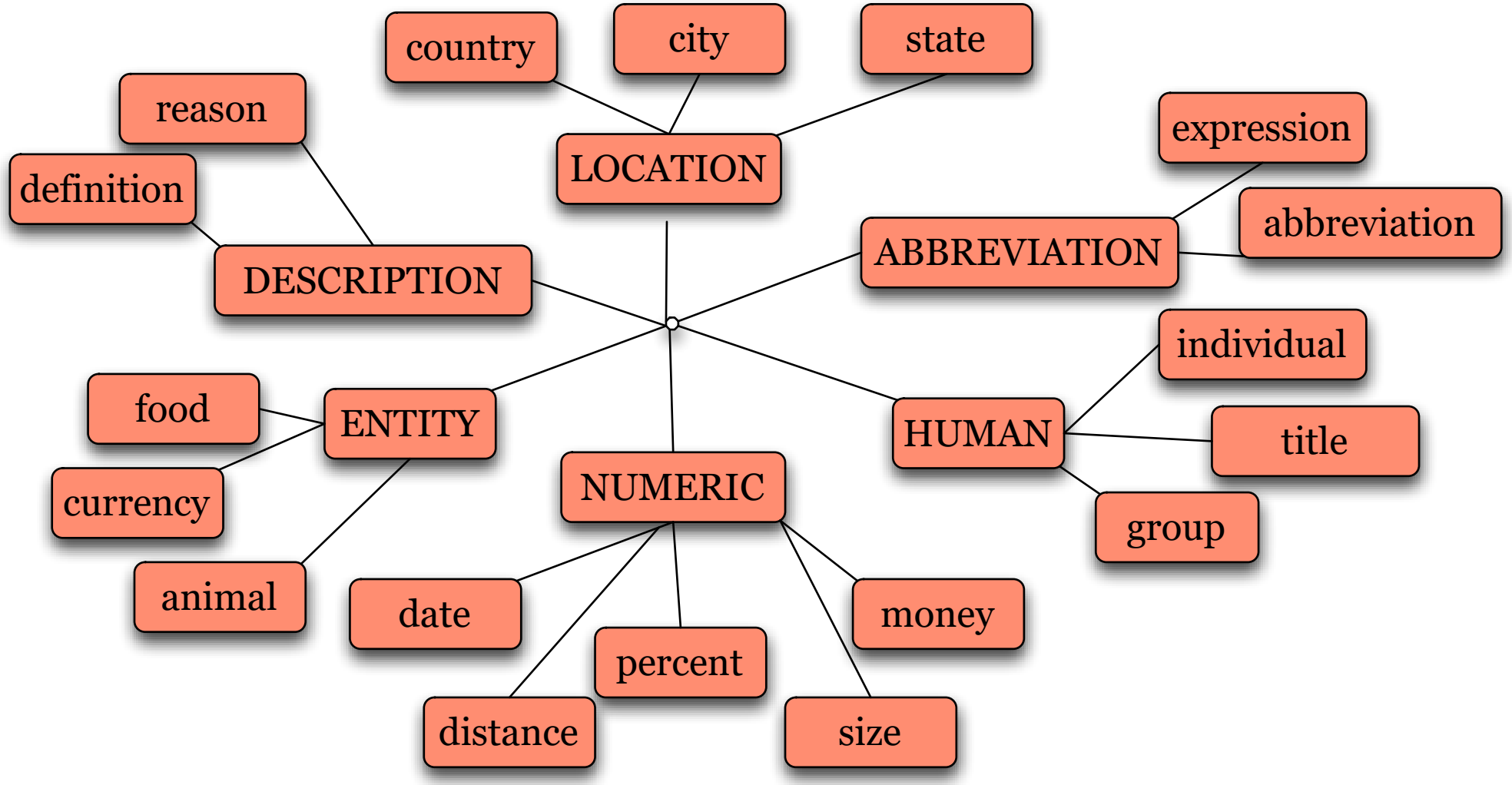
Answer type detection or Question classification

- Question classification is the task of finding the answer type
  - Rule based approaches (pattern matching, handcrafted rules)
    - Who {is|are|was|were} <PERSON>
      - The word Who is person tag
      - The word Where is location tag
      - The word How many is number tag
      - The group words Which, What, How do not give clear answer types. Each words can represent more than one answer types
    - Using headword of the first noun phrase after the wh-word
      - For example, Which city in China has the largest number of foreign financial companies?
  - Learning based approaches (Maximum entropy, SVM, CRF, LSTM classifier)
    - Features extraction: question word, lexicon word, head words, word gram, semantically relates words, part-of-speech tags, and name entities
  - Hybrid approaches

Answer type taxonomy

- A larger hierarchical set of answer types called an answer type taxonomy
- 6 coarse-grained classes
  - ABBEVIATION, ENTITY, DESCRIPTION, HUMAN, LOCATION, NUMERIC
- 50 fine-grained classes
  - LOCATION: city, country, mountain…
  - HUMAN: group, individual, title, description…
  - ENTITY: animal, body, color, currency…

Answer type taxonomy (cont'd)



Passages retrieval

- Step 1: IR engine retrieves documents using query terms
- Step 2: Segment the top *n* documents into smaller passages
  - Sections, paragraphs, or sentences
- Step 3: Passage ranking
  - Use answer type to help rerank passages

Features for passage retrieval

- The number of named entities of the right type in passage
- The number of question keywords in passage
- The longest exact sequence of question keywords that occurs in the passage
- The rank of the document from which the passage was extracted
- The proximity of the keywords from the original query to each other in passage
- The number of n-grams that overlap between the passage and the question

Answer processing

- Named entity tagger
- Featured-based answer extraction
- N-gram tilling answer extraction
- Neural answer extraction
- Reinforcement answer extraction
  - Giving reasonable feedback for outcomes, similar to a carrot-and-stick strategy to reinforce learning outcomes

Answer processing (cont'd)

Absolute Location

Cross-Curricular Focus: History/Social Sciences

Where on Earth are you? Navigators use lines of **latitude** and lines of **longitude** to locate places. Lines of latitude run east and west around Earth. On a map or globe, these lines appear as running sideways or horizontally. Lines of longitude run north and south around Earth. These lines go up and down or vertically on a map or globe. These lines create an imaginary graph paper on the Earth. They make it possible to find an absolute, or exact, location on Earth. They even allow us to give an absolute location to a place out in the middle of the ocean.

Lines of latitude tell us how far north or south of the Equator we are. Sailors have used primitive navigation tools, like astrolabes, since ancient times. The astrolabe uses the sun and stars to find an approximate location. Using such tools, they have been able to approximate their distance from the equator. Although their instruments may not have been the high quality we have now, they were incredibly accurate for their time.

Lines of longitude tell us how far east or west of the prime meridian we are. Sailors constantly looked for new ways to increase their navigation skills. Still, it wasn't until the 18th century they were able to measure degrees of longitude. They would have been very envious of the technology available to us today.

When we use lines of latitude and longitude together, we can get a very precise location. If we want to identify the absolute location of a point, we look where the latitude and longitude lines cross nearest to that point. We use the coordinates for that point as its address. Many maps today include degrees of latitude and longitude.

Another tool that helps us navigate is the **magnetic compass.** The magnetic compass was developed in China. In medieval times, sailors brought it from China to Europe during their regular trade **expeditions** to Asia. This technology made worldwide travel easier and encouraged more exploration.

Name: Key

Answer the following questions based on the reading passage. Don't forget to go back to the passage whenever necessary to find or confirm your answers.

Actual wording of answers may vary.

1) What is the function of lines of latitude and longitude? to allow us to find an absolute location of a point on Earth

2) Which imaginary lines run north and south?
longitude

3) Which imaginary lines are based on the Equator?
latitude

4) Explain what is meant by an absolute location.
It is an address of longitude and latitude of a place on Earth

5) In your opinion, which invention was more important: the astrolabe or the magnetic compass? Why? student's choice

Teach machine to reading comprehension

Answer extraction: named entity tagger

- Run an answer-type named-entity tagger on the candidate passages
  - Each answer type requires a named-entity tagger that detects it
  - If answer type is CITY, tagger has to tag CITY
    - Can be full NER, simple regular expressions, or hybrid
- Return the string with the right type:
  - **Question:** Who is the prime minister of India? (PERSON)
    - **Manmohan Singh**, Prime Minister of India, had told left leaders that the deal would not be renegotiated
  - **Question:** How tall is Mt. Everest? (LENGTH)
    - The official height of Mount Everest is **29035 feet**

Answer extraction: feature-based for ML

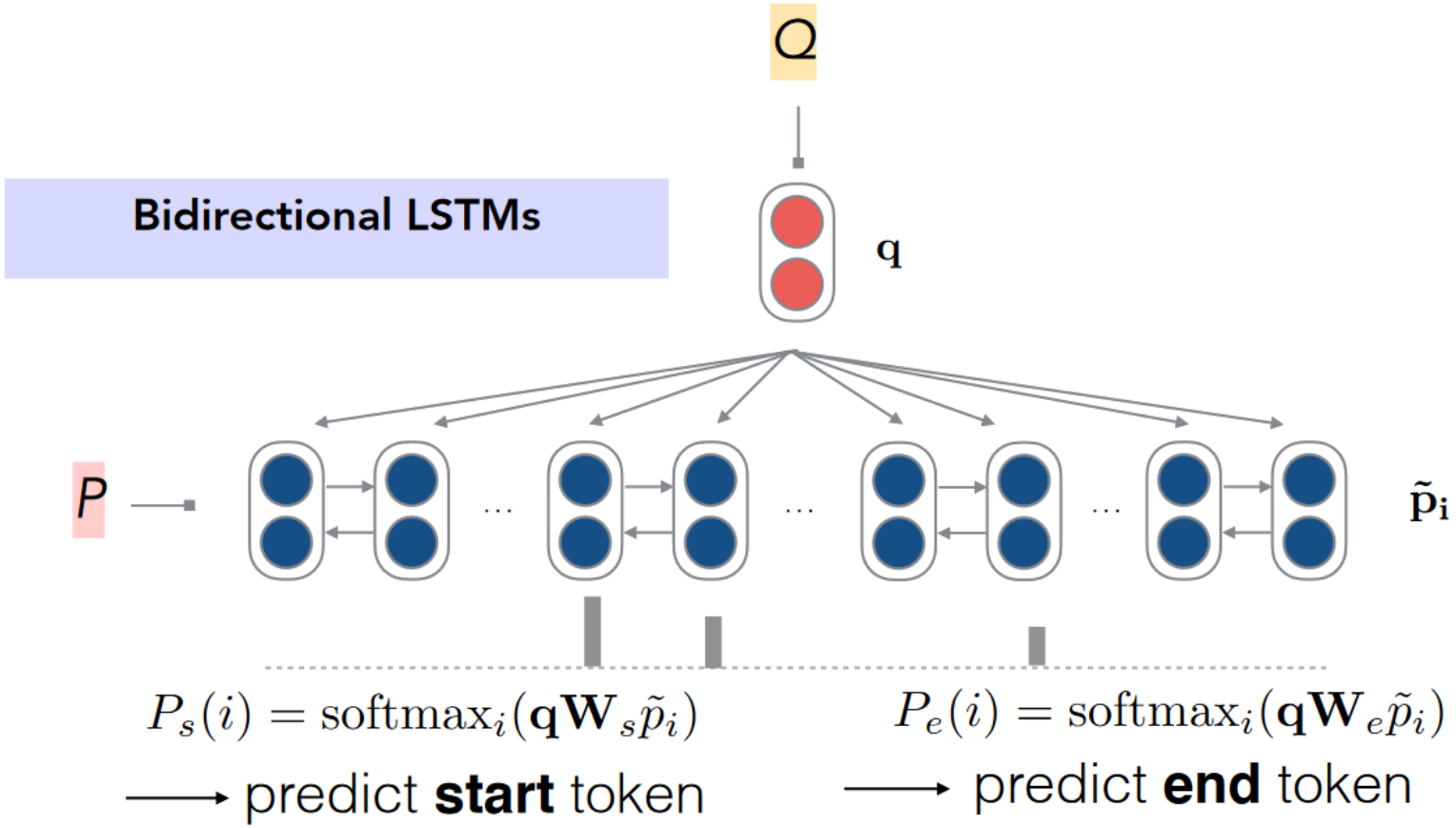
- If there are multiple candidate answers
  - **Question:** Who was Queen Victoria's second son?
  - **Answer type:** Person
  - **Passage:** The Marie biscuit is named after Marie Alexandrovna, the daughter of Czar Alexander II of Russia and wife of Alfred, the second son of Queen Victoria and Prince Albert

Answer extraction: feature-based for ML (cont'd)

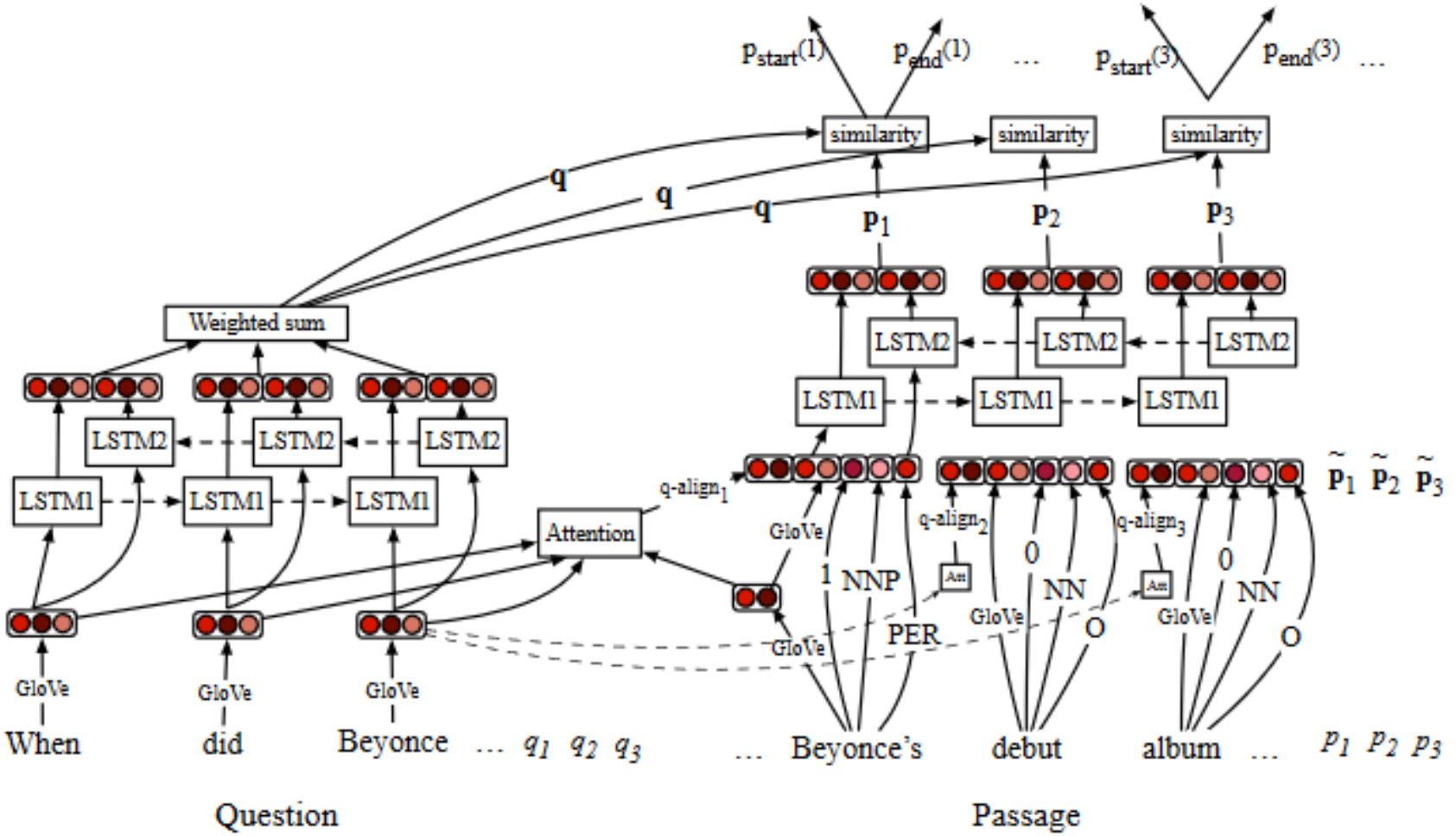
- Answer type match
  - Candidate contains a phrase with the correct answer type.
- Pattern match
  - Regular expression pattern matches the candidate.
- Question keywords
  - Number of question keywords in the candidate.
- Keyword distance
  - Distance in words between the candidate and query keywords
- Novelty factor
  - A word in the candidate is not in the query.
- Apposition features
  - The candidate is an appositive to question terms
- Punctuation location
  - The candidate is immediately followed by a comma, period, quotation marks, semicolon, or exclamation mark
- Sequences of question terms
  - The length of the longest sequence of question terms that occurs in the candidate answer

Answer extraction: neural network

- Task: Given paragraph P and question Q, the goal is to find a span answer in the paragraph which answers the question
- Model: Attention weight computes similarity score between word embeddings of the question and passage

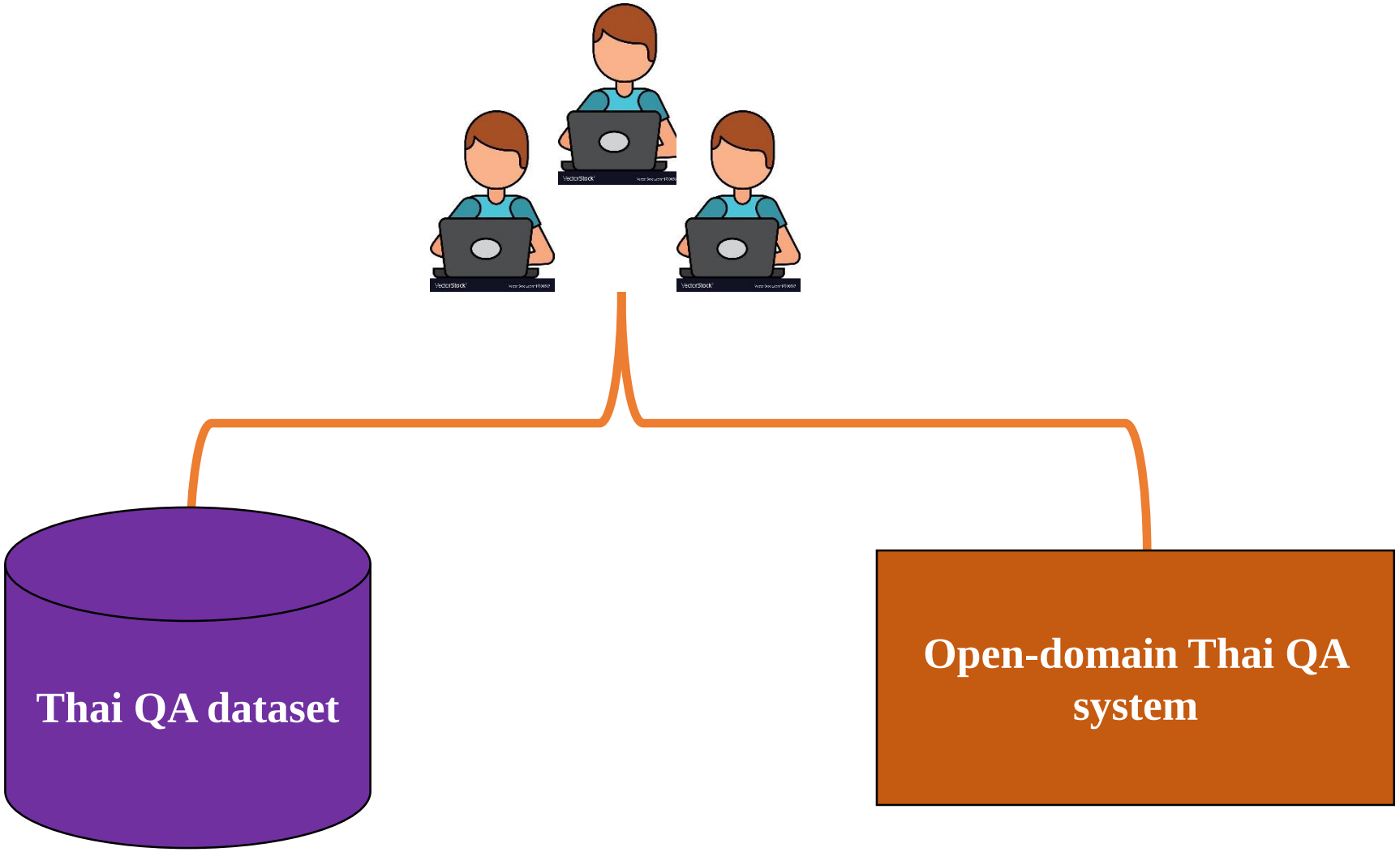


Answer extraction: neural network



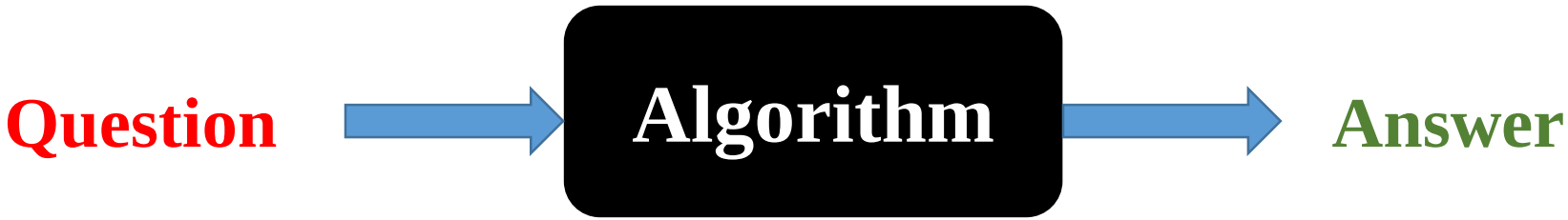
This figure shows the question answering system of Chen et al. (2017), considering part of the question When did Beyonce release Dangerously in Love? and the passage starting Beyonce's debut album, Dangerously in Love (2003) …

Our tasks



Objectives of Thai QA dataset task

- To build a first Thai QA dataset
- To encourage the QA corpus to be available as the standard corpus for research and development of QA algorithms
- Participates in NSC 2019 develop an algorithm to answer a question from Thai Wikipedia
  - Retrieving small snippet of text contained an answer
  - Finding an exact answer



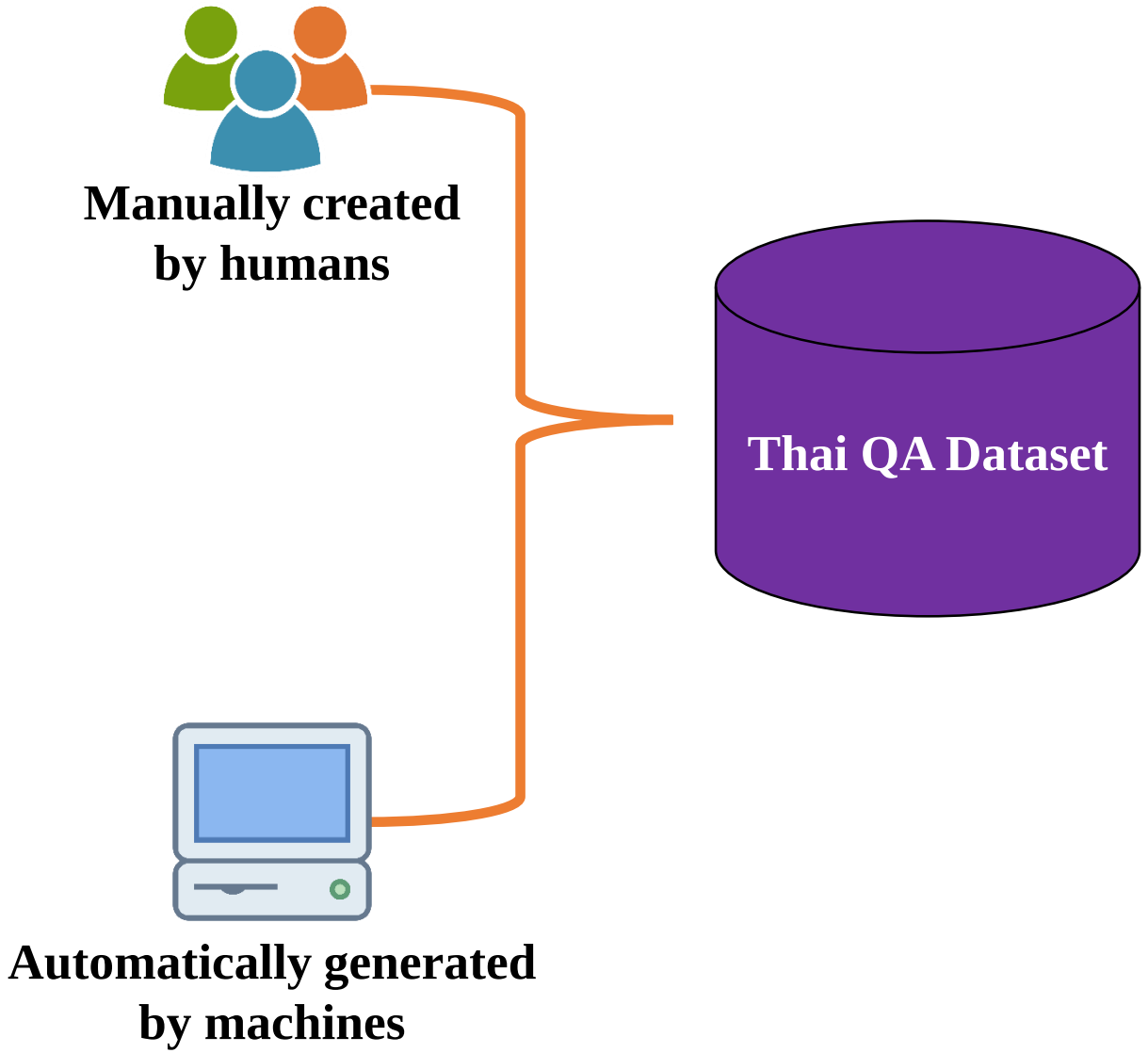
Existing QA datasets

- TriviaQA (2017)
  - 650,000 question-answer-evidence triples
- NewsQA (2017)
  - 100,000 human-generated question-answer pairs
- SQuAD (2016)
  - 100,000+ questions
- MS MACRO (2016)
  - 100,000 queries with their corresponding answers
- WikiQA (2015)
  - 3,047 questions

Thai QA dataset

- Question-answer pairs created by crowdworkers on a set of Wikipedia articles
  - A answer is a word, segment of text, or span appearing on a part of the corresponding reading passage
- Dataset is approximately 10,000 question-answer pairs
  - Simple dataset is 100 question-answer pairs
  - Development dataset is 4,000 question-answer pairs
  - Validation dataset is 1,000 question-answer pairs
  - Evaluation dataset is 5,000 question-answer pairs

Thai QA construction



Process of building Thai QA dataset

- Step 1: A user creates both question and answer
  - Questions and answers are in Thai Wikipedia article
  - Questions are written by human in natural language
  - An answer of a question is always a part of the context
- Step 2: Linguists verify the question-answer pairs
  - Checking the valid questions
  - Checking both the questions and answers corresponding each other
  - Checking the answers being a part of the context

Text example of Thai QA dataset

Context: หลังจากการปฏิรูปเมจิ กีฬาตะวันตกก็เริ่มเข้ามาในญี่ปุ่นและแพร่หลายไปทั่วประเทศด้วยระบบการศึกษาในญี่ปุ่น กีฬานับเป็นกิจกรรมยามว่างที่ดีต่อสุขภาพ ช่วยพัฒนาวินัย การเคารพกฎกติกา และช่วยสังสมน้ำใจนักกีฬา ชาวญี่ปุ่นทุกวัยให้ความสนใจกับกีฬาทั้งในฐานะผู้ชมและผู้เล่น กีฬาที่ได้รับความนิยมในญี่ปุ่น ได้แก่ ซูโม่เป็นกีฬาประจำชาติของญี่ปุ่นที่มีประวัติอันยาวนาน และเป็นกีฬาที่ได้รับความนิยมอย่างมากในญี่ปุ่น ศิลปะป้องกันตัวของญี่ปุ่น เช่น ยูโด คาราเต้ และเคนโด้ ก็เป็นกีฬาที่มีผู้เล่นและผู้ชมมากเช่นเดียวกัน

Question: กีฬาประจำชาติแห่งแดนอาทิตย์อุทัยที่มีประวัติยาวนานคือกีฬาอะไร

Answer: ซูโม่

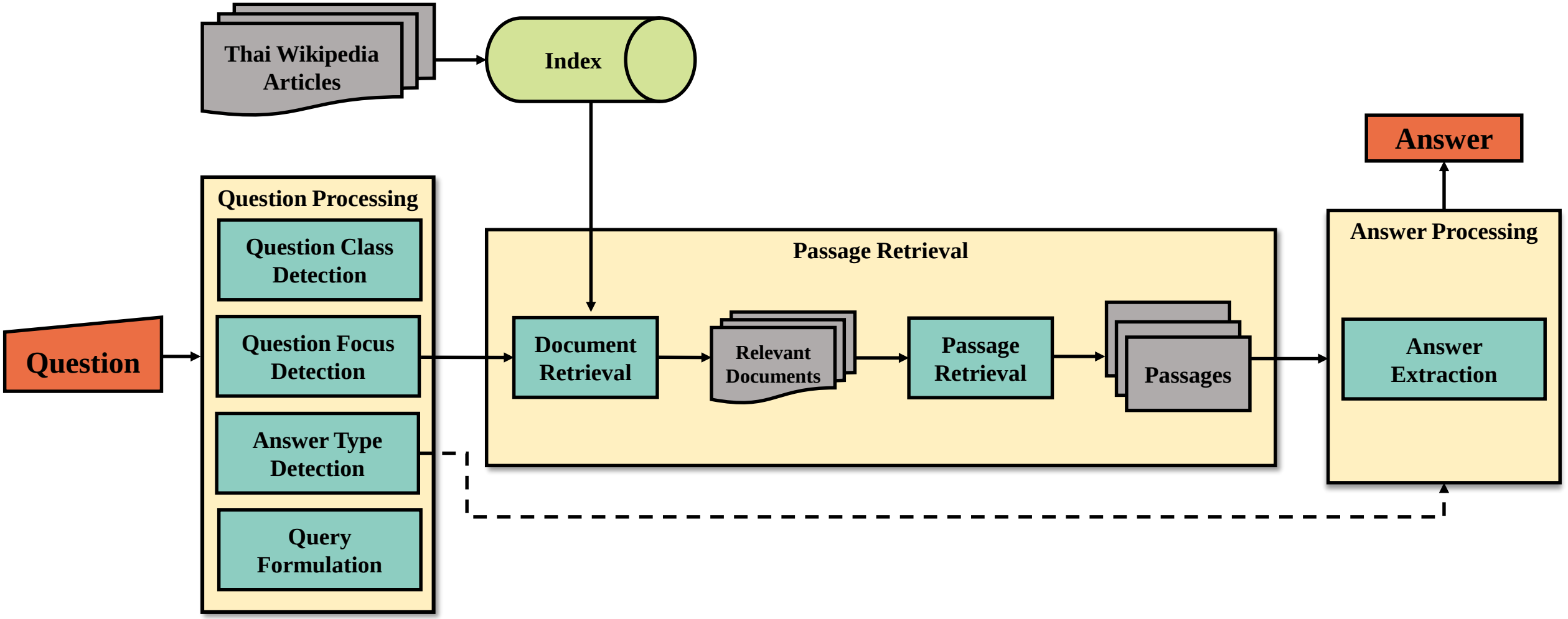
Snippet having right answer

Exact answer

Thai QA GUI


Question Answering Corpus (คลังถามตอบโดยใช้ข้อมูลจากวิกิพีเดียภาษาไทย)
NECTEC
a member of NSTDA
MainPage
DocSuccess: 1
QA-Total: 1
QA-Exact: 1
QA-Modify: 0
Cost: 10
tagger_nung
ค้นหาชื่อเรื่อง (วิกิพีเดีย)
สุ่มเอกสาร
เอกสาร (ปลาช่อนบานคาน)
ปลาช่อนบานคาน
ปลาช่อนบานคาน หรือ ปลาช่อนบังกา (;) ปลาน้ำจืดชนิดหนึ่ง ในวงศ์ปลาช่อน (Channidae) ปลาช่อนบานคาน เป็นปลาช่อนชนิดหนึ่ง มีลักษณะเหมือนปลากระสง (C. lucius) แต่ส่วนหัวไม่เรียวแหลมเหมือนปลากระสง และรูปทรงลำตัวค่อนข้างจะกลมเป็นทรงกระบอกมากกว่า มีขนาดโตเต็มที่ประมาณ 23.5 เซนติเมตร พบกระจายพันธุ์ในตอนใต้ของมาเลเซีย, เกาะสุมาตรา, เกาะกาลิมันตัน และเกาะบังกาในอินโดนีเซีย ในแหล่งน้ำในป่าพรุที่มีค่าพีเอช (pH) ไม่เกิน 4 เป็นปลาสวยงามที่หาได้ยาก ในประเทศไทยเคยมีเข้ามาจำหน่ายเพียงครั้งเดียวเท่านั้น โดยการเลี้ยงในตู้ปลาสามารถปรับค่าพีเอชของน้ำให้อยู่ที่ราว 5.5-6 ได้ สามารถเพาะขยายพันธุ์ได้แล้วในตู้กระจก โดยวางไข่แบบไข่ลอย แม่ปลาคอยดูแลลูกปลา กินอาหารจำพวกแมลง และปลาขนาดเล็ก มีอุปนิสัยดุร้าย
คำถาม-คำตอบ
Q: ปลาช่อนบานคานมีขนาดโตเต็มที่ประมาณเท่าไร
A: 23.5 เซนติเมตร
A-Ext:
T: Exact Modify

Our proposed system



Question classification

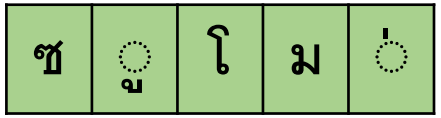
- Head words (HW)
  - Head words are considered as the keyword or the central word in a question
- Focus word (FW)
  - Focus word is string of words in a question that are likely to be replaced by the answer

Question patterns		Expected answer types
ใคร + {เป็น, คือ} + <Noun>	ใครเป็น$นายกรัฐมนตรี_{FW}$คนแรกของไทย	Human:Person
<Noun> + อะไร + ...	$กีฬา_{FW}$อะไรที่ได้รับความนิยมมากที่สุดในญี่ปุ่น	Entity:Sport
<Noun> + ... +อะไร	$กีฬา_{FW}$ที่ได้รับความนิยมมากที่สุดในญี่ปุ่นคืออะไร	Entity:Sport

Evaluation of QA system

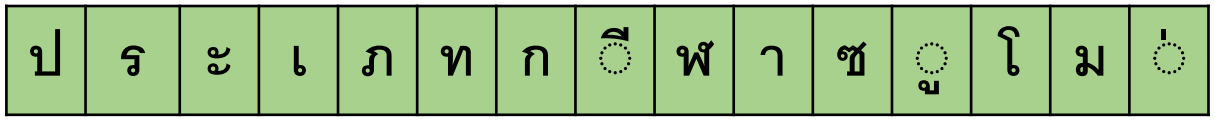
- We use exact atch (EM) and Partial Match metrics, computed on longest common substring of character level between the predicted answer and the gold answer

Gold standard#1

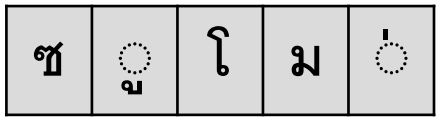


Gold standard#2

Gold standard#3

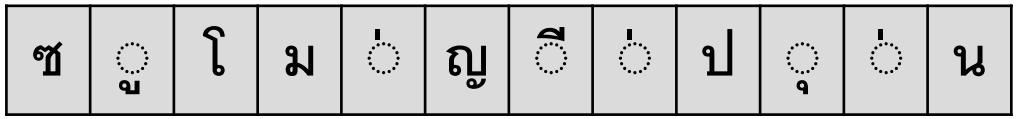


Prediction#1



EM = 1, PM = 5/5

Prediction#2



EM = 0, PM = 5/12

Future works

- We plan to increase the number of question answer pairs in various types of questions
- We plan to construct a Thai conversational chatbot dataset
- We plan to develop the Thai question answering system using RNN and Reinforcement Learning (RL) in part of machine reading comprehension