amazon alexa prize

Maris Basha

Pierpaolo Sepe

What is the Alexa Prize?

The Alexa Prize is a global competition for **university students** dedicated to accelerating the field of **conversational AI.**

Teams of university students will develop a **socialbot**, an Alexa skill that converses with users on popular societal topics.

The grand challenge for the Alexa Prize is to create a socialbot that can engage in a fun, high quality conversation on popular societal topics for **20-minutes** and achieve an average rating of at least **4.0/5.0**.

Prizes

1. $500,000
2. $100,000
3. $ 50,000

The Grand Prize

A prize of a **$1 million** research grant will be awarded to the winning team's university if their socialbot achieves the grand challenge of conversing coherently and engagingly with humans for 20 minutes with a 4.0 or higher rating.

No one has actually won the grand prize.

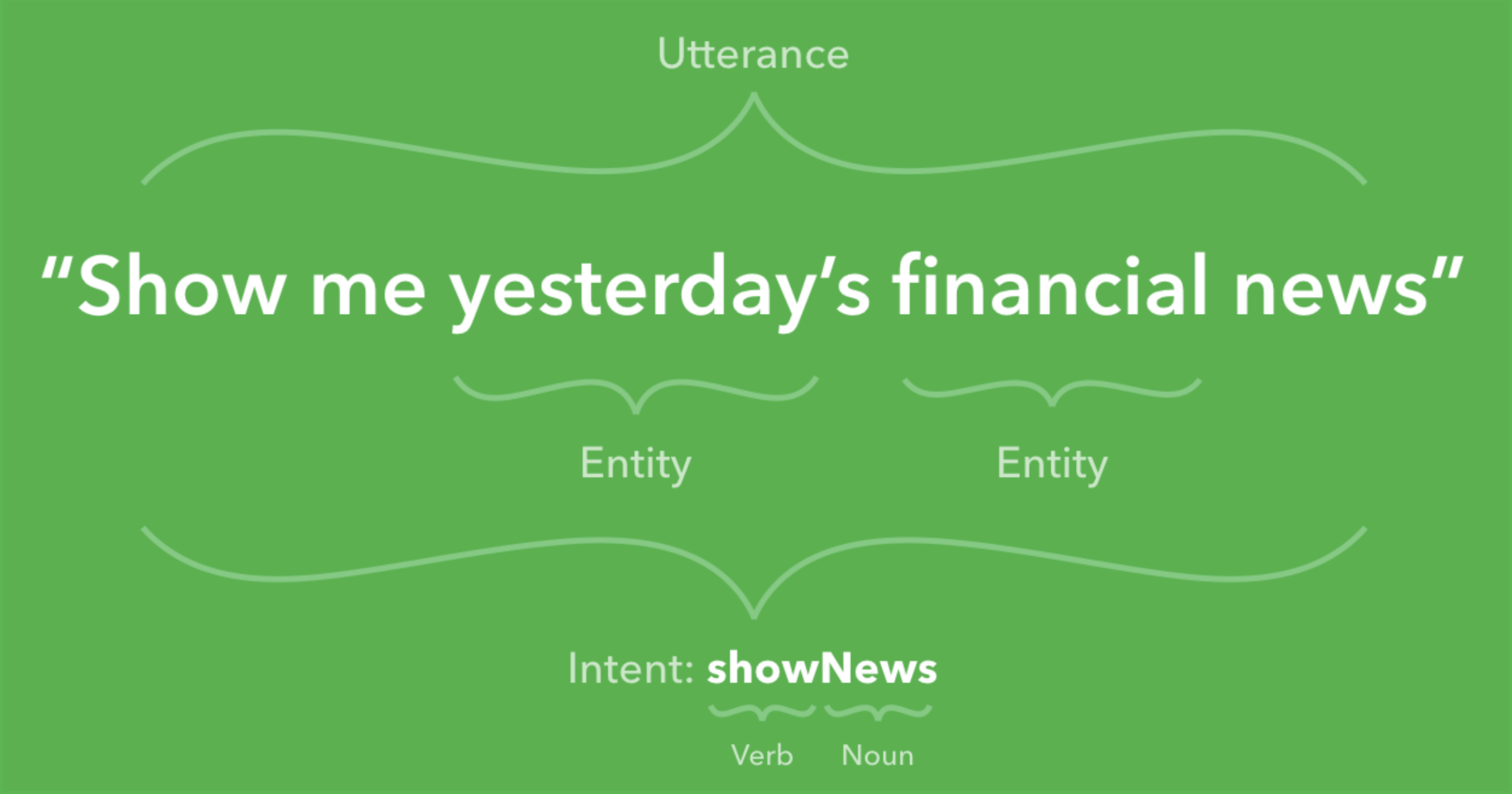
The 2018 Alexa Prize Winner

Average score : **3.1**
Average duration: **9 minutes and 59 seconds.**

Team **Gunrock** will share the $500,000 prize among the student team members.

Gunrock

Recap



Utterance

"Show me yesterday's financial news"

Entity

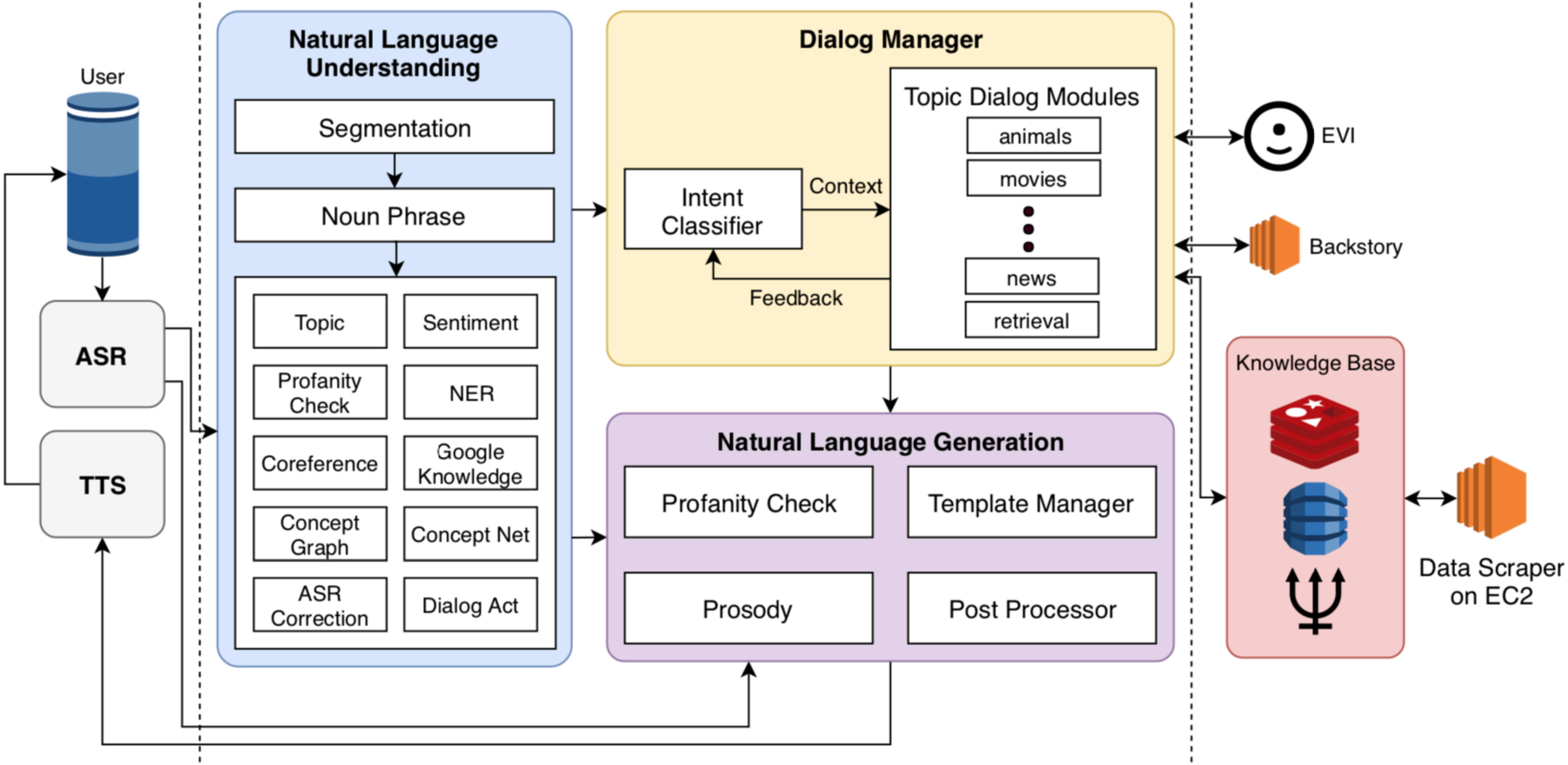
Entity

Intent: **showNews**

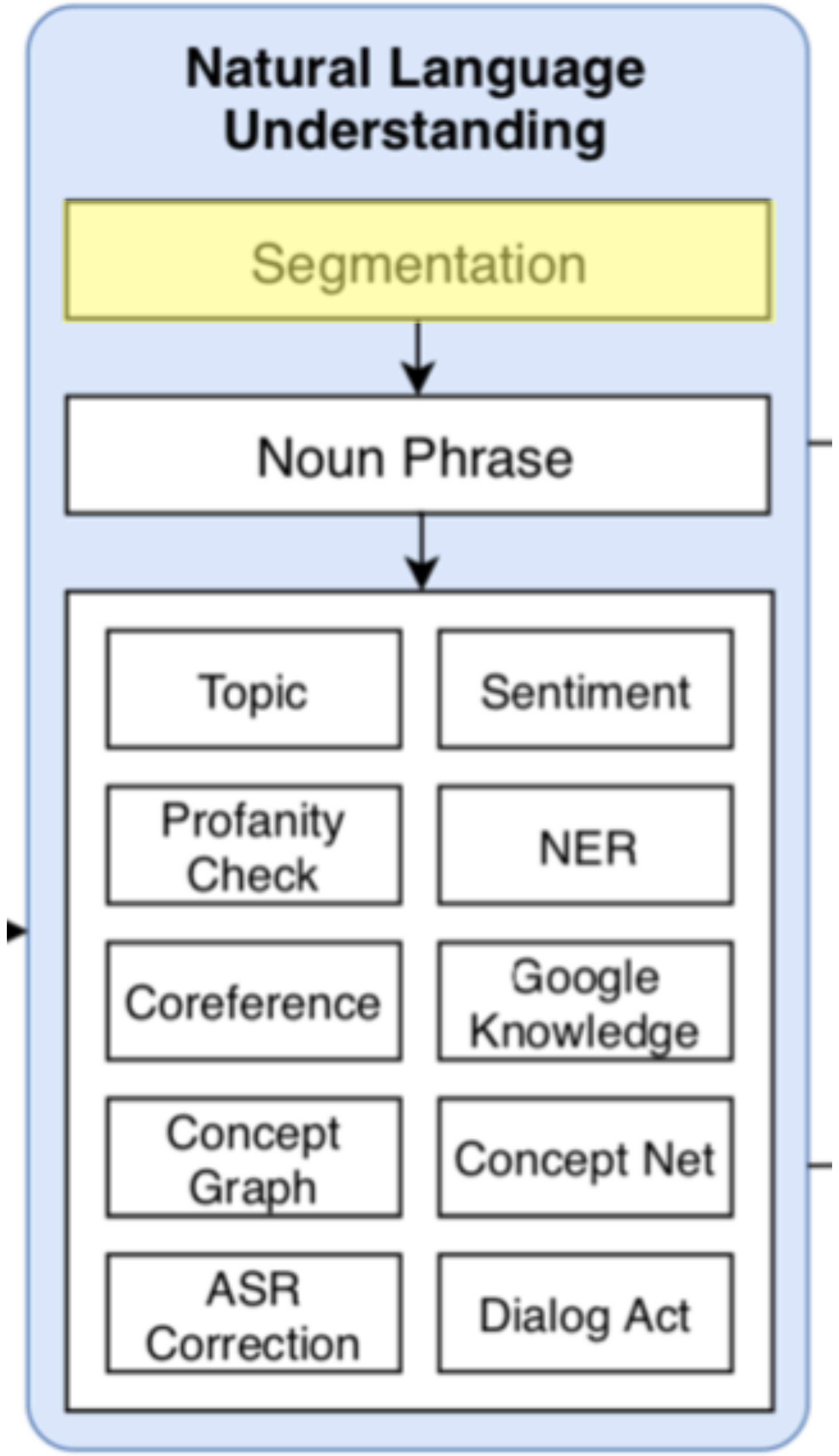
Verb

Noun

Gunrock: Architecture



Sentence Segmentation



Sequence to Sequence model
using the Cornell Movie-Quotes Corpus.

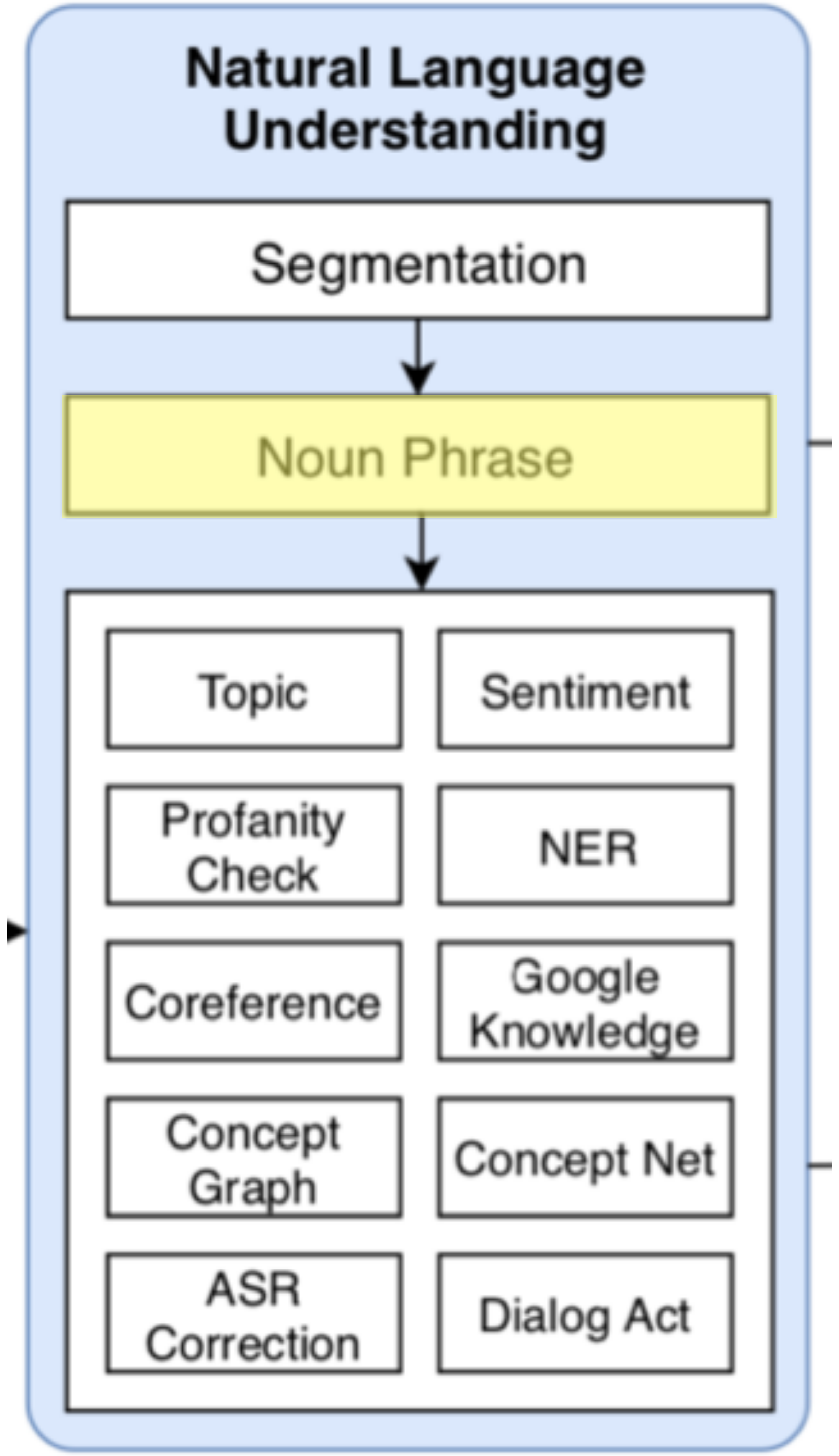
It uses a 2-layer bidirectional LSTM as the encoder and a 2-layer RNN decoder with input feed and global attention.

“Alexa that is cool what do you think of the Avengers”

is segmented into

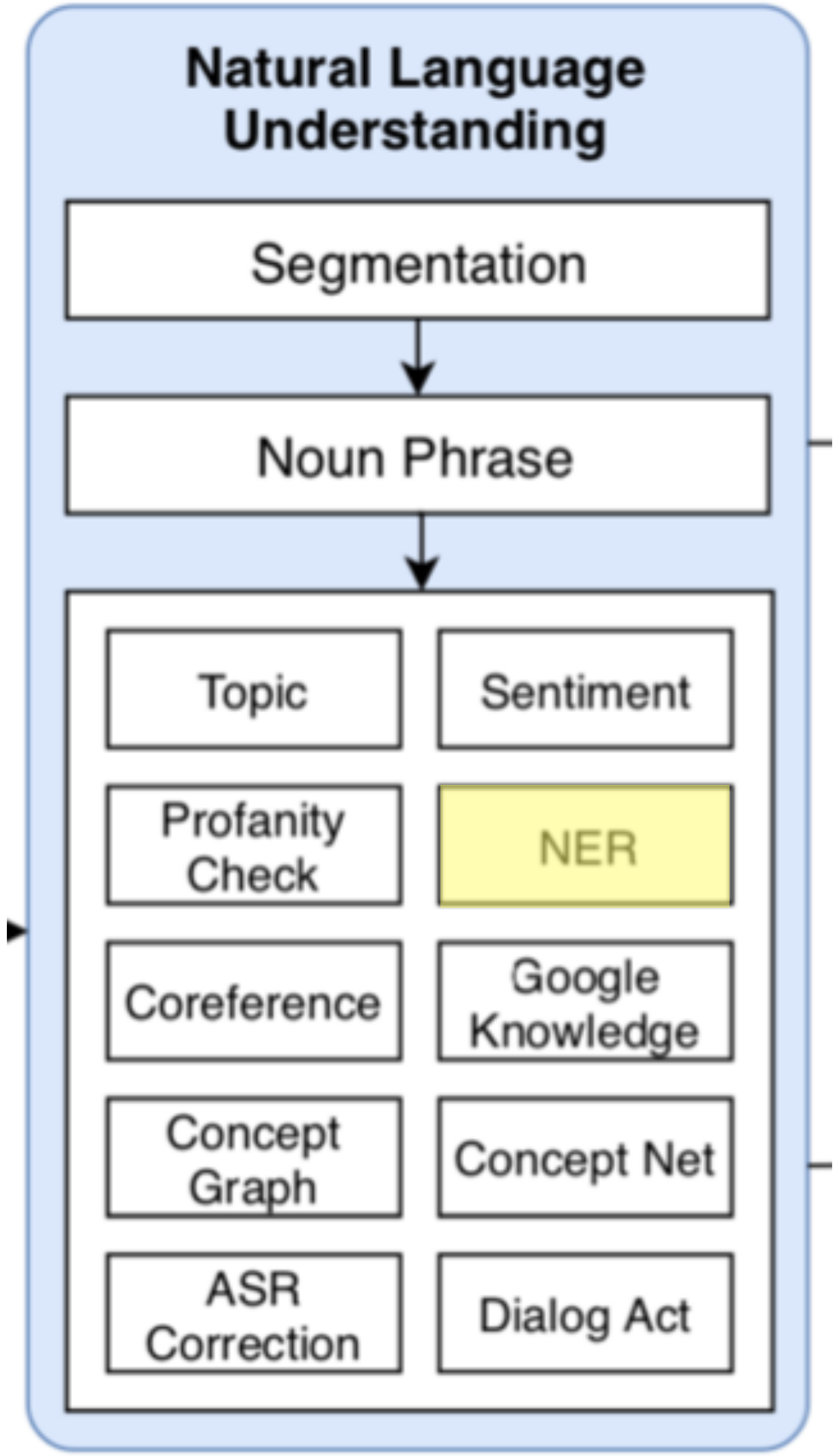
“Alexa <BRK> that is cool <BRK> what do you think of the Avengers <BRK>”

Noun Phrase Extraction



They used the **Stanford CoreNLP** constituency parser to extract noun phrases and local noun phrases from the input sentence.

Entity Recognition

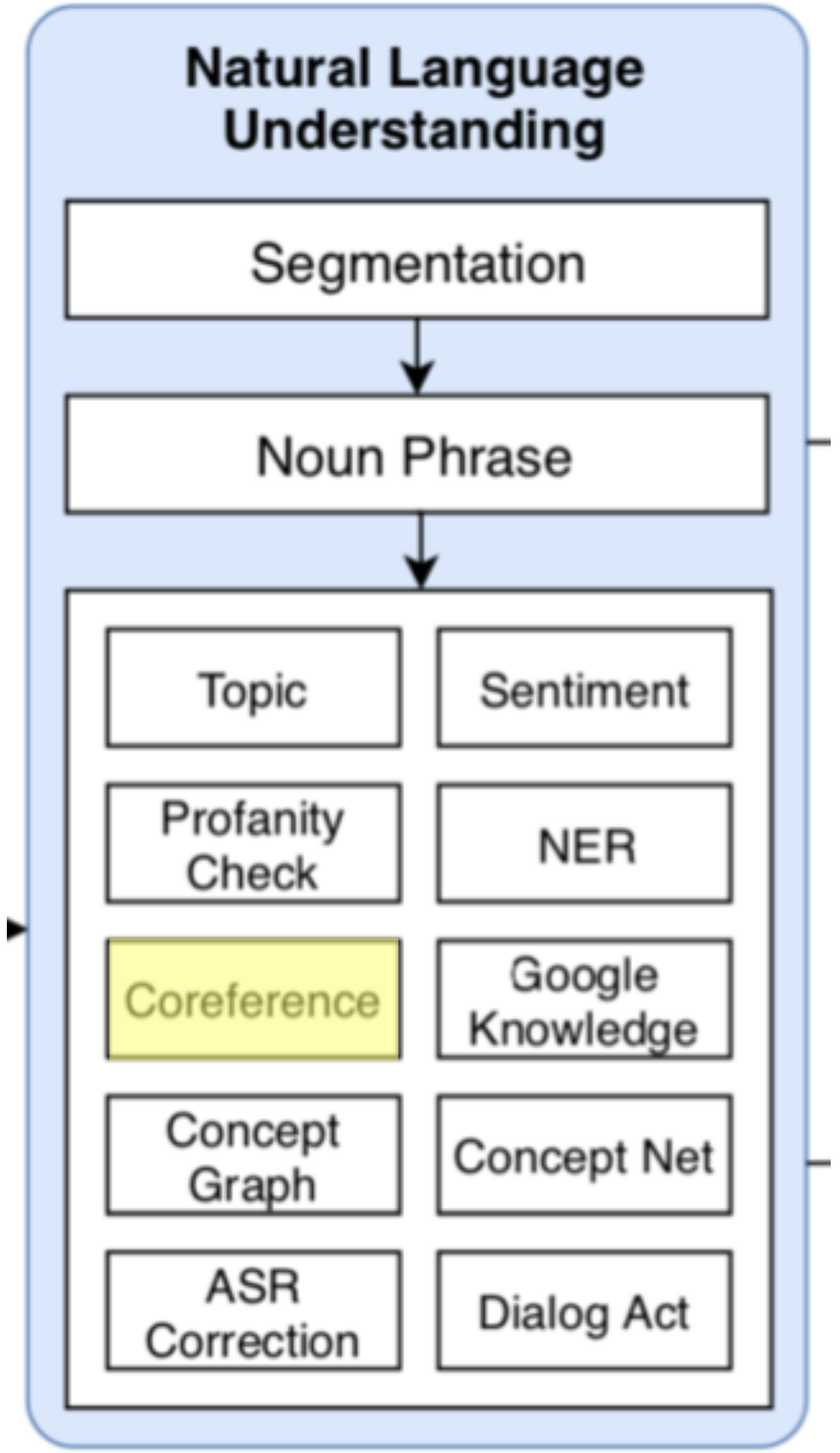


NER tools such as **Stanford CoreNLP** and **spaCy** heavily rely on the letter case of the words in the sentence (e.g. capital letters), that are not available.

They have three recognizers running in parallel with the extracted noun phrases:

- **Google Knowledge Graph**
- **Microsoft Concept Graph**
- **ASR Correction**

Coreference Resolution



Natural Language Understanding

Segmentation

Noun Phrase

Topic

Sentiment

Profanity Check

NER

Coreference

Google Knowledge

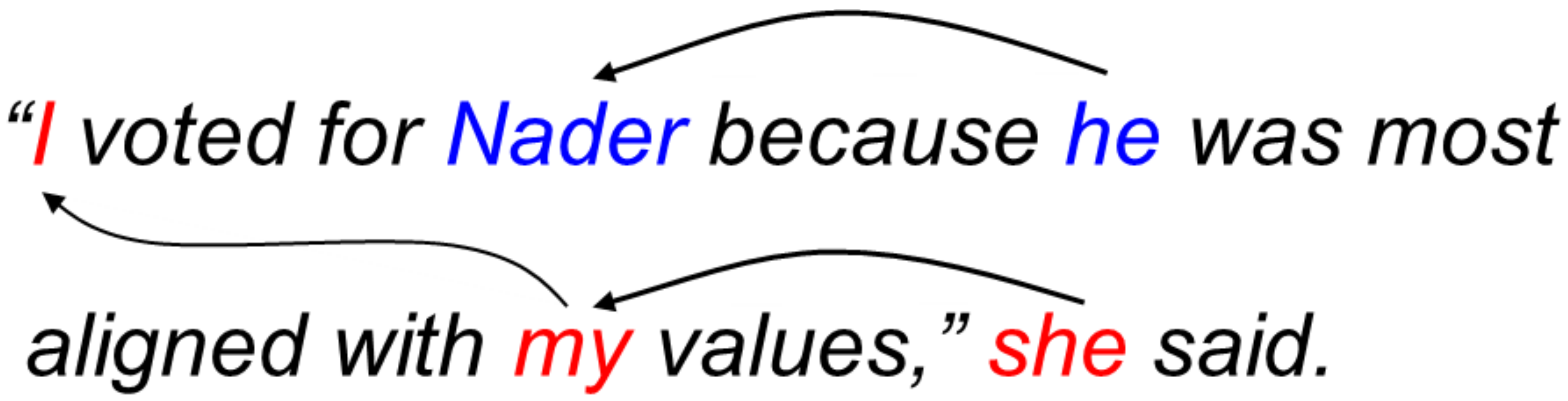
Concept Graph

Concept Net

ASR Correction

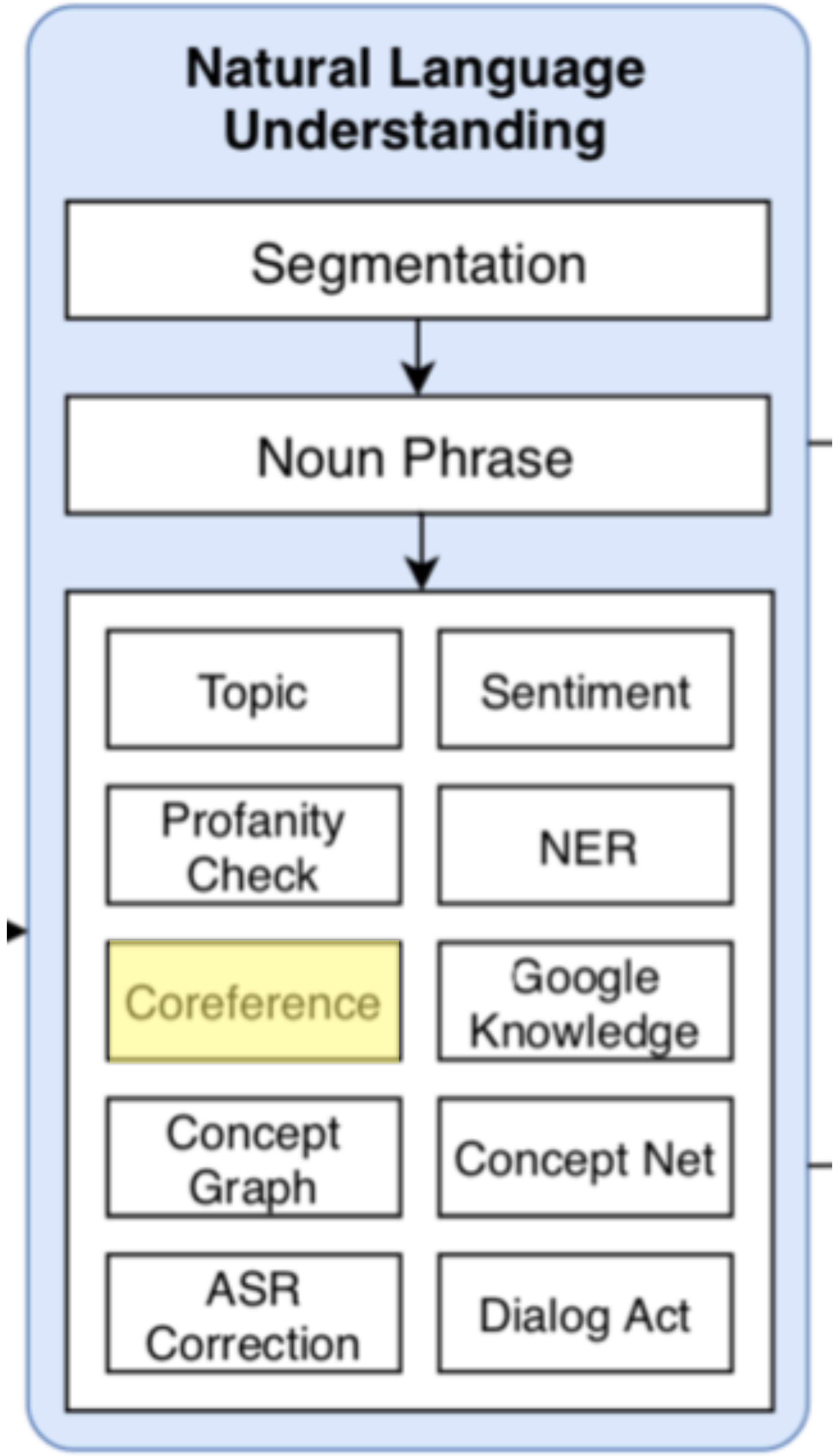
Dialog Act

Coreference resolution is the task of finding **all expressions** that refer to the **same entity** in a text.



“I voted for Nader because he was most aligned with my values,” she said.

Coreference Resolution

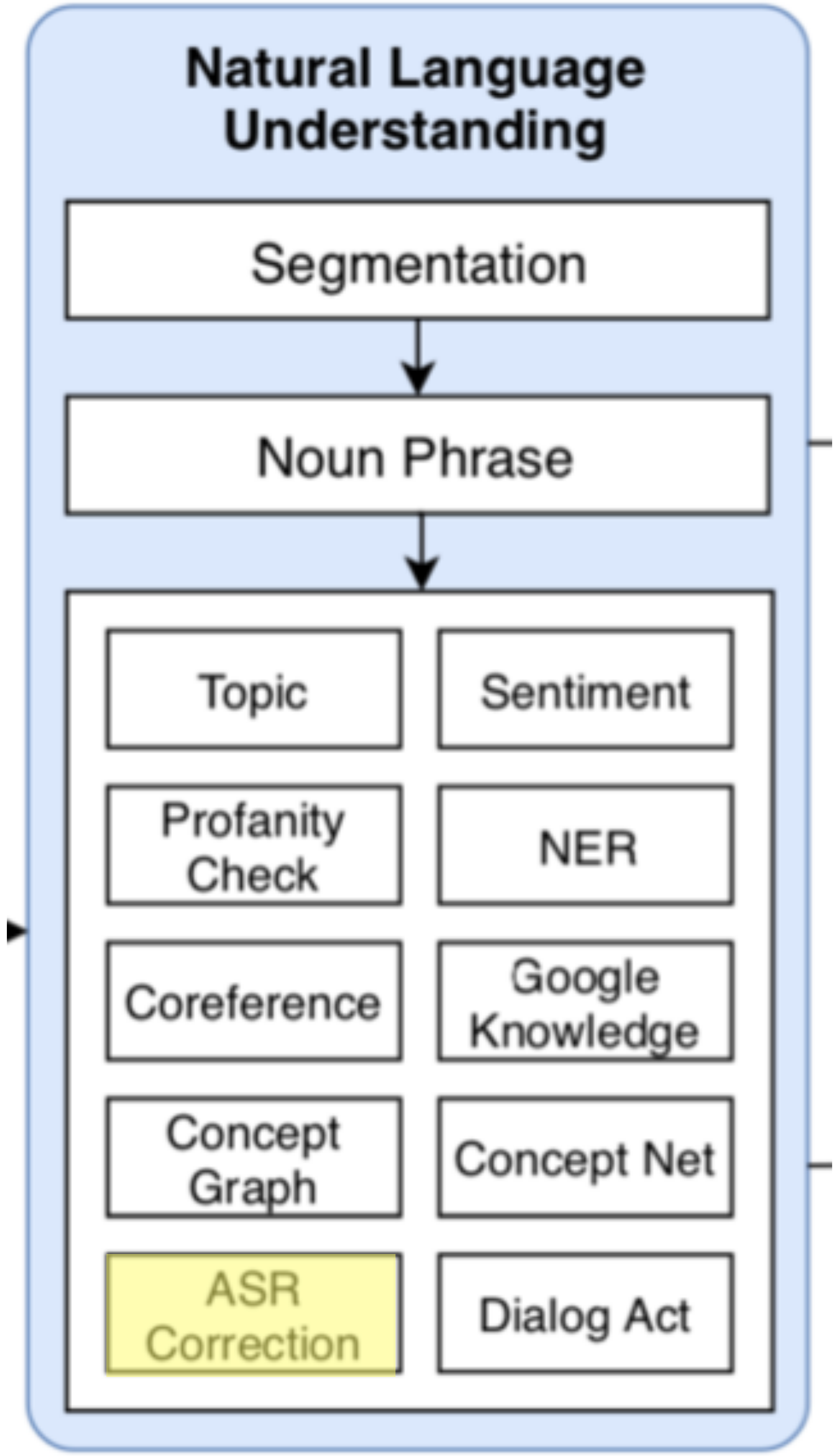


The state-of-the-art models by Stanford CoreNLP and NeuralCoref are trained on **non-conversational** data and do not work well for **dialog conversations.**

They labeled words such as “more” and “one” to undergo coreference.
They replaced such words by considering both the user utterance and the system response.

Depending on what the user refers to (ex. person or event, male or female), they provide the corresponding coreference solution.

ASR Correction



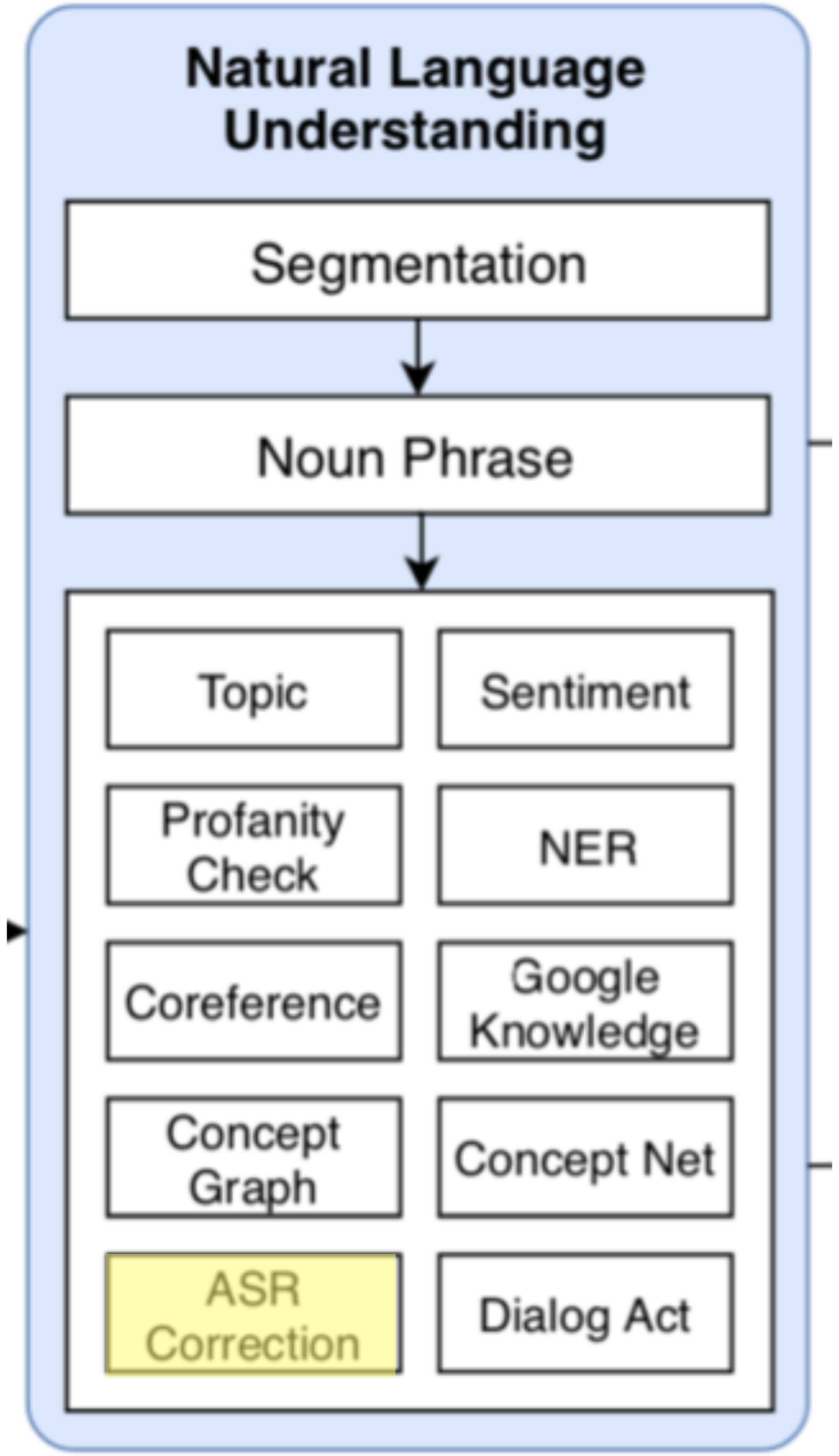
They define three ASR error responses based on the confidence score range:

- **Critical Range**: Overall confidence score and each word confidence score are below 0.1.
The system directly interrupts the overall pipeline and asks users to repeat or rephrase their utterance or request.

- **Warning Range**: Overall confidence score is lower than 0.4 but is not in a Critical Range, it is allowed to pass through to ASR correction.

- **Safe Range**: For other cases.

ASR Correction

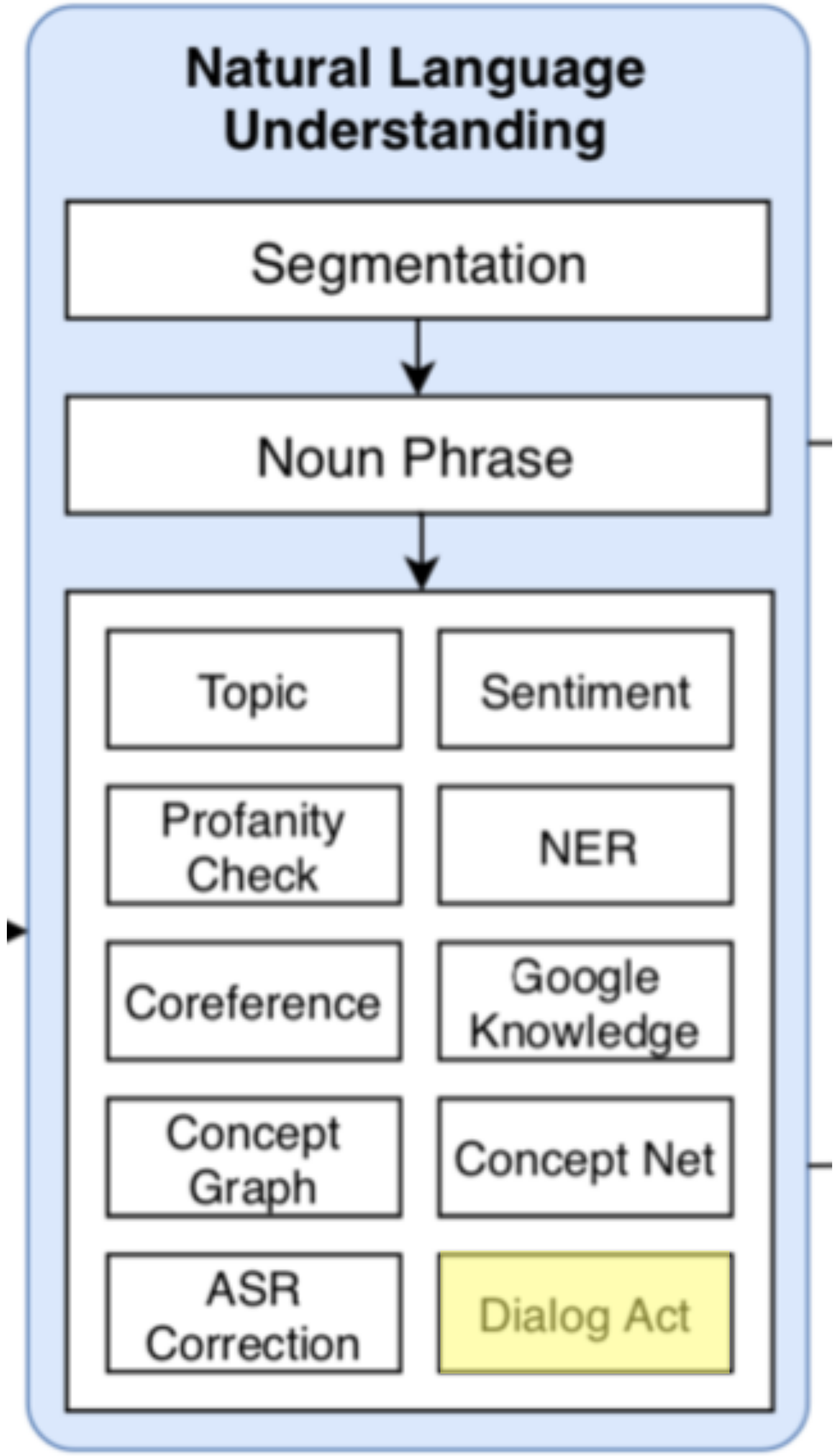


Through the **double metaphone algorithm**, they improved the very important issue of homophone words (words that sound the same but have different spelling).

Using phonetics to find a match **increases the accuracy of NER** in specific domains.

For instance, the user input “mama mia” from ASR is more likely referring to “Mamma Mia,” the movie.

Dialog Act

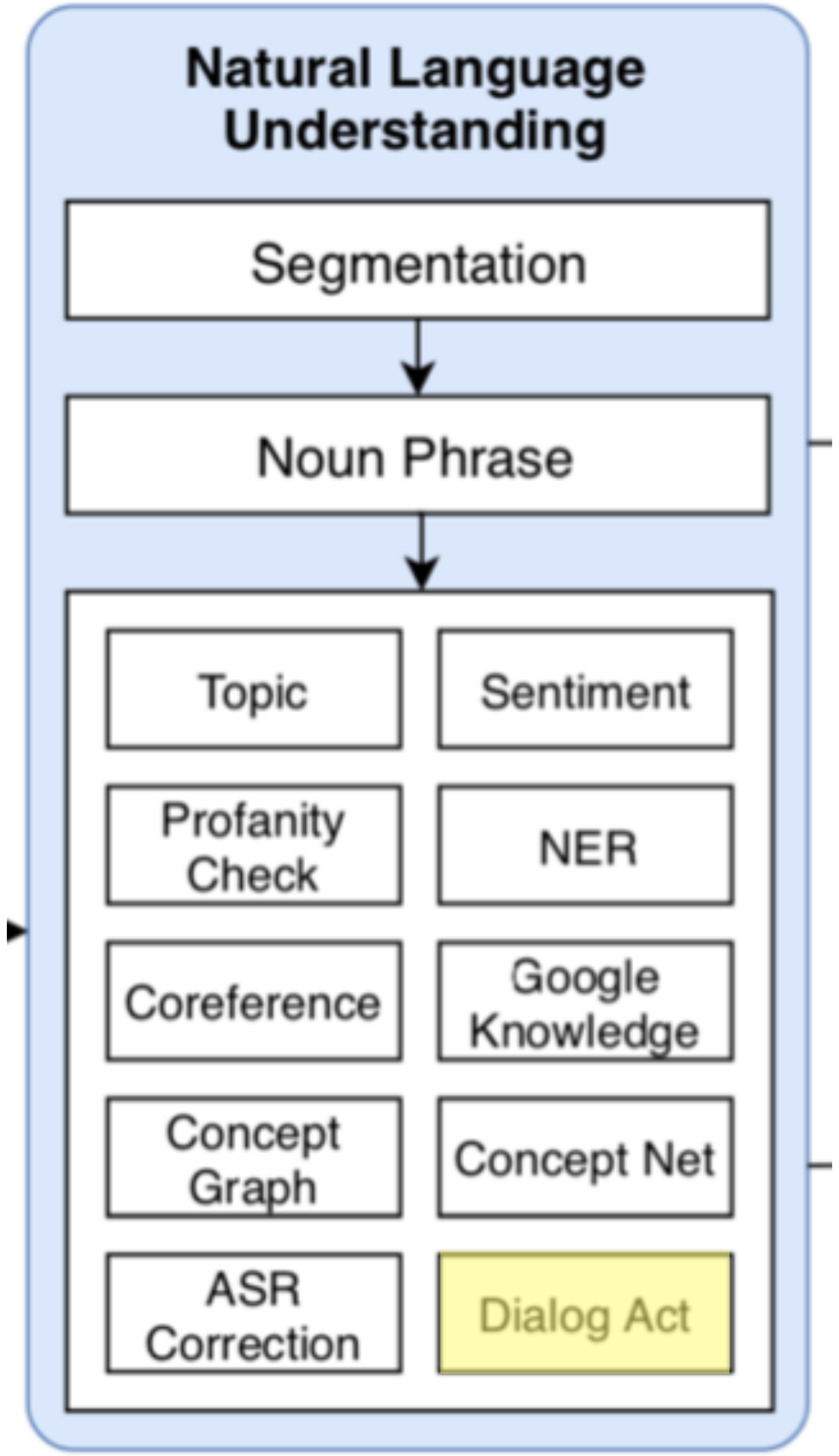


The dialog act is the function in the dialog given the context of the conversation i.e opinion, statement. Each segmented sentence from NLU is associated with a dialog act.

They trained an LSTM and a CNN model to predict the dialog act:

- **2-layer bi-LSTM model** pre-trained with fastText with an embedding size of 300 and a hidden size of 500.
- **2-layer CNN model** also pre-trained with fastText on a kernel width of 3.

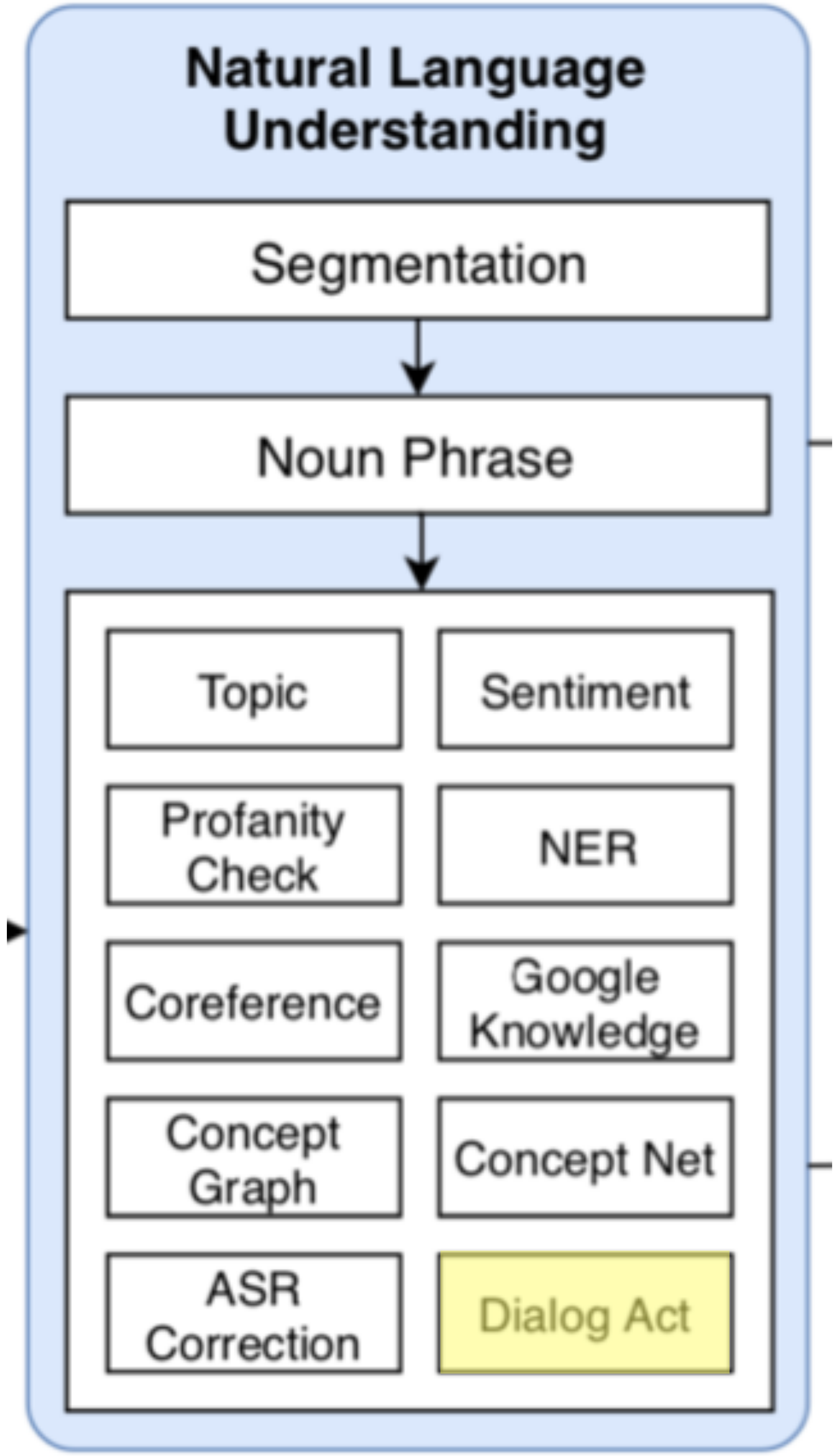
Dialog Act

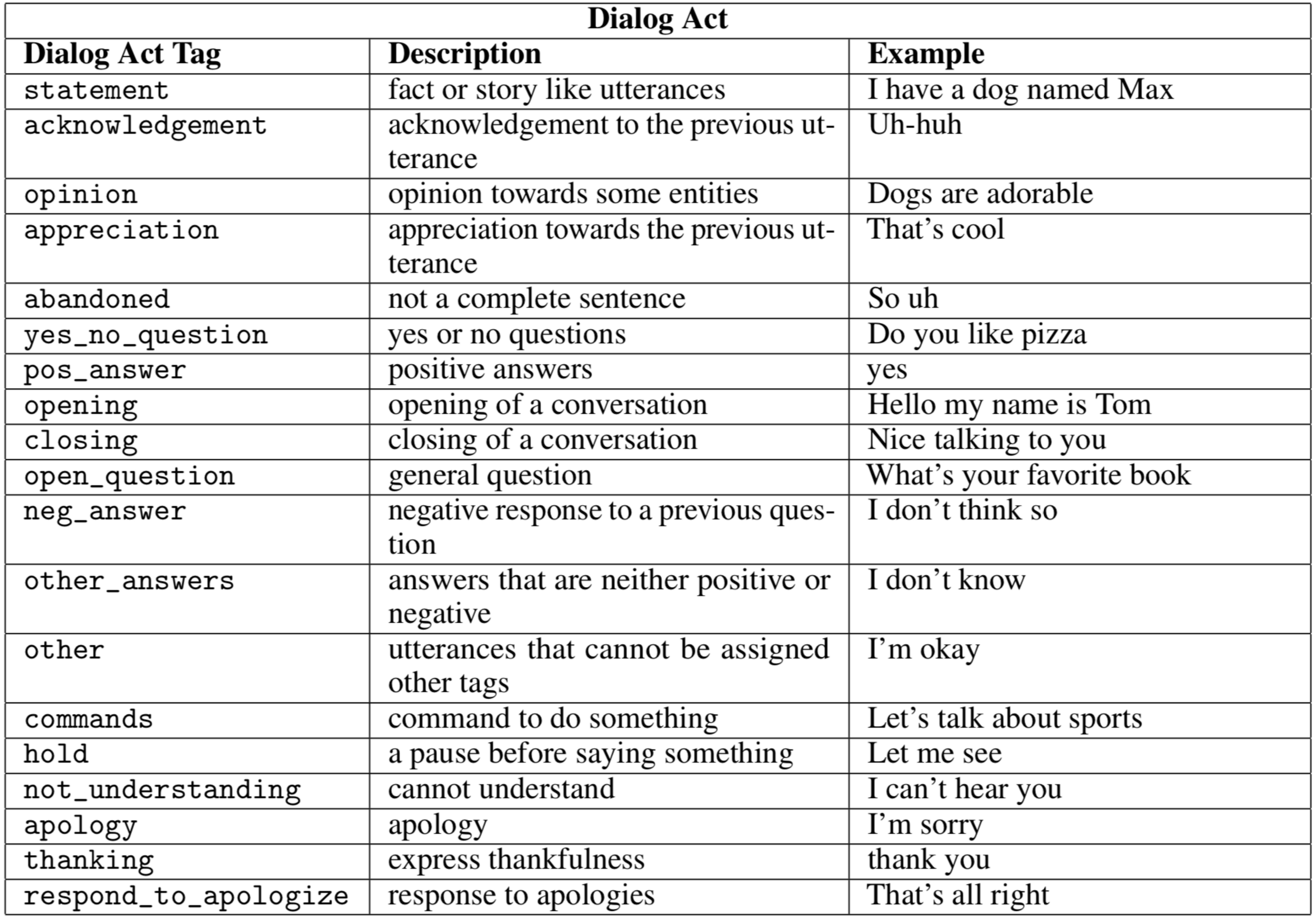


They use the **Switchboard Dialog Act Corpus (SWDA)**. The SWDA dataset collects 205,000 utterances of telephone conversations in open domain and has 60 dialog act tags.

Then they reduced the target to **19**.

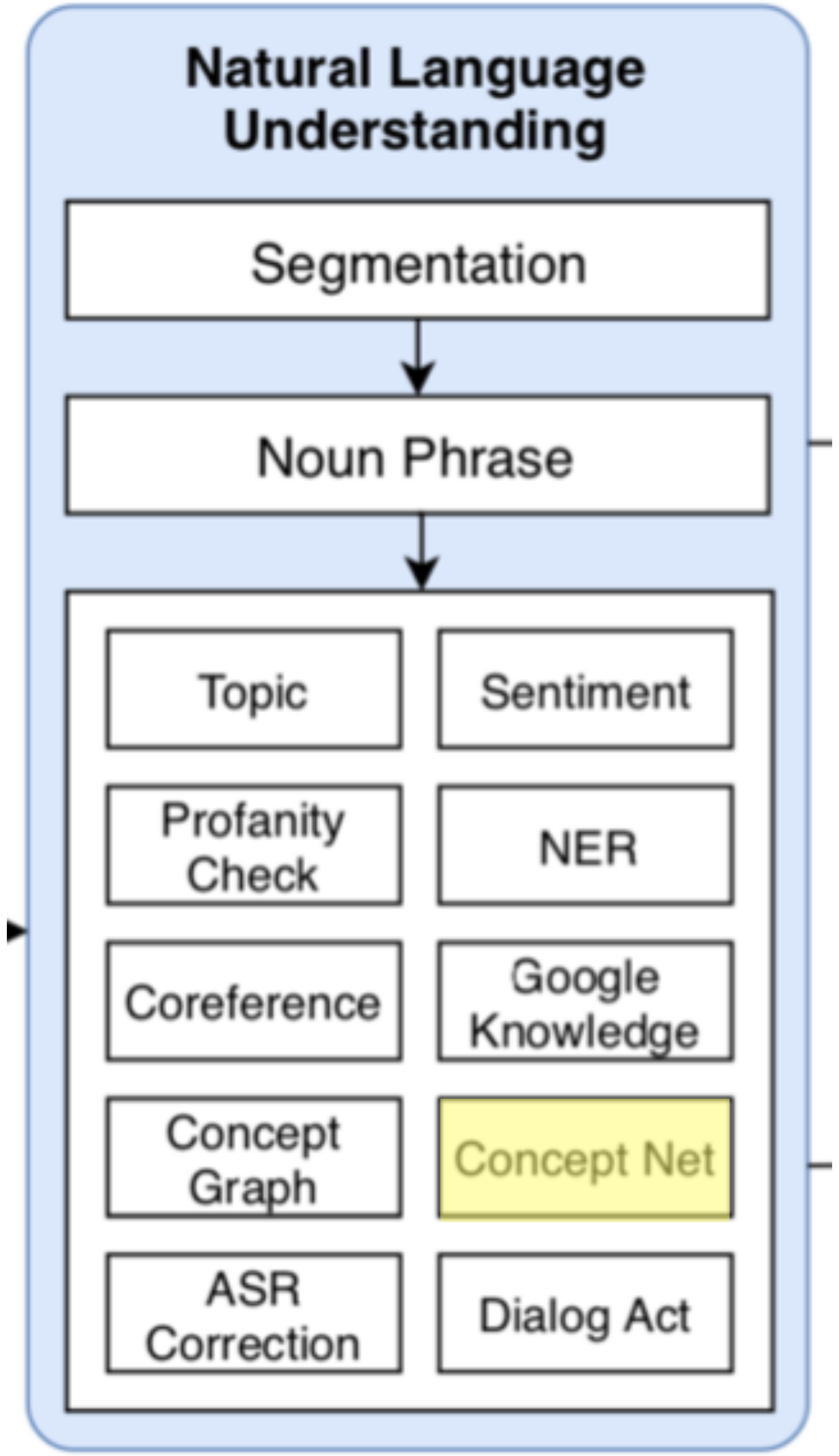
Dialog Act





Dialog Act		
Dialog Act Tag	**Description**	**Example**
`statement`	fact or story like utterances	I have a dog named Max
`acknowledgement`	acknowledgement to the previous utterance	Uh-huh
`opinion`	opinion towards some entities	Dogs are adorable
`appreciation`	appreciation towards the previous utterance	That's cool
`abandoned`	not a complete sentence	So uh
`yes_no_question`	yes or no questions	Do you like pizza
`pos_answer`	positive answers	yes
`opening`	opening of a conversation	Hello my name is Tom
`closing`	closing of a conversation	Nice talking to you
`open_question`	general question	What's your favorite book
`neg_answer`	negative response to a previous question	I don't think so
`other_answers`	answers that are neither positive or negative	I don't know
`other`	utterances that cannot be assigned other tags	I'm okay
`commands`	command to do something	Let's talk about sports
`hold`	a pause before saying something	Let me see
`not_understanding`	cannot understand	I can't hear you
`apology`	apology	I'm sorry
`thanking`	express thankfulness	thank you
`respond_to_apologize`	response to apologies	That's all right

Topic Expansion

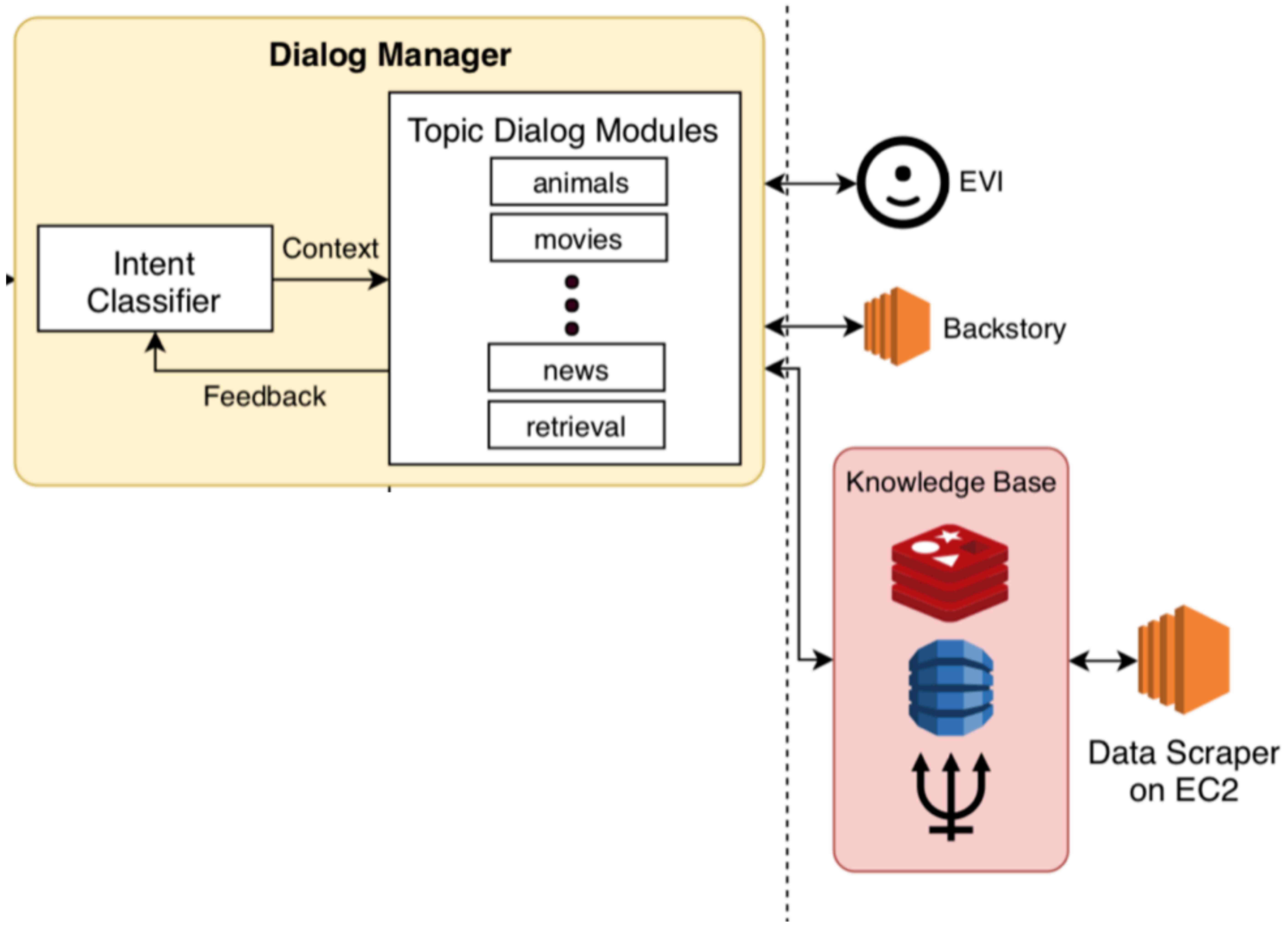


They use **ConceptNet** as a knowledge graph for entity expansion on the extracted noun phrases.

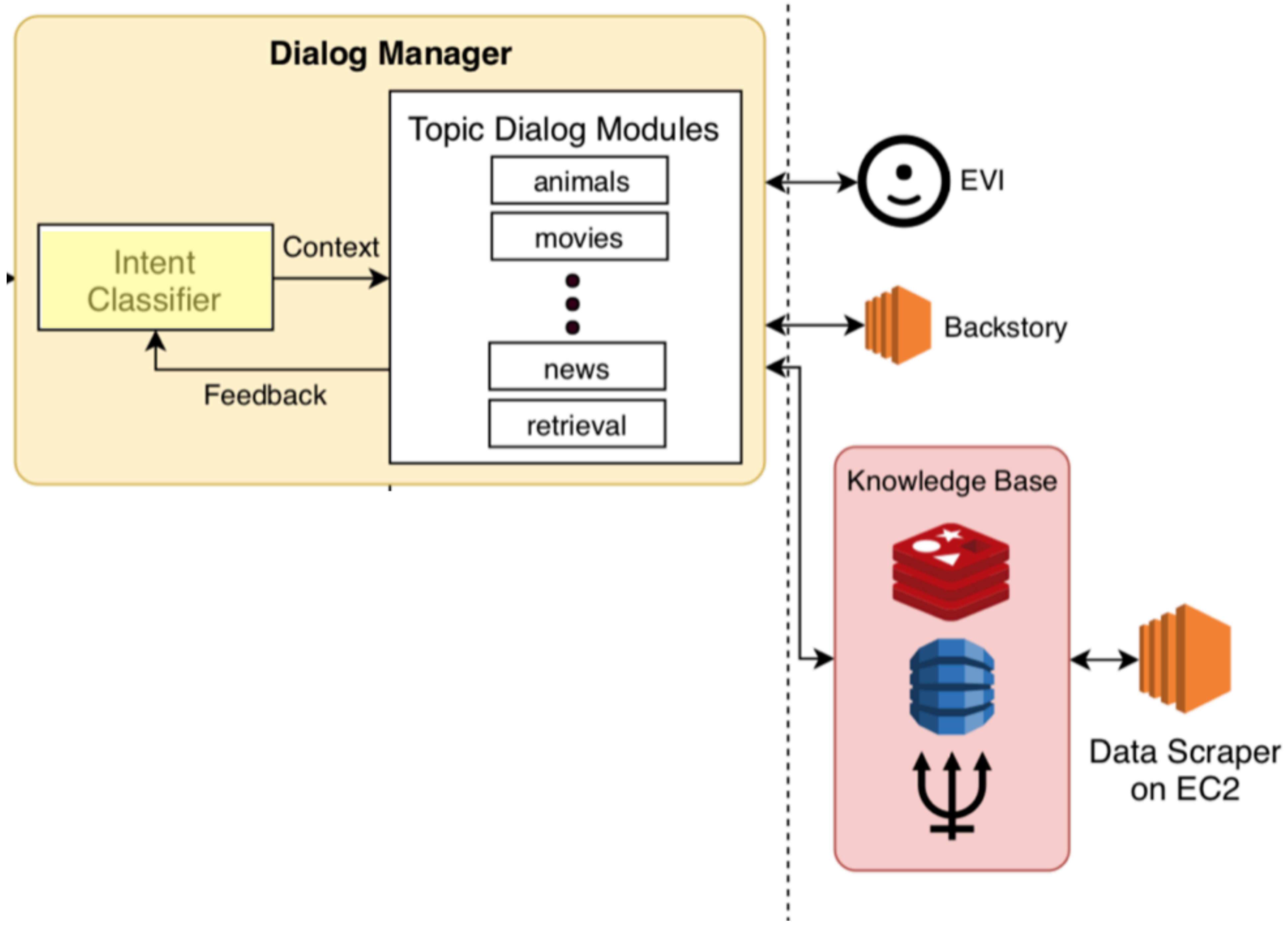
Apart from asking the users to share more and store the information in their database as a learning process, they can talk about similar topics.

For example, if a user wants to talk about **cars**, they can retrieve from ConceptNet a list of different car types (such as a **Volvo**).

Dialog Manager



Intent Classifier



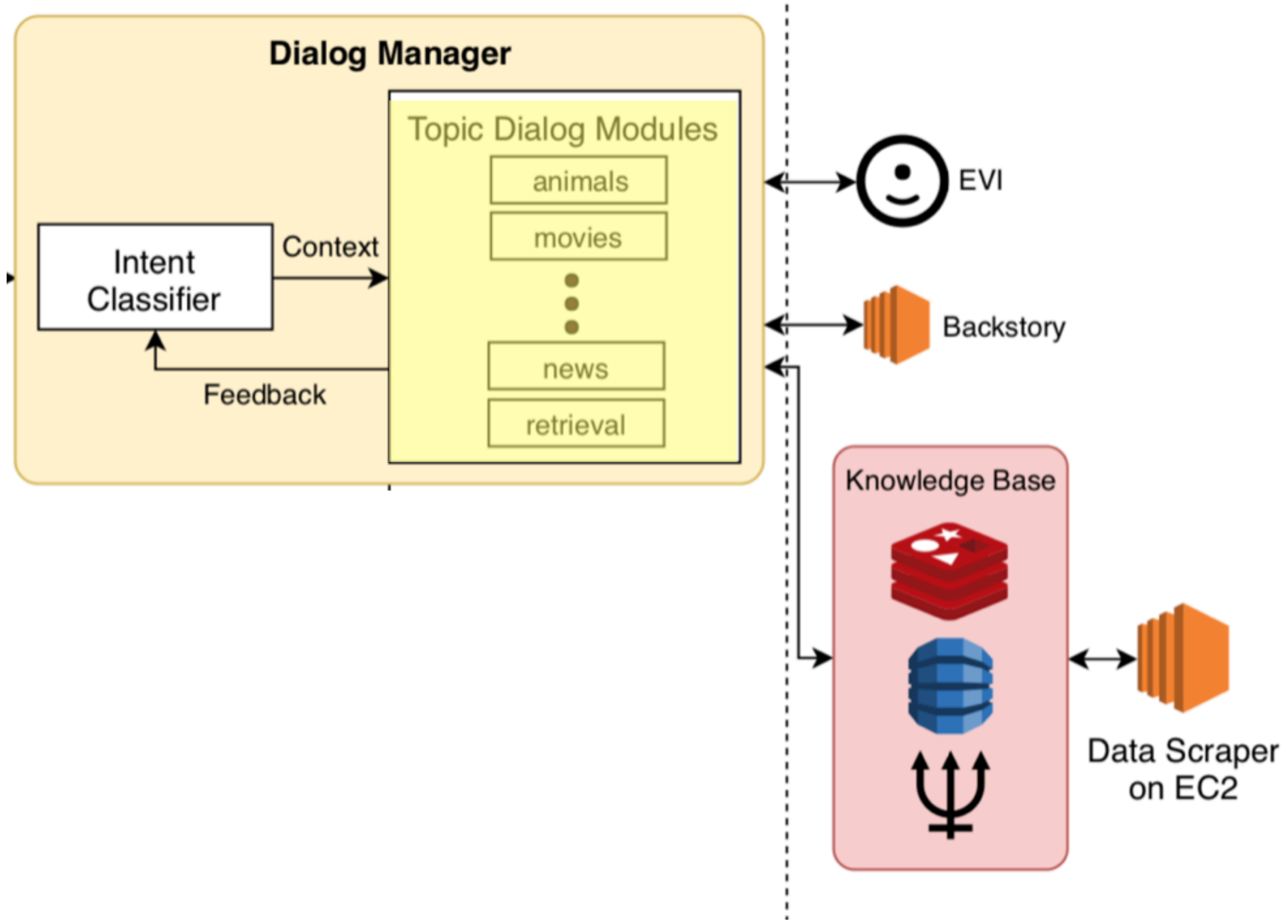
They defined three levels of user intents.

1. They first handle social chat domain system requests. For these requests, the system would explain to the user how to exit social mode.

2. Next, they detect topic intents from user utterances. They also tune the confidence thresholds of these three detectors based on the heuristic and dialog data they collected.

3. The last level is called lexical intents. They use the regular expression to analyze user requests, such as whether the user is asking about the preference or opinion of our social bot.

Dialog Module Selector

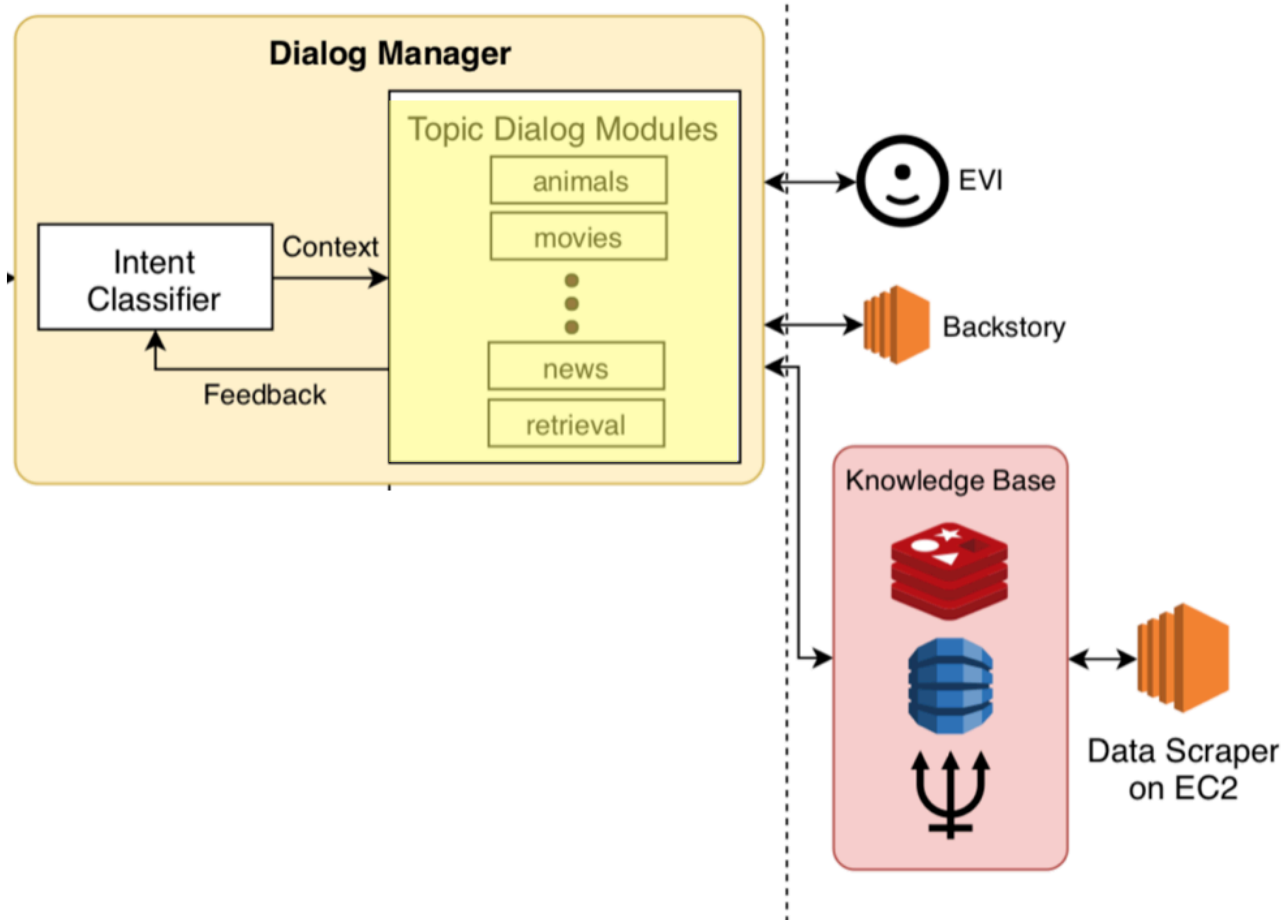


Our dialog module selector first picks a topic dialog module responsible for the topic intent detected by our intent classifier.

The priority of the modules to be proposed are tuned based on their everyday performance.

For example, if the utterance is “let’s talk about movies”, the dialog module selector would select the movie module immediately.

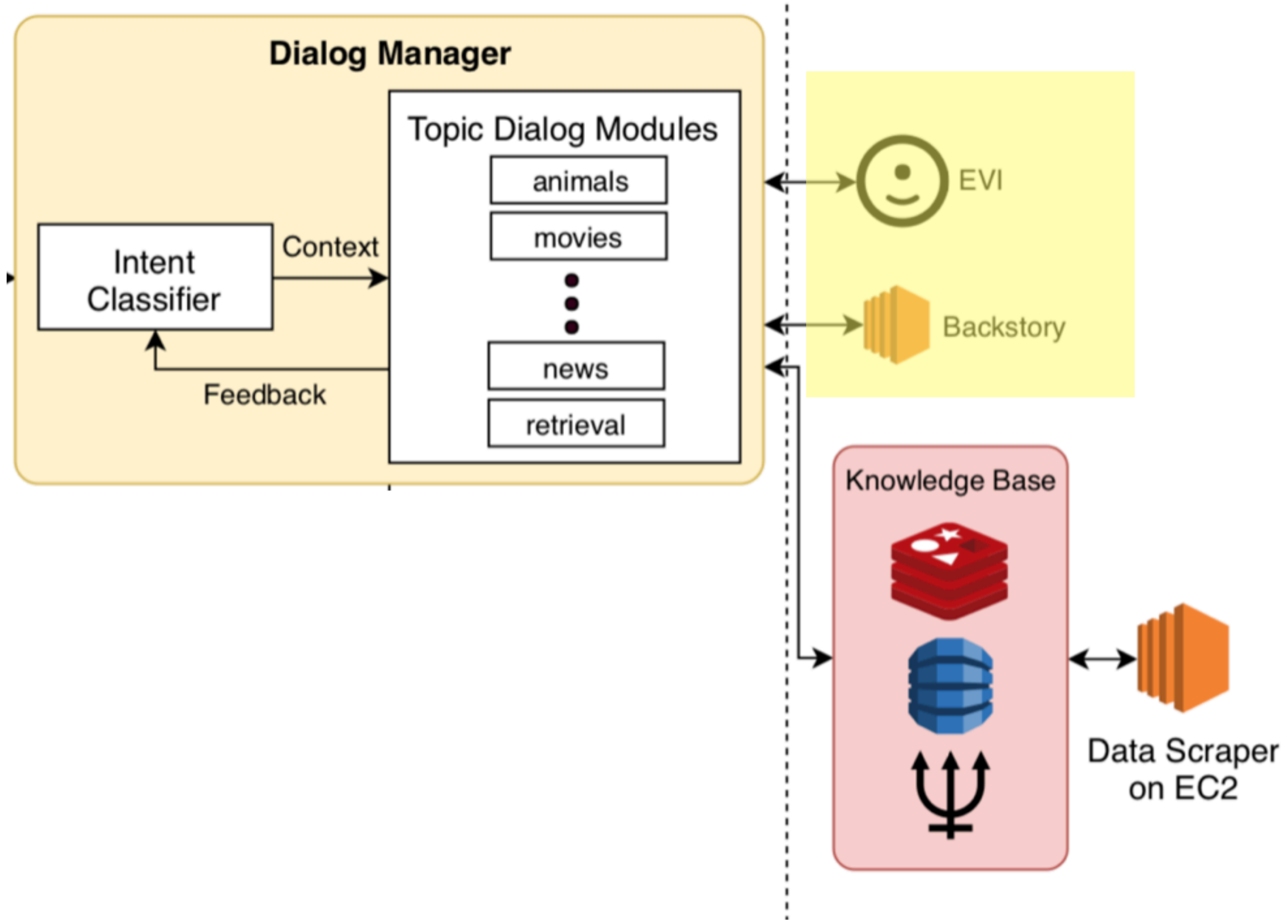
Topic Dialog Modules



Each topic dialog module has its own dialog flow design

- Animals
- Movies
- Music
- News
- Retrieval
- Sport
- Game
- Psychology
- Technology
- Travel

Low Level Dialog Management

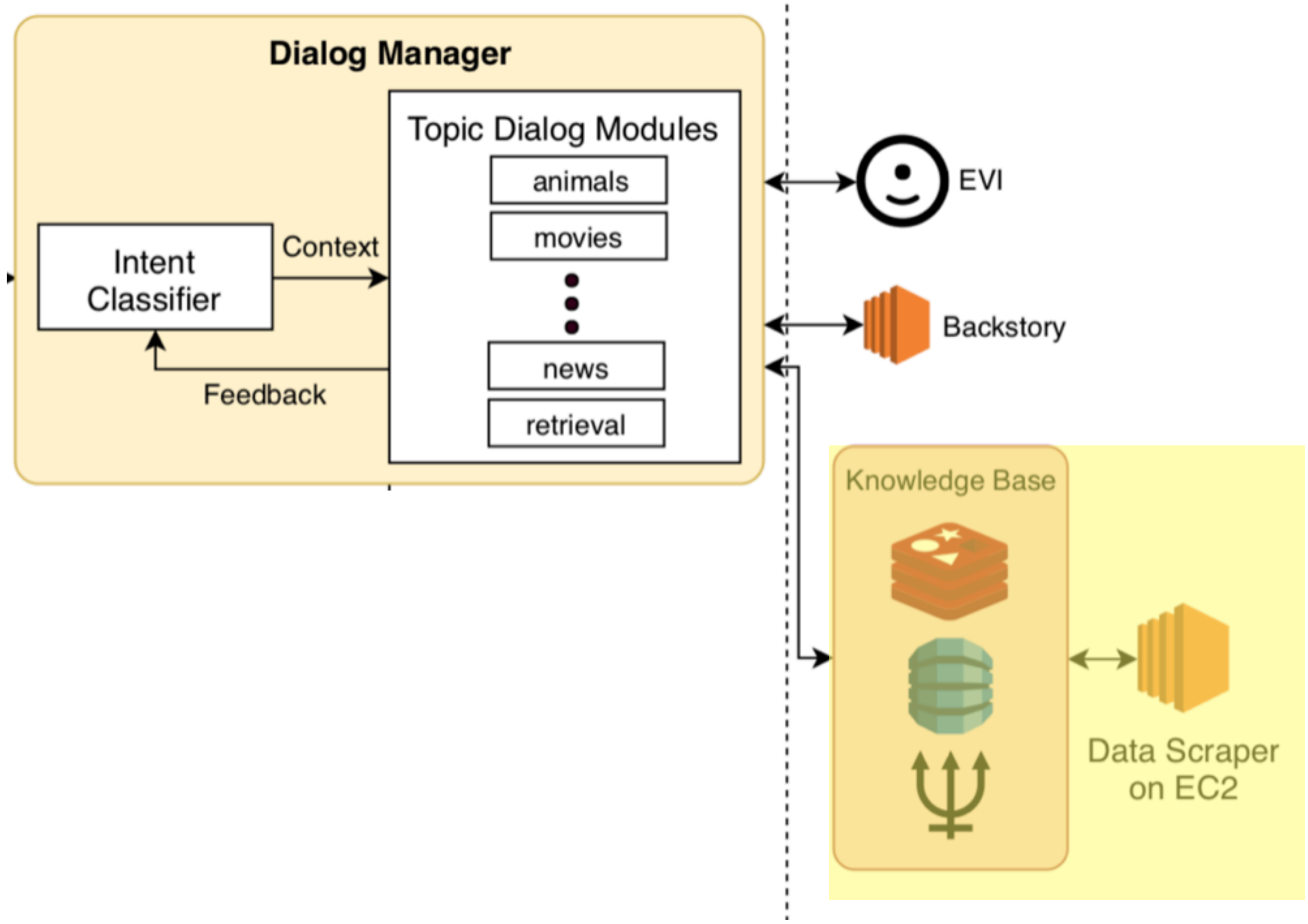


They built two APIs, which are *Backstory* and *EVI* to answer general facts and background questions about our chatbot.

- **Backstory**: A service designed to retrieve responses for questions related to our chatbot’s background and preferences

- **EVI**: A service provided by Amazon. It can answer factual questions such as "how old is Lebron James?". EVI will return "I don’t have opinion on that" or "I don’t know about that" if it does not have a corresponding answer.

Knowledge Base



Our knowledge base consists of unified datasets stored in DynamoDB tables by topics.

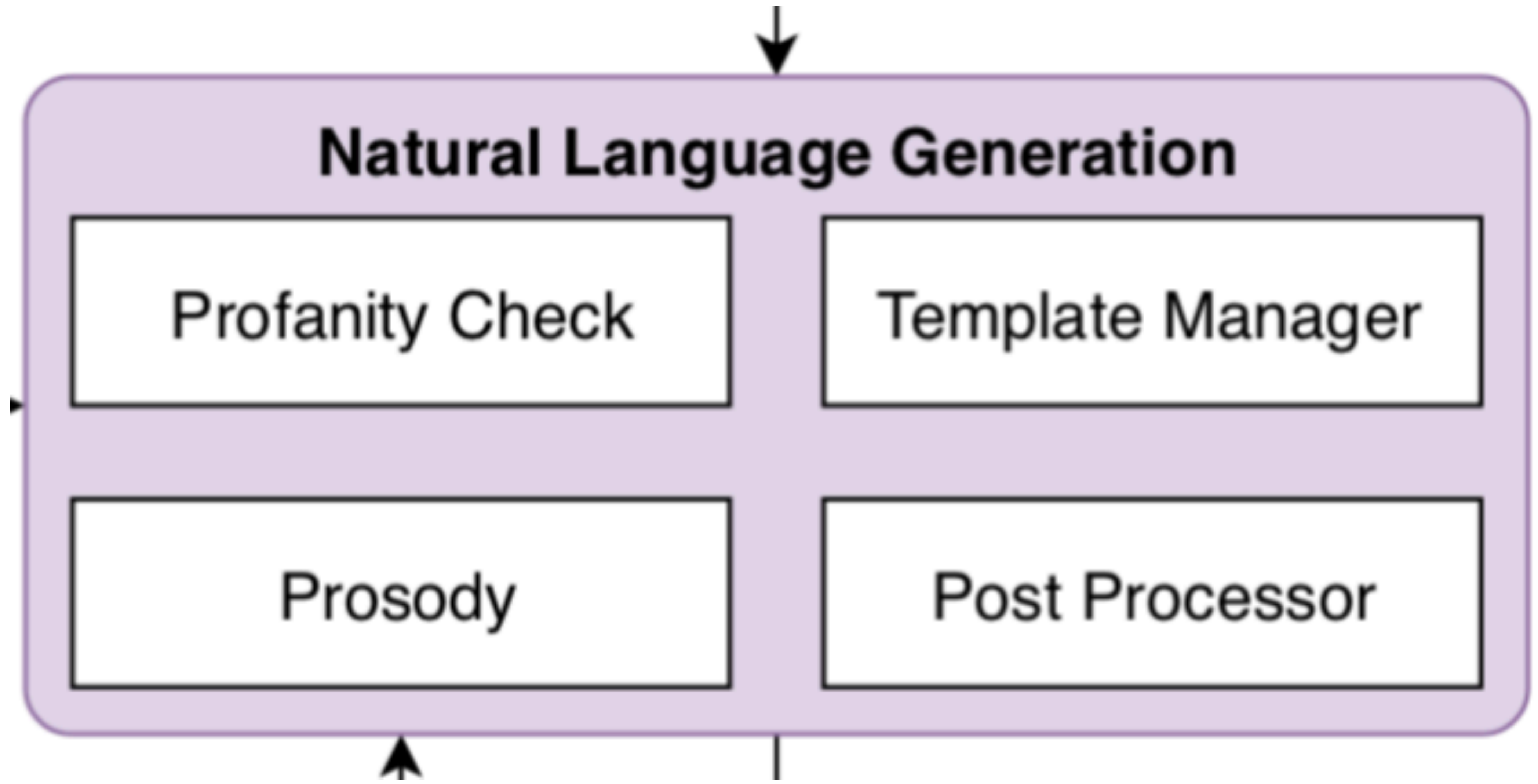
- Factual Content

  – Reddit: We compose a large collection of events daily from various subreddits.
  – Twitter Moments: Twitter moments to help users keep up with what the world is talking about in real time.
  – General Information: For general information on movies and music, they use IMDB database dumps and Spotify's One Million Playlist dataset.

- Opinionated Content

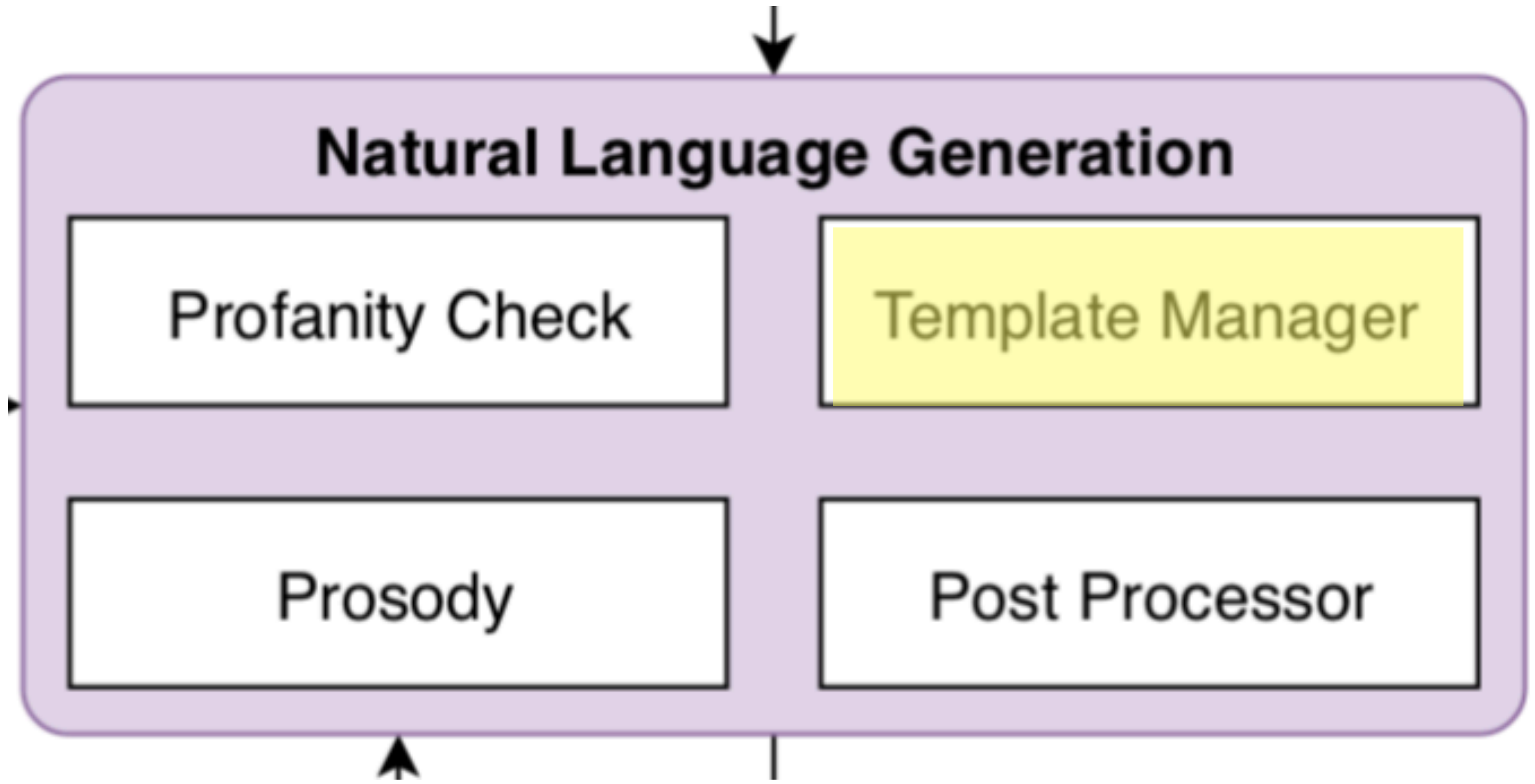
  – Twitter Opinions: They accompany the Twitter Moment with mined opinions.
  – Debate Opinions Gunrock attempts to match statements and opinions with over 71, 000 topics and 460, 000 opinions using a universal sentence encoder.

Natural Language Generator



Their system’s natural language generation module is **template-based**. It selects a manually designed template and fills out specific slots with information retrieved from the knowledge base by the dialog manager.

Template Manager

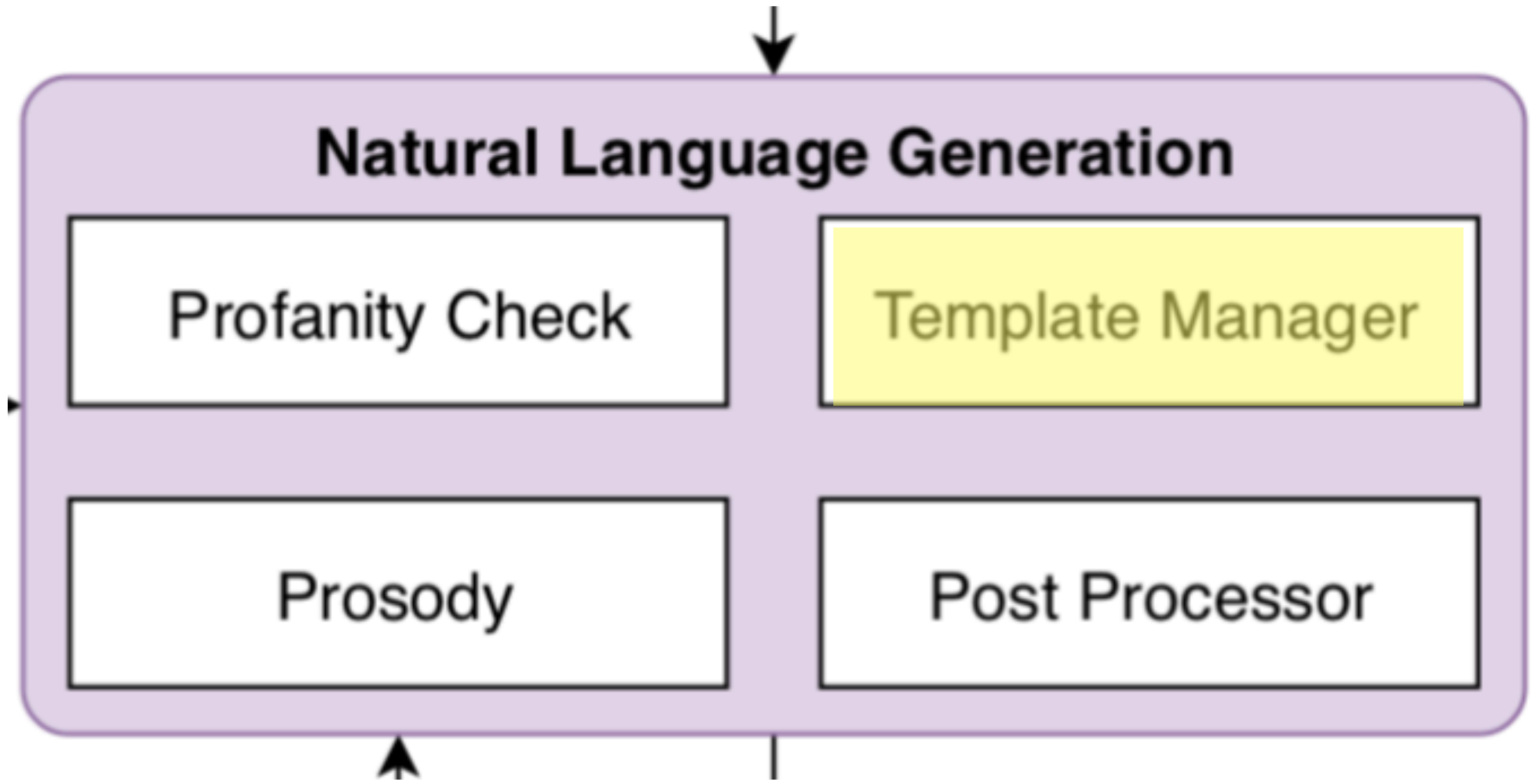


The template manager module stores and parses response templates used by the system. It centralizes all response templates from their system's several parallel dialog flow components.

One of the main goals in using a template manager is to avoid duplicate responses.

By varying the responses, our system can sound more natural and human-like, and avoid situations where the system has only one response to the user's request.

Prosody Synthesis



Their system utilizes Amazon Alexa’s speech synthesis system for speech synthesis.

They use Amazon SSML format to enhance their templates, such as when reading out phone numbers or correctly pronouncing homographs and acronyms.

Thanks for the attention