

**FITUR ROBOT BANTUAN EDUKASI : SISTEM *QUESTION*  
*ANSWERING* BERBASIS LISAN DALAM DOMAIN  
PENDIDIKAN ANAK USIA DINI**

***EDUCATIONAL ASSISTANCE ROBOT FEATURES: ORAL-  
BASED QUESTION ANSWERING SYSTEM IN THE DOMAIN  
OF EARLY CHILDHOOD EDUCATION***

**ANDRI DWI UTOMO**



**SEKOLAH PROGRAM PASCASARJANA  
UNIVERSITAS HASANUDDIN  
MAKASSAR  
2020**

**FITUR ROBOT BANTUAN EDUKASI : SISTEM *QUESTION*  
*ANSWERING* BERBASIS LISAN DALAM DOMAIN  
PENDIDIKAN ANAK USIA DINI**

Tesis

Sebagai Salah Satu Syarat Untuk Mencapai Gelar Magister

Program Studi

Teknik Elektro

Disusun Dan Diajukan Oleh

**ANDRI DWI UTOMO**

Kepada

**PROGRAM PASCASARJANA  
UNIVERSITAS HASANUDDIN  
MAKASSAR  
2020**

# TESIS

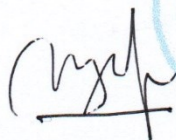
## FITUR ROBOT BANTUAN EDUKASI: SISTEM *QUESTION ANSWERING* BERBASIS LISAN DALAM DOMAIN PENDIDIKAN ANAK USIA DINI

Disusun dan diajukan oleh

**ANDRI DWI UTOMO**  
Nomor Pokok D032171001

Telah dipertahankan di depan Panitia Ujian Tesis  
Pada tanggal 14 Agustus 2020  
dan dinyatakan telah memenuhi syarat

Menyetujui  
Komisi Penasehat,



Prof. Dr. Ir. Syafruddin Syarif, M.T

Ketua



Dr. Ir. Zahir Zainuddin, M.Sc.

Sekretaris

Ketua Program Studi S2  
Teknik Elektro,



Prof. Dr. Eng. Syafaruddin, ST, M.Eng.  
NIP. 197405301999031003

Dekan Fakultas Teknik  
Universitas Hasanuddin,



Prof. Dr. Ir. Muh. Arsyad Thaha, M.T.  
NIP. 196012311986091001

## PERNYATAAN KEASLIAN TESIS

Yang bertanda tangan dibawah ini :

Nama : ANDRI DWI UTOMO

Nomor Pokok : D032171001

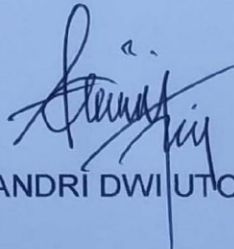
Program Studi : Teknik Elektro

Program Studi : Teknik Informatika

Menyatakan dengan sebenarnya bahwa tesis yang saya tulis ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambilan tulisan atau pemikiran orang lain. Apabila di kemudian hari terbukti atau dapat dibuktikan bahwa sebagian atau keseluruhan tesis ini hasil karya orang lain, saya bersedia menerima sanksi atas perbuatan tersebut.

Makassar, 14 Agustus 2020

Yang menyatakan,



ANDRI DWI UTOMO

## KATA PENGANTAR

Puji syukur kami panjatkan kehadirat Allah SWT. atas berkah dan Rahmat-Nya sehingga penelitian dengan judul “**Fitur robot bantuan edukasi : sistem *question answering* berbasis lisan dalam domain pendidikan anak usia dini**” Tak lupa pula shalawat dan salam kepada Nabi Muhammad SAW yang telah menyinari dunia ini dengan keindahan ilmu dan akhlak yang diajarkan kepada seluruh umatnya.

Dalam penyusunan tesis ini, penulis menyadari banyak mengalami tantangan dan hambatan namun berkat dukungan serta kerja sama yang baik dengan berbagai pihak sehingga tesis ini dapat diselesaikan. Untuk itu penulis mempersembahkan ucapan terima kasih teristimewa dan tak terhingga kepada Ibunda yang terkasih **Almh. Nanik S**, Ayahanda **Sugeng Raharjo**, kakak **Agung Ari Purwanto** dan istri penulis **Nurlinda** atas segala dedikasi yang tak terhingga, kasih sayang, doa restu beserta dukungan moril maupun materil kepada penulis dalam menyelesaikan tesis ini

Ucapan terima kasih pun penulis hanturkan kepada kepada pembimbing, **Prof. Dr. Ir. Syafruddin Syarif M.T** dan **Dr. Ir. Zahir Zainuddin, M.Sc.** yang telah meluangkan waktunya kepada penulis untuk membimbing dan berkonsultasi tentang materi dalam tesis ini dan juga kepada seluruh dosen dan staf Departemen Teknik Elektro, Universitas

Hasanuddin yang telah membantu dalam hal keilmuan maupun administrasi pada tahap tesis ini.

Penulis menyadari bahwa tesis ini masih belum sempurna. Dengan demikian, penulis tetap mengharapkan saran dan kritik dengan harapan semoga tulisan ini bisa memberikan manfaat kepada seluruh pihak.

Makassar, 14 Agustus 2020

Penulis

## ABSTRAK

**ANDRI DWI UTOMO.** Fitur robot bantu edukasi: sistem *question answering* berbasis lisan dalam domain pendidikan anak usia dini.

Penelitian ini menciptakan media bantu bagi tenaga pengajar di sekolah pendidikan anak usia dini yaitu perangkat lunak untuk fitur robot bantu edukasi sistem penjawab pertanyaan. Sistem penjawab pertanyaan dibangun dari 3 komponen utama yaitu *speech recognition*, sistem *question answering* dan *text to speech*. Ucapan digunakan sebagai data masukan dan keluaran dari sistem yang dibangun. Sistem penjawab pertanyaan dibangun menggunakan *Recurrent Neural Network* (RNN) dengan memanfaatkan mekanisme gerbang *Long-Short Term Memory* (LSTM). Dataset yang digunakan merupakan dataset percakapan dalam pendidikan anak usia dini yang terdiri dari pasangan pertanyaan dan jawaban. Pada proses pelatihan didapatkan hasil nilai metrik cenderung naik yang berarti proses pelatihan berlangsung dengan baik. Hasil dari uji generasi respon dari 115 pertanyaan yang diajukan 23 siswa secara acak dihasilkan nilai 93 jawaban yang benar, 20 salah dan 2 pertanyaan yang tidak terdeteksi.

Kata kunci: RNN lstm, *seq2seq*, sistem penjawab pertanyaan, pendidikan anak usia dini

## ABSTRACT

**ANDRI DWI UTOMO.** Education Assistance Robot Features: Question-Based Answer System Questions In Early Education Domains In Children.

This research creates aids for teaching staff in early childhood education schools, namely software for robots to help educate question answering systems. The question answering system is built from 3 main components, namely speech recognition, question answering system and text to speech. Speech is used as input and output data from the system being built. The question answering system was built using the Recurrent Neural Network (RNN) by utilizing the Long-Short Term Memory gate mechanism (LSTM). The dataset used is a conversation dataset in early childhood education consisting of pairs of questions and answers. In the training process, the results of the metric values tend to increase, which means the training process is going well. The results of the response generation test of 115 questions raised by 23 students were randomly generated 93 correct answers, 20 incorrect questions and 2 undetected questions.

Keywords: RNN LSTM, Seq2seq, question answering system, early childhood education programs



## DAFTAR ISI

<b>KATA PENGANTAR .....</b>	<b>iv</b>
<b>ABSTRAK .....</b>	<b>vi</b>
<b>ABSTRACT .....</b>	<b>vii</b>
<b>DAFTAR ISI .....</b>	<b>viii</b>
<b>DAFTAR GAMBAR.....</b>	<b>x</b>
<b>DAFTAR TABEL .....</b>	<b>xi</b>
<b>BAB I PENDAHULUAN</b>	
<b>A. Latar Belakang .....</b>	<b>1</b>
<b>B. Rumusan Masalah.....</b>	<b>6</b>
<b>C. Tujuan Penelitian .....</b>	<b>6</b>
<b>D. Manfaat Penelitian.....</b>	<b>7</b>
<b>E. Batasan Masalah .....</b>	<b>7</b>
<b>F. Sistematika Penulisan .....</b>	<b>8</b>
<b>BAB II TINJAUAN PUSTAKA</b>	
<b>A. Landasan Teori.....</b>	<b>10</b>
<b>B. Penelitian Terkait .....</b>	<b>31</b>
<b>C. State of the Art .....</b>	<b>34</b>
<b>D. Kerangka Pikir .....</b>	<b>36</b>
<b>E. Kamus Kata .....</b>	<b>36</b>

**BAB III METODE PENELITIAN**

<b>A. Tahapan Penelitian .....</b>	<b>40</b>
<b>B. Waktu &amp; Lokasi Penelitian.....</b>	<b>41</b>
<b>C. Jenis Penelitian .....</b>	<b>42</b>
<b>D. Perancangan Sistem.....</b>	<b>42</b>
<b>E. Sumber Data .....</b>	<b>58</b>
<b>F. Instrumen Penelitian.....</b>	<b>59</b>
<b>G. Tolak Ukur Akurasi Sistem .....</b>	<b>60</b>

**BAB IV HASIL DAN PEMBAHASAN**

<b>A. Hasil Pelatihan .....</b>	<b>61</b>
<b>B. Hasil Uji dan Prediksi Jawaban.....</b>	<b>65</b>

**BAB V KESIMPULAN DAN SARAN**

<b>A. Kesimpulan.....</b>	<b>72</b>
<b>B. Saran.....</b>	<b>73</b>

**DAFTAR PUSTAKA**

## DAFTAR GAMBAR

<b>Nomor</b>		<b>Halaman</b>
1.	Arsitektur RNN Sederhana .....	27
2.	Unit LSTM .....	29
3.	Arsitektur <i>Seq2Seq</i> .....	31
4.	<i>State Of The Art</i> .....	35
5.	Kerangka Pikir .....	36
6.	Gambaran Umum sistem berbasis komputer pada sekolah PAUD	43
7.	gambaran umum sistem QA pada sekolah PAUD .....	44
8.	susunan dataset <i>question answer</i> .....	46
9.	Processing Dataset .....	47
10.	Kode <i>preprocessing dataset</i> .....	47
11.	Kode proses <i>Tokenizing</i> .....	48
12.	Kode proses <i>Encode Sequences</i> .....	50
13.	Kode proses <i>Encode Output</i> .....	51
14.	Kode proses prediksi sequences .....	52
15.	Flowchart Proses Pelatihan .....	55
16.	Flowchart Proses Prediksi Jawaban .....	57
17.	Grafik hasil pelatihan RNN Istm .....	64
18.	Hasil uji prediksi jawaban .....	71

## DAFTAR TABEL

<b>Nomor</b>	<b>Halaman</b>
1. Dataset Hasil Preprocessing .....	48
2. Dataset Hasil Tokenizing .....	49
3. Hasil <i>Encode Sequence</i> .....	50
4. Struktur Layer Model LSTM.....	53
5. Lama Waktu Pelatihan .....	62
6. Nilai Metrik Pelatihan Model Seq2Seq RNN .....	63
7. Nilai Metrik Pelatihan Model Seq2Seq RNN LSTM.....	63
8. Ukuran Hasil Bobot Pelatihan.....	64
9. Hasil Uji Generasi Respon .....	66

# BAB I

## PENDAHULUAN

### A. Latar Belakang

Sistem penjawab pertanyaan atau *Question Answering System*(QAS) merupakan sistem yang jauh lebih didepan dibanding sistem mesin pencari untuk menghasilkan suatu jawaban dari pertanyaan yang diajukan(Garg and Kumar, 2016). Dimana sistem ini memanfaatkan teknik *Information Retrieval*(IR) dan *Natural Language Processing* (NLP) untuk menghitung dan menampilkan jawaban yang terbaik dari sumber atau dokumen yang tersedia (Pudaruth et al., 2016). Pada sebuah sistem QA, awalnya akan menerima *query* terlebih dahulu dalam bentuk pertanyaan dengan menggunakan Bahasa alami yang kemudian mengekstrak dan mencari jawaban pada sekumpulan dokumen atau pada sebuah domain basis pengetahuan(Bendi, n.d.).

Pesatnya perkembangan ilmu pengetahuan dan teknologi, mengakibatkan pergeseran model pembelajaran yang kini memanfaatkan teknologi sebagai media bantu untuk menciptakan suasana pendidikan yang menyenangkan, kreatif dan dinamis sesuai dengan yang tertuang pada UU Sisdiknas No. 20 tahun 2003 pasal 40;2. Robot edukasi merupakan salah satu media bantu yang efektif atau alat inovatif (“Impact of Using an Educational Robot-Based Learning System on Students’ Motivation in Elementary Education - IEEE Journals & Magazine,” n.d.)

yang melayani kegiatan pendidikan melalui interaksi sosial. Anak usia dini merupakan usia awal untuk membentuk perkembangan kognitif, motorik, dan bahasa dari seorang anak yang merupakan suatu tujuan diadakannya pendidikan anak usia dini. Rendahnya kemampuan berbahasa anak disebabkan karena dalam proses pembelajaran anak cenderung kurang percaya diri, pasif dan kurang kreatif sehingga anak tidak mampu melakukan kegiatan menarik pada saat proses pembelajaran (Pebriana, 2017). Saat ini semakin banyak perangkat atau fitur dari robot yang telah dikembangkan untuk mendukung perkembangan seorang anak dan diterapkan di dunia pendidikan sehingga mampu membantu kegiatan pembelajaran lebih menarik untuk anak. Robot edukasi yang ditanamkan sistem *question answering* dengan masukan pertanyaan menggunakan ucapan dan mengeluarkan jawaban berupa ucapan akan membantu seorang anak untuk meningkatkan pengetahuan dan melatih kemampuan komunikasi sehingga tujuan pendidikan anak usia dini dapat tercapai. Penelitian media bantu pembelajaran untuk anak usia dini masih tergolong jarang terutama sistem *question answering* yang disematkan ke dalam robot bantu edukasi, dikarenakan sulitnya dalam penyusunan bahan informasi yang hanya boleh diketahui oleh seorang anak usia dini dan yang akan berpengaruh untuk membentuk perkembangan seorang anak. Domain pertanyaan menjadi hal yang sangat harus dipertimbangkan untuk membangun sistem *question answering* pada anak usia dini.

Penelitian terkait sistem *question answering* dengan memanfaatkan suara sebagai data masukan dilakukan oleh Dessi PL dan Riandy RN pada tahun 2017 mengenai sistem *question answering* berbasis lisan untuk layanan rute kereta dengan menggunakan pendekatan *frame-based*. Ada 3 komponen utama dari sistem yang dirancang yaitu *automatic speech recognizer*(ASR) yang mentranskripsikan masukan ucapan pengguna ke urutan kata-kata kueri, sistem *question answering* yang bertugas menerjemahkan teks yang ditranskripsikan ke dalam kueri dan mencari jawaban yang relevan, dan *automatic speech synthesizer*(ASS) yang bertugas menghasilkan jawaban yang relevan menggunakan ucapan. Untuk sistem *question answering* digunakan beberapa teknik seperti *frame-based*, *keyword spotting*, *data fetching*, dan *canned answer* menghasilkan akurasi hingga 100% (Lestari and Nugraha, 2017).

Waheeb Ahmed dkk pada tahun 2017 juga melakukan penelitian tentang sistem pertanyaan jawaban untuk pertanyaan factoid menggunakan Bahasa arab. Penelitian yang dilakukan dimulai dengan mengumpulkan data-data pertanyaan factoid Bahasa arab yang diambil dari web. Model sistem pertanyaan jawaban yang dibangun terdiri dari question analyzer, knowledge retriever, answer generator, dan external knowledge-base. Keempat model sistem pertanyaan jawaban dibangun menggunakan *Recurrent Neural Network*. Penelitian yang dilakukan menghasilkan akurasi sebesar 53% (Ahmed and Anto, 2017).

Sementara itu, R. Poonguzhali dan K. Lakshmi melakukan penelitian tentang performance dari *Recurrent Neural Network* dalam menangani sistem penjawab pertanyaan. Dataset yang digunakan dalam penelitian yang dilakukan adalah bAbI. bAbI mempunyai 20 *task* dan penelitian ini menggunakan task-1 (*Basic factoid QA with single supporting fact*) dan task-20 (*Reasoning about agent's motivation*) untuk menguji *accuracy* dari sistem yang telah dirancang. Hasil menunjukkan untuk task-1 menghasilkan nilai *accuracy* 0.4830 dan task-20 menghasilkan *accuracy* sebesar 0.9360 (Dr. K. Lakshmi, 2020).

Perbandingan antara algoritma *Long Short Term-Memory* dan *simple Recurrent Neural Network* dilakukan oleh Yulius Denny Prabowo dkk. Algoritma diimplementasikan pada kasus chatbot dalam bidang bisnis layanan transportasi. Data pelatihan yang digunakan sebanyak 50 data percakapan kepada pelanggan. Dari hasil penelitian dari 9 pertanyaan yang diajukan, algoritma *simple* RNN menghasilkan 5 (55.6%) jawaban tepat dan algoritma *Istm* menghasilkan 7 (77.8%) jawaban tepat (Prabowo et al., 2018).

Guangjie Zhang dkk juga melakukan penelitian sistem penjawab pertanyaan dengan menggunakan model berbasis CNN attention mechanism dengan menambahkan model LSTM yang digunakan untuk menyelesaikan masalah ketergantungan jangka panjang, korelasi symbol dalam suatu urutan, terutama urutan panjang. Dataset yang digunakan adalah dataset pertanyaan jawaban domain terbuka Microsoft sebanyak



230.324 pasangan jawaban dan pertanyaan dalam Bahasa china dan non china. Hasil yang didapatkan menunjukkan hasil yang baik untuk Bahasa china top-1 accuracy 57.1% dan MRR 70.6%, sedangkan untuk non-china menghasilkan top-1 accuracy 60.1% dan MRR 73.0% (Zhang et al., 2019).

Penelitian yang menggunakan *Convolutional Recurrent Neural Network* (C-RNN) untuk sistem penjawab pertanyaan dilakukan M.M Arefin Zaman dan Sadia Zaman Mishu. Model yang digunakan mengumpan output RNN ke lapisan *convolutional* untuk mendeteksi ketergantungan jangka pendek antara kata yang lebih baik. Hasil yang ditunjukkan pada dataset pertanyaan bAbl mencapai akurasi 74.95% (Zaman and Mishu, 2017).

Oleh karena itu, penelitian ini mengusulkan perancangan perangkat lunak sistem penjawab pertanyaan yang ditanamkan ke dalam robot bantuan edukasi sebagai medianya untuk membantu perkembangan kognitif dan bahasa seorang anak. Penelitian dimulai dengan membuat dataset percakapan tentang pendidikan anak usia dini yang terdiri dari pasangan pertanyaan dan jawaban. Jenis pertanyaan yang diolah merupakan jenis pertanyaan factoid. Sistem *question answering* dibangun menggunakan *Recurrent Neural Network*(RNN) menggunakan mekanisme gerbang *Long-Short Term Memory* (LSTM). Sistem dibangun dengan menggunakan ucapan sebagai masukan system yang ditransformasikan ke dalam bentuk teks menjadi sebuah pertanyaan. Output system berupa jawaban dari pertanyaan yang sudah ditransdormasikan ke dalam bentuk

ucapan. Diharapkan metode-metode yang digunakan akan memberikan hasil yang optimal sehingga robot bantu edukasi dengan fitur sistem *question answering* yang dibangun akan membantu tumbuh kembang anak usia dini.

### **B. Rumusan Masalah**

Berdasarkan latar belakang maka rumusan masalah dari usulan penelitian ini adalah

1. Bagaimana *accuracy*, *precision*, dan *recall* dari algoritma *simple RNN* dan algoritma RNN menggunakan mekanisme gerbang *Long-Short Term Memory (LSTM)* ?
2. Bagaimana ketepatan hasil jawaban dari pertanyaan yang didapatkan menggunakan algoritma *Recurrent Neural Network* dengan mekanisme gerbang *Long-Short Term Memory(LSTM)* ?

### **C. Tujuan Penelitian**

Tujuan yang akan dicapai dari usulan penelitian ini adalah

1. Untuk membandingkan hasil *accuracy*, *precision* dan *recall* dari algoritma *simple RNN* dan algoritma RNN dengan menggunakan mekanisme gerbang *Long-Short Term Memory (LSTM)*.
2. Untuk mengukur tingkat ketepatan jawaban yang dihasilkan dari pertanyaan menggunakan algoritma *Recurrent Neural Network* dengan mekanisme gerbang *Long-Short Term Memory(LSTM)*.

#### **D. Manfaat Penelitian**

Manfaat dari usulan penelitian adalah

1. Membantu kemampuan komunikasi atau berbicara pada anak usia dini
2. Membantu mengembangkan kemampuan kognitif anak usia dini sehingga mempunyai kesiapan yang optimal sebelum memasuki ke jenjang pendidikan selanjutnya.

#### **E. Batasan Masalah**

Adapun batasan-batasan masalah dalam penelitian ini :

1. Domain sistem *question answering* adalah pendidikan anak usia dini
2. Masukan sistem *question answering* adalah pertanyaan dalam bentuk ucapan.
3. Proses pengajuan pertanyaan dengan menggunakan lisan dilakukan pada kondisi sunyi dan tidak bising
4. Algoritma yang digunakan untuk sistem *question answering* adalah *Recurrent Neural Network*(RNN) dengan mekanisme gerbang *Long-Short Term Memory*(LSTM)..
5. Pertanyaan yang dapat diajukan hanya 1 pertanyaan dengan 1 jawaban
6. Menghasilkan jawaban berupa teks yang dikonversi menjadi ucapan

## **F. Sistematika Penulisan**

Adapun sistematika penulisan dalam usulan penelitian ini adalah

### **BAB I Pendahuluan**

Bab I berisikan penjelasan tentang latar belakang penelitian dari fitur robot bantu edukasi sistem *question answering* dengan menggunakan masukan dan keluaran dalam bentuk ucapan menggunakan algoritma yang tepat dan dapat langsung diterapkan sebagai media bantu dalam kegiatan pembelajaran Pendidikan anak usia dini.

### **BAB II Landasan Teori dan Kerangka Pemikiran**

Bab II berisikan tentang tinjauan pustaka terkait landasan teori dari sistem *question answering* dan metode-metode yang digunakan, penelitian terkait dan kerangka pikir penelitian.

### **BAB III Metodologi Penelitian**

Bab III ini merupakan penjelasan tentang tahapan penelitian, bagaimana penerapan algoritma *Recurrent Neural Network* untuk sistem *question answering* dan rancangan sistem.

### **BAB IV Hasil dan pembahasan**

Bab IV menjelaskan tentang hasil dan pembahasan dari penelitian yang dilakukan. Hasil penelitian yang telah dilakukan dengan menggunakan algoritma yang digunakan.

BAB V Kesimpulan dan saran

Pada bab V ini, berisi ringkasan tahapan penelitian, proses penggunaan algoritma yang digunakan dalam sistem *question answering*. Dan berisi saran dari penulis, untuk tahap pengembangan sistem selanjutnya.

## BAB II

### TINJAUAN PUSTAKA

#### A. Landasan Teori

#### 1. Natural Language Processing

##### 1.1 Pengertian

*Natural Language Processing* atau Pemrosesan Bahasa Alami adalah serangkaian teknik komputasi yang termotivasi secara teoritis untuk menganalisis dan mewakili teks yang terjadi secara alami pada satu atau lebih tingkat analisis linguistik untuk tujuan mencapai pemrosesan bahasa yang menyerupai manusia untuk berbagai tugas atau aplikasi (Liddy, 2001).

##### 1.2 Sejarah

Penelitian dalam pemrosesan bahasa alami telah berlangsung selama beberapa dekade sejak tahun 1940-an. *Machine Translation* (MT) adalah aplikasi berbasis komputer pertama yang terkait dengan bahasa alami. Sementara *Weaver* and *Booth* memulai salah satu proyek MT awal pada tahun 1946 pada penerjemahan komputer berdasarkan keahlian dalam memecahkan kode musuh selama Perang Dunia II, umumnya disepakati bahwa itu adalah memorandum *Weaver* tahun 1949 yang membawa gagasan MT ke pemberitahuan umum dan menginspirasi banyak orang. proyek. Dia menyarankan menggunakan ide-ide dari kriptografi dan teori informasi untuk

penerjemahan bahasa. Penelitian dimulai di berbagai lembaga penelitian di Amerika Serikat dalam beberapa tahun.

Pekerjaan awal di MT mengambil pandangan sederhana bahwa satu-satunya perbedaan antara bahasa tinggal dalam kosa kata mereka dan perintah kata yang diijinkan. Sistem yang dikembangkan dari perspektif ini hanya menggunakan kamus pencarian untuk kata-kata yang tepat untuk terjemahan dan mengatur ulang kata-kata setelah terjemahan agar sesuai dengan aturan kata-urutan dari bahasa target, tanpa memperhitungkan ambiguitas leksikal yang melekat dalam bahasa alami. Ini menghasilkan hasil yang buruk. Kegagalan yang nyata membuat para peneliti menyadari bahwa tugas itu jauh lebih sulit daripada yang diantisipasi, dan mereka membutuhkan teori bahasa yang lebih memadai. Namun, tidak sampai 1957 ketika Chomsky menerbitkan *Struktur Sintaksis* memperkenalkan gagasan tata bahasa generatif, apakah bidang tersebut mendapatkan wawasan yang lebih baik mengenai apakah atau bagaimana linguistik arus utama dapat membantu MT.

Selama periode ini, area aplikasi NLP lainnya mulai muncul, seperti pengenalan suara. Komunitas pemrosesan bahasa dan komunitas pidato kemudian dibagi menjadi dua kubu dengan komunitas pemrosesan bahasa yang didominasi oleh perspektif teoritis dari tata bahasa generatif dan bermusuhan dengan metode

statistik, dan komunitas pidato didominasi oleh teori informasi statistik dan linguistik teoritis yang bermusuhan.

Karena perkembangan teori sintaksis bahasa dan parsing algoritma, ada over-antusiasme pada tahun 1950-an yang orang percaya bahwa sistem penerjemahan berkualitas tinggi otomatis akan mampu menghasilkan hasil yang tidak dapat dibedakan dari terjemahan manusia, dan sistem harus beroperasi dalam beberapa tahun. Itu tidak hanya tidak realistis mengingat pengetahuan linguistik dan sistem komputer yang tersedia saat itu, tetapi juga tidak mungkin pada prinsipnya.

Ketidakcukupan sistem yang sudah ada, dan mungkin disertai dengan overenthusiasm, mengarah ke ALPAC (*Automatic Language Processing Advisory Committee of National Academy of Science - National Research Council*) laporan tahun 1966. Laporan menyimpulkan bahwa MT tidak segera dapat dicapai dan direkomendasikan agar tidak didanai. Ini memiliki efek menghentikan MT dan sebagian besar bekerja dalam aplikasi lain NLP setidaknya di Amerika Serikat.

Meskipun ada penurunan besar dalam pekerjaan NLP selama bertahun-tahun setelah laporan ALPAC, ada beberapa perkembangan yang signifikan, baik dalam isu-isu teoritis dan dalam pembangunan sistem prototipe. Dokumen teoritis pada akhir 1960-an dan awal 1970-an difokuskan pada masalah bagaimana



merepresentasikan makna dan mengembangkan solusi komputasi yang dapat dituntaskan bahwa teori tata bahasa yang ada saat itu tidak dapat dihasilkan. Pada tahun 1965, *Chomsky* memperkenalkan model transformasional kompetensi linguistik. Namun, tata bahasa generatif transformasional terlalu berorientasi sintaktis untuk memungkinkan kekhawatiran semantik. Mereka juga tidak memberikan kemudahan dalam melakukan implementasi komputasi. Sebagai reaksi terhadap teori *Chomsky* dan karya generativists transformasional lainnya, *case grammar of Fillmore*, jaringan semantik Quillian, dan teori ketergantungan konseptual ofchank, dikembangkan untuk menjelaskan anomali sintaksis, dan memberikan semantik representasi. Jaringan transisi Augmented Woods, memperluas kekuatan tata bahasa frase-struktur dengan menggabungkan mekanisme dari pemrograman languagesuch sebagai LISP. Formalisme representasi lainnya termasuk semantik kesukaan Wilks, dan tata bahasa fungsional Kay.

Bersamaan dengan pengembangan teoritis, banyak sistem prototipe dikembangkan untuk menunjukkan efektivitas prinsip-prinsip tertentu. Weizenbaum's ELIZA dibangun untuk mereplikasi percakapan antara seorang psikolog dan seorang pasien, hanya dengan mengubah atau menggemakan input pengguna. Winograd's SHRDLU menyimulasikan robot yang memanipulasi blok di atas meja. Meskipun keterbatasannya, itu menunjukkan

bahwa pemahaman bahasa alami memang mungkin untuk komputer. PARRY berusaha untuk mewujudkan teori paranoia dalam suatu sistem. Alih-alih kata kunci tunggal, itu digunakan sekelompok kata kunci, dan digunakan sinonim jika kata kunci tidak ditemukan. LUNAR dikembangkan oleh Woods sebagai sistem antarmuka ke database yang terdiri dari informasi tentang sampel batuan bulan menggunakan jaringan transisi ditambah dan semantik prosedural.

Pada akhir tahun 1970-an, perhatian beralih ke masalah semantik, fenomena wacana, dan tujuan dan rencana komunikatif. Grosz menganalisis dialog yang berorientasi pada tugas dan mengusulkan teori untuk mempartisi diskursus menjadi unit berdasarkan temuannya tentang hubungan antara struktur tugas dan struktur dialog berorientasi tugas. Mann dan Thompson mengembangkan Teori Struktur Retorika, menghubungkan struktur hierarkis dengan wacana. Peneliti lain juga telah memberikan kontribusi signifikan, termasuk Hobbs dan Rosenschein, Polanyi dan Scha, dan Reichman.

Periode ini juga melihat banyak karya pada generasi bahasa alami. Perencana wacana McKeown TEXT dan responsorator McDonald MUMMBLE menggunakan predikat retorik untuk menghasilkan deskripsi deklaratif dalam bentuk teks pendek, biasanya paragraf. Kemampuan TEXT untuk menghasilkan

tanggapan yang koheren secara online dianggap sebagai pencapaian besar.

Pada awal 1980-an, termotivasi oleh ketersediaan sumber daya komputasi kritis, kesadaran yang berkembang di masing-masing komunitas tentang keterbatasan solusi terisolasi untuk masalah NLP, dan dorongan umum menuju aplikasi yang bekerja dengan bahasa di dunia nyata yang luas. konteks, peneliti mulai memeriksa kembali pendekatan non-simbolis yang telah kehilangan popularitas pada hari-hari awal. Pada akhir 1980-an, pendekatan simbolis telah digunakan untuk mengatasi banyak masalah signifikan dalam NLP dan pendekatan statistik terbukti menjadi pelengkap dalam banyak hal pendekatan simbolis.

Dalam sepuluh tahun terakhir milenium, bidang itu berkembang pesat. Hal ini dapat dikaitkan dengan: a) peningkatan ketersediaan teks elektronik dalam jumlah besar; b) ketersediaan komputer dengan peningkatan kecepatan dan memori; dan c) munculnya Internet. Pendekatan statistik berhasil dalam menangani banyak masalah umum dalam linguistik komputasional seperti identifikasi bagian-ucapan, disambiguasi kata rasa, dll, dan telah menjadi standar di seluruh NLP. Peneliti NLP sekarang mengembangkan sistem NLP generasi berikutnya yang berhubungan cukup baik dengan teks umum dan bertanggung jawab atas sebagian besar variabilitas dan ambiguitas bahasa.

### 1.3 Area NLP

Pustejovsky dan Stubbs menjelaskan bahwa ada beberapa area utama penelitian pada field NLP, diantaranya:

- a. Question Answering Systems (QAS). Kemampuan komputer untuk menjawab pertanyaan yang diberikan oleh user. Daripada memasukkan keyword ke dalam browser pencarian, dengan QAS, user bisa langsung bertanya dalam bahasa natural yang digunakannya, baik itu Inggris, Mandarin, ataupun Indonesia.
- b. Summarization. Pembuatan ringkasan dari sekumpulan konten dokumen atau email. Dengan menggunakan aplikasi ini, user bisa dibantu untuk mengkonversikan dokumen teks yang besar ke dalam bentuk slide presentasi.
- c. Machine Translation. Produk yang dihasilkan adalah aplikasi yang dapat memahami bahasa manusia dan menterjemahkannya ke dalam bahasa lain. Termasuk di dalamnya adalah Google Translate yang apabila dicermati semakin membaik dalam penterjemahan bahasa. Contoh lain lagi adalah BabelFish yang menterjemahkan bahasa pada real time.
- d. Speech Recognition. Field ini merupakan cabang ilmu NLP yang cukup sulit. Proses pembangunan model untuk digunakan telpon/komputer dalam mengenali bahasa yang diucapkan

sudah banyak dikerjakan. Bahasa yang sering digunakan adalah berupa pertanyaan dan perintah.

- e. Document classification. Sedangkan aplikasi ini adalah merupakan area penelitian NLP Yang paling sukses. Pekerjaan yang dilakukan aplikasi ini adalah menentukan dimana tempat terbaik dokumen yang baru diinputkan ke dalam sistem. Hal ini sangat berguna pada aplikasi spam filtering, news article classification, dan movie review (“Natural Language Processing,” n.d.)

#### 1.4 Terminologi NLP

Perkembangan NLP menghasilkan kemungkinan dari interface bahasa natural menjadi *knowledge base* dan penterjemahan bahasa natural. Poole dan Mackworth (2010) menjelaskan bahwa ada 3 (tiga) aspek utama pada teori pemahaman mengenai natural language:

- a. *Syntax*: menjelaskan bentuk dari bahasa. Syntax biasa dispesifikasikan oleh sebuah grammar. Natural *language* jauh lebih daripada formal *language* yang digunakan untuk logika kecerdasan buatan dan program komputer
- b. *Semantics*: menjelaskan arti dari kalimat dalam satu bahasa. Meskipun teori semantics secara umum sudah ada, ketika membangun sistem *natural language understanding* untuk

aplikasi tertentu, akan digunakan representasi yang paling sederhana.

- c. *Pragmatics*: menjelaskan bagaimana pernyataan yang ada berhubungan dengan dunia. Untuk memahami bahasa, agen harus mempertimbangan lebih dari hanya sekedar kalimat. Agen harus melihat lebih ke dalam konteks kalimat, keadaan dunia, tujuan dari speaker dan listener, konvensi khusus, dan sejenisnya.

### 1.5 Information Retrieval

*Information Retrieval*(IR) adalah pekerjaan untuk menemukan dokumen yang relevan dengan kebutuhan informasi yang dibutuhkan oleh user. Contoh sistem IR yang paling populer adalah *search engine* pada *World Wide Web*. Seorang pengguna Web bisa menginputkan query berupa kata apapun ke dalam sebuah *search engine* dan melihat hasil dari pencarian yang relevan. Karakteristik dari sebuah sistem IR (Russel & Norvig, 2010) diantaranya adalah:

- a. *A corpus of documents*. Setiap sistem harus memutuskan dokumen yang ada akan diperlakukan sebagai apa. Bisa sebagai sebuah paragraf, halaman, atau teks multipage.
- b. *Queries posed in a query language*. Sebuah *query* menjelaskan tentang apa yang user ingin peroleh. *Query language* dapat

berupa list dari kata-kata, atau bisa juga menspesifikasikan sebuah frase dari kata-kata yang harus berdekatan

- c. *A result set*. Ini adalah bagian dari dokumen yang dinilai oleh sistem IR sebagai yang relevan dengan *query*.
- d. *A presentation of the result set*. Maksud dari bagian ini adalah tampilan list judul dokumen yang sudah di ranking.

## 2. Question Answering System

### 2.1 Pengertian

*Question answering system*(QAS) adalah sebuah sistem yang memungkinkan user menyatakan kebutuhan informasinya dalam bentuk yang lebih spesifik dan alami, yaitu dalam bentuk natural language question, dan tidak mengembalikan daftar dokumen yang harus disaring oleh user untuk menentukan apakah dokumendokumen tersebut mengandung jawaban atas pertanyaan, tetapi mengembalikan kutipan teks singkat atau bahkan frase sebagai jawaban (Monz, 2003).

Bila dilihat dari jenis pertanyaan yang diberikan dan jawaban yang dikembalikan, QAS dapat dibagi menjadi lima jenis [5], yaitu:

- a. Factoid, pertanyaan dengan jawaban berupa sebuah fakta singkat yang biasanya terdiri dari beberapa kata saja.

Contoh: pertanyaan : “siapakah presiden pertama Republik Indonesia?”

Jawaban: "Soekarno"

b. Non factoid, pertanyaan dengan jawaban fakta yang cukup panjang yang biasanya terdiri dari satu klausa atau lebih.

c. Definition, pertanyaan yang menanyakan definisi dari sesuatu.

Contoh: pertanyaan : "apakah ular itu?"

Jawaban : "ular adalah reptil yang tak berkaki dan bertubuh panjang"

d. How, pertanyaan yang menanyakan cara.

Contoh: pertanyaan : "bagaimana cara menggunakan sedotan?"

Jawaban : "cara menggunakan sedotan adalah dengan memasukan sedotan ke air, kemudian hisap dari ujung lubangnya"

e. Why, pertanyaan mengenai penyebab.

Contoh: pertanyaan : "mengapa indonesia miskin?"

Jawaban : "banyak pejabat yang koruptor"

f. Hypothetical, pertanyaan hipotesis yaitu penjelasan yang belum bisa dibuktikan.

Contoh: pertanyaan: "bila seluruh hutan habis menjadi pemukiman manusia, apa yang akan terjadi?"

Jawaban: "aneka ragam flora dan fauna yang berhabitat di hutan tersebut akan punah"



- g. List, pertanyaan dengan jawaban berupa daftar atau beberapa jawaban lebih dari satu.

Contoh: pertanyaan : “apa saja tim yang memenangkan liga champion 3 tahun terakhir?”

Jawaban : “Barcelona, inter milan, Bayern Muenchen”

- h. Yes / No, pertanyaan yang jawabannya terdiri dari dua jawaban “ya” atau “tidak”.

Contoh: pertanyaan : “apakah Soeharto adalah presiden ke 2 Republik Indonesia?”

Jawaban : “ya”

- i. Opinion, pertanyaan yang mengharapkan jawaban berupa pendapat.

Contoh: pertanyaan : “apakah mengerjakan skripsi itu susah?”

Jawaban : “menurut beberapa orang, mengerjakan skripsi itu sangat susah karena banyak yang tidak lulus karena belum sanggup mengerjakannya”

## 2.2 Sejarah

Dua sistem QA awal adalah BASEBALL dan LUNAR. BASEBALL menjawab pertanyaan tentang liga bisbol AS selama satu tahun. LUNAR, pada gilirannya, menjawab pertanyaan tentang analisis geologi batuan yang dikembalikan oleh misi bulan Apollo. Kedua sistem QA sangat efektif dalam domain pilihan mereka.

Faktanya, LUNAR didemonstrasikan pada konvensi ilmu bulan pada tahun 1971 dan mampu menjawab 90% pertanyaan dalam domain yang diajukan oleh orang yang tidak terlatih pada sistem. Selanjutnya sistem QA domain terbatas dikembangkan pada tahun-tahun berikutnya. Fitur umum dari semua sistem ini adalah bahwa mereka memiliki database inti atau sistem pengetahuan yang dituliskan oleh para ahli dari domain yang dipilih. Kemampuan bahasa BASEBALL dan LUNAR menggunakan teknik yang mirip dengan ELIZA dan DOCTOR, program chatterbot pertama.

SHRDLU adalah program penjawab pertanyaan yang sangat sukses yang dikembangkan oleh Terry Winograd di akhir tahun 60an dan awal tahun 70an. Ini simulasi operasi robot di dunia mainan ("blok dunia"), dan itu menawarkan kemungkinan menanyakan pertanyaan robot tentang keadaan dunia. Sekali lagi, kekuatan sistem ini adalah pilihan dari domain yang sangat spesifik dan dunia yang sangat sederhana dengan aturan fisika yang mudah dikodekan dalam program komputer.

Pada 1970-an, basis pengetahuan dikembangkan yang menargetkan domain pengetahuan yang lebih sempit. Sistem QA yang dikembangkan untuk berinteraksi dengan sistem pakar ini menghasilkan lebih banyak tanggapan yang dapat diulang dan valid untuk pertanyaan dalam suatu area pengetahuan. Sistem pakar ini sangat mirip dengan sistem QA modern kecuali dalam arsitektur

internal mereka. Sistem pakar sangat bergantung pada basis pengetahuan yang dibangun oleh ahli dan terorganisir, sedangkan banyak sistem QA modern bergantung pada pengolahan statistik dari korpus teks bahasa yang besar, tidak terstruktur, dan alami.

1970-an dan 1980-an melihat perkembangan teori komprehensif dalam linguistik komputasional, yang mengarah pada pengembangan proyek ambisius dalam pemahaman teks dan penjawaban pertanyaan. Salah satu contoh sistem semacam itu adalah *Unix Consultant*(UC), yang dikembangkan oleh Robert Wilensky di U.C. Berkeley pada akhir 1980-an. Sistem menjawab pertanyaan yang berkaitan dengan sistem operasi Unix. Itu memiliki basis pengetahuan yang dikerjakan secara komprehensif dari domainnya, dan itu bertujuan untuk mengutarakan jawaban untuk mengakomodasi berbagai jenis pengguna. Proyek lain adalah LILOG, sistem pemahaman teks yang beroperasi pada domain informasi pariwisata di kota Jerman. Sistem yang dikembangkan di proyek UC dan LILOG tidak pernah melewati tahap demonstrasi sederhana, tetapi mereka membantu pengembangan teori pada linguistik dan penalaran komputasi.

Baru-baru ini, sistem QA bahasa alami khusus telah dikembangkan, seperti EAGLi untuk ilmuwan kesehatan dan kehidupan.

### 2.3 Skema Question Answering

Pada umumnya sistem QA terdiri atas tiga komponen utama, yaitu *question analyzer*, *passage retriever*, dan *answer finder*.

#### a. Question analyzer

Komponen ini berfungsi untuk menentukan beberapa informasi yang dapat diperoleh dari kalimat pertanyaan masukan, antara lain: tipe jawaban yang diharapkan / EAT (*Expected answer type*) dan kata-kata kunci. Kata-kata kunci yang dihasilkan oleh komponen ini selanjutnya akan digunakan oleh komponen *passage retriever* untuk mengambil dokumen-dokumen yang diperkirakan mengandung jawaban dari pertanyaan.

#### b. Passage retriever

Komponen ini berfungsi untuk menemukan / mengambil dokumen-dokumen (atau bagian dari dokumen) yang kemungkinan mengandung jawaban dari pertanyaan yang diberikan berdasarkan kata kunci yang dihasilkan oleh komponen *question analyzer*. Komponen ini menggunakan metode *information retrieval*(IR) untuk mengambil dokumen-dokumen tersebut.

c. Answer finder

Komponen ini berfungsi untuk menentukan kandidat-kandidat jawaban yang diperoleh dari dokumen-dokumen yang telah ditemukan sebelumnya oleh komponen passage retriever. Biasanya jawaban ditentukan dengan mencocokkan kandidat jawaban dengan EAT yang diperoleh sebelumnya oleh komponen question analyzer.

3. Pendidikan Anak Usia Dini

Pendidikan anak usia dini (PAUD) adalah jenjang pendidikan sebelum jenjang pendidikan dasar yang merupakan suatu upaya pembinaan yang ditujukan bagi anak sejak lahir sampai dengan usia enam tahun yang dilakukan melalui pemberian rangsangan pendidikan untuk membantu pertumbuhan dan perkembangan jasmani dan rohani agar anak memiliki kesiapan dalam memasuki pendidikan lebih lanjut, yang diselenggarakan pada jalur formal, nonformal, dan informal (Pendidikan and No, n.d.).

Pendidikan usia dini merupakan pendidikan yang dilaksanakan sebelum jenjang pendidikan dasar. Pendidikan anak usia dini pada jalur pendidikan formal yaitu Taman Kanak-kanak (TK), Raudatul Athfal (RA), atau bentuk lain yang sederajat.

Ada dua tujuan mengapa perlu diselenggarakan pendidikan anak usia dini, yaitu:

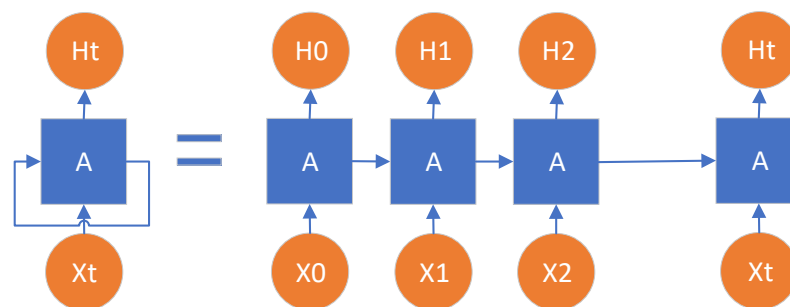
- a. Tujuan utama: untuk membentuk anak Indonesia yang berkualitas, yaitu anak yang tumbuh dan berkembang sesuai dengan tingkat perkembangannya sehingga memiliki kesiapan yang optimal di dalam memasuki pendidikan dasar serta mengarungi kehidupan di masa dewasa.
- b. Tujuan penyerta: untuk membantu menyiapkan anak mencapai kesiapan belajar (akademik) di sekolah.

#### 4. *Recurrent Neural Network*

*Recurrent Neural Network*(RNN) merupakan salah satu kelas dari *deep neural network* yang diawasi. Pelatihan RNN dalam tingkat yang diawasi membutuhkan *dataset* pelatihan dari pasangan input-target dengan tujuan meminimalkan perbedaan nilai *loss* pasangan itu dengan mengoptimalkan bobot jaringan(Salehinejad et al., 2017). RNN dibentuk dari neuron buatan dengan satu atau lebih umpan balik yang berulang. Pada setiap langkah waktu, neuron akan menerima data, melakukan komputasi, dan menghasilkan keluaran. RNN menangkap dinamika yang kaya dari keadaan tersembunyi untuk konteks jangka panjang, sehingga membentuk model yang ekspresif dan sangat kuat untuk tugas-tugas yang sekuens, seperti pengenalan suara, sintesis ucapan, visi mesin, generasi deskripsi video, dan rangkaian teks.

RNN memiliki tiga lapisan yaitu layer *input*, layer tersembunyi yang berulang, dan layer *output*(Salehinejad et al., 2017). Layer input memiliki unit input, terkoneksi penuh ke unit tersembunyi yang ada di

layer tersembunyi. Unit tersembunyi itu terhubung satu sama lain secara berulang. Layer tersembunyi bisa didefinisikan sebagai “memori” atau ruang keadaan yang berdimensi tinggi dengan dinamika non-linier untuk mengingat dan memproses informasi masa lalu. Keadaan tersembunyi akan merangkum semua informasi unik yang diperlukan sebagai keadaan terakhir dari jaringan, melalui serangkaian langkah waktu. Informasi itu lalu terintegrasi, sehingga mampu menentukan perilaku jaringan di masa depan dan melakukan prediksi yang akurat di layer *output*.



Gambar 1. Arsitektur RNN sederhana

## 5. Mekanisme Gerbang RNN

Salah satu kelemahan RNN adalah pembelajaran jangka panjang dengan *gradient descent* bisa menghasilkan masalah menghilang atau meledaknya gradien (*vanishing/explode gradient*). Pada saat pembelajaran, ketika nilai gradien dipropagasikan kembali, nilai itu terus dikalikan dengan bobot yang lebih kecil dari satu ( $<1.0$ ). Oleh karena itu, nilai gradien yang dihasilkan saat iterasi lama kelamaan cenderung mengecil atau mengalami kehilangan nilai. Nilai

gradien secara eksponensial juga bisa membesar, karena nilai gradien dikalikan dengan bobot yang lebih besar dari satu ( $>1.0$ ). Maka perlu mencegah nilai gradien tersebut menjadi angka-angka yang terlalu kecil atau terlalu besar.

Salah satu cara mengatasi masalah menghilang atau meledaknya gradien adalah memodifikasi arsitektur model dengan memasukkan unit gerbang yang dirancang khusus untuk menyimpan informasi selama waktu periode yang lama. Mekanisme gerbang yang paling dikenal saat ini adalah *Long-Short Term Memory* dan *Gated Recurrent Unit*(GRU). LSTM mampu menangani penghafalan dan pengingatan kembali untuk jangka Panjang, khususnya data yang sangat besar. LSTM pada prinsipnya dapat menggunakan unit memorinya untuk mengingat informasi yang jaraknya jauh dan melacak berbagai atribut teks yang sedang diproses.

Persamaan variable dari LSTM yaitu  $X_t$  sebagai input vector ke LSTM,  $F_t$  sebagai vector aktivasi *forget gate*,  $I_t$  sebagai vector aktivasi *input gate*,  $O_t$  sebagai vector aktivasi *output gate*.  $H_t$  sebagai vector *hidden state* atau dikenal sebagai vector keluaran dari unit LSTM,  $C_t$  sebagai vector *cell state*, W-U-b sebagai matriks bobot dan parameter vector bias yang perlu dipelajari selama pelatihan. Bentuk persamaan LSTM yang ringkas adalah sebagai berikut.

$$F_t = \sigma_g(W_f X_t + U_f h_{t-1} + b_f) \quad (1)$$

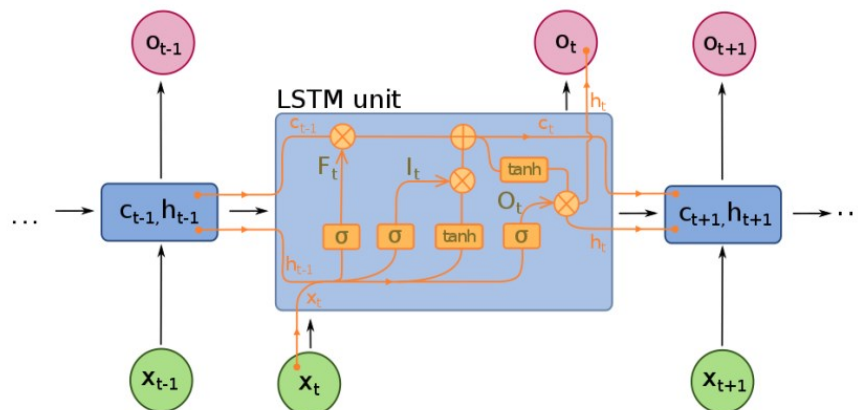
$$I_t = \sigma_g(W_i X_t + U_i h_{t-1} + b_i) \quad (2)$$



$$O_t = \sigma_g(W_o X_t + U_o h_{t-1} + b_o) \quad (3)$$

$$C_t = f_t \circ \sigma_{t-1} + i_t \circ \sigma_g(W_c X_t + U_c h_{t-1} + b_c) \quad (4)$$

$$H_t = O_t \circ \sigma_h(C_t) \quad (5)$$



Gambar 2. Unit LSTM

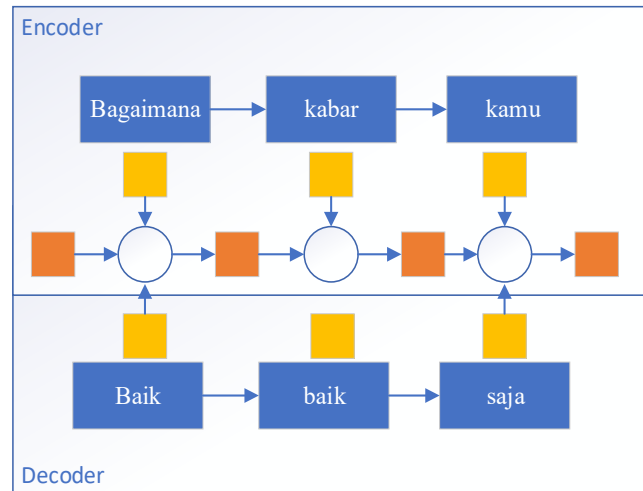
## 6. *Sequential to Sequential*(SEQ2SEQ)

*Recurrent Neural Network*(RNN) sejauh ini berguna untuk pemodelan informasi yang sekuensial seperti pada tugas-tugas Bahasa alami. Model RNN yang populer saat ini adalah *sequential-to-sequential*(SEQ2SEQ). Model seq2seq dibentuk dari dua jaringan RNN, yaitu *encoder* dan *decoder*. Jaringan RNN pada *encoder* melakukan *encoding* untuk sekuens input ke dalam vector konteks dengan ukuran tetap, sedangkan *decoder* menggunakan vector konteks tadi sebagai “benih” untuk menghasilkan sekuens target. Oleh

karena itu, model seq2seq sering juga disebut sebagai model *encoder-decoder*.

Aplikasi di bidang Bahasa alami mulai berkembang karena kemampuan ingatan RNN untuk konteks jangka Panjang. Model percakapan seq2seq mampu memprediksi kata berikutnya dengan diberikan kata atau kalimat konteks sebelumnya dari suatu percakapan. Hal itu akan menghasilkan model generative yang merangkai kata per kata dan membentuk kalimat yang benar-benar baru, dibandingkan dengan model *retrieval* yang hanya mengambil kalimat utuh pada pra konstruksi repositori.

Penelitian arsitektur seq2seq mengalami kemajuan untuk menghasilkan model dengan akurasi yang tinggi. Perkembangan model seq2seq ada yang memakai gabungan dua gerbang yang berbeda (LSTM dan GRU) pada *encoder* model. Ada juga yang memakai mekanisme gerbang yang *multilayer*, contohnya 2 layer LSTM pada *encoder* dan 2 layer LSTM pada *decoder*. Arsitektur sederhana seq2seq bias dilihat pada ilustrasi gambar 3



Gambar 3. Arsitektur *seq2seq*

## B. Penelitian Terkait

Beberapa penelitian terkait tentang sistem question answering berbasis lisan dalam beberapa tahun terakhir.

Penelitian terkait sistem *question answering* dengan memanfaatkan suara sebagai data masukan dilakukan oleh Dessi PL dan Riandy RN pada tahun 2017 mengenai sistem *question answering* berbasis lisan untuk layanan rute kereta dengan menggunakan pendekatan *frame-based*. Ada 3 komponen utama dari sistem yang dirancang yaitu *automatic speech recognizer*(ASR) yang mentranskripsikan masukan ucapan pengguna ke urutan kata-kata kueri, sistem *question answering* yang bertugas menerjemahkan teks yang ditranskripsikan ke dalam kueri dan mencari jawaban yang relevan, dan *automatic speech synthesizer*(ASS) yang bertugas menghasilkan jawaban yang relevan menggunakan ucapan. Untuk sistem *question answering* digunakan beberapa teknik seperti *frame-*

*based*, *keyword spotting*, *data fetching*, dan *canned answer* menghasilkan akurasi hingga 100% (Lestari and Nugraha, 2017).

Waheeb Ahmed dkk pada tahun 2017 juga melakukan penelitian tentang sistem pertanyaan jawaban untuk pertanyaan factoid menggunakan Bahasa arab. Penelitian yang dilakukan dimulai dengan mengumpulkan data-data pertanyaan factoid Bahasa arab yang diambil dari web. Model sistem pertanyaan jawaban yang dibangun terdiri dari question analyzer, knowledge retriever, answer generator, dan external knowledge-base. Keempat model sistem pertanyaan jawaban dibangun menggunakan *Recurrent Neural Network*. Penelitian yang dilakukan menghasilkan akurasi sebesar 53% (Ahmed and Anto, 2017).

Sementara itu, R. Poonguzhali dan K. Lakshmi melakukan penelitian tentang performance dari *Recurrent Neural Network* dalam menangani sistem penjawab pertanyaan. Dataset yang digunakan dalam penelitian yang dilakukan adalah bAbl. bAbl mempunyai 20 *task* dan penelitian ini menggunakan task-1 (*Basic factoid QA with single supporting fact*) dan task-20 (*Reasoning about agent's motivation*) untuk menguji *accuracy* dari sistem yang telah dirancang. Hasil menunjukkan untuk task-1 menghasilkan nilai *accuracy* 0.4830 dan task-20 menghasilkan *accuracy* sebesar 0.9360 (Dr. K. Lakshmi, 2020).

Perbandingan antara algoritma *Long Short Term-Memory* dan *simple Recurrent Neural Network* dilakukan oleh Yulius Denny Prabowo dkk. Algoritma diimplementasikan pada kasus chatbot dalam bidang bisnis

layanan transportasi. Data pelatihan yang digunakan sebanyak 50 data percakapan kepada pelanggan. Dari hasil penelitian dari 9 pertanyaan yang diajukan, algoritma *simple* RNN menghasilkan 5 (55.6%) jawaban tepat dan algoritma *Istm* menghasilkan 7 (77.8%) jawaban tepat (Prabowo et al., 2018).

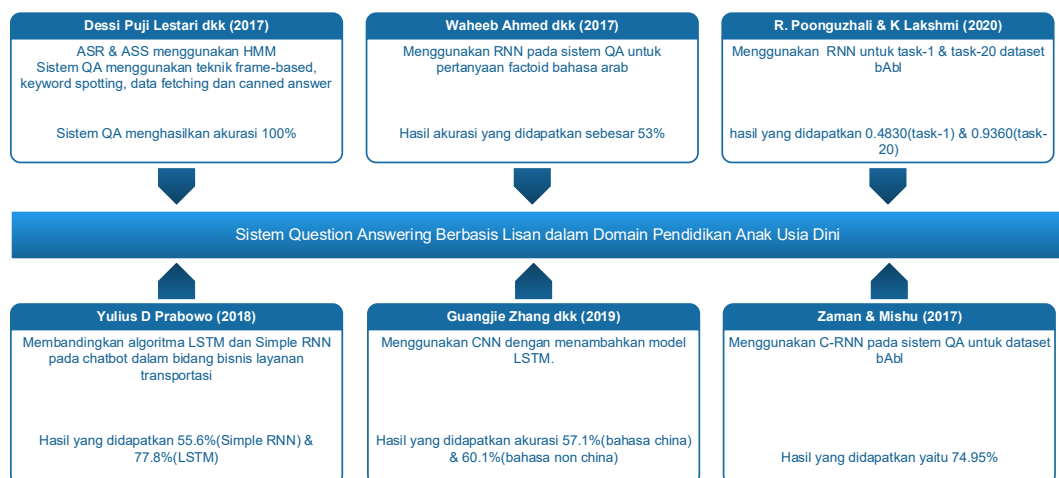
Guangjie Zhang dkk juga melakukan penelitian sistem penjawab pertanyaan dengan menggunakan model berbasis CNN attention mechanism dengan menambahkan model LSTM yang digunakan untuk menyelesaikan masalah ketergantungan jangka panjang, korelasi symbol dalam suatu urutan, terutama urutan panjang. Dataset yang digunakan adalah dataset pertanyaan jawaban domain terbuka Microsoft sebanyak 230.324 pasangan jawaban dan pertanyaan dalam Bahasa china dan non china. Hasil yang didapatkan menunjukkan hasil yang baik untuk Bahasa china top-1 accuracy 57.1% dan MRR 70.6%, sedangkan untuk non-china menghasilkan top-1 accuracy 60.1% dan MRR 73.0% (Zhang et al., 2019).

Penelitian yang menggunakan *Convolutional Recurrent Neural Network* (C-RNN) untuk sistem penjawab pertanyaan dilakukan M.M Arefin Zaman dan Sadia Zaman Mishu. Model yang digunakan mengumpan output RNN ke lapisan *convolutional* untuk mendeteksi ketergantungan jangka pendek antara kata yang lebih baik. Hasil yang ditunjukkan pada dataset pertanyaan bAbI mencapai akurasi 74.95% (Zaman and Mishu, 2017).

### C. State of the art

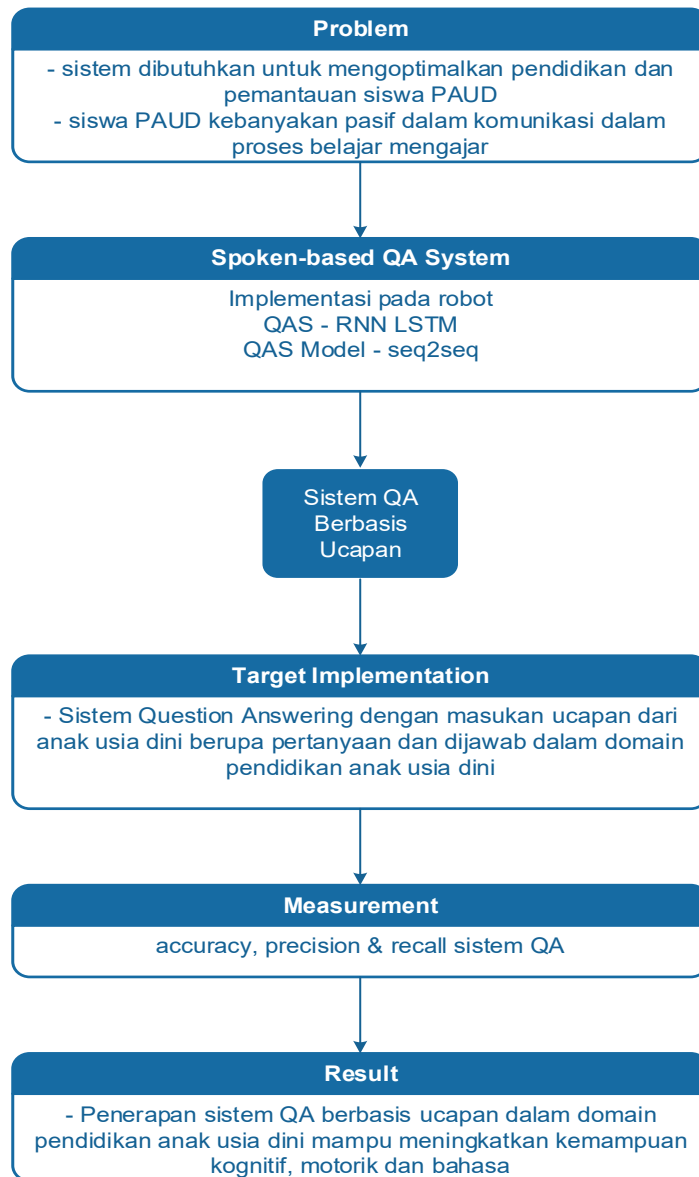
Pada gambar 4 menampilkan penelitian-penelitian yang telah dilakukan sebelumnya mengenai sistem *question answering*. Dessi PL dan Riandy RN pada tahun 2017 melakukan penelitian sistem *question answering* berbasis lisan untuk layanan rute kereta dengan menggunakan pendekatan *frame-based*. Untuk sistem *question answering* digunakan beberapa teknik seperti *frame-based*, *keyword spotting*, *data fetching*, dan *canned answer* menghasilkan akurasi hingga 100% (Lestari and Nugraha, 2017). Waheeb Ahmed dkk pada tahun 2017 juga melakukan penelitian tentang sistem pertanyaan jawaban untuk pertanyaan factoid Bahasa arab dengan menggunakan algoritma RNN dan menghasilkan akurasi sebesar 53% (Ahmed and Anto, 2017). Sementara itu, R. Poonguzhali dan K. Lakshmi melakukan penelitian tentang performance dari algoritma *Recurrent Neural Network* dalam menangani sistem penjawab pertanyaan dengan hasil yang menunjukkan untuk task-1 menghasilkan nilai *accuracy* 0.4830 dan task-20 menghasilkan *accuracy* sebesar 0.9360 (Dr. K. Lakshmi, 2020). Perbandingan antara algoritma *Long Short Term-Memory* dan *simple Recurrent Neural Network* dilakukan oleh Yulius Denny Prabowo dkk. Algoritma diimplementasikan pada kasus chatbot dalam bidang bisnis layanan transportasi yang menghasilkan 5 (55.6%) jawaban tepat dan algoritma *Istm* menghasilkan 7 (77.8%) jawaban tepat dari 9 pertanyaan yang diajukan (Prabowo et al., 2018). Guangjie Zhang dkk juga melakukan penelitian sistem penjawab pertanyaan dengan menggunakan

model berbasis CNN attention mechanism dengan menambahkan model LSTM yang digunakan untuk menyelesaikan masalah ketergantungan jangka panjang, korelasi symbol dalam suatu urutan, terutama urutan panjang. Hasil yang didapatkan menunjukkan hasil yang baik untuk Bahasa china top-1 accuracy 57.1% dan MRR 70.6%, sedangkan untuk non-china menghasilkan top-1 accuracy 60.1% dan MRR 73.0% (Zhang et al., 2019). Penelitian yang menggunakan algoritma convolutional Recurrent Neural Network (C-RNN) untuk sistem penjawab pertanyaan dilakukan M.M Arefin Zaman dan Sadia Zaman Mishu. Model yang digunakan mengumpukan output RNN ke lapisan convolutional untuk mendeteksi ketergantungan jangka pendek antara kata yang lebih baik. Hasil yang ditunjukkan pada dataset pertanyaan bAbI mencapai akurasi 74.95% (Zaman and Mishu, 2017). Oleh karena itu, diusulkan penelitian untuk dapat memberi masukan pertanyaan lebih beragam tetapi jawaban tetap pada satu domain.



Gambar 4. *State Of The Art*

## D. Kerangka Pikir



Gambar 5. Kerangka Pikir

## E. KAMUS KATA

Berikut merupakan kamus kata untuk kata akronim yang berada pada tesis ini :



- **RNN** (*Recurrent Neural Network*) atau jaringan saraf berulang merupakan jenis arsitektur jaringan saraf tiruan yang pemrosesannya dipanggil berulang-ulang untuk memproses masukan yang biasanya adalah data sekuensial.
- **LSTM** (*Long Short Term-Memory*) merupakan salah satu jenis dari *Recurrent Neural Network*, dimana dilakukan modifikasi pada RNN dengan menambahkan *memory cell* yang dapat menyimpan informasi untuk jangka waktu yang lama.
- **Seq2Seq** (*Sequence to Sequence*) merupakan metode penerjemahan mesin berbasis encoder-decoder yang memetakan input sekuens ke output sekuens dengan tag dan *attention value*.
- **Pertanyaan Factoid** merupakan pertanyaan dengan jawaban berupa sebuah fakta singkat yang biasanya terdiri dari beberapa kata saja.
- **System QA** (*System Question Answering*) merupakan sistem yang menerima masukan berupa pertanyaan dan menghasilkan keluaran berupa jawaban
- **Speech Recognition** atau yang biasa disebut dengan speech to text merupakan suatu pengembangan teknik dan sistem yang memungkinkan komputer untuk menerima masukan berupa kata yang diucapkan.
- **Text to Speech** merupakan sistem yang dapat melakukan konversi dari teks menjadi ucapan.

- **Preprocessing** merupakan proses perubahan bentuk data yang belum terstruktur menjadi data yang terstruktur sesuai dengan kebutuhan, untuk proses mining yang lebih lanjut.
- **Tokenizing** merupakan proses pemotongan string masukan berdasarkan tiap kata yang menyusunnya.
- **Encode Sequence** merupakan proses penyandian kata-kata ke dalam bentuk bilangan integer (kamus kata)
- **Accuracy** merupakan tingkat kedekatan antara nilai prediksi dengan nilai aktual
- **Precision** merupakan tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem
- **Recall** merupakan tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi
- **Encode Output** merupakan tahapan mengubah vector kelas integer dari target ke bentuk metrik kelas biner
- **Prediction sequences** merupakan tahapan untuk memprediksi dan menghasilkan kalimat target dengan diberikan konteks yang telah disandikan dalam bentuk sekuens.
- **Vanishing Gradient** merupakan keadaan pada Neural Network yang memungkinkan *gradient* mengecil saat bobotnya kecil, karena perkalian matriks yang berulang kali saat propagasi balik.

- ***Exploding Gradient*** merupakan keadaan Neural Network yang memungkinkan nilai gradient meledak karena perkalian bobotnya menjadi lebih besar dari batas norma gradient saat propagasi balik.
- **Generasi Respon** merupakan tahapan untuk menguji sistem dengan memasukkan data pertanyaan yang akan dikeluarkan dalam bentuk pertanyaan secara aktual.