# BAB II KAJIAN PUSTAKA

## Pengertian *Natural Language*

*Natural Language* atau bahasa alami adalah suatu bahasa yang diucapkan, ditulis, atau diisyaratkan (secara visual atau isyarat lain) oleh manusia untuk berkomunikasi. Singkatnya *Natural Language* adalah bahasa yang sering kita gunakan untuk berkomunkasi sehari-hari, seperti bahasa Indonesia, bahasa Inggris, bahasa isyarat dan bahasa lainnya sesuai letak geografisnya (Putra *et al.*, 2017). Bahasa alami menjadi topik yang hangat diperbincangkan akhir-akhir ini, banyak peneliti yang berlomba-lomba untuk menciptakan teknologi sehingga interaksi manusia dan komputer menjadi lebih mudah lagi, salah satu penerapan teknologi adalah *Natural Language Processing* (NLP). Karena erat sekali dengan kaidah atau aturan, maka setidaknya ada tiga aspek utama pada *Natural Langugage*, yaitu:

1. Sintaks: menjelaskan bentuk atau struktur dari sebuah bahasa. Sintaks biasa direpresentasikan oleh sebuah *grammar* atau tata bahasa. Sebagai contoh, untuk membentuk sebuah kalimat yang valid dalam bahasa kita memakai struktur: Subjek + Predikat + Objek. *Natural language* jauh melebii daripada *formal language* yang digunakan untuk logika kecerdasan buatan dan program komputer
2. Semantik: menggambarkan hubungan antara sintaks dan model komputasi. Meskipun teori Semantik secara umum sudah ada, ketika membangun sistem *natural language understanding* untuk aplikasi tertentu, akan digunakan representasi yang paling sederhana.
3. *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.

## Pengertian *Natural Language Processing*

Ketika seseorang melihat atau membaca sebuat tulisan, orang tersebut akan menggunakan seluruh pengetahuan dan wawasan yang ia miliki untuk memahami tulisan tersebut. Tidak hanya sebatas tata bahasa atau *grammar*, namun lebih dari itu manusia akan mengolah informasi yang ia dapatkan dan menganalisa substansi atau konteks dari tulisan tersebut sehingga didapatkanlah pengetahuan yang baru. Maka, dikembangkan sebuah teknologi Artificial Intelligence (AI) yang berfokus pada bahasa alami, yaitu *Natural Language Processing* (NLP). Menurut Liddy (2001) NLP adalah teknik-teknik komputasi yang didorong secara teoritis untuk menganalisa dan merepresentasikan bahasa alami pada tingkat analisis linguistik untuk mencapai pemrosesan bahasa seperti manusia. NLP hadir sebagai teknologi agar komputer dapat memahami dan memproses bahasa alami tanpa menghilangkan makna yang terkandung di dalamnya. Selain itu, NLP hadir untuk mempermudah interaksi antara manusia dan komputer. Banyak aplikasi yang menerapkan prinsip-prinsip NLP, salah satunya adalah fitur “Oke, Google” pada smartphone, pengguna kini dapat melakukan perintah-perintah seperti memutar lagu, menyetel alarm, hingga melakukan navigasi hanya cukup dengan perintah suara. Contoh lainnya adalah sentimen analisis, dimana komputer dapat menentukan sentimen dari sebuah respon yang diberikan, apakah itu berupa respon yang positif, negatif, ataupun netral. Contoh lainnya adalah aplikasi penerjemah, dimana komputer dapat menerjemahkan suatu bahasa kedalam bahasa lainnya secara otomatis.

*Natural Language Processing* (NLP) mempelajari tentang bagaimana memproses dan mengubah bahasa alami kedalam bentuk yang lebih sederhana dan terstruktur (simbol atau numerik) sehingga proses komputasi menjadi lebih mudah. Lebih dari itu, NLP bertujuan agar informasi yang kita dapatkan dari bahasa alami didapatkan juga oleh komputer, atau dengan kata lain komputer dapat memhami bahasa alami kita dengan memperhatikan batasan sintaks, semantik, gramatikal dan konteks.

Liddy (2001) mengungkapkan, setidaknya ada enam istilah yang sering digunakan dalam NLP:

1. *Part-of-speech tagging*:Sangatlah sulit untuk menandai istilah-istilah dalam suatu teks yang terkait dengan bagian tertentu dari suatu naskah (misalnya kata benda, kata kerja, kata sifat atau kata keterangan), karena bagian dari naskah tidak hanya bergantung pada definisi istilah tetapi juga pada konteks dimana teks digunakan.
2. *Text segmentation*: Beberapa bahasa tulisan, seperti bahasa mandarin, jepang, dan thai, tidak memiliki batasan kata. Dalam contoh ini, tugas *text-parsing* memerlukan idetifikasi terhadap batasan kata, yang seringkali merupakan tugas yang sangat sulit. Tantangan serupa dalam segmentasi naskah muncul ketika menganalisa bahasa verbal, karena suara menyajikan rangkaian huruf dan kata yang bercampur satu sama lain.
3. *Word sense disambiguation*:Banyak kata yang memiliki lebih dari satu arti. Memilih arti yang paling masuk akal hanya bisa dicapai dengan mempertimbangkan konteks di mana kata digunakan.
4. *Syntactic ambiguity*:Tata bahasa dalam bahasa alami seringkali ambigu yang artinya, ada berbagai struktur kalimat yang memungkinakan yang perlu dipertimbangkan. Memilih struktur yang paling tepat biasanya memerlukan paduan informasi kontekstual dan semantik.
5. *Imperfect or irregular input*:Aksen asing atau lokal dan berbagai hambatan vokal dalam pidato dan kesalahan ketik dan tata bahasa dalam teks-teks menyebabkan pengolahan bahasa bahkan lebih sulit.
6. *Speech acts*:Suatu kalimat seringkali dianggap sebagai suatu aksi oleh si pembicara. Struktur kalimatnya sendiri mungkin tidak berisi cukup informasi untuk mendefinisikan tindakan ini.

Liddy (2001) juga mengungkapkan bahwa penerapan-penerapan NLP seringkali tidak akan terlepas dari bidang-bidang berikut ini:

1. *Information Retrieval*: Ilmu untuk melakukan pencarian terhadap berbagai dokumen yang relevan, menemukan informasi tertentu didalamnya, dan menghasilkan metadata untuk isinya.
2. *Information Extraction* (IE): Sejenis ‘*information retrieval’* yang tujuannya adalah untuk mengekstrak secara otomatis informasi terstruktur, seperti data yang sudah terdefinisi dengan baik secara semantik dan secara kontekstual yang sudah terkelompok dari domain tertentu, dengan menggunakan berbagai dokumen tak-terstruktur yang bisa terbaca oleh mesin.
3. *Question-Answering*: Pekerjaan menjawab secara otomatis suatu pertanyaan yang diajukan dalam bahasa alami; yaitu, menghasilkan jawaban bahasa manusia ketika diberi pertanyaan bahasa manusia. Untuk mendapatkan jawaban terhadap pertanyaan, program computer bisa menggunakan baik database pra-terstruktur atau kumpulan dokumen bahasa alami (suatu ‘*text corpus*’ seperti *world wide web*).
4. *Summarization*: Tingkat NLP yang lebih lanjut, menghasilkan sebuah rangkuman singkat dari sebuah dokumen (dengan teks dalam jumlah besar) dengan waktu yang cepat.
5. *Machine Translation*: Penerjemahan otomatis dari satu bahasa manusia ke bahasa manusia yang lain.
6. *Dialogue Systems*: Berkomunikasi dengan komputer layaknya dengan manusia, yang memungkinkan dikembangkan lebih lanjut mengingat potensi yang dimiliki sistem ini sangat besar.
7. *Speech Recognition*: Mengubah kata-kata verbal menjadi *input* yang bisa terbaca oleh mesin. Dengan adanya sound clip dari orang yang sedang berbicara, sistem menghasilkan dikte dari teks.
8. *Natural Language Generation*: Sistem mengubah informasi dari database komputer (simbolik atau numerik) menjadi bahasa manusia yang bisa dibaca.
9. *Natural Language Understanding*: Sistem mengekstrak informasi yang mewakili makna dari suatu sumber teks (dokumen atau rangkuman).
10. *Speech Synthesis*: Perangkat yang mampu berbicara atau membaca teks.

## Pengertian *Natural Language Generation*

Berbeda dengan *Natural Language Processing* (NLP), *Natural Language Generation* (NLG) merupakan sistem yang mampu membangkitkan bahasa alami sebagai keluaran. NLG dapat diartikan juga sebagai proses penyusunan teks bahasa alami untuk memenuhi tujuan komunikatif tertentu. Sedangkan Bateman & Zock (2012) menuturkan bahwa NLG adalah sistem yang mampu menghasilkan informasi dalam bentuk text (linguistic) dengan berdasarkan data non-linguistic (data raw atau mentah yang terukur atau berasal dari serangkaian kejadian) agar mudah dipahami oleh manusia. Selain itu, hal yang paling membedakan antara NLP dan NLG adalah pemilihan informasi untuk keluaran, dimana sistem NLG harus memilih beberapa pilihan berupa teks yang akan disampaikan, singkatnya jika terdapat pemilihan informasi untuk keluaran berupa teks maka sistem tersebut lebih mendekati ciri-ciri sistem NLG dibandingkankan dengan NLP (Reiter, 2010). Seringkali, pemilihan teks keluaran dapat mempermudah proses pada sistem NLG, contohnya ada pilihan teks keluaran sebagai berikut:

1. Saya bertemu dengan Lira di jalan, dan pergi ke sekolah bersamanya.
2. Saya bertemu dengan Lira di jalan, dan pergi ke sekolah bersama Lira.

Kedua pilihan tersebut jika dilihat secara makna, tentu memiliki makna yang sama. Tetapi secara kasat mata tentu kita akan memilih pilihan a, dikarenakan lebih sederhana dan tidak terkesan kaku. Namun dalam perspektif komputer, pilihan a akan lebih lama pemrosesannya dibandingkan pilihan b. Karena, komputer harus mendefinisikan konteks dari kata bersamanya pada pilihan a, tentu sistem harus memiliki pemahaman untuk mendefinisikan konteks dari kata tersebut. Berbeda dengan pilihan b, sistem akan lebih mudah mendefinisikan bahwa subjek pulang pulang bersama Lira.

Maka tak heran, salah satu indikator kesuksesan dalam membangun sistem NLG adalah adanya pengetahuan teknik (menguraikan, merepresentasikan, dan mengatur proses informasi dari input) serta pengetahuan tentang keadaan dan kendala dari pengguna (prosesor informasi yang diterima) (Bateman & Zock, 2012). Bateman & Zock (2012) juga mengungkapkan bahwa setidaknya ada tiga definisi NLG, yaitu:

1. NLG sebagai *mapping problem*
2. NLG sebagai *problem of choice*
3. NLG sebagai *planning problem*

Dimana setiap definisi tersebut mewakili setiap layer dalam arsitektur utamanya. Berikut adalah penjelasan arsitektur utama dalam sistem NLG menurut (Bateman & Zock, 2012):

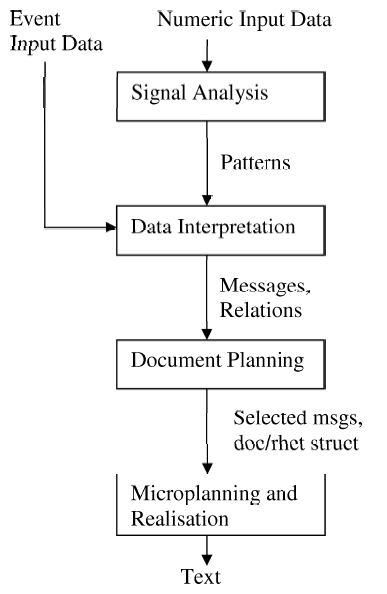
1. *Macroplanning:*  Penentuan konten, tujuan, dan *knowledge source* dari data mentah, lalu mengorganisir dan melakukan perencaan untuk membangun sebuah teks keluaran.
2. *Microplanning:* Mendeskripsikan kejadian-kejadian lalu mengelompokkan materi yang berkaitan sehingga dapat membangun informasi yang lebih terintegrasi dan ringkas.
3. *Surface realization:* Pengkontruksian gramatikal yang akan dipilih (pemilihan sintaksis), penandaan *Part-of-Speech* (POS), penambahan atribut-atribut seperti preposisi, dan penentuan bentuk akhir sehingga menjadi kata-kata yang tersusun (morfologi).
4. *Physical presentation:* Proses penambahan artikulasi, pungtuasi, dan layout mana yang akan digunakan.

## Arsitektur sistem *Data-to-Text*

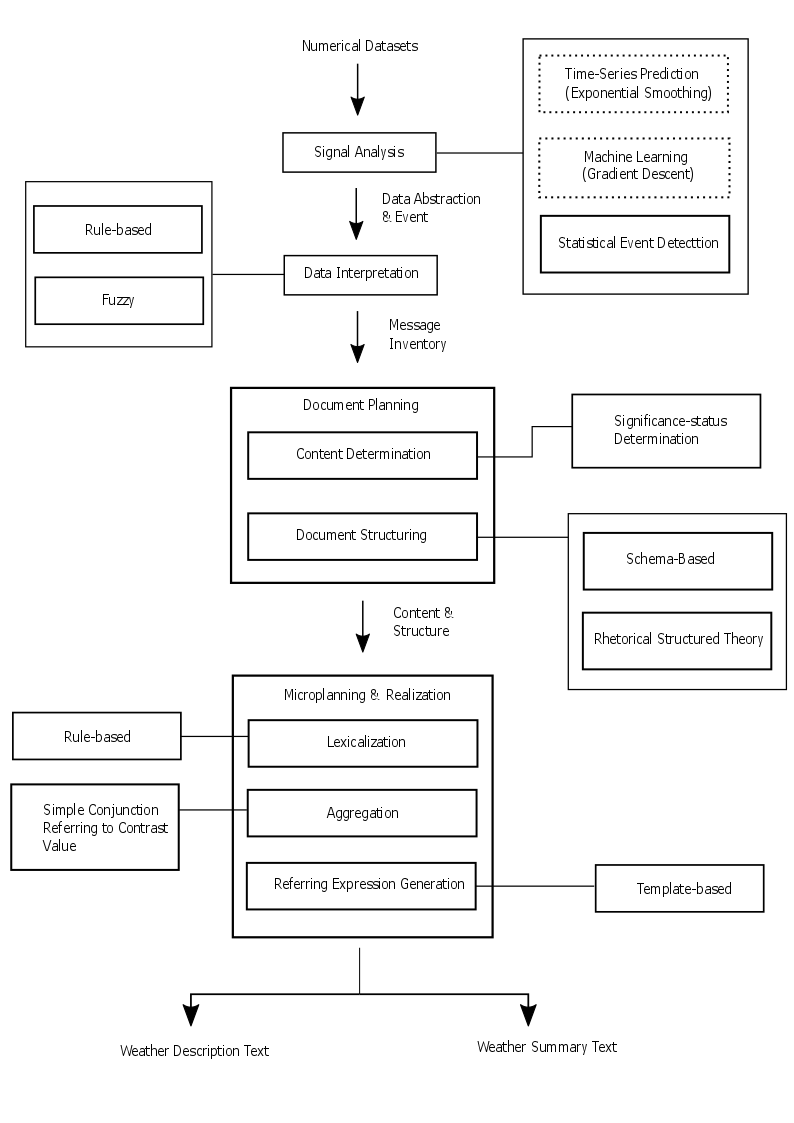
*Data-to-text* (D2T) adalah sistem *Natural Language Generation* (NLG) yang mampu menghasilkan teks dari input data *non-linguistic*, seperti data sensor dan serangkaian kejadian (Reiter, 2011). Seperti namanya, sistem D2T mengkonversikan data raw (data mentah) dari sensor baik berupa data numerik, data log, ataupun data ­ *non-linguistic* menjadi sebuah teks yang mudah dipahami oleh pembaca data memuat informasi serta *knowledge* sesuai dengan data masukannya.

Berbagai pengaplikasian D2T sudah banyak dgunakan dalam berbagai bidang, seperti pada bidang peramalan cuaca misalnya, Putra *et al.* (2017) membuat sistem bernama *Data-to-text Weather Prediction* (DWP) yang mampu menghasilkan ringkasan berita klimatologis dan cuaca selama satu bulan serta memberikan informasi prediksi untuk satu hari berikutnya. Setelah itu, sistem DWP dikembangkan lagi oleh Abidin *et al.* ( 2018) , sehingga sistem D2T yang dibangun mampu menerima masukan berupa data *streaming.* Pada bidang lain, seperti bidang kesehatan dibangun sistem seperti *BabyTalk* yang mampu menghasilkan ringkasan teks dari data neonatal selama 45 menit kemudian ringkasan tersebut digunakan sebagai bahan pendukung keputusan presentasi modalitas yang terjadi saat itu (Gatt et al., 2009), selain itu terdapat sistem BT-Nurse yang mampu meringkas kejadian selama pergantian shift keperawatan berlangsung, berdasarkan hasil rekaman medis elektronik pasien (Hunter et al., 2011). Pada bilang ekonomi, terdapat *Knowledge-Based Report Generator* yang mampu menghasilkan laporan stok berdasarkan data stok produk (*non-linguistic*) suatu pasar (Kukich, 1983). Beberapa contoh tersebut membuktikan bahwa D2T menjadi pilihan yang tepat dalam berbagai bidang.

Seperti yang digambarkan pada Gambar 2.1, Reiter (2011) memaparkan bahwa setidaknya ada empat elemen utama yang diperlukan untuk membangun sebuah sisten D2T, yaitu *Signal Analysis*, *Data Interpretation*, *Document Planning*, dan *Microplanning and Realisation*. Elemen-elemen ini hampir mirip dengan arsitektur sistem NLG yang sudah dijelaskan pada sub-bab sebelumnya. Reiter (2011) memaparkan bahwasannya, perbedaan terbesar antara sistem D2T dan sistem NLG adalah sistem D2T harus menganalisis dan menginterpretasikan data masukannya, begitu juga dengan menentukan bagaimana proses penyampaiannya dari segi linguistik.



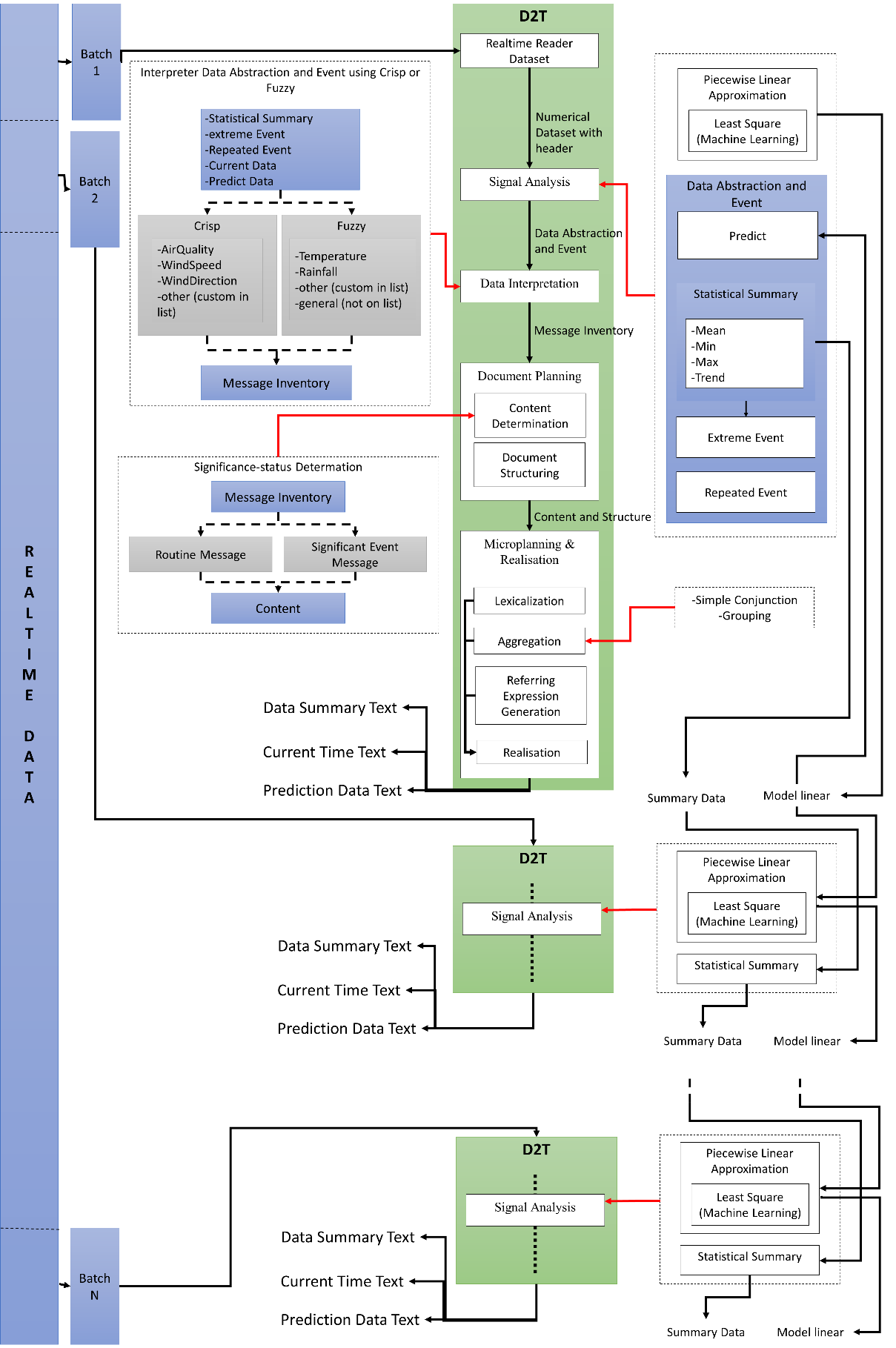
Gambar 2.1 Arsitektur sistem *Data-to-Text* (Reiter, 2011)



Gambar 2.2 Arsitektur sistem DWP (Putra *et al.*, 2017)

Dalam penelitian DWP, dilakukan pengembangan model dengan memodifikasi beberapa elemen, sehingga output yang dihasilkan dibagi menjadi dua kategori berupa *Weather Description Text* dan *Weather Summary Text*. Gambar 2.2 menjelaskan bagaimana sistem DWP yang dikembangkan oleh Putra *et al.*, (2017), dimana ada beberapa penambahanan yaitu penggunaan *Machine Learning* yaitu *Gradient Descent* dan pendekatan *Time Series* yaitu *Exponential Smoothing*, dan juga terdapat pendeteksian *Statistical Event*. Proses *Signal Analysis* menghasilkan *Data Abstraction and Event,* yang menjadi masukan bagi proses *Data Interpretation.* Proses yang diklaim oleh Reiter (2011) sebagai pembeda antara sistem D2T dan sistem NLG ini, merupakan proses untuk menginterpretasikan data masukan. DWP menggunakan 2 cara untuk menginterpretasikan datanya, yang pertama yaitu *Rule-Based* dan *Fuzzy* sehingga dihasilkanlah *Message Inventory* yang akan diproses pada proses selanjutnya yaitu *Document Planning.* Pada proses *Document Planning*, terdapat *Content Determination* dan *Document Structuring*. *Content Determination* atau pemilihan konten dilakukan dengan memilah seberapa pentingnya status suatu pesan atau nama lainnya yaitu *Significance Status Determination.* Setelah ditentukan seberapa pentingnya status suatu pesan, maka pesan-pesan tersebut digabungkan pada sebuah proses bernama *Document Structuring*. Setelah ditentukan statusnya, maka pesan tersebut diproses pada proses *Document Planning,* sehingga dihasilkan *Content* dan *Structure*, yang nantinya akan diproses pada tahap *Microplanning* dan *Realization,* sehingga dihasilkanlah sebuah teks yang terdiri dari *Weather Description Text* dan *Weather Summary Text.*

Pada sistem D2T yang dikembangkan oleh (Abidin *et al.*, 2018) . Diperkenalkan sebuah sistem yang dapat menerima masukan berupa data *streaming*. Seperti yang bisa kita lihat pada gambar 2.3 terdapat perbedaan pada tahap *Signal Analysis* dan terdapatnya proses bernama *Realtime Reader Dataset,* dimana data yang masuk akan terlebih dahulu dihandle pada tahap *Realtime Reader Dataset* sebelum diproses pada tahap *Signal Analysis*. Setelah data masukan diproses pada tahap *Realtime Reader Dataset*, data masukan selanjutnya akan diproses pada tahap *Signal Analysis,* dimana data *streaming* akan diproses dan dipecah berdasarkan *batch-batch*. Sehingga proses prediksi data dilakukan menggunakan teknik Piecewise Linear Approximation (PLA) dengan metode Least Square yang diperkenalkan oleh Palpanas *et al.*, (2004). Abidin *et al.*, (2018) menuturkan bahwa pada proses ini dihasilkan model linear dari setiap batch data dan model linear untuk keseluruhan data (batch pertama hingga batch ke-N), hal ini dilakukan karena data yang digunakan merupakan data Time Series sehingga antara batch satu dan lainnya memiliki keterikatan. Penerapan PLA dilakukan untuk merepresentasikan data streaming lampau dan menggabungkannya menjadi sebuah garis linear, mengingat data yang digunakan hanya dalam satu kali proses dan selanjutnya data terhapus untuk mengatasi masalah storage berlebihan dikarenakan kemunculan data sangat cepat dan memiliki ukuran yang tidak sedikit bila kita menggabungkan setiap data stream.



Gambar 2.3 Arsitektur D2T untuk data *streaming* (Abidin *et al.*, 2018)

1. *Signal Analysis*

Tahapan pertama dalam membangun sistem D2T adalah *Signal Analysis.* Tahapan ini bertujuan untuk mendeteksi pola-pola yang terdapat pada data masukan. Sehingga sistem dapat memproses data dengan bentuk simbolis (diskrit) dibandingkan memprosesnya dalam bentuk numerik. Namun, dalam beberapa kasus tahapan ini menjadi opsional, ketika data masukan sudah terstruktur dalam bentuk kejadian-kejadian diskrit, contohnya seperti data rekord kejadian medis, atau *log file*. Jika data masukan sudah berbentuk simbolis (diskrit) sepeti yang sudah dijelaskan sebelumnya, maka proses *Signal* *Analysis* ini tidak perlu dilakukan (McKeown *et al.*, 1994).

Pada dasarnya tahap *Signal Analysis* merupakan proses untuk menganalisa data masukan yang berupa numerik sehingga dihasilkan sebuah informasi berbentuk simbolis yang akan disampaikan. Contohnya untuk kasus pembangkitan berita cuaca pada DWP, data masukan berupa seluruh data klimatologi dan kualitas udara selama satu tahun, lalu data tersebut dianalisa dan dihasilkanlah sinyal-sinyal dari data tersebut seperti curah hujan terbesar, kondisi kualitas udara hari ini, serta bagaimana kondisi curah hujan dan temperatur bulan ini dibandingkan dengan curah hujan dan temperatur satu tahun penuh, dan lain-lain. Pada penmbangunan sistem DWP (Putra *et al.*, 2017) , dilakukan pendekatan *Time Series* dan pendekatan *Statistical Tools* dalam menganalisa data masukan sehingga dihasilkan sebuah pesan diskrit untuk satu bulan. Contoh penggunaan statistik (*mean, min, max, sum*) dalam bahasa R digambarkan pada pada gambar 2.4 di bawah ini.

|  |
| --- |
| ...  **for(**i **in** i**:**n**){** #Loping sebanyak jumlah parameter  max\_amt**[**i**]** **<-** max**(**LM**[**i**])** #Nilai Max pd parameter ke-i  min\_amt**[**i**]** **<-** min**(**LM**[**i**])** #Nilai Min pd parameter ke-i  sum\_amt**[**i**]** **<-** sum**(**LM**[**,i**])** #Total nilai dari parameter ke-i  ....  **}**  #Rata-rata 1 bulan terakhir untuk setiap parameter  LMmean\_result **<-** colMeans**(**xLM**)**  .... |

Gambar 2.4 Contoh implementasi *Signal Analysis* (Putra et al., 2017)

1. *Data Interpretation*