Natural Language Generation

An Overview, Some Explorations

Overview

- An Introduction to Natural Language Generation

- Specific Research Themes
  - Machine Learning for NLG
  - Soft Computing for Lexicalization

Context: Collaborative work with Accenture Labs

What is NLG?

Natural language generation is the process of deliberately constructing a natural language text in order to meet specified communicative goals.

[McDonald 1992]

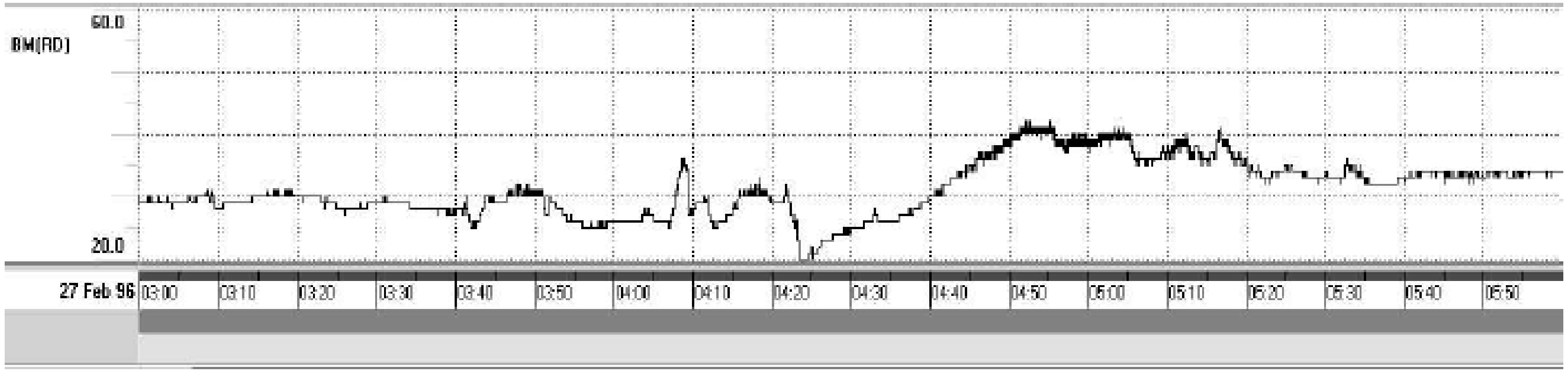
SUMTIME

MAUSAM

data-text

Input Data Interpretation

- Example Input data to SumTime-Neonate



Output Text for the above data set

"Initially BP is stable around 30 kpa until 4:23:14. In the next 28 minutes it gradually rises to 41 kpa. It gradually falls to 34 kpa by 5:59:59."

Why is NLG hard?

The problem of non-conscious knowledge

-- Interaction of language with memory

Symbol grounding

-- When is a match “tense”?

-- symbolic approaches

-- subsymbolic approaches

-- When is a movie “romantic”?

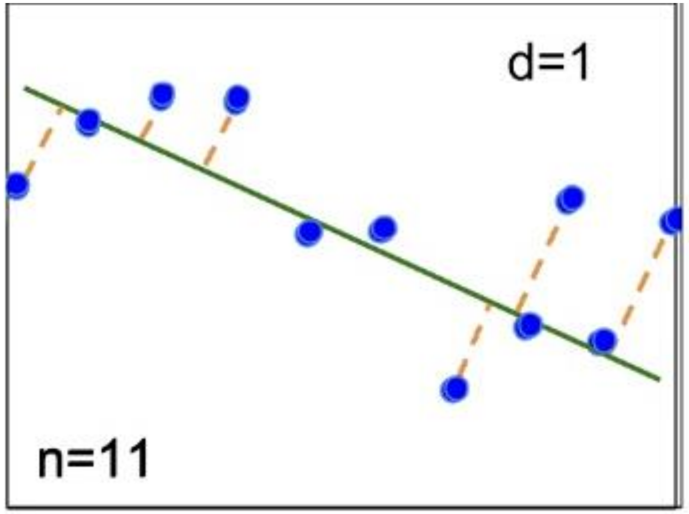
Integration of communication and problem solving

More …

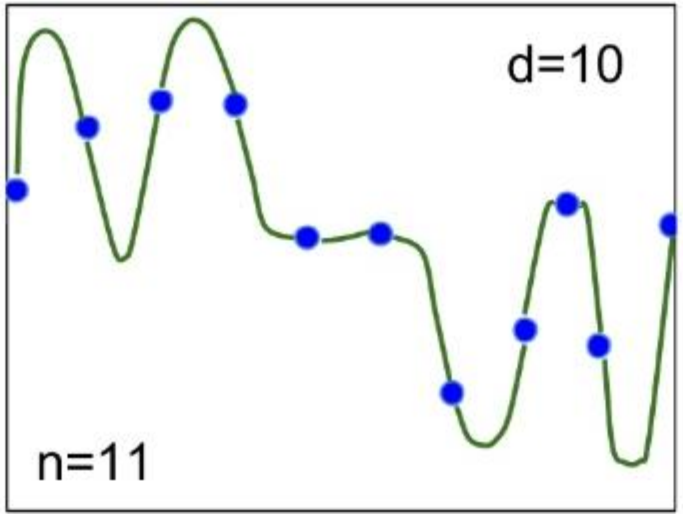
- How does a child acquire language generation skills?
- Language for compression – Minimum Description Length (MDL)
- Levels of abstraction and MDL
- Is the communication goal achieved? The pragmatic goal of language

- Can machines come up with a language of their own ?
  - At present, machines and humans do not share the same evolutionary goals

Minimum Description Length



VS.



$$\min L(M) \quad + \quad L(D|M)$$

bits for M

bits for the data using M

$a_1 x + a_0$

errors

$a_{10} x^{10} + a_9 x^9 + \ldots + a_0 \quad \{\ \}$

simple & good explanations

Ack : Danai Koutra (CMU)

NLG as Machine Translation

- Parallel Corpus: Collection of pairs of input data & their corresponding human written texts
- Useful to learn how humans map quantitative information to linguistic descriptions
- Similar to the idea of parallel corpus in Machine Translation

Quantitative Information

Human Written Linguistic Description

Shakespeare a écrit la hameau.

Shakespeare wrote Hamlet.

7	NO	NO
5	YES	YES
3	NO	YES
6	NO	NO
4	YES	YES
6	YES	NO
4	NO	???

Domain : Term papers

Days Late	Medical Certificate?	Accepted?
7	NO	NO
5	YES	YES
3	NO	YES
6	NO	NO
4	YES	YES
6	YES	NO
4	NO	???

Domain : Term papers

Days Late	Medical Certificate?	Accepted?
7	NO	NO
5	YES	YES
3	NO	YES
6	NO	NO
4	YES	YES
6	YES	NO
4	NO	???

Bottom-line: "Domain independent NLG" is a misnomer

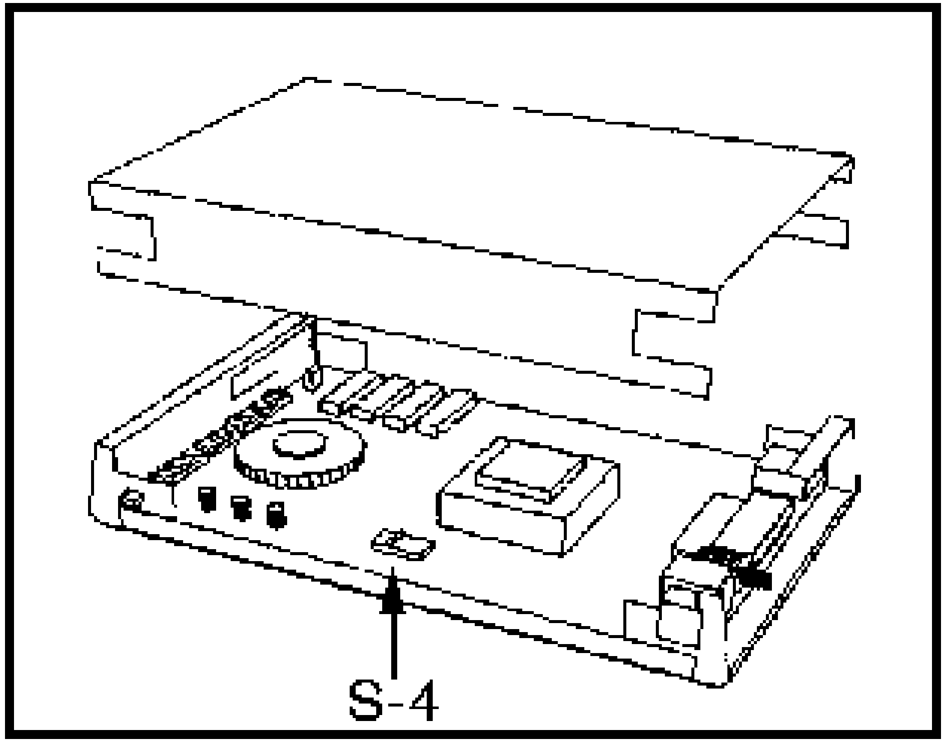
A central NLG challenge

- Integrating the "top-down" with the "bottom up"

NLG as a means of Human Computer Interaction

- Weak structure function correspondence
  - Gas chamber analogy
- Search : Google makes humans do most of the hard work

Combining Text with other forms of Visualization : An Example

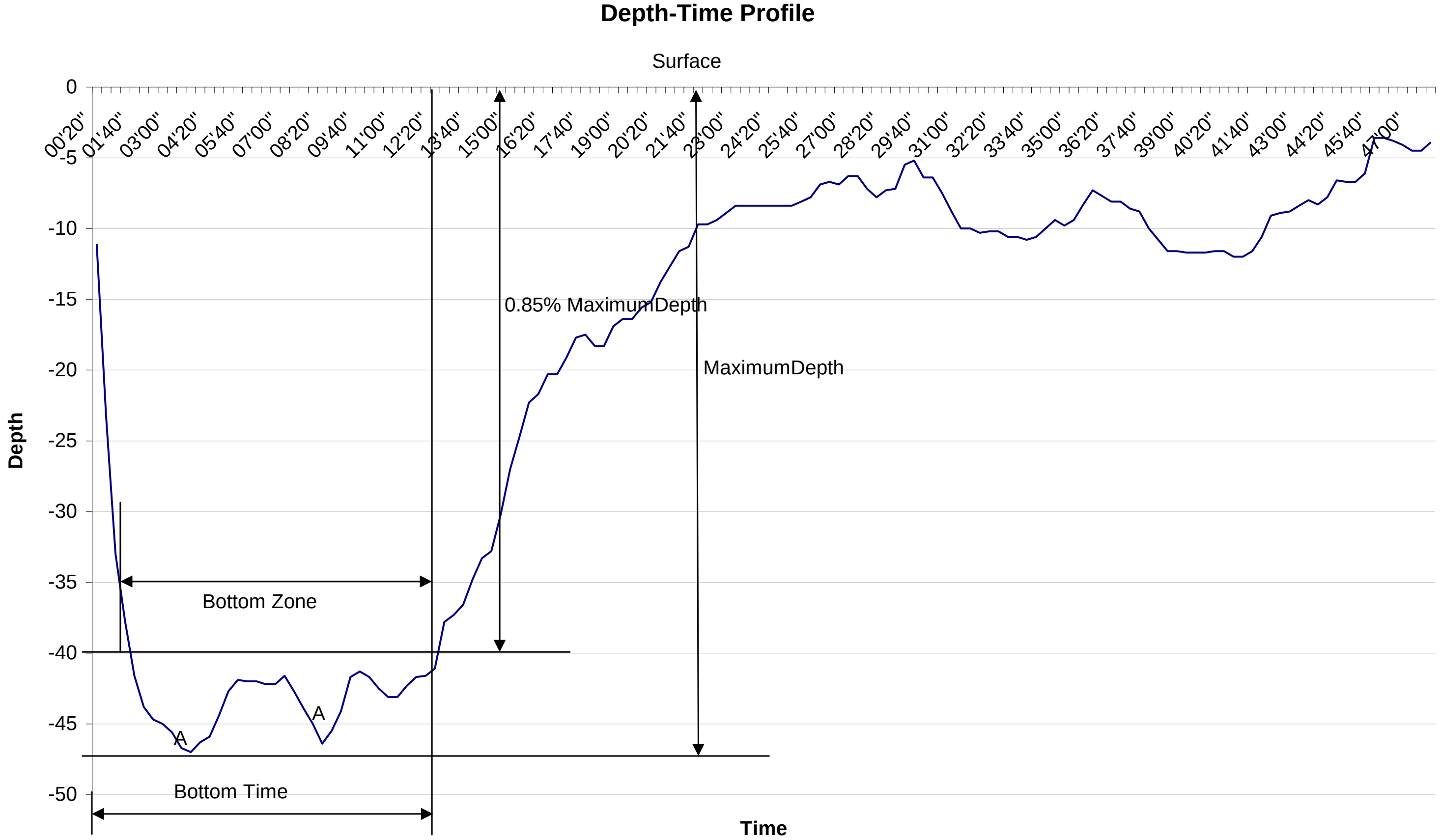


Push the code switch S-4 to the right.

Information Visualization

- Using computer graphics to facilitate humans gain new insights from an underlying information source
- Natural way of presenting quantitative information?
- Very successful field of research
  - Several fielded applications
  - Exploits computer graphics technology and superior human perception capabilities
- IBM's many eyes
  - http://services.alphaworks.ibm.com/manyeyes/app

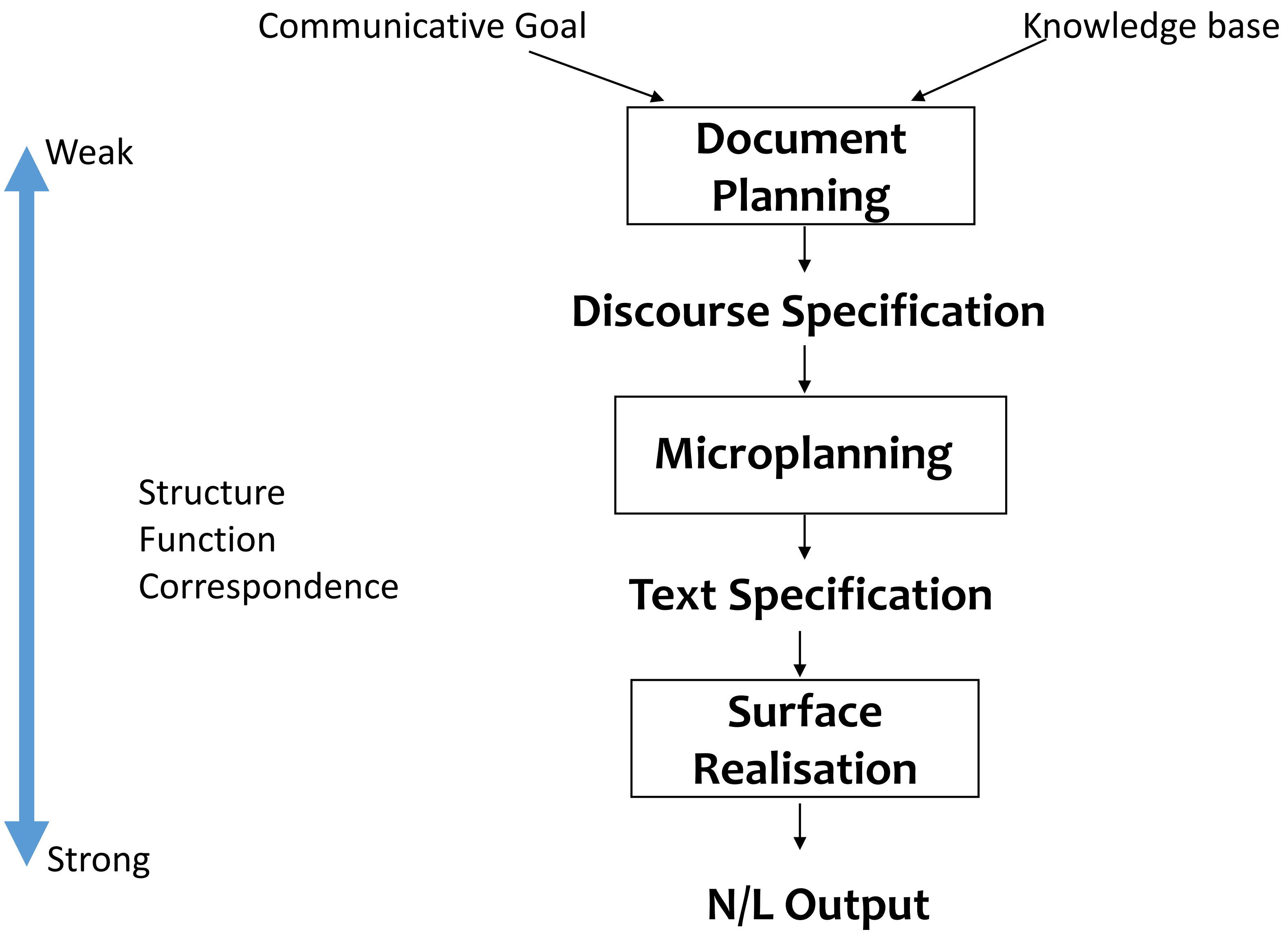
Text+Annotated Graphics



Risky dive with some minor problems. Because your bottom time of 12.0min exceeds no-stop limit by 4.0min this dive is risky. But you performed the ascent well. Your buoyancy control in the bottom zone was poor as indicated by 'saw tooth' patterns marked 'A' on the depth-time profile.

Browsability ?

Architecture for Generation



NL Subtasks

Weak

Structure
Function
Correspondence

Strong

- Content Determination
- Document Structuring

Document Planning

- Aggregation
- Lexicalisation
- Referring Expression Generation

Micro-planning

- Linguistic Realisation
- Structure Realisation

Surface Realisation

Content Determination

(1) a.

$$\begin{bmatrix} \text{message-id: msg01} \\ \text{relation: IDENTITY} \\ \text{arguments: } \begin{bmatrix} \text{arg1: NEXT-TRAIN} \\ \text{arg2: CALEDONIAN-EXPRESS} \end{bmatrix} \end{bmatrix}$$

b. The next train is the Caledonian Express

(2) a.

$$\begin{bmatrix} \text{message-id: msg02} \\ \text{relation: DEPARTURE} \\ \text{arguments: } \begin{bmatrix} \text{departing-entity: CALEDONIAN-EXPRESS} \\ \text{departure-location: ABERDEEN} \\ \text{departure-time: 1000} \end{bmatrix} \end{bmatrix}$$

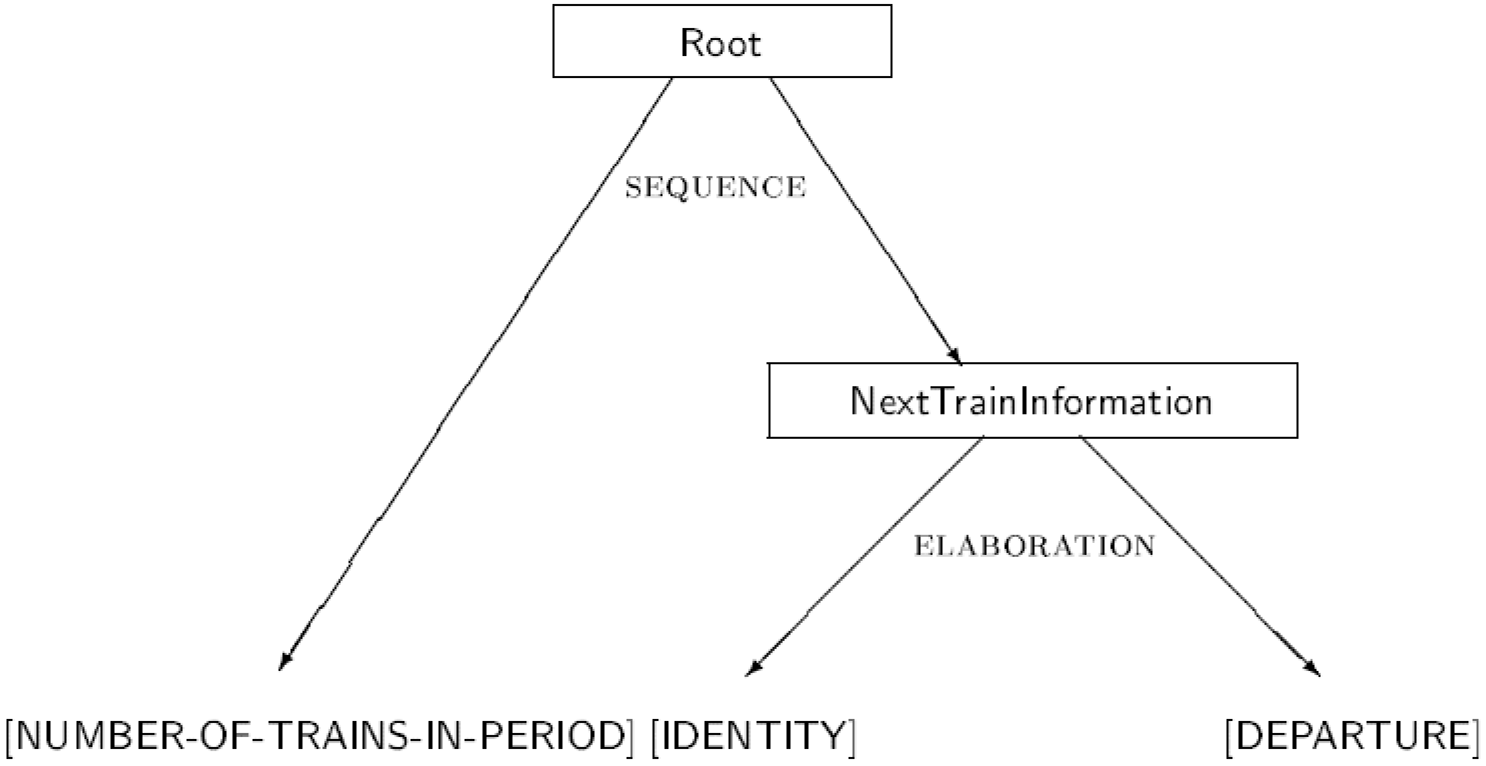
b. The Caledonian Express leaves Aberdeen at 10am

Content Determination (contd.)

(3) a. $$\begin{bmatrix} \textsf{message-id: msg03} & \\ \textsf{relation: NUMBER-OF-TRAINS-IN-PERIOD} & \\ \textsf{arguments:} & \begin{bmatrix} \textsf{source: ABERDEEN} \\ \textsf{destination: GLASGOW} \\ \textsf{number: 20} \\ \textsf{period: DAILY} \end{bmatrix} \end{bmatrix}$$

b. There are 20 trains each day from Aberdeen to Glasgow

Document Structuring



An example Discourse Structure Tree

Sentence aggregation

- Process of grouping messages together into sentences
- Can increase fluency and readability of text
- In our example :
  - IDENTITY and DEPARTURE messages can be combined into a single sentence, which can be realized as

    The next train, which leaves at 10 am, is the Caledonian Express

Lexicalization

- The process of deciding which words and phrases should be chosen to express the domain concepts and relations which appear in the messages

- How should the event represented in the DEPARTURE message be expressed?
  - The words *leave* and *depart* are both possibilities.

Referring Expression Generation

- The task of selecting words or phrases to identify domain entities.
- Example: In the text below, *Caledonian Express* and *it* refer to the domain entity CALEDONIAN-EXPRESS.
  - *The next train is the Caledonian Express; it leaves Aberdeen at 10 am*
- Closely related to lexicalization, but makes use of the DISCOURSE HISTORY

Linguistic Realization

- The process of applying grammar rules to produce a text which is syntactically, morphologically and orthographically correct

a.

$$
\begin{bmatrix}
\text{message-id: msg03} & \\
\text{relation: NUMBER-OF-TRAINS-IN-PERIOD} & \\
\text{arguments:} & \begin{bmatrix} \text{source: ABERDEEN} \\ \text{destination: GLASGOW} \\ \text{number: 20} \\ \text{period: DAILY} \end{bmatrix}
\end{bmatrix}
$$

b. There are 20 trains each day from Aberdeen to Glasgow.

Schank, A theory of Reminding, and Content Generation for Accenture Alerts

There have been a total of 6 alerts generated today. The day ended with two metrics in red status, two in amber and the rest in green.

The number of defects logged is extremely high, however most (more than 50%) have been logged in the last test iteration.

The number of high severity defects is more than 25%. This may point to serious system issues, please analyze.

Analogy with reproducing a newspaper report
Results in Neuroscience

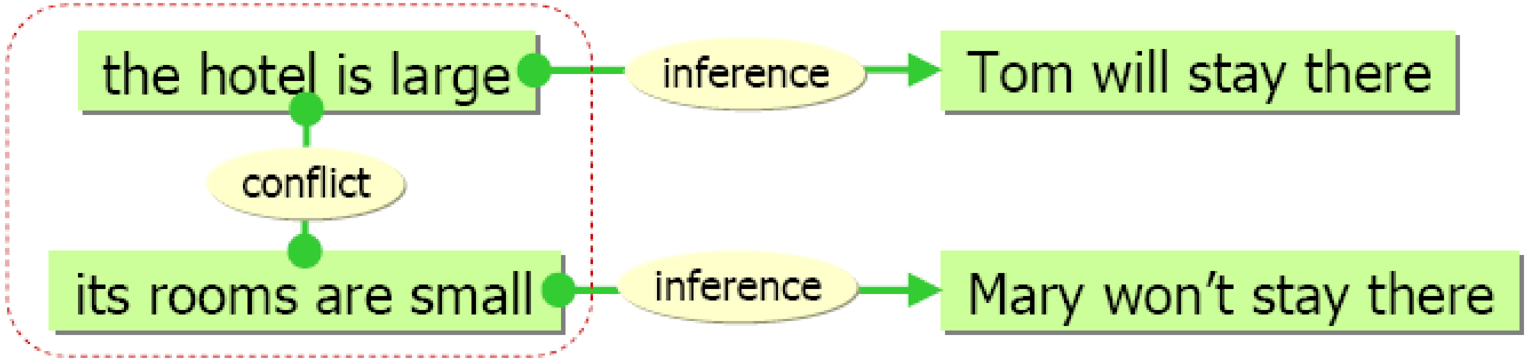
The notion of surprise

Is it surprising with respect to the mental model of the intended recipient?

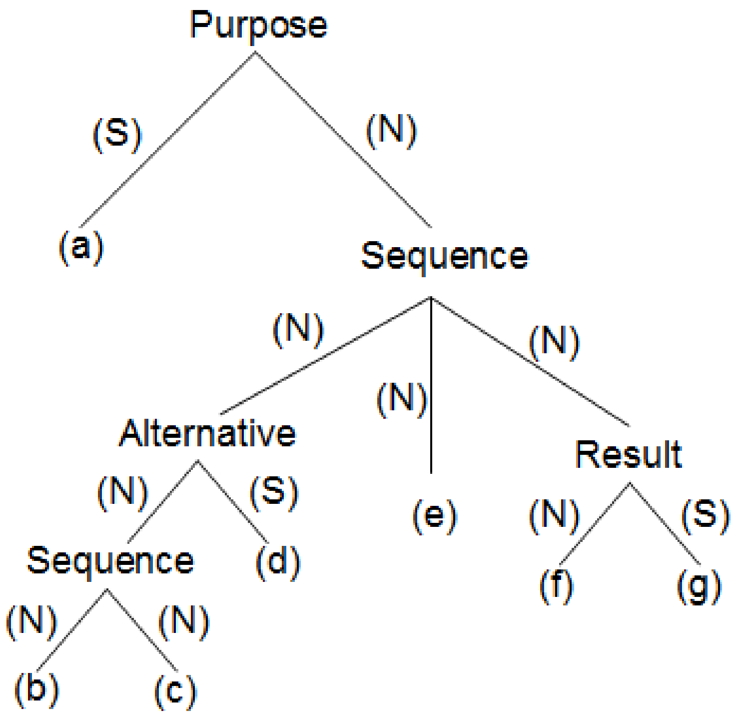
https://30daysofautism.wordpress.com/2011/04/16/welcome-to-interactive-post-day-or-more-advise-from-the-refrigerator-zone/

Rhetorical Structure Theory

- The hotel is large though their rooms are small. So Tom will stay there.
- Although the hotel is large, its rooms are small. So Mary won't stay there.



RST



(a) In order to reach the conference venue from Hotel Raintree and attend conference

(b) Take bus 5E to IIT Madras IN gate, and (c) take the IIT bus from IN gate to ICSR building

OR (d) Take a cab from hotel to ICSR building

(e) Register for the conference at ICSR reception

(f) Choose the session you want to attend and ask for help at reception

(g) The volunteers will help you with directions to auditoriums and lecture halls.

A Case Study in Applied NLG

- Each month an institutional newsletter publishes a summary of the month’s weather
- The summaries are based on automatically collected meteorological data
- The person who writes these summaries will no longer be able to
- The institution wants to continue publishing the reports and so is interested in using NLG techniques to do so

A Weather Summary

MARSFIELD (Macquarie University No 1)

On Campus, Square F9

TEMPERATURES (C)

Mean Max for Mth:	18.1 Warmer than average
Mean Max for June (20 yrs):	17.2
Highest Max (Warmest Day):	23.9 on 01
Lowest Max (Coldest Day):	13. On 12
Mean Min for Mth:	08.2 Much warmer than ave
Mean Min for June (20 yrs):	06.4
Lowest Min (Coldest Night):	02.6 on 09
Highest Min (Warmest Night):	13.5 on 24

RAINFALL (mm) (24 hrs to 09:00)

Total Rain for Mth:	90.4 on 12 days. Slightly below average.
Wettest Day (24h to 09:00):	26.4 on 11
Average for June (25 yrs):	109.0 on 10
Total for 06 mths so far:	542.0 on 72 days. Very depleted.
Average for 06 mths (25 yrs):	762.0 on 71 days
Annual Average Rainfall (25 yrs):	1142.8 on 131 days

WIND RUN (at 2m height) (km) (24 hrs to 09:00)

Total Wind Run for Mth:	1660
Windiest Day (24 hrs to 09:00):	189 on 24, 185 on 26, 172 on 27
Calmest Day (24 hrs to 09:00):	09 on 16

SUNRISE & SUNSET

Date	Sunrise	Sunset	Difference
01 Jun	06:52	16:54	10:02
11 Jun	06:57	16:53	09:56
21 Jun	07:00	16:54	09:54
30 Jun	07:01	16:57	09:56

(Sunset times began to get later after about June 11)
(Sunrise times continue to get later until early July)
(Soon we can take advantage of the later sunsets)

SUMMARY

The month was warmer than average with average rainfall, but the total rain so far for the year is still very depleted. The month began with mild to warm maximums, and became cooler as the month progressed, with some very cold nights such as June 09 with 02.6. Some other years have had much colder June nights than this, and minimums below zero in June are not very unusual. The month was mostly calm, but strong winds blew on 23, 24 and 26, 27. Fog occurred on 17, 18 after some rain on 17, heavy rain fell on 11 June.

Output: A Weather Summary

The month was warmer than average with average rainfall, but the total rain so far for the year is still very depleted. The month began with mild to warm maximums, and became cooler as the month progressed, with some very cold nights such as June 09 with 02.6. Some other years have had much colder June nights than this, and minimums below zero in June are not very unusual. The month was mostly calm, but strong winds blew on 23, 24 and 26, 27. Fog occurred on 17, 18 after some rain on 17, heavy rain fell on 11 June.

The Input Data

- A set of 16 data elements collected automatically every 15 minutes: air pressure, temperature, wind speed, rainfall …
- Preprocessed to construct DailyWeatherRecords:

```
((type dailyweatherrecord)
    (date ((day ...)
           (month ...)
           (year ...)))
    (temperature ((minimum ((unit degrees-centigrade)
                            (number ...)))
                  (maximum ((unit degrees-centrigrade)
                            (number ...)))))
    (rainfall ((unit millimetres)
               (number ...))))
```

Other Available Data

- Historical Data: Average temperature and rainfall figures for each month in the Period of Record (1971 to present)
- Historical Averages: Average values for temperature and rainfall for the twelve months of the year over the period of record

Requirements Analysis

The developer needs to:

- understand the client's needs
- propose a functionality which addresses these needs

Corpus-Based Requirements Analysis

A corpus:

- consists of examples of output texts and corresponding input data
- specifies ‘by example’ the functionality of the proposed NLG system
- is a very useful resource for design as well as requirements analysis

Corpus-Based Requirements Analysis

Four steps:

- assemble an initial corpus of (human-authored) output texts and associated input data
- analyse the information content of the corpus texts in terms of the input data
- develop a target text corpus
- create a formal functional specification

Step 1: Creating an Initial Corpus

- Collect a corpus of input data and associated (human-authored) output texts
- One source is archived examples of human-authored texts
- If no human-authored examples of the required texts exist, ask domain experts to produce examples
- The corpus should cover the full range of texts expected to be produced by the NLG system

Initial Text (April 1995)

SUMMARY

The month was rather dry with only three days of rain in the middle of the month. The total for the year so far is very depleted again, after almost catching up during March. Mars Creek dried up again on 30th April at the waterfall, but resumed on 1st May after light rain. This is the fourth time it dried up this year.

Step 2: Analyzing the Content of the Texts

- Goal:
  - to determine where the information present in the texts comes from, and the extent to which the proposed NLG system will have to manipulate this information
- Result:
  - a detailed understanding of the correspondences between the available input data and the output texts in the initial corpus

Information Types in Text

- Unchanging text
- Directly-available data
- Computable data
- Unavailable data

Unchanging Text

SUMMARY

The month was rather dry with only three days of rain in the middle of the month. The total for the year so far is very depleted again, after almost catching up during March. Mars Creek dried up again on 30th April at the waterfall, but resumed on 1st May after light rain. This is the fourth time it dried up this year.

Directly Available Data

SUMMARY

The month was rather dry with only three days of rain in the middle of the month. The total for the year so far is very depleted again, after almost catching up during March. Mars Creek dried up again on 30th April at the waterfall, but resumed on 1st May after light rain. This is the fourth time it dried up this year.

Computable Data

SUMMARY

The month was rather dry with only three days of rain in the middle of the month. **The total for the year so far is very depleted again, after almost catching up during March.** Mars Creek dried up again on 30th April at the waterfall, but resumed on 1st May after light rain. This is the fourth time it dried up this year.

Unavailable Data

SUMMARY

The month was rather dry with only three days of rain in the middle of the month. The total for the year so far is very depleted again, after almost catching up during March. **Mars Creek dried up again on 30th April at the waterfall, but resumed on 1st May after light rain. This is the fourth time it dried up this year.**

Solving the Problem of Unavailable Data

- More information can be made available to the system: this may be expensive
  - add sensors to Mars Creek?
- If the system is an authoring-aid, a human author can add this information
  - system produces the first two sentences, the human adds the second two
- The target corpus can be revised to eliminate clauses that convey this information
  - only produce the first two sentences

Step 3: Building the Target Text Corpus

Mandatory changes:

- eliminate unavailable data
- specify what text portions will be human-authored

Optional changes:

- simplify the text to make it easier to generate
- improve human-authored texts
- enforce consistency between human authors

Target Text

The month was rather dry with only three days of rain in the middle of the month. The total for the year so far is very depleted again.

Step 4: Functional Specification

- Based on an agreed target text corpus
- Explicitly states role of human authoring, if present at all
- Explicitly states structure and range of inputs to be used

Initial Text #2

The month was our driest and warmest August in our 24 year record, and our first 'rainless' month. The 26th August was our warmest August day in our record with 30.1, and our first 'hot' August day (30). The month forms part of our longest dry spell 47 days from 18 July to 02 September 1995. Rainfall so far is the same as at the end of July but now is very deficient.

Target Text #2

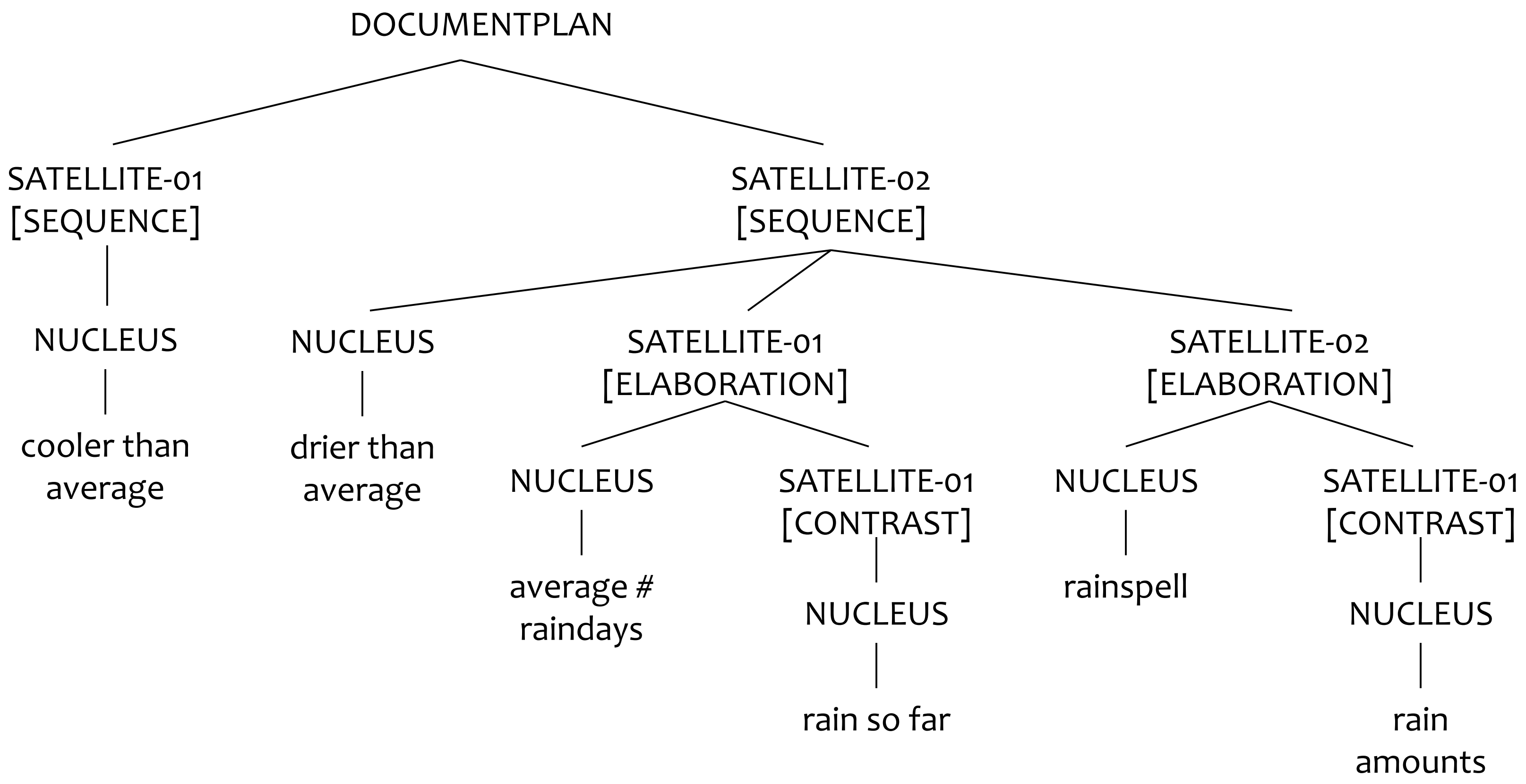
The month was the driest and warmest August in our 24 year record, and the first rainless month of the year. 26th August was the warmest August day in our record with 30.1, and the first hot day of the month. Rainfall for the year is now very deficient.

The Case Study So Far

We’ll assume that:

- We have located the source data
- We have preprocessed the data to build the DailyWeatherRecords
- We have constructed an initial corpus of texts
- We have modified the initial corpus to produce a set of target texts

The Document Plan



One Message per Sentence

- The Result:

  The month was cooler than average.
  The month was drier than average.
  There were the average number of rain days.
  The total rain for the year so far is well below average.
  There was rain on every day for 8 days from 11th to 18th.
  Rainfall amounts were mostly small.

- The Target Text:

  The month was cooler and drier than average, with the average number of rain days, but the total rain for the year so far is well below average. Although there was rain on every day for 8 days from 11th to 18th, rainfall amounts were mostly small.

Relatively strong structure function correspondence

Mapping Accenture Domain to Cricket

Inferencing for Generating Computable Data

Snapshot of data from cricsheet

ball	2	0.1	Sri Lanka	MD Gunathilaka	TM Dilshan	R Ashwin	0	2		
ball	2	0.2	Sri Lanka	TM Dilshan	MD Gunathilaka	R Ashwin	0	0	stumped	TM Dilshan
ball	2	0.3	Sri Lanka	S Prasanna	MD Gunathilaka	R Ashwin	0	0		
ball	2	0.4	Sri Lanka	S Prasanna	MD Gunathilaka	R Ashwin	0	0		
ball	2	0.5	Sri Lanka	S Prasanna	MD Gunathilaka	R Ashwin	1	0		
ball	2	0.6	Sri Lanka	MD Gunathilaka	S Prasanna	R Ashwin	0	0		
ball	2	0.7	Sri Lanka	MD Gunathilaka	S Prasanna	R Ashwin	0	0		
ball	2	1.1	Sri Lanka	S Prasanna	MD Gunathilaka	A Nehra	0	0	caught	S Prasanna

Ball by ball data

(wicket-state, current RR/required RR, wickets left) → (0.8, 1.0206,9)

Crisp input to FIS

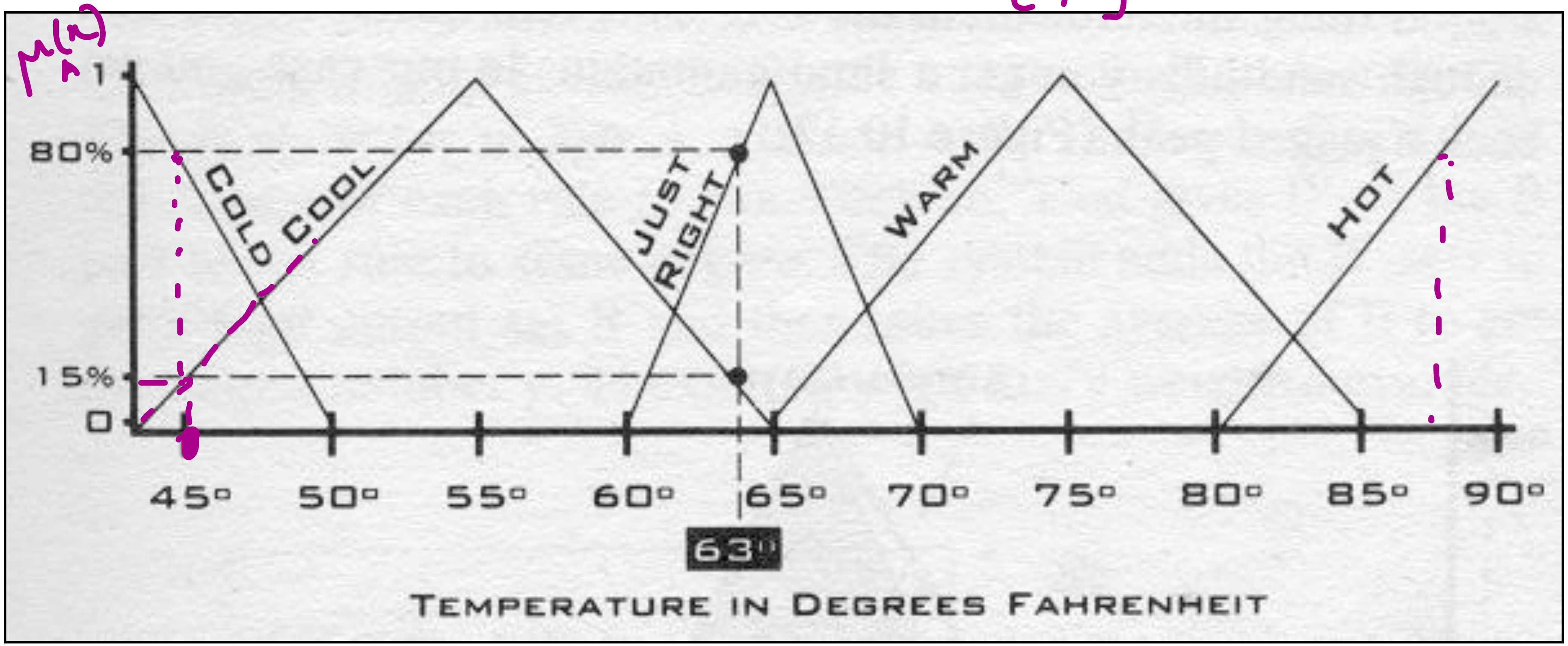
$\mu_{cold}(t) =$

$\mu_{cold}(45^\circ) = 0.8$

$\mu_{cool}(45^\circ) = 0.15$

$[0,1]$

$\langle 0.8, 0.15, 0, 0, 0 \rangle$



Linguistic Hedges

Concentration: CON(A)

$$\mu_{\mathrm{CON}_{(A)}}(u) = (\mu_A(u))^2 .$$

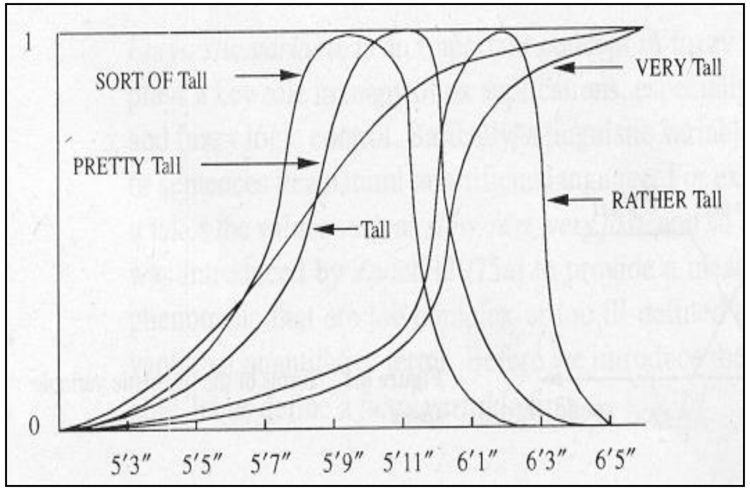
Dilation: DIL(A)

$$\mu_{\mathrm{DIL}_{(A)}}(u) = (\mu_A(u))^{1/2} .$$

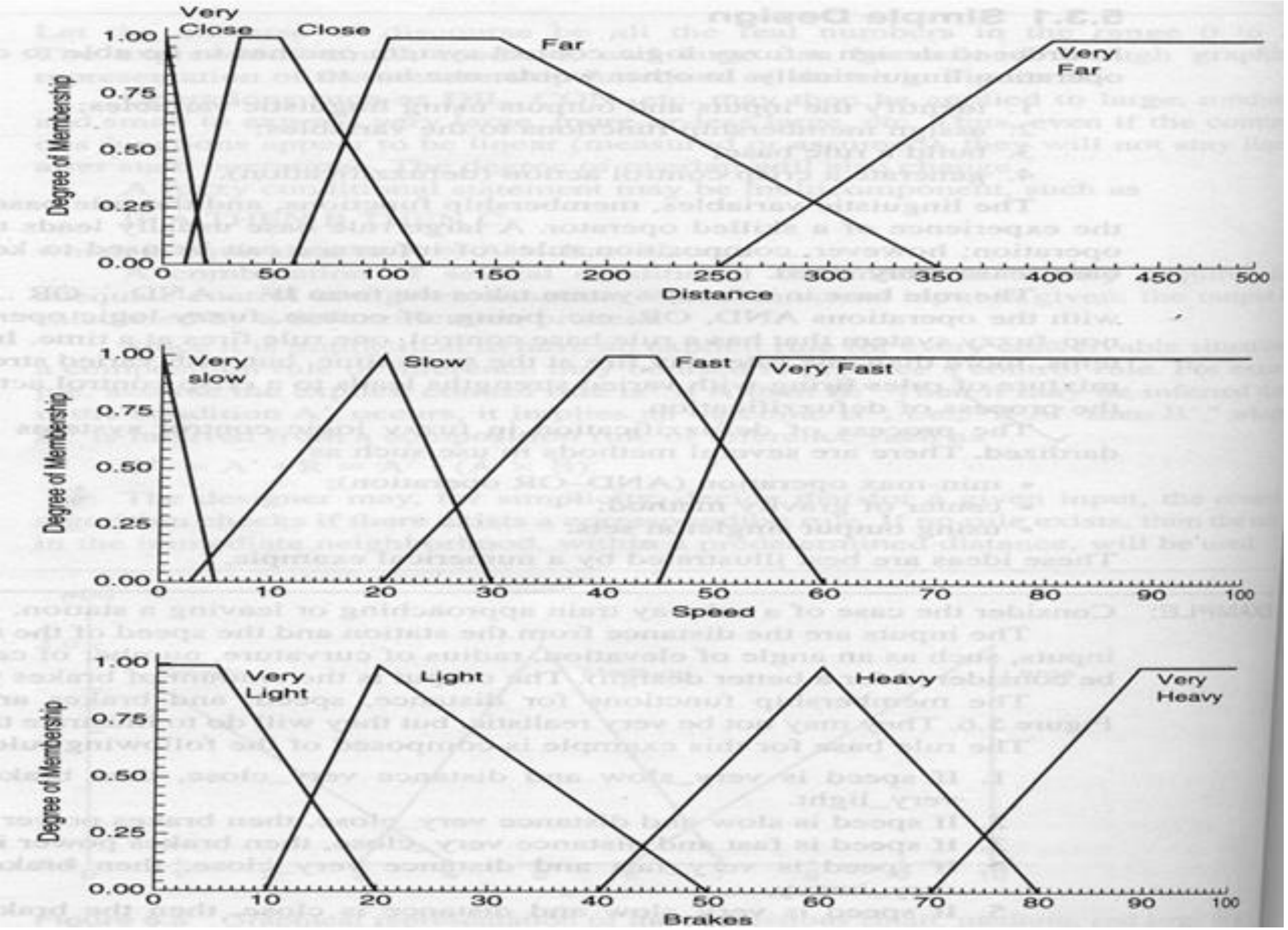
Intensification: INT(A)

$$\mu_{\mathrm{INT}_{(A)}}(u) = \begin{cases} 2(\mu_A(u))^2 , & \mu_A(u) \in [0, 0.5] \\ 1 - 2(1 - \mu_A(u))^2 , & \text{otherwise.} \end{cases}$$

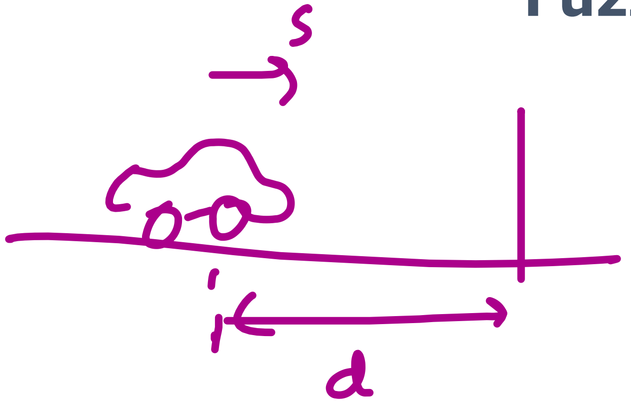
$$\begin{aligned}
\mathrm{VERY}(A) &= \mathrm{CON}(A) = \mathrm{A}^2, \\
\mathrm{HIGHLY}(A) &= \mathrm{A}^3, \\
\mathrm{FAIRLY\ (MORE\ OR\ LESS)}(A) &= \mathrm{DIL}(A) = \mathrm{A}^{1/2}, \\
\mathrm{ROUGHLY}(A) &= \mathrm{DIL}[\mathrm{DIL}(A)], \\
\mathrm{PLUS}(A) &= \mathrm{A}^{1.25} \\
\mathrm{MINUS}(A) &= \mathrm{A}^{0.75}, \\
\mathrm{RATHER}(A) &= \mathrm{INT}[\mathrm{CON}(A)]\ \mathrm{AND\ NOT}[\mathrm{CON}(A)], \\
\mathrm{SLIGHTLY}(A) &= \mathrm{INT}[\mathrm{PLUS}(A)\ \mathrm{AND\ NOT}(\mathrm{VERY}(A))], \\
\mathrm{SORT\ OF}(A) &= \mathrm{INT}[\mathrm{DIL}(A)]\ \mathrm{AND\ INT}[\mathrm{DIL}(\mathrm{NOT}(A))], \\
\mathrm{PRETTY}(A) &= \mathrm{INT}(A)\ \mathrm{AND\ NOT}[\mathrm{INT}(\mathrm{CON}(A))].
\end{aligned}$$



Input and Output Variables



Fuzzy Logic Control: a concrete example (contd.)



Distance \ Speed	Very_Slow	Slow	Fast	Very_Fast
Very_Close	Light	Heavy	Very_Heavy	Very_Heavy
Close	Light	Light	Heavy	Very_Heavy
Far	Light	Very_Light	Light	Heavy
Very_Far	Very_Light	Very_Light	Light	Light

Example of inferencing

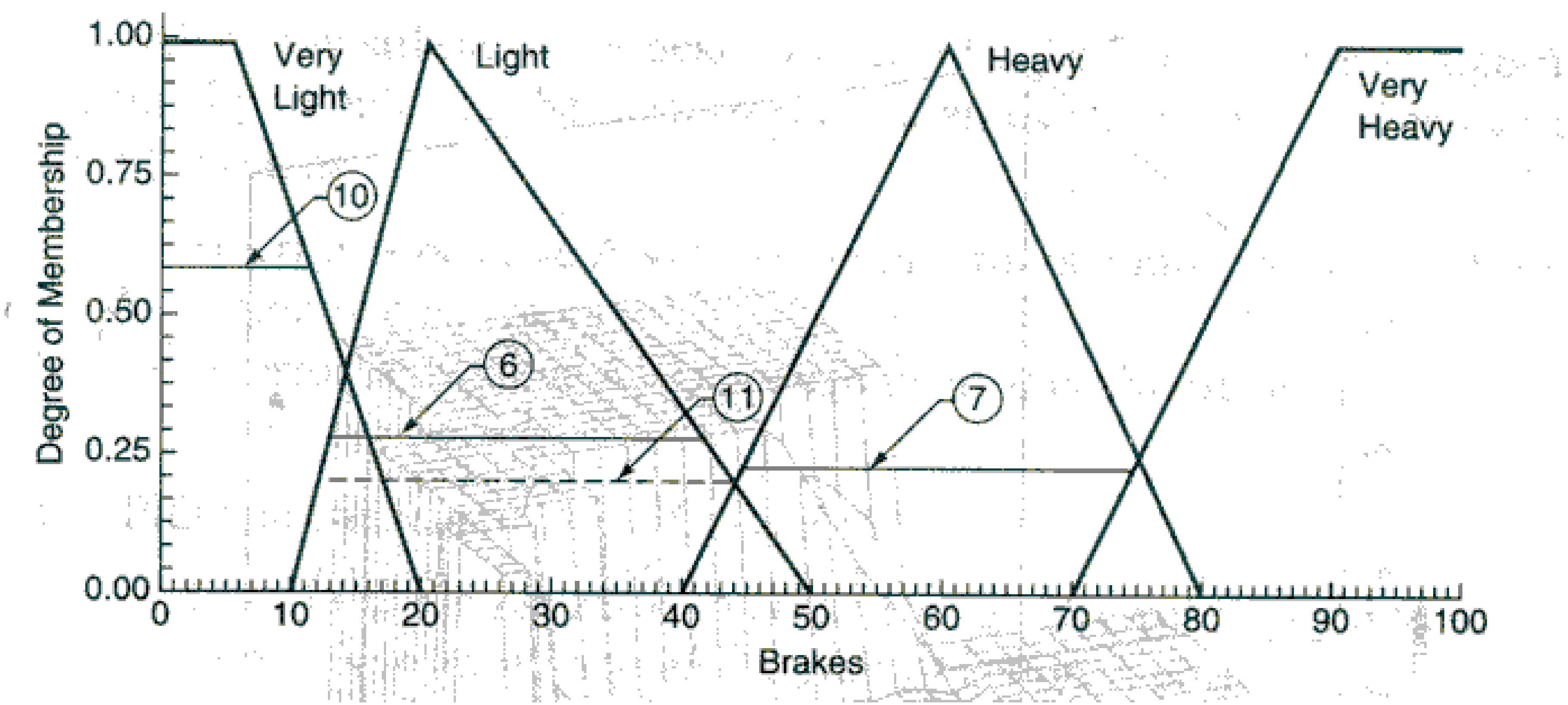
- Distance : 100m
- Speed : 24.6 km/h
- What should be the braking power?

- Speed: {0,0.58, 0.21, 0}
- Distance: {0,0.29,0.88, 0}

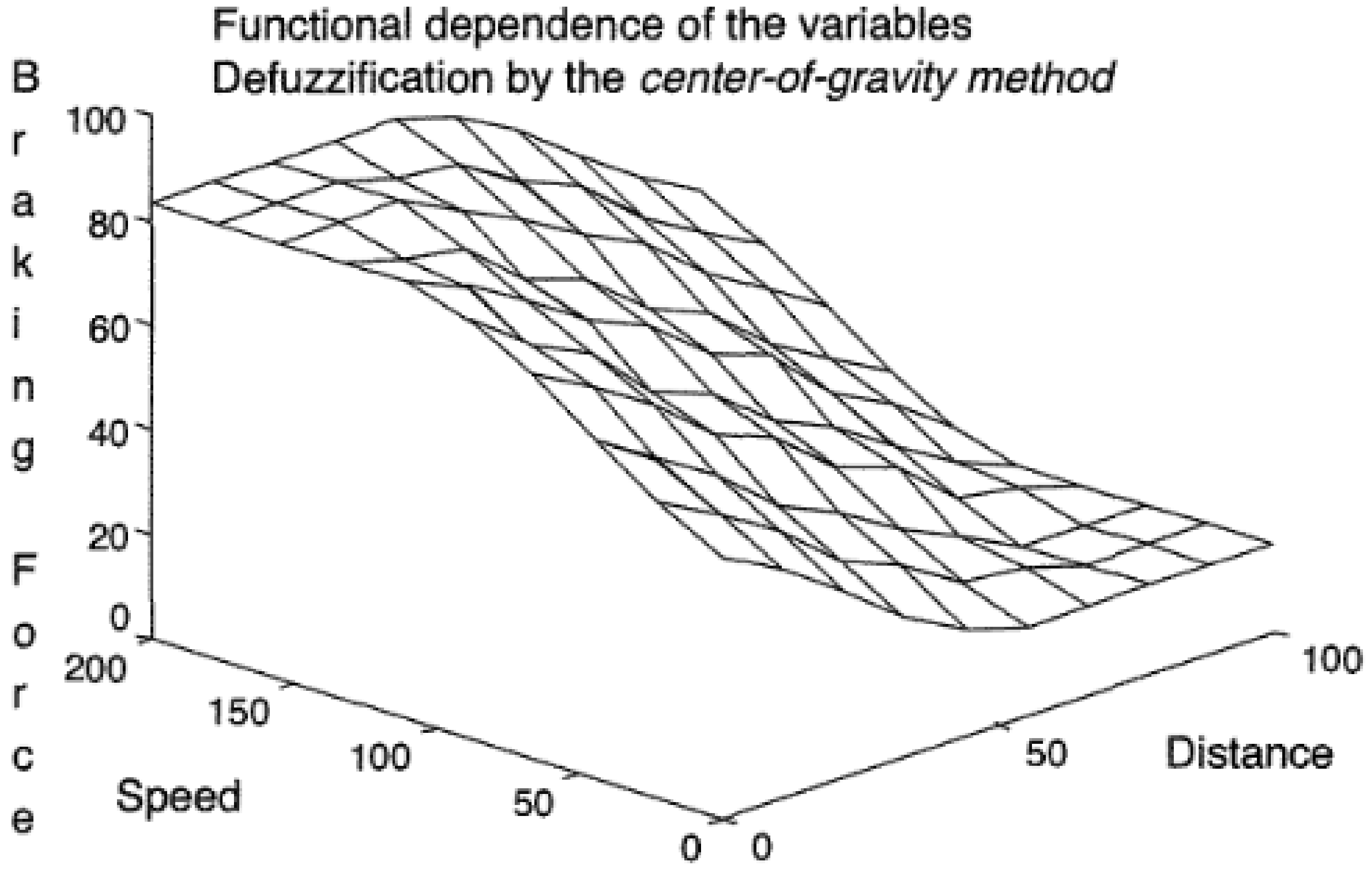
Rule activations

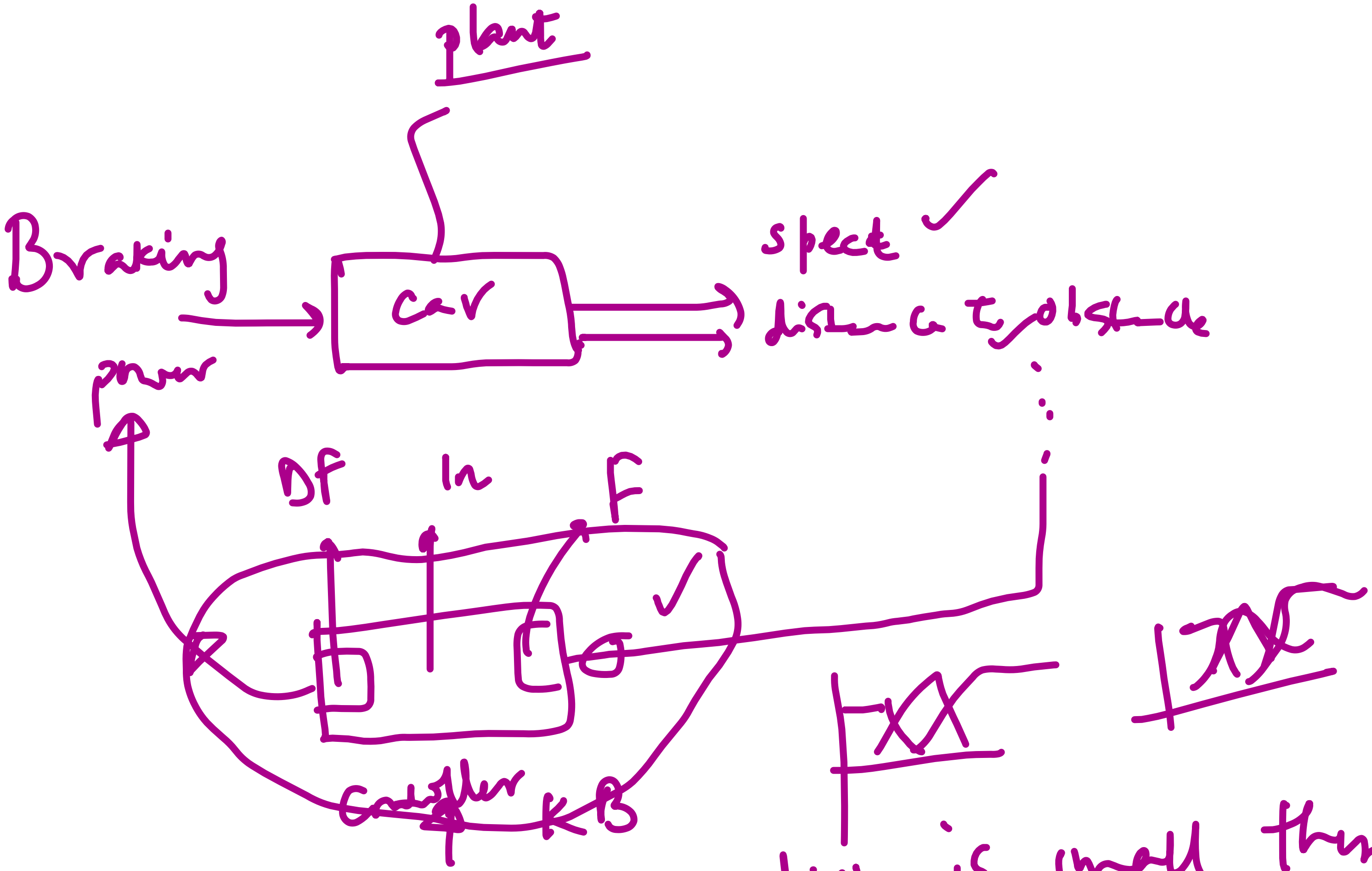
		0	0.58	0.21	0
	Distance / Speed	*Very_Slow*	*Slow*	*Fast*	*Very_Fast*
0	Very_Close	1 Light	2 Heavy	3 Very_Heavy	4 Very_Heavy
0.29	Close	5 Light	6 Light	7 Heavy	8 Very_Heavy
0.88	Far	9 Light	10 Very_Light	11 Light	12 Heavy
0	Very_Far	13 Very_Light	14 Very_Light	15 Light	16 Light

Fuzzy Logic Control: a concrete example (contd.)



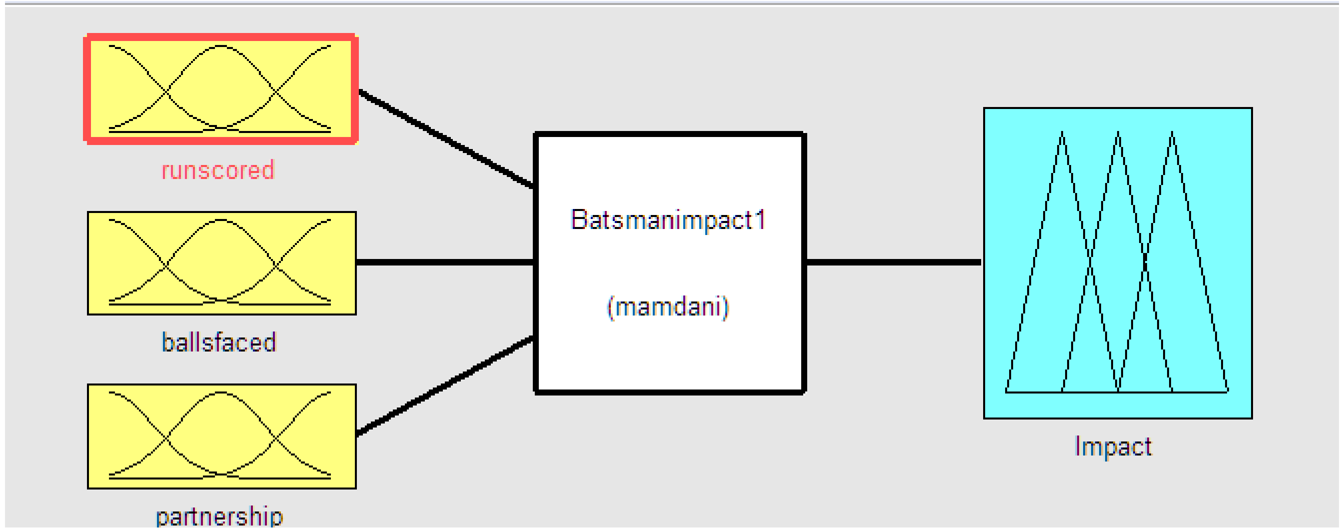
Control surface

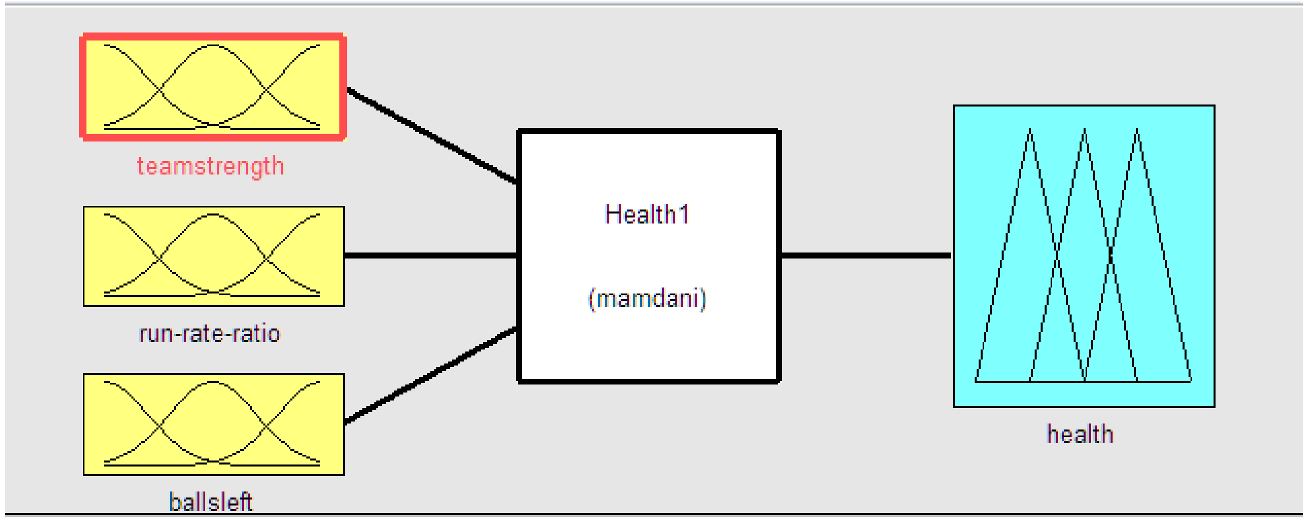


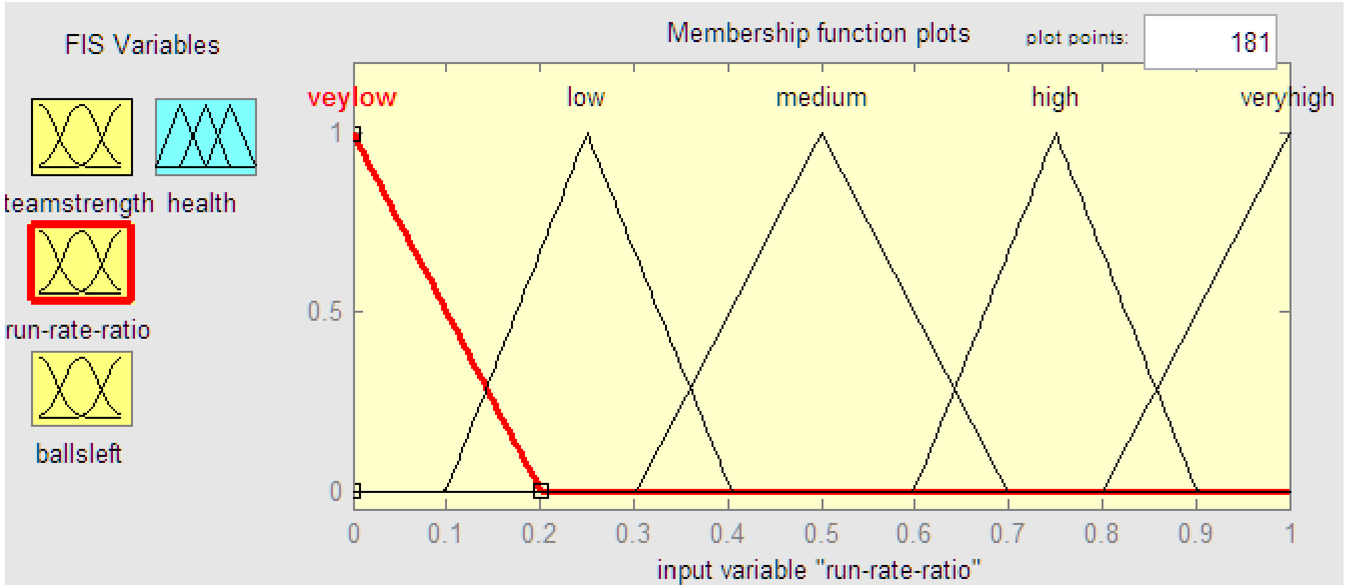


if speed is high & ~~brea~~ dist. is small then { bp is high

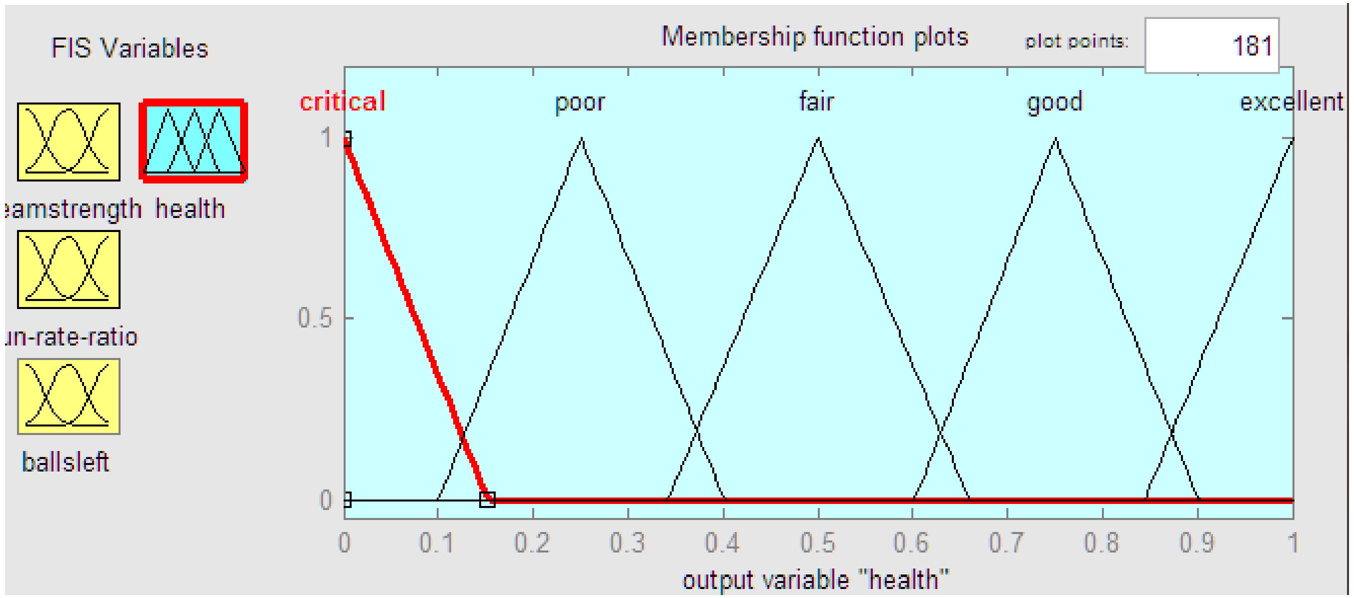
Explorations in the Cricket domain







FIS Variables
teamstrength
health
run-rate-ratio
ballsleft
Membership function plots
plot points:
181
veylow
low
medium
high
veryhigh
1
0.5
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
input variable "run-rate-ratio"

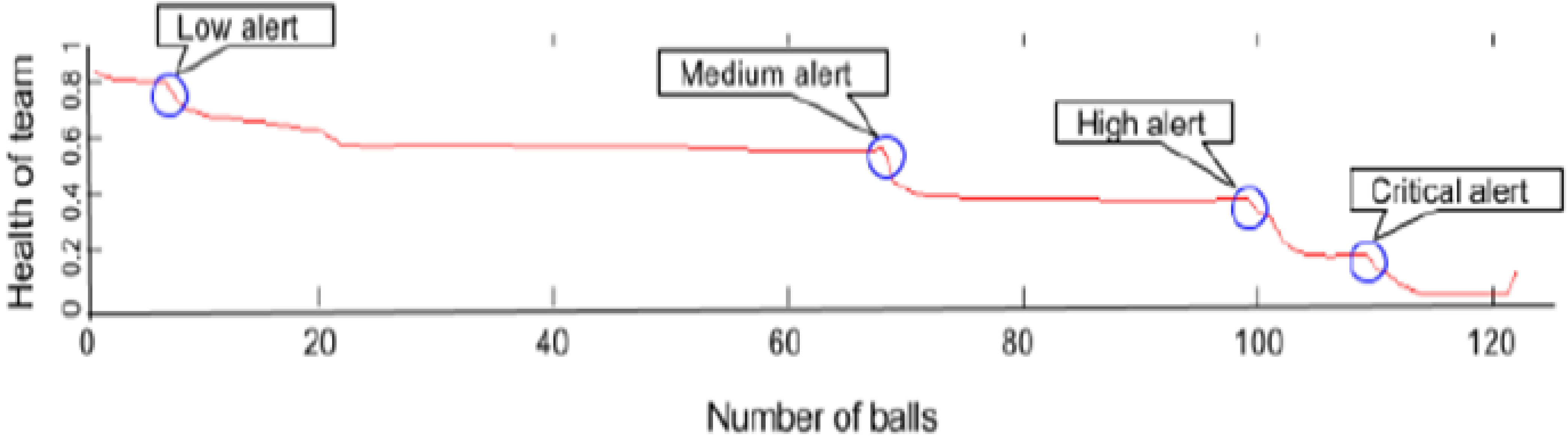


FIS Variables
eamstrength
health
un-rate-ratio
ballsleft
Membership function plots
plot points:
181
critical
poor
fair
good
excellent
1
0.5
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
output variable "health"

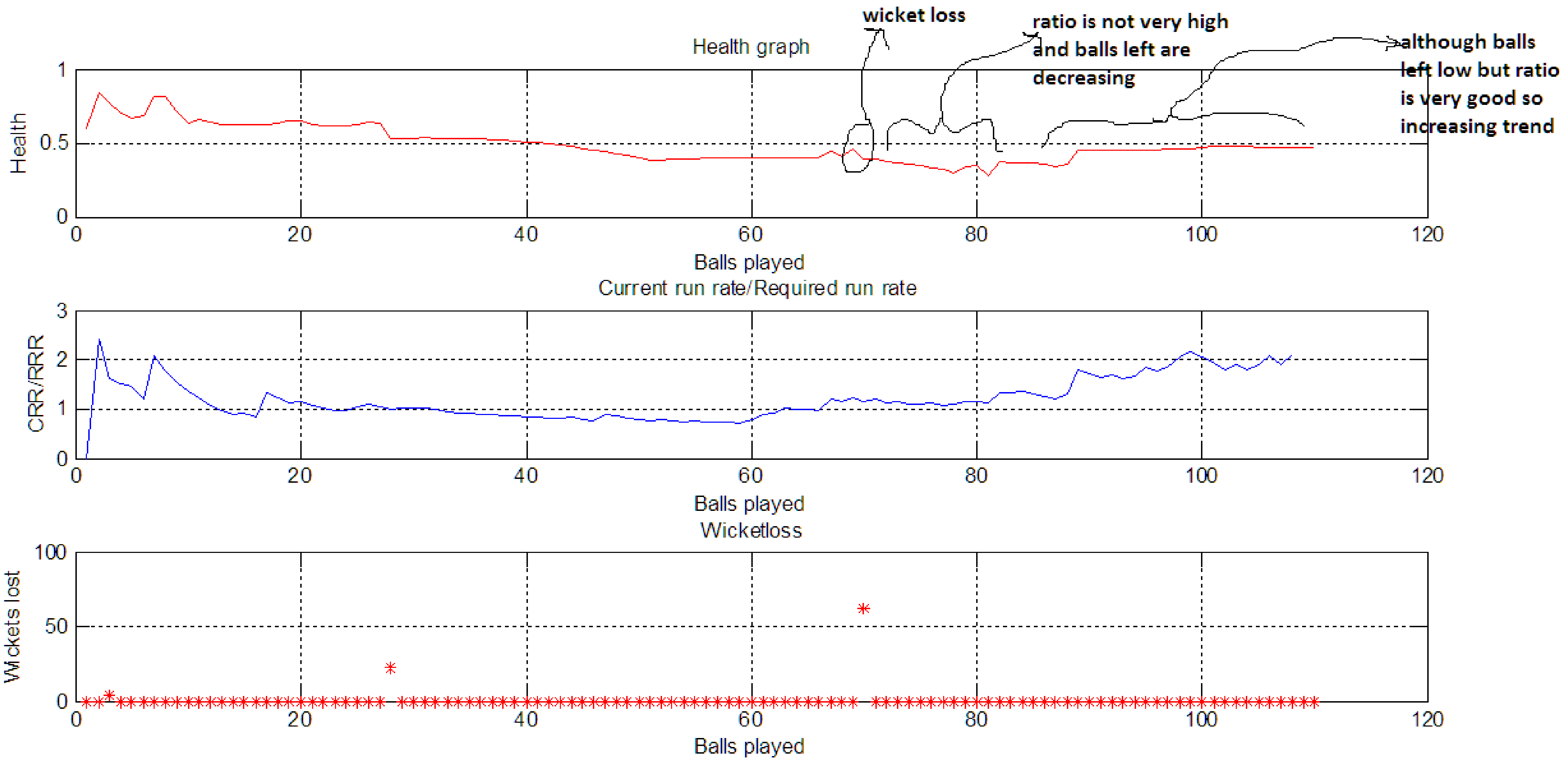
Inference Rules

1. If (teamstrength is verypoor) and (run-rate-ratio is veylow) and (ballsleft is veryless) then (health is critical) (1)
2. If (teamstrength is verypoor) and (run-rate-ratio is veylow) and (ballsleft is less) then (health is poor) (1)
3. If (teamstrength is verypoor) and (run-rate-ratio is veylow) and (ballsleft is average) then (health is fair) (1)
4. If (teamstrength is verypoor) and (run-rate-ratio is veylow) and (ballsleft is high) then (health is fair) (1)
5. If (teamstrength is verypoor) and (run-rate-ratio is veylow) and (ballsleft is veryhigh) then (health is good) (1)
6. If (teamstrength is verypoor) and (run-rate-ratio is low) and (ballsleft is veryless) then (health is critical) (1)
7. If (teamstrength is verypoor) and (run-rate-ratio is low) and (ballsleft is less) then (health is poor) (1)
8. If (teamstrength is verypoor) and (run-rate-ratio is low) and (ballsleft is average) then (health is poor) (1)
9. If (teamstrength is verypoor) and (run-rate-ratio is low) and (ballsleft is high) then (health is fair) (1)
10. If (teamstrength is verypoor) and (run-rate-ratio is low) and (ballsleft is veryhigh) then (health is fair) (1)
11. If (teamstrength is verypoor) and (run-rate-ratio is medium) and (ballsleft is veryless) then (health is critical) (1)
12. If (teamstrength is verypoor) and (run-rate-ratio is medium) and (ballsleft is less) then (health is poor) (1)
13. If (teamstrength is verypoor) and (run-rate-ratio is medium) and (ballsleft is average) then (health is poor) (1)
14. If (teamstrength is verypoor) and (run-rate-ratio is medium) and (ballsleft is high) then (health is fair) (1)

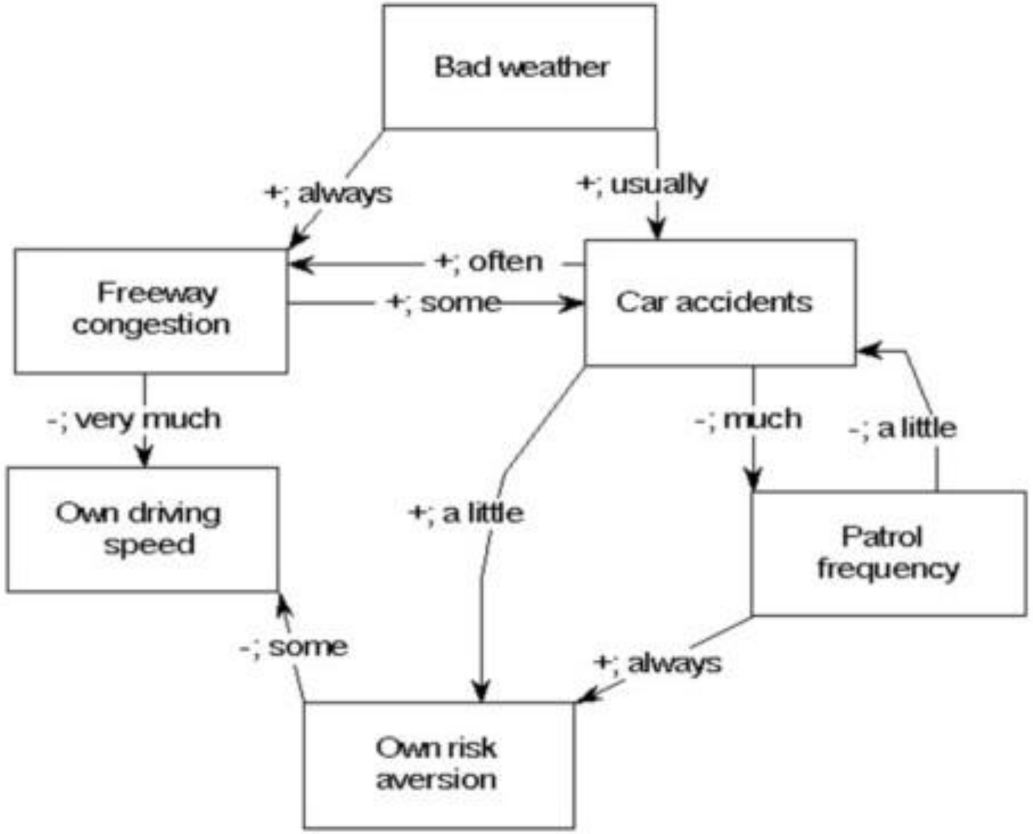
Generating computable data for summarization



Generating computable data for summarization



Fuzzy Cognitive Maps



http://www.ochoadeaspuru.com/fuzcogmap/badweatherdriving.php

Machine Learning for NLG

- Two sub-problems :
  - Content selection
  - Lexicalization

- Work done in collaboration with University of Aberdeen

Content Selection

- Task: To compute an appropriate subset of the input data set for communication
- E.g. the wind description selects the subset of {4,18} from the input wind speed values {4,6,7,10,12,16,18}

Input:

hour	windDir	windSpeed
6	S	4
9	S	6
12	S	7
15	S	10
18	S	12
21	S	16
0	S	18

Output:

S 02-06 INCREASING 16-20 BY EVENING

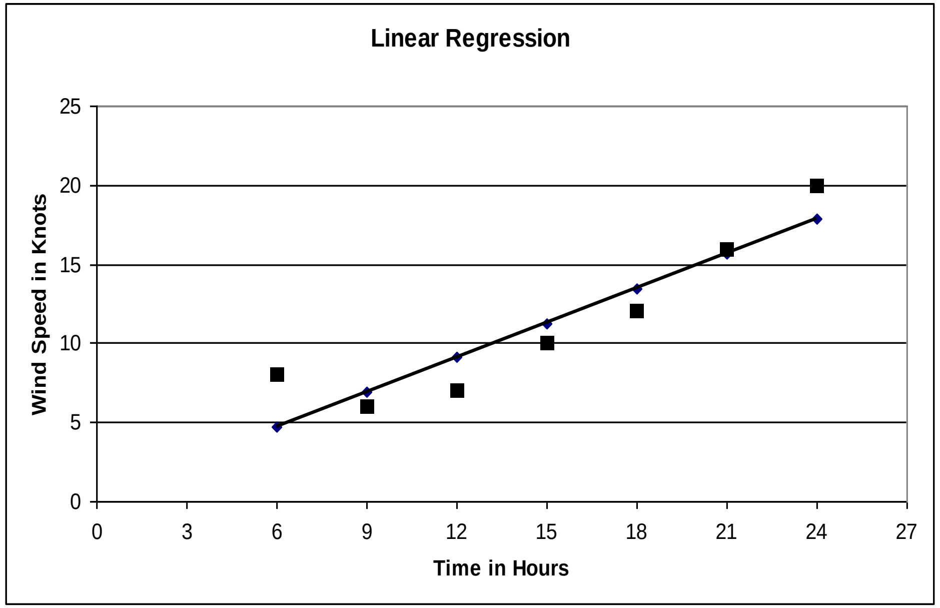
02-06 is a range created around 04 and

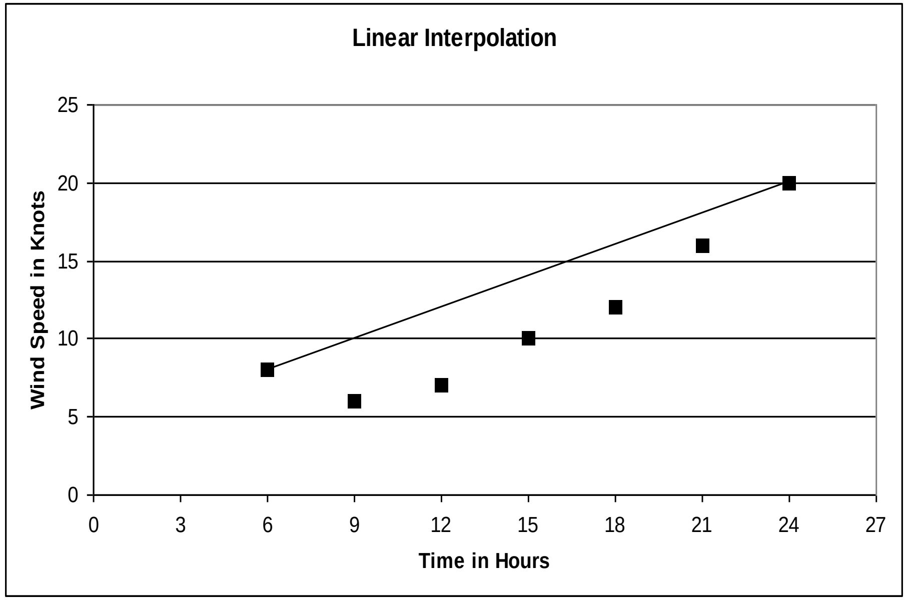
16-20 is a range created around 18

Selecting Modelling Method: Sample Data

Hour	Wind Direction	Wind Speed
0600	S	8
0900	S	6
1200	S	7
1500	S	10
1800	S	12
2100	S	16
0000	S	20

Linear Regression Vs Linear Interpolation





Lexical Selection

- Task: To select an appropriate word/phrase to express a portion of the input information
- E.g. the wind rises to 18 at 00 hours which is expressed as 'by evening'

Input:

hour	windDir	windSpeed
6	S	4
9	S	6
12	S	7
15	S	10
18	S	12
21	S	16
0	S	18

Output:

S 02-06 INCREASING 16-20 BY EVENING

02-06 is a range created around 04 and

16-20 is a range created around 18

Example: SumTime-Meteo

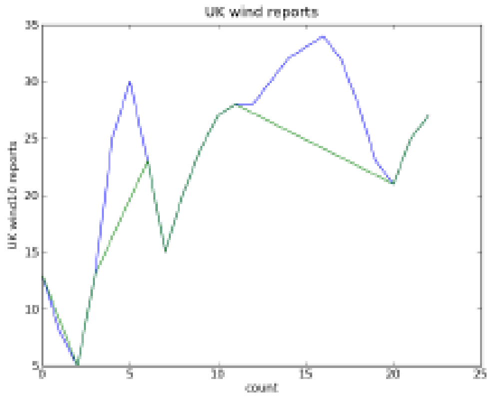
Hour	Wind Dir	Wind Speed
00	WSW	12
03	WSW	15
06	WSW	19
09	WSW	18
12	W	17
15	W	15
18	WSW	13
21	WSW	11
24	WSW	11

WSW 10-15 increasing 17-22 by early morning, then gradually easing 9-14 by midnight.

Segmentation for Content Selection

Given a time series T, produce the best representation either

- using exactly K segments,
- error for any segment does not exceed some user-specifi threshold, or
- combined error of all segments is less than some user-specified threshold.



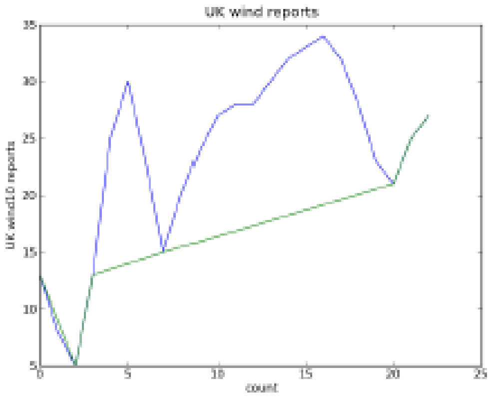
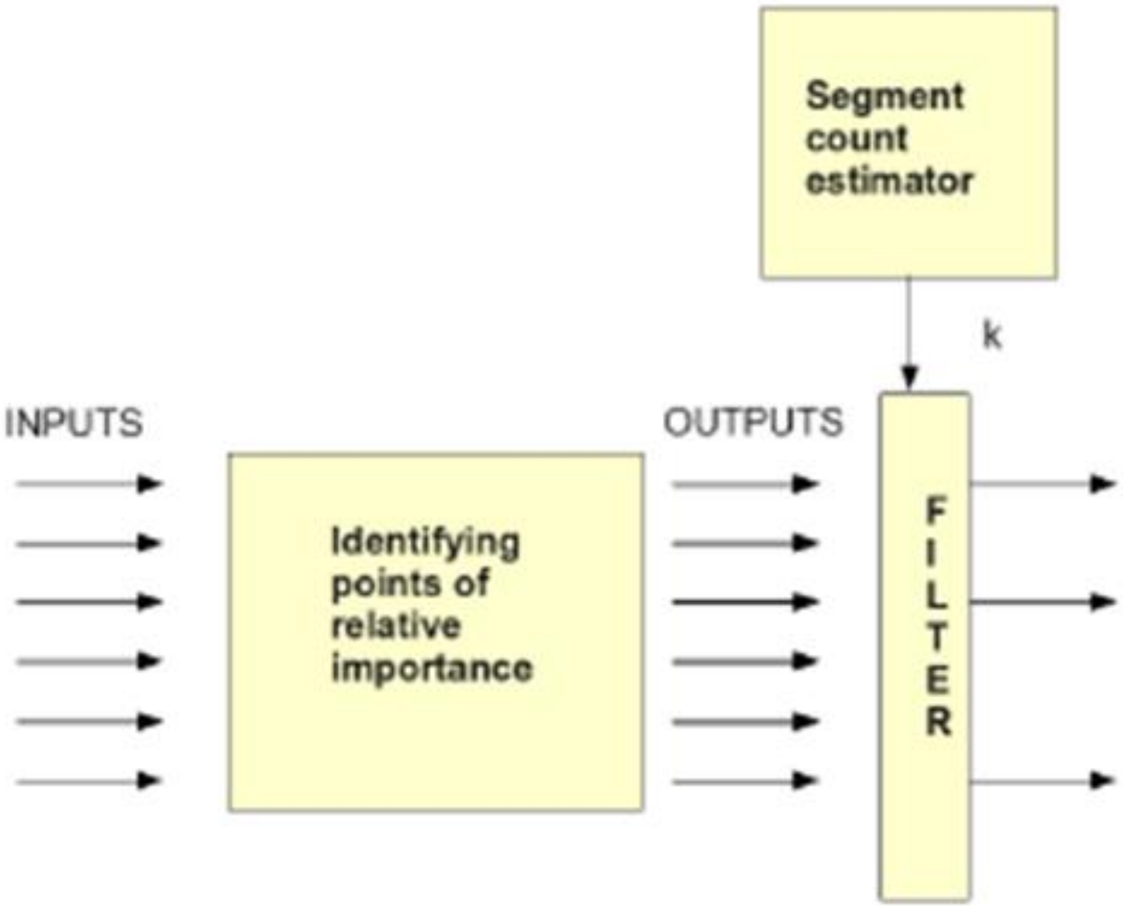


Figure : In figure, blue : original data; green : compressed data.

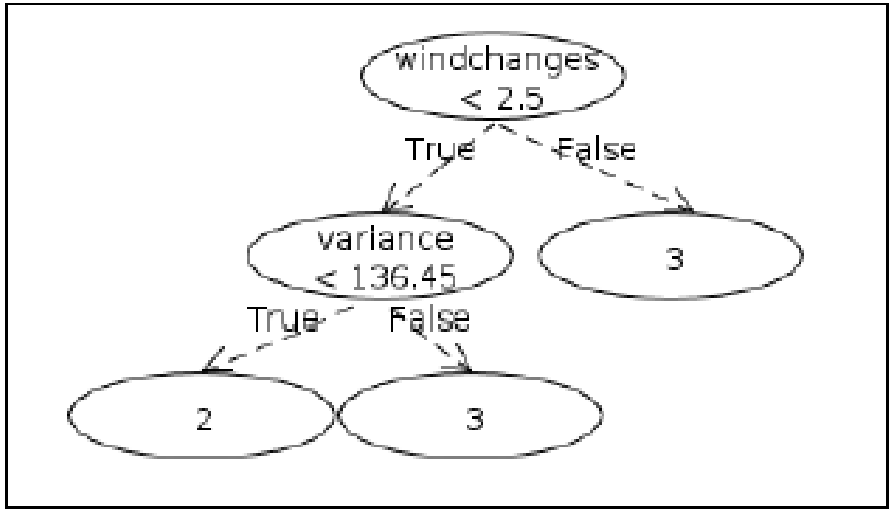
ML for NLG

- A set of simple rules can explain part of the manual generation process.
- Change in wind direction might be reported at high wind speeds but not necessarily if wind speed is slow
- An equal change in wind speed might be significant at higher speeds but not at lower speeds
- Channels such as wind speed and direction interact with each other

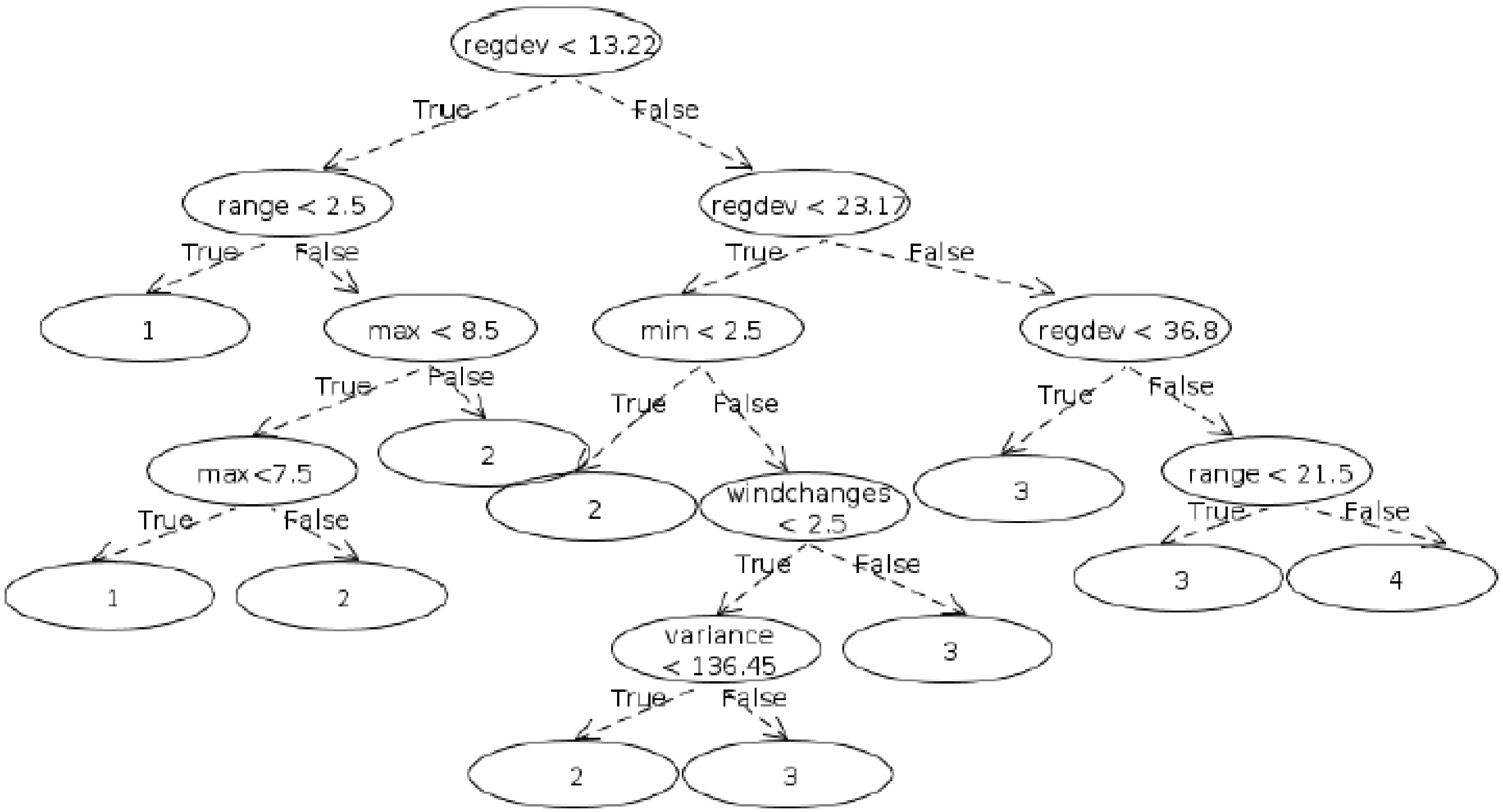
The overall scheme for Content Selection



An example decision tree output



A subtree depicting the coordination between wind speed and wind direction parameters.



A decision tree learned to estimate the number of segments.

Confusion matrix for lexical selection

FRESHENING	INC	RISING	EASING	DEC	FALLING	System/*Actual*
7	2	3	0	0	0	*FRESHENING*
27	37	24	1	0	0	*INCREASING*
9	10	14	0	1	0	*RISING*
4	0	0	62	19	20	*EASING*
0	0	0	20	6	5	*DECREASING*
0	0	0	4	3	10	*FALLING*

Temporal summaries

Test cycle 1: 20 defects raised, of which 5 were fixed, and 8 are high severity.

Test cycle 2: 15 were raised, of which 10 have been fixed and 4 are high severity.

Test cycle 3: 4 were raised, of which 10 have been fixed and 4 are high severity.

We have found that the defects raised are maximum for the first cycle and minimum for the 2nd.

The Trend is overall a decreasing trend.

Next Generation Text Generation ?

Acknowledgments

The entire team of Accenture collaborators/researchers (Sanjay, Harsha, Shubhashis, Roshni) for providing us the motivation

Dr. Yaji Sripada, University of Aberdeen (for sharing his wisdom and slides)

Ahmad Ibrahim (for Fuzzy Inferencing illustrations)

The IIT team: Prof. Deepak Khemani, Prof. Rupesh Nasre, Harshita Jhavar, Neha Dubey, Ravishekhar, Pranay

Sentence Planning Language

```
(S1/exist
 :object (O1/train
         :cardinality 20
         :relations ((R1/period
                      :value daily)
                     (R2/source
                      :value Aberdeen)
                     (R3/destination
                      :value Glasgow))))
```

An SPL Expression

Discourse Planning: Schema based approaches

a. Inform-Next-Train-Schema $\longrightarrow$
Sequence(Message:NUMBER-OF-TRAINS-IN-PERIOD,
Next-Train-Information-Schema)

b. Next-Train-Information-Schema $\longrightarrow$
Elaboration(Message:IDENTITY, Message:DEPARTURE)

Aggregation

a. there are 20 trains each day from Aberdeen to Glasgow
b. the next train is the Caledonian Express
c. the Caledonian Express leaves at 10am

There are 20 trains each day from Aberdeen to Glasgow. The next train is the Caledonian Express. It leaves at 10am.

a. There are 20 trains each day from Aberdeen to Glasgow. The next train is the Caledonian Express, which leaves at 10am.
b. There are 20 trains each day from Aberdeen to Glasgow. The next train, which leaves at 10am, is the Caledonian Express.

There are 20 trains each day from Aberdeen to Glasgow, and the next train is the Caledonian Express. It leaves at 10am.

There are 20 trains each day from Aberdeen to Glasgow, and the next train is the Caledonian Express, which leaves at 10am.

Aggregation

Simple conjunction

Ellipsis

Set formation

Embedding

			syntagmatic			
	The	cat	sat	on	the	mat.
paradigmatic	His	dog	slept	under	that	table.
	Our	parrot	perched	in	its	cage.

Appendix

NLG versus MailMerge

NLG is useful when

-- better quality text is needed
-- content and structure of text needs to be updated regularly

Important research groups, people, systems

Appendix

Gricean Maxims

- **Maxim of Quality**
  - **Be Truthful**
    - **Only say what you believe to be true.**
    - **Only say what you have evidence for.**
  - Example:
    A: Should I buy my son this new sports car?
    B: I don't know if that's such a good idea. His record isn't so great.
    vs.
    B: No, for he has totaled two cars since he got his license last year.

Gricean Maxims

- **Maxim of Quantity**
  - **Quantity of Information**
  - Make your contribution as informative as is required for the current purposes of the exchange.
  - Do not make your contribution more informative than is required.

  - Example:

  A: Where is the post office?

  B: Down the road, about 50 metres past the second left.

  vs.

  B: Not far.

Gricean Maxims

Maxim of Relation
Relevance
Make your contribution relevant to the interaction.
Indicate any way that it is not.

Example:
A: How are you doing in school?
B: Not too well, actually. I'm failing two of my classes.
vs.
B: What fine weather we're having lately!

Example 2:
A:(Waving at B, who is driving a taxi) Taxi!
B:(Waving at A, who is walking along the side of the road) Pedestrian!

Gricean Maxims

- **Maxim of Manner**
- **Be Clear**

  - Avoid unnecessary prolixity.
  - Avoid ambiguity.
  - Be brief.
  - Be orderly.

- Example:
  A: What did you think of that movie?
  B: I liked the creative storyline. The ending was really a surprise!
  vs.
  B: It was interestingly done, sir .

- Departure from Conventional Machine Learning : the column names are important – the domain model establishes relations between column names
- Spin : creative NLG : Glenn McGrath
- NLG as Machine Translation (parallel corpus)
- Requirements Analysis
- Complement graphics (HCI)
- Takeoff and touchdown
- Rhetorical Structures
- Surprises : Roshni's sheet
- The cricket commentary domain
- Fuzzy FCMs : capturing qualitative relationships between variables
- Lexicalization: fuzzy systems (lexical semantics)
- Pranay's slides on content selection and lexical semantics
- Mapping the Accenture domain to cricket domain

Narrative Science

- “WISCONSIN appears to be in the driver’s seat en route to a win, as it leads 51-10 after the third quarter. Wisconsin added to its lead when Russell Wilson found Jacob Pedersen for an eight-yard touchdown to make the score 44-3 ... . ”

Tradeoff between Interpretability and Accessability

Approaches to Lexical Selection

- When input information is represented using a KR scheme
  - Words/phrases are selected based on mappings defined in the knowledge base
  - In other words, language is grounded in the symbolic KR scheme
- When working with quantitative information
  - Language grounding information cannot be assumed to be available
- Two aspects of language grounding
  - NLG view
    - NLG maps parts of input information to words/phrases
    - Knowledge of such mapping relates to language grounding
  - User view
    - Reader of the output text should be able to associate the speaker intended meaning to the words/phrases used in the output text