

Textual Summarization of Time Series using Case-Based Reasoning

Neha Dubey

Guide: Dr. Sutanu Chakraborti, Prof. Deepak Khemani

Dept. of Computer Science & Engineering, IIT Madras

Table of contents

1. Introduction
2. Problem Definition
3. Time Series Summarization
 - Content Selection
 - Text Generation
4. Experiments
5. Conclusion

Natural Language Generation (NLG)

NLG systems are computer systems that can produce understandable texts in English from underlying nonlinguistic data sources

- Input
 - Image, Video, Time Series
- Output
 - Text

Natural Language Generation (NLG)

NLG systems are computer systems that can produce understandable texts in English from underlying nonlinguistic data sources

- Input
 - Image, Video, Time Series
- Output
 - Text

Time Series Summarization

Time series is a series of values of a quantity obtained at successive time instants

Time series summarization is the task of describing time series using natural language

Examples

- **Economy and Financial domain:** Stock market summary
- **Meteorology domain:** Weather report
- **Medical domain:** Summaries of patient information in clinical context

Time Series Summarization

Time series data: Complex, hard to interpret, multiple attributes to link

- Textual summary: Simpler, more effective, and understandable
- Less bandwidth requirement than graphical plots and videos
- Suitable for small screen devices

NLG Subtasks

- Content Selection
 - Deciding which information to include in the text under construction
- Microplanning
 - Finding right words and phrases to express information
- Linguistic Realisation
 - Combining all words and phrases in well formed sentences

NLG Subtasks : Time Series Summarization

- Content Selection
 - Extraction of relevant information from a time series (Abstraction of time series)
 - Interpretation of a time series: Requirement of domain expertise, for example, which spikes and trends are important to report in text
 - Relevance of information depends on end user
 - Same weather time series can have different textual summaries for marine forecast, public forecast, and for farmers

NLG Subtasks : Time Series Summarization

- Microplanning
 - Finding right words and phrases to express information
 - Increasing trend can be described using Increasing, gradually increasing, rising ?
- Linguistic Realisation
 - Combining all words and phrases in well formed sentence

- **Knowledge acquisition bottleneck**
 - Domain experts may not have procedural or algorithmic knowledge of how do they write summary
 - Might justify/explain the fluctuations in the summary based on his broader knowledge based on past experiences

Challenges

- **Variation in text:** People write differently!
 - For a given input, different writers will write differently (Idiosyncrasy)
 - For a given input, the same writer might write differently over a period of time
- **Evaluation of the system**
 - Multiple possible solution for a given input

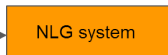
Problem: Time Series Summarization

Weather domain

Wind Time Series

Time	Wind Speed	Wind Direction
0600	4	S
0900	6	S
1200	12	S
1500	15	S
1800	18	S
2100	21	S
2400	18	S

Input



Text

S 02-06 increasing 16-20 by evening



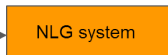
Problem: Time Series Summarization

Weather domain

Wind Time Series

Time	Wind Speed	Wind Direction
0600	4	S
0900	6	S
1200	12	S
1500	15	S
1800	18	S
2100	21	S
2400	18	S

Input



Text

S 02-06 increasing 16-20 by evening



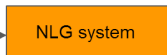
Problem: Time Series Summarization

Weather domain

Wind Time Series

Time	Wind Speed	Wind Direction
0600	4	S
0900	6	S
1200	12	S
1500	15	S
1800	18	S
2100	21	S
2400	18	S

Input

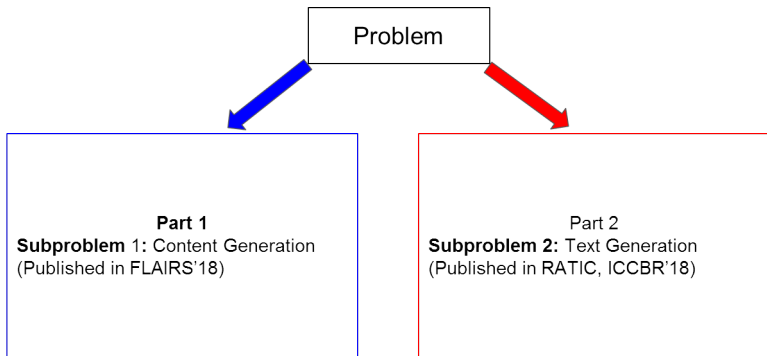


Text

S 02-06 increasing 16-20 by evening



Problem: Time Series Summarization



Subproblem 1 - Content Selection

Content Selection

- Selection of representative points from time series

Time	Wind Speed	Wind Direction
0600	4	S
0900	6	S
1200	12	S
1500	15	S
1800	18	S
2100	21	S
2400	18	S

Subproblem 1 - Content Selection

Content Selection

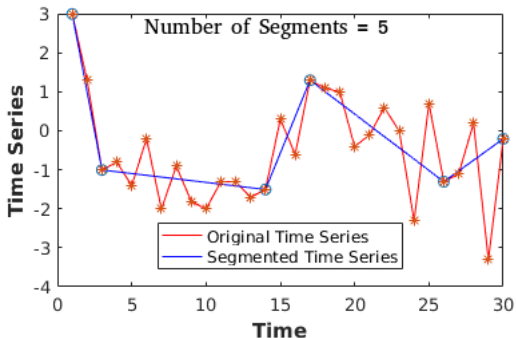
- Selection of representative points from time series
- Segmentation

Time	Wind Speed	Wind Direction
0600	4	S
0900	6	S
1200	12	S
1500	15	S
1800	18	S
2100	21	S
2400	18	S

Subproblem 1 - Content Selection

Segmentation

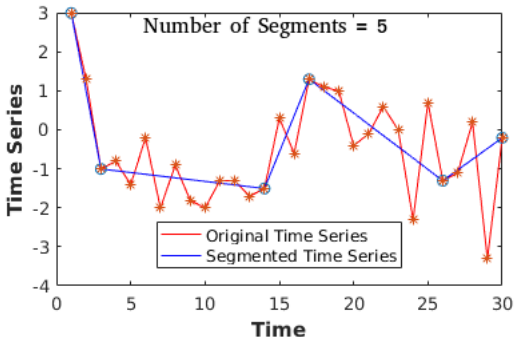
- Approximation of a time series with straight line segments



Subproblem 1 - Content Selection

Segmentation

- Approximation of a time series with straight line segments
- Requires number of segments as input



Our Approach

We propose a Case-based Reasoning (CBR) approach for content selection (learn number of segments) from time series

Our Approach - CBR

Case-based Reasoning (CBR) solves new problem by adapting previously successful solutions to similar problem

- Knowledge acquisition bottleneck

Our Approach - CBR

Case-based Reasoning (CBR) solves new problem by adapting previously successful solutions to similar problem

- Knowledge acquisition bottleneck
 - CBR does not require an explicit domain model and so elicitation becomes a task of gathering case histories (Past experiences)

Our Approach - CBR

Case-based Reasoning (CBR) solves new problem by adapting previously successful solutions to similar problem

- Knowledge acquisition bottleneck
 - CBR does not require an explicit domain model and so elicitation becomes a task of gathering case histories (Past experiences)
- Incremental learning

Our Approach - CBR

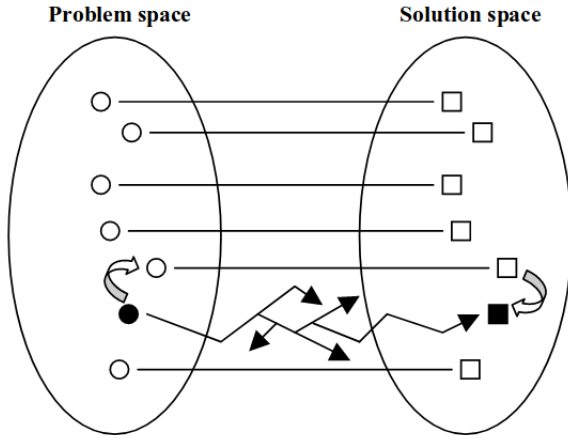
Case-based Reasoning (CBR) solves new problem by adapting previously successful solutions to similar problem

- Knowledge acquisition bottleneck
 - CBR does not require an explicit domain model and so elicitation becomes a task of gathering case histories (Past experiences)
- Incremental learning
- Inspired by human reasoning and memory organization

CBR Working Style

- Case retrieval: The best matching case is searched in the case base and an approximate solution is retrieved
- Case adaptation: The retrieved solution is adapted to fit the new problem
- Solution evaluation: The adapted solution can be evaluated either before the solution is applied to the problem or after the solution has been applied
- Case-base updating: If the solution was verified as correct, the new case may be added to the the case base

CBR Problem Solving



Acknowledgment:

<https://ibug.doc.ic.ac.uk/media/uploads/documents/courses/syllabus-CBR.pdf>

Hypothesis

Observation

Days with similar weather conditions have similar level of abstraction for wind time series

Hypothesis

Observation

Days with similar weather conditions have similar level of abstraction for wind time series

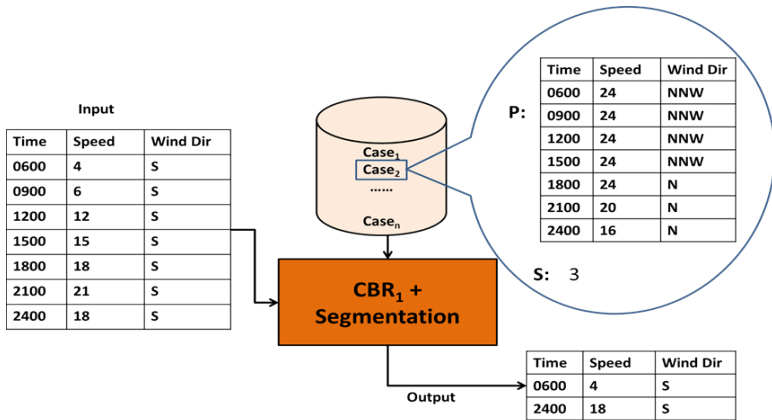
Hypothesis

Similar wind time series have the similar number of segments in the forecast text

Similarity measure for multivariate time series

- Interaction between channels like wind speed and wind direction
- Two time series that look very similar to a non-expert can have different interpretations

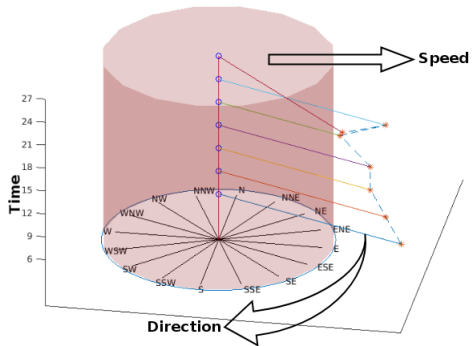
Our Approach: System Architecture



Our Approach: Case Representation

Case Representation

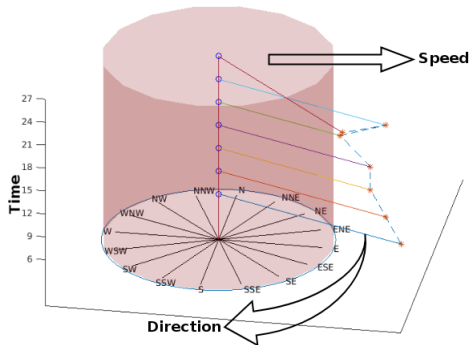
- Problem Component
 - Wind vector representation (Polar)



Our Approach: Case Representation

Case Representation

- Problem Component
 - Wind vector representation (Polar)
- Solution Component
 - Number of Segments



Proposed System

The System works in 3 steps:

- Retrieval of *similar* time series to input query time series
- Estimation of the segment count of query time series
- Segmentation of the query time series to get representative points

Proposed System: Retrieval

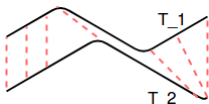
A case is relevant to a query if it has the patterns similar to that in the query time series and the similar error tolerated in approximating a case and the query

- Pattern matching
 - Dynamic Time Warping
- Error matching
 - *Error*_{Distance}

Proposed System: Retrieval

Dynamic Time Warping

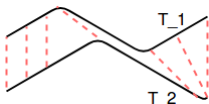
- Shape matching by aligning two time series by scaling/shrinking on time axis



Proposed System: Retrieval

Dynamic Time Warping

- Shape matching by aligning two time series by scaling/shrinking on time axis

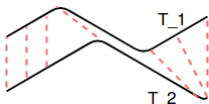


- $DTW(i, j) = \min\{DTW(i - 1, j), DTW(i, j - 1), DTW(i - 1, j - 1)\} + vector_{dist}(i, j)$

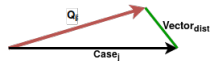
Proposed System: Retrieval

Dynamic Time Warping

- Shape matching by aligning two time series by scaling/shrinking on time axis



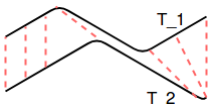
- $DTW(i, j) = \min\{DTW(i - 1, j), DTW(i, j - 1), DTW(i - 1, j - 1)\} + vector_{dist}(i, j)$



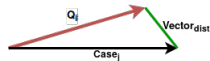
Proposed System: Retrieval

Dynamic Time Warping

- Shape matching by aligning two time series by scaling/shrinking on time axis



- $DTW(i, j) = \min\{DTW(i - 1, j), DTW(i, j - 1), DTW(i - 1, j - 1)\} + vector_{dist}(i, j)$



- $DTW_{dist}(Case, Query) = DTW(n, n)$; where $n = 7$

Proposed System: Retrieval

ErrorDistance

- $Case_{error}$ is the error tolerated in approximating a time series by an expert

Proposed System: Retrieval

ErrorDistance

- $Case_{error}$ is the error tolerated in approximating a time series by an expert
- $Query_{error}$ is the error in approximating a query time series similar to the case time series, i.e., with same segment number

Proposed System: Retrieval

*Error*_{Distance}

- *Case*_{error} is the error tolerated in approximating a time series by an expert
- *Query*_{error} is the error in approximating a query time series similar to the case time series, i.e., with same segment number

- *Error*_{distance} =
$$\left\{ \begin{array}{ll} \alpha * |case_{error} - query_{error}| & \text{if } case_{error} \geq query_{error} \\ \text{Case is relevant} & \alpha < 1 \\ \beta * |case_{error} - query_{error}| & case_{error} < query_{error} \\ \text{Case is irrelevant} & \beta > 1 \end{array} \right.$$

Proposed System: Retrieval

Final distance measure:

- $dist(Case, Query) =$
 $DTW_{dist}(Case, Query) + Error_{distance}(Case, Query)$

Proposed System: Retrieval

Final distance measure:

- $dist(Case, Query) = DTW_{dist}(Case, Query) + Error_{distance}(Case, Query)$
- $Similarity(Case, Query) = \frac{1}{1+dist(Case, Query)}$

Proposed System: Step 2, 3

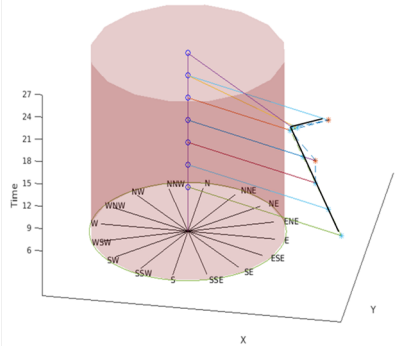
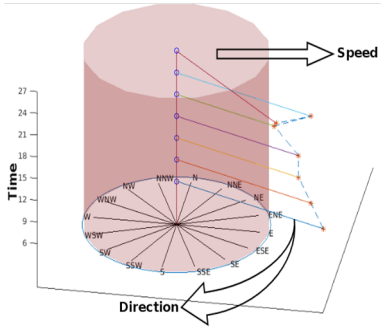
Estimation of segment count

- CBR regression problem

Selection of points (Segmentation of time series)

- Segment the time series by using optimal segmentation algorithm
 - Input to the algorithm is the estimated segment count

Proposed System: Step 2,3



Data Set: SUMTIME-MAUSAM parallel corpus of 1045 numerical weather data and human written forecasts

- **Evaluation measure**

Data Set: SUMTIME-MAUSAM parallel corpus of 1045 numerical weather data and human written forecasts

- **Evaluation measure**

- Accuracy of segment prediction (Classification)
- Error in segment prediction (Regression), K_{error}

Methods Compared:

- Elbow method: In the plot of error of approximation against varying number of segments for a time series, the elbow point is chosen as the number of segments

Methods Compared:

- Elbow method: In the plot of error of approximation against varying number of segments for a time series, the elbow point is chosen as the number of segments
- Decision tree: As a classification problem, where features extracted are minimum, maximum, range, end to end slope, regression error, standard deviation of speed and direction, respectively

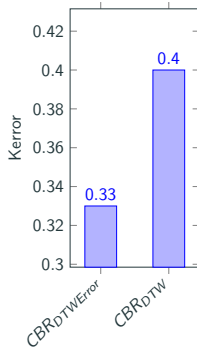
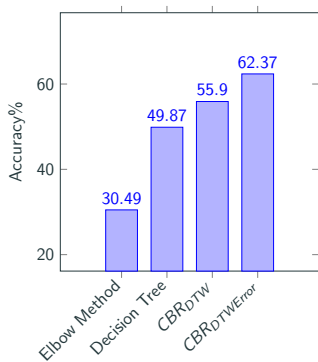
Methods Compared:

- Elbow method: In the plot of error of approximation against varying number of segments for a time series, the elbow point is chosen as the number of segments
- Decision tree: As a classification problem, where features extracted are minimum, maximum, range, end to end slope, regression error, standard deviation of speed and direction, respectively
- CBR system: Similarity based on DTW

Methods Compared:

- Elbow method: In the plot of error of approximation against varying number of segments for a time series, the elbow point is chosen as the number of segments
- Decision tree: As a classification problem, where features extracted are minimum, maximum, range, end to end slope, regression error, standard deviation of speed and direction, respectively
- CBR system: Similarity based on DTW
- CBR system: Similarity based on DTW and $Error_{distance}$

Results



Analysis

- Only shape matching is not enough!
 - If the change in wind happens at the end of day, forecasters try to ignore it

Analysis

- Only shape matching is not enough!
 - If the change in wind happens at the end of day, forecasters try to ignore it
- Among 38% of misclassified cases, 57% of the cases are consistently misclassified in all of the above methods
 - We suspect these cases need extra domain knowledge

Analysis

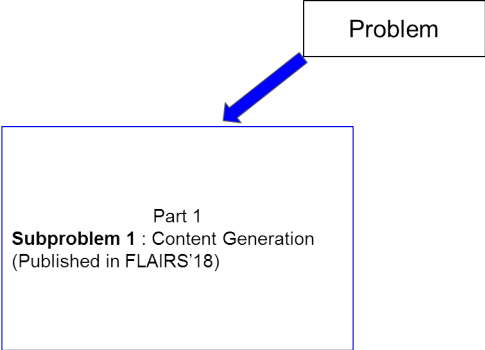
- Only shape matching is not enough!
 - If the change in wind happens at the end of day, forecasters try to ignore it
- Among 38% of misclassified cases, 57% of the cases are consistently misclassified in all of the above methods
 - We suspect these cases need extra domain knowledge
- Given a raw time series, it is often difficult even for a human to decide the exact segment count
 - For example, even a forecaster may find it hard to determine whether a time series should have 2 or 3 segments

- Only shape matching is not enough!
 - If the change in wind happens at the end of day, forecasters try to ignore it
- Among 38% of misclassified cases, 57% of the cases are consistently misclassified in all of the above methods
 - We suspect these cases need extra domain knowledge
- Given a raw time series, it is often difficult even for a human to decide the exact segment count
 - For example, even a forecaster may find it hard to determine whether a time series should have 2 or 3 segments

Conclusion Content selection algorithm is useful even when number of segments is not identical to that in the ground truth, but close to it

Story so far...

Problem



```
graph TD; Problem[Problem] --> Subproblem["Part 1  
Subproblem 1 : Content Generation  
(Published in FLAIRS'18)"]
```

Part 1

Subproblem 1 : Content Generation
(Published in FLAIRS'18)

The main issue in our system is evaluation

- Since we are evaluating at the intermediate level, there is no guarantee that the final generated text will be correct

The main issue in our system is evaluation

- Since we are evaluating at the intermediate level, there is no guarantee that the final generated text will be correct
- It is possible that

The main issue in our system is evaluation

- Since we are evaluating at the intermediate level, there is no guarantee that the final generated text will be correct
- It is possible that
 - Even if a time series have wrong segment count, the actual forecast text might be similar to the generated text

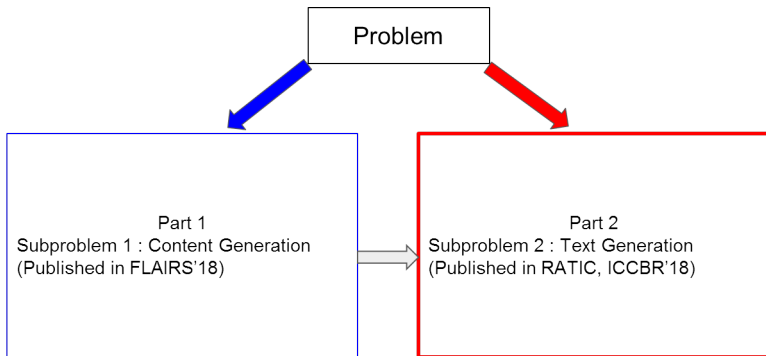
The main issue in our system is evaluation

- Since we are evaluating at the intermediate level, there is no guarantee that the final generated text will be correct
 - It is possible that
 - Even if a time series have wrong segment count, the actual forecast text might be similar to the generated text
- OR

The main issue in our system is evaluation

- Since we are evaluating at the intermediate level, there is no guarantee that the final generated text will be correct
- It is possible that
 - Even if a time series have wrong segment count, the actual forecast text might be similar to the generated text
OR
 - Even if the time series have correct segment count, but the points which are generated are not same as mentioned in actual text (our segmentation is not same as human-authored segmentation)

Way ahead...



Way ahead...

- **Knowledge acquisition bottleneck**
- **Variation in text:** People write differently!
- **Evaluation of the system**

Way ahead...

- **Knowledge acquisition bottleneck**
- **Variation in text:** People write differently!
- **Evaluation of the system**

Way ahead...

- **Knowledge acquisition bottleneck**
- **Variation in text:** People write differently!
- **Evaluation of the system**

We propose

- End-to-End CBR system for time series summarization in weather domain
- Multiple textual summaries to assist the user in decision making
- Evaluation measure adapted according to domain

Our Approach

Observation

Days with similar weather conditions have similar forecast text and the similar level of abstraction for wind time series

Our Approach

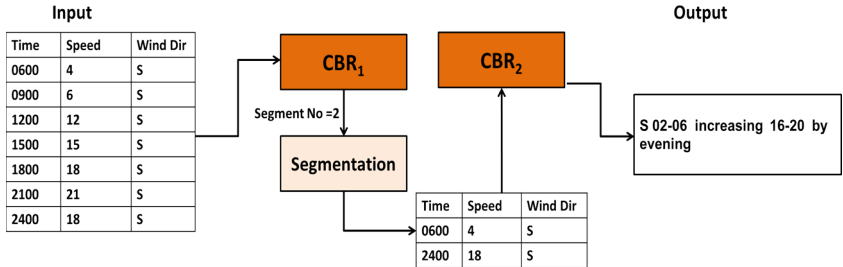
Observation

Days with similar weather conditions have similar forecast text and the similar level of abstraction for wind time series

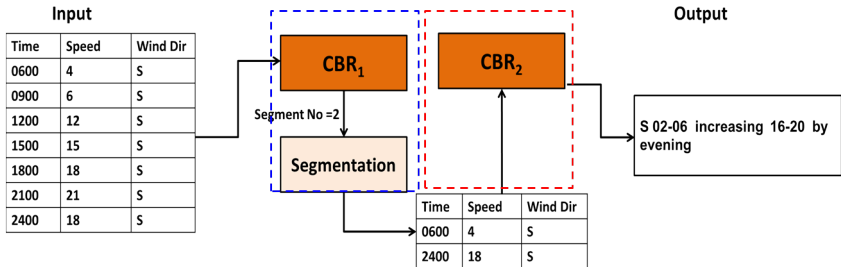
Hypothesis

Similar wind time series have the similar number of segments in the forecast text and hence similar forecast text

System Architecture

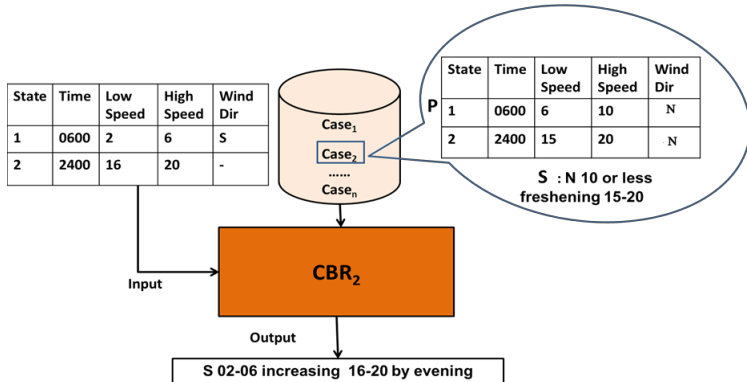


System Architecture



- - - - Content Generation
- - - - Text Generation

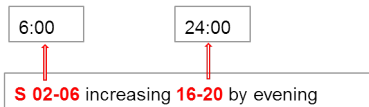
Text Generation: CBR₂



CBR₂: Case Representation

- Problem Component
 - Intermediate representation of time series as mentioned in textual summary

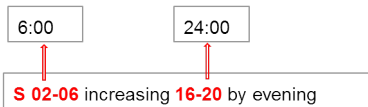
Time	Low wind speed	High Wind Speed	Direction
0600	2	6	S
2400	16	20	S



CBR₂: Case Representation

- Problem Component
 - Intermediate representation of time series as mentioned in textual summary
- Solution Component
 - Textual summary of time series

Time	Low wind speed	High Wind Speed	Direction
0600	2	6	S
2400	16	20	S



Proposed System

The proposed system works in 2 steps:

- Retrieval: Retrieve the most similar case to the query
- Case Reuse: Reuse the text of the retrieved case to generate the textual summary for the query

Proposed System: Step 1

Retrieval

1. Segment number matching

Proposed System: Step 1

Retrieval

1. Segment number matching
2. Pattern matching

Proposed System: Step 1

Retrieval

1. Segment number matching
2. Pattern matching
 - Speed pattern matching; Speed patterns: Increasing, Decreasing, Stable

Proposed System: Step 1

Retrieval

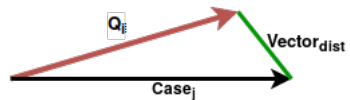
1. Segment number matching
2. Pattern matching
 - Speed pattern matching; Speed patterns: Increasing, Decreasing, Stable
 - Direction pattern matching; Direction patterns: Backing(clockwise), Veering (anti-clockwise), Stable

Proposed System: Step 1

Retrieval

1. Segment number matching
2. Pattern matching
 - Speed pattern matching; Speed patterns: Increasing, Decreasing, Stable
 - Direction pattern matching; Direction patterns: Backing(clockwise), Veering (anti-clockwise), Stable
3. Case with the minimum Euclidean distance with the query

$$\text{dist}(Q, C) = \sum_{i=1}^n \text{vectordist}(Q_i, C_i) / n,$$



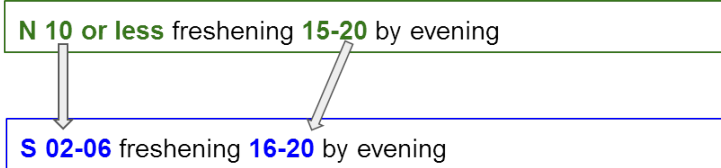
Case Reuse

- Retrieved case is reused to generate the final summary

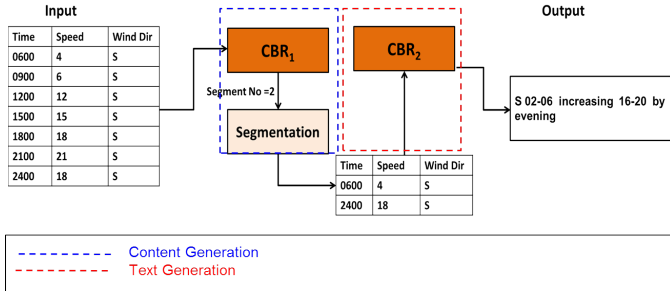
Proposed System: Step 2

Case Reuse

- Retrieved case is reused to generate the final summary

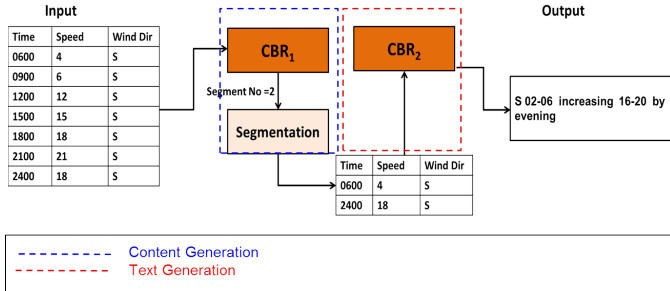


Content Selection Revisited



- Experts can form different views for a time series

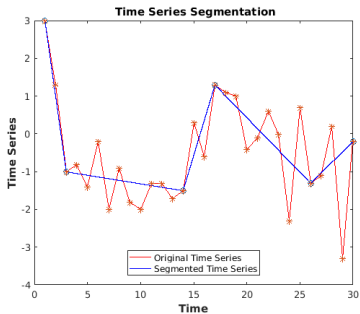
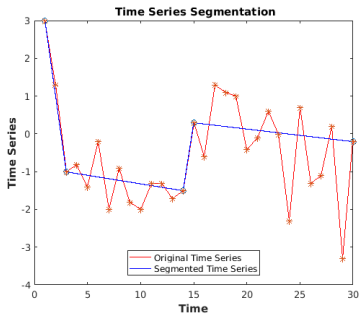
Content Selection Revisited



- Experts can form different views for a time series
 - Multiple representations of a query, each with a certain confidence value

Content Selection Revisited

Multiple representations of a query



Multiple representations of a query

- Confidence value for a query representation with k segments;

$$\text{confidence}(k) = \frac{\sum_{i=1}^m s_i * I(k_i=k)}{\sum_{i=1}^m s_i}$$

Multiple representations of a query

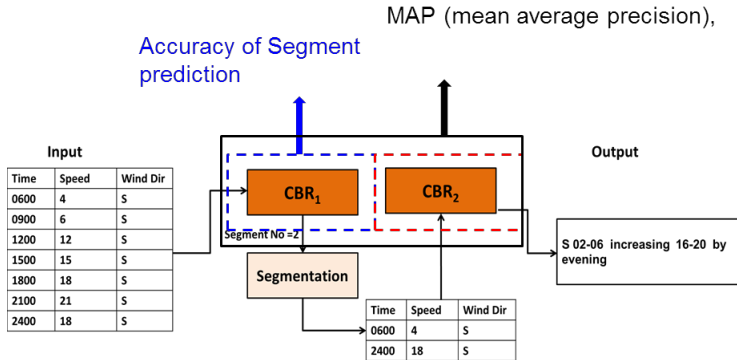
- Confidence value for a query representation with k segments;
$$\text{confidence}(k) = \frac{\sum_{i=1}^m s_i * I(k_i=k)}{\sum_{i=1}^m s_i}$$
- Confidence value associated with summary is
 - $\text{confidence}(k) * \text{similarity}(Q_k, \text{Case}_{CBR_2})$

- Change the similarity measure in the first module, i.e., CBR_1 to generate the intermediate representation of a query time series
 - The evaluation measure for CBR_1 is the accuracy of correctly predicting the number of segments

Experiment Design

- Change the similarity measure in the first module, i.e., CBR_1 to generate the intermediate representation of a query time series
 - The evaluation measure for CBR_1 is the accuracy of correctly predicting the number of segments
- Keep the configuration of CBR_1 fixed, change the output configuration of second CBR, i.e., CBR_2
 - System generates one textual summary
 - System generates a ranked list of summaries
 - The evaluation measures used here is Mean Average Precision (MAP)

Experiments



Formulation

Semantic similarity of generated text with ground truth text,

$$\text{sim}(text_{generated}, text_{groundtruth})$$

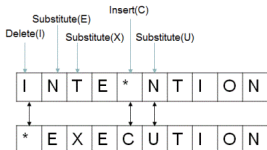
Formulation

Semantic similarity of generated text with ground truth text,

$sim(text_{generated}, text_{groundtruth})$

- Edit distance between two words,

$$L(i, j) = \min\{L(i-1, j) + 1, L(i, j-1) + 1, L(i-1, j-1) + c_l(a_i, b_j)\}$$



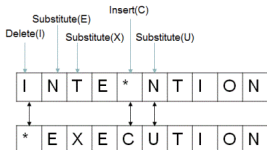
Formulation

Semantic similarity of generated text with ground truth text,

$sim(text_{generated}, text_{groundtruth})$

- Edit distance between two words,
- Edit distance between two sentences

$$L(i, j) = \min\{L(i-1, j) + 1, L(i, j-1) + 1, L(i-1, j-1) + c_l(a_i, b_j)\}$$



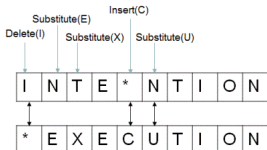
Formulation

Semantic similarity of generated text with ground truth text,

$sim(text_{generated}, text_{groundtruth})$

- Edit distance between two words,

$$L(i, j) = \min\{L(i-1, j) + 1, L(i, j-1) + 1, L(i-1, j-1) + c_l(a_i, b_j)\}$$



- Edit distance between two sentences
- $c_l(a_i, b_j) = 1 - sim(a_i, b_j)$; a_i, b_j are semantic units in the sentence
- Semantic units: speed, direction, speed-direction verb phrase, time phrase

Formulation

Similarity between the semantic units of text

- Speed and direction similarity: similarity is high if the generated value of speed or direction falls in the range given in ground truth summary

Formulation

Similarity between the semantic units of text

- Speed and direction similarity: similarity is high if the generated value of speed or direction falls in the range given in ground truth summary
- Patterns similarity (text variability)

Formulation

Similarity between the semantic units of text

- Speed and direction similarity: similarity is high if the generated value of speed or direction falls in the range given in ground truth summary
- Patterns similarity (text variability)
 - Synonym similarity

Word Phrase1\Word Phrase2	Increasing	Rising	Gradually Increasing
Increasing	1	0.8	0.8
Rising		1	0.8
Gradually Increasing			1

Formulation

Similarity between the semantic units of text

- Speed and direction similarity: similarity is high if the generated value of speed or direction falls in the range given in ground truth summary
- Patterns similarity (text variability)
 - Synonym similarity

Word Phrase1\Word Phrase2	Increasing	Rising	Gradually Increasing
Increasing	1	0.8	0.8
Rising		1	0.8
Gradually Increasing			1

- $sim(text_{generated}, text_{groundtruth}) = 1 - L(m, n) / \max(m, n)$

Evaluation of end-to-end CBR system

- Multiple generated summaries are ranked using the associated confidence values
 - In a ranked sequence of summaries, a summary is relevant if $\text{sim}(\text{text}_{\text{generated}}, \text{text}_{\text{groundtruth}}) > \text{Threshold}$

Evaluation of end-to-end CBR system

- Multiple generated summaries are ranked using the associated confidence values
 - In a ranked sequence of summaries, a summary is relevant if $sim(text_{generated}, text_{groundtruth}) > Threshold$
 - Evaluation measure used is Mean Average Precision (MAP)

Table 1: Result for various configurations of the CBR system

SI No.	CBR ₁ Configuration		CBR ₂ Configuration	
	Similarity	Accuracy of Segment Prediction	Single Output	Multiple Output
1	DTW	55.90	0.63	0.60
2	DTW + $Error_{distance}$	62.37	0.65	0.68

Analysis of generated textual summary with respect to the ground truth textual summary based on the level of abstraction for a time series, i.e., number of segments and the similarity with the ground truth summary

- *Estimated number of segments is same as of ground truth and the generated text is not similar to the ground truth summary*
 - **Observation:** Humans elide verb information and their text is shorter. Therefore, the correct level of abstraction, i.e., number of segments does not guarantee the high similarity of generated text with the ground truth text

Analysis

- *Estimated number of segments is not same as of ground truth and the generated text is similar to the ground truth summary*
- **Observation** People view time series differently!

Analysis

- *Estimated number of segments is not same as of ground truth and the generated text is similar to the ground truth summary*
 - **Observation** People view time series differently!
- *Estimated number of segments is not same as of ground truth and the generated text is similar to the ground truth*
 - **Observation** Cases are harder cases and need more domain knowledge!

General observations:

- Data analysis techniques to summarize time series need to be adapted according to the domain and end-user requirements
 - For example, in the medical domain, artefacts, anomalous spikes are more important, while in the weather domain, trends are more important
 - This knowledge can be integrated into various forms:
 - Estimation of number of segments using available data
 - Use of distributional measures like word2vec and wordnet on a large parallel corpus to get the synonyms similarity to evaluate the system

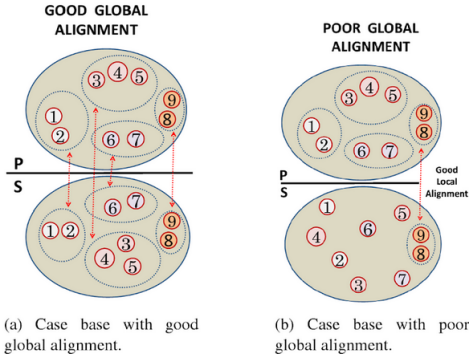
Way ahead...

Store some specific cases as exceptional cases in the case base

- **Identification of exceptional cases**
 - Casebase alignment

Way ahead...

Casebase alignment: extent to which similar problems in a given casebase correspond to similar solutions



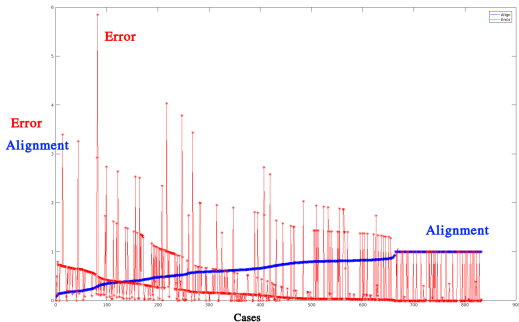
Hypothesis

Can higher error in text generation be associated with the poorly aligned region of casebase?

Casebase Alignment

Hypothesis

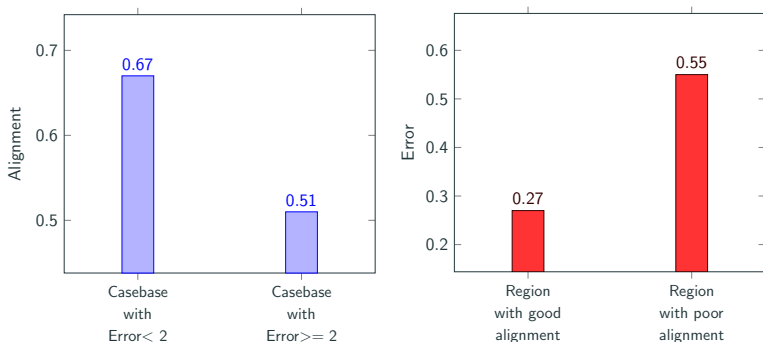
Can higher error in text generation be associated with the poorly aligned region of casebase?



Casebase Alignment

Hypothesis

Can higher error in text generation be associated with the poorly aligned region of casebase?



Expert intervention

Can we simulate an expert who decides whether a case is noisy case or an informative case?

Table 2: User interaction with the CBR system

CBR System	Performance Accuracy of Segment prediction	MAP			
		Single Output Relevance Threshold (0.3, 0.4)		Multiple Output Relevance Threshold (0.3, 0.4)	
Without noise removal	62.37	0.75	0.65	0.77	0.68
With noise removal (0.04% cases removed)	63.30	0.77	0.64	0.81	0.67

Conclusion

- End-to-end CBR System for time series summarization to assist the user in decision making
- Evaluated the system by using a proposed measure based on domain constraints
- Simulated expert intervention to improve the performance of the system using a measure called casebase alignment

Contributions

- Dubey, N.; Chakraborti, S.; and Khemani, D. 2018. “Content Selection for Time Series Summarization using Case-Based Reasoning”. In Proceedings of the Thirty First International Florida Artificial Intelligence Research Society Conference, FLAIRS 2018, Melbourne, Florida, USA. May 21-23 2018. 395398
- Dubey, N.; Chakraborti, S.; and Khemani, D. 2018. “Textual Summarization of Time Series using Case-Based Reasoning: A Case Study”. In Workshop Proceedings of the 26th International Conference on Case-Based Reasoning, ICCBR 2018, Stockholm, Sweden. July 09-12 2019. 164-174

Acknowledgement

- Dr. Rupesh Nasre, Dept. of Computer Science, IITM for insights on Segmentation algorithm
- G.Devi, Ditty Mathew, Harshita Jhavar, K.V.S. Dileep, Swapnil Hingmire, Abhijit Sahoo AIDB Lab, IITM for their feedbacks



I. Adeyanju.

Generating Weather Forecast Texts with Case based Reasoning.

International Journal of Computer Applications, 45(10):35–40, 2012.



E. Reiter.

An architecture for data-to-text systems.

In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, pages 97–104. Association for Computational Linguistics, 2007.



E. Reiter, S. Sripada, J. Hunter, J. Yu, and I. Davy.

Choosing words in computer-generated weather forecasts.

Artificial Intelligence, 167(1-2):137–169, 2005.



S. Sripada, E. Reiter, J. Hunter, and J. Yu.

Segmenting time series for weather forecasting.

Applications and Innovations in Intelligent Systems X, pages 105–118, 2002.



S. Sripada, E. Reiter, J. Hunter, and J. Yu.

Sumtime-meteo: Parallel corpus of naturally occurring forecast texts and weather data.

Computing Science Department, University of Aberdeen, Aberdeen, Scotland, Tech. Rep. AUCS/TR0201, 2002.

Thank you for listening!

Questions?

Segmentation

Input: Time Series \mathbb{T} , length N , Number of Segments K , Error function $E()$

Output: Piecewise linear approximation of a time series of length with K segments Seg_TS

```
for  $i = 1$  to  $N$  do
    | //initialize first Row  $A[1, i] = E(T(1, \dots, i))$  //Error when everything is
    |   in one cluster
end
for  $k = 1$  to  $K$  do
    | //Initialize diagonal  $A[k, k] = 0$ 
end
for  $k = 2$  to  $K$  do
    |   for  $i = k + 1$  to  $N$  do
    |       |  $A(k, i) = \min_{j < i} \{A[k - 1, j] + E(T[j + 1, \dots, i])\}$ 
    |   end
end
end
```

To recover original Segmentation, we store the minimizing values of j as well

Algorithm 1: Optimal Segmentation

Segmentation

$$E(S[1, n], k)$$

$$= \underbrace{\min_{k \leq j \leq n-1}}_{\text{Minimum over all possible placements of the last boundary point } b_{k-1}} \left\{ \underbrace{E(S[1, j], k-1)}_{\text{Error of optimal segmentation } S[1, j] \text{ with } k-1 \text{ segments}} + \underbrace{\sum_{j+1 \leq t \leq n} (t - \mu_{[j+1, n]})^2}_{\text{Error of } k\text{-th (last) segment when the last segment is } [j+1, n]} \right\}$$

Minimum over all possible placements of the last boundary point b_{k-1}

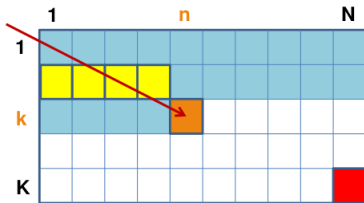
Error of optimal segmentation $S[1, j]$ with $k-1$ segments

Error of k -th (last) segment when the last segment is $[j+1, n]$

Segmentation

- Two-dimensional table $A[1 \dots K, 1 \dots N]$

$$A[k, n] = E(S[1, n], k)$$

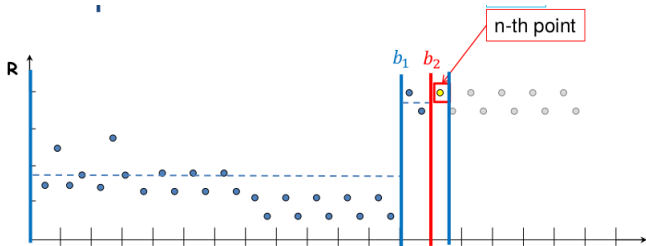


$$E(S[1, n], k) = \min_{k \leq j \leq n-1} \left\{ E(S[1, j], k-1) + \sum_{j+1 \leq t \leq n} (t - \mu_{[j+1, n]})^2 \right\}$$

- Fill the table top to bottom, left to right.

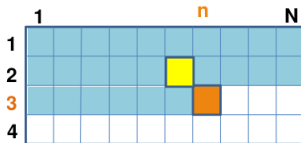
Error of optimal K-segmentation

Segmentation



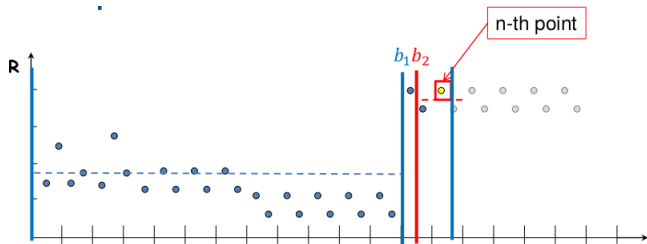
$$E(S[1, n], k)$$

$$= \min_{k \leq j \leq n-1} \left\{ E(S[1, j], k-1) \right. \\ \left. + \sum_{j+1 \leq t \leq n} (t - \mu_{[j+1, n]})^2 \right\}$$



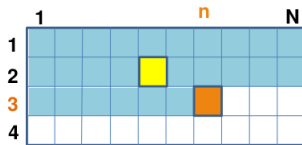
Where should we place boundary b_2 ?

Segmentation



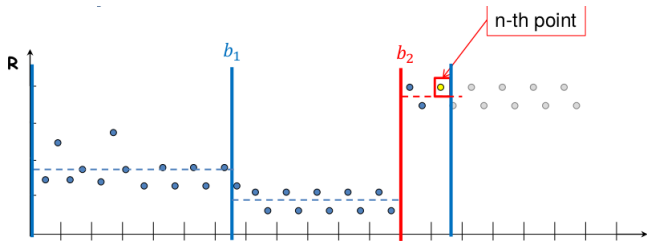
$$E(S[1, n], k)$$

$$= \min_{k \leq j \leq n-1} \left\{ E(S[1, j], k-1) \right. \\ \left. + \sum_{j+1 \leq t \leq n} (t - \mu_{[j+1, n]})^2 \right\}$$



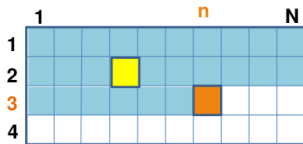
Where should we place boundary b_2 ?

Segmentation



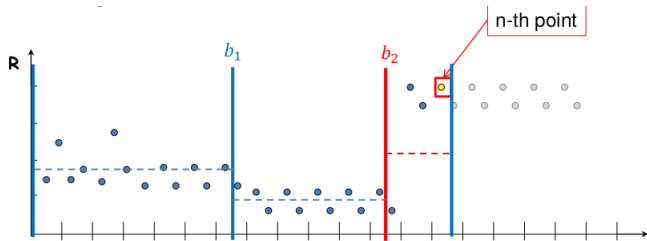
$$E(S[1, n], k)$$

$$= \min_{k \leq j \leq n-1} \left\{ E(S[1, j], k-1) + \sum_{j+1 \leq t \leq n} (t - \mu_{[j+1, n]})^2 \right\}$$



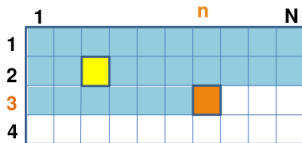
Where should we place boundary b_2 ?

Segmentation



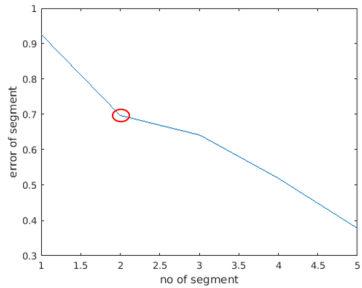
$$E(S[1, n], k)$$

$$= \min_{k \leq j \leq n-1} \left\{ E(S[1, j], k-1) + \sum_{j+1 \leq t \leq n} (t - \mu_{[j+1, n]})^2 \right\}$$



Where should we place boundary b_2 ?

Elbow method



Misclassified Cases

- Case 2: *SE 25-30 gusts 40 backing SE-ESE 20-25* and the generated text is *SSE 25.0 easing later to 23.0*
- Case 3: *NE-NNE 08-12, E-NE 06-12 backing NNE 08-13 later* are the two possible summaries for the same time series by two different authors
- Case 4: *SW 30-35 rising 38-42 by afternoon/evening and later veering W'LY 25-30* and the final text is *SW 31.0 veering W 26.0 in the evening*