Time Extraction from Real-time Generated Football Reports

Markus Borg
D02, Lund Institute of Technology
d02mbr@student.lth.se

Abstract

This paper describes a method to extract events and time information from football match reports generated through minute-by-minute reporting. By using regular expressions to find the events and dividing them into different types, one can determine in which order they occured. Time expressions are also detected and a way to structure the collected data using XML is presented.

1 Introduction

Football reports written in real-time is an increasingly popular way to report what happens during a football game. A reporter working on covering the match continually writes usually one or two sentences at a time. Whenever an interesting event happens (a goal scoring opportunity, an injury, a booking etc.), a brief description is presented and often the time is given. As the sentences are produced, people interested in what happens can for instance conveniently follow this on the Internet. Aftonbladet [1] is a Swedish newspaper that offers this service, the UEFA homepage [2] is an example in English with so called minute-by-minute reporting.

There are also other ways to present the report. People can subscribe to a game and get text messages directly to the mobile phone, the Swedish newspaper Helsingborgs Dagblad [3] provides this, but there are numerous other examples. Traditionally many viewers have also followed the latest results through teletext using a normal TV-set in a similar way.

By discovering the various events within the texts and analysing them, one can order them. With those chains of events found, this information could be submitted in addition to the text. It would then be possible to receive the information and import it to an application according to taste on an arbitrary platform. Some users maybe interested in viewing a short graphical version of the action on the display of the mobile phone, while others might want to collect statistical data on a PC.

My work has been inspired by the CarSim project [4], a system that converts textual descriptions of accidents into animated three-dimensional scenes. This paper presents a way to extract time information from one source of football reports generated as mentioned above. The texts I worked with come from an online football management game called Hattrick [5]. It is currently the biggest game of its kind with close to one million active players in January 2007. Every player takes on the role as manager for a team and plays one or two games each week, which results in a huge amount of available reports in the database. At this time, the reports are available in 39 langauges.

I have chosen these texts because of the availability and the fact that the variety of expressions fits very well to test the theory. The texts are not too simple however, since sentences are generated as results from 170 various events. Each event has on average five different wordings, resulting in a vocabulary suitable for this project. Hattrick offered vivid texts, but the time consuming work of teaching the system enough football related expressions was still limited. I used 25 different reports in the training set and the results I obtained while evaluating my work against texts annotated by hand were acceptable. An example of a match report is shown in the appendix.

There are some limitations of the project. Firstly, all texts I have worked with are in Swedish. Secondly, I only consider events that have to do with goal scoring opportunities. Thirdly, since the nature of real-time generated reports means that the events in the current sentence happen after the previously reported events in prior sentences, I only construct partial orderings. In this case it means I only look at chains of events within sentences.

In section 2 I describe how the output data is structured. Section 3 presents how the time expressions and events are found, section 4 shows how the links between those are determined. In section 5 I present the results of my evaluation. Finally, section 6 draws some conclusions and outlines directions for future work.

2 The output XML format

To get a useful information exchange, it is important to structure the data in a good way. For this task I have used a subset of TimeML [6] with some modifications. It is a robust specification language for events and temporal expressions in natural language. My system annotates absolute time expressions, events and time links to represent the necessary information.

The absolute time expressions are represented by TIMEX3 elements. Each element contains two attributes: tid (unique ID number) and type (so far always TIME). Example:

```
<TIMEX3 tid="t3" type="TIME">
den 19:e matchminuten</TIMEX3>
```

The various events are annotated as <EVENT> elements. Each element has three attributes: eid (unique ID number), class (OCCURENCE or STATE) and type (IDLEBALL, PREFINISH, FINISH, SAVE or OTHER). The elements of class STATE have type OTHER. The elements of class OCCURENCE have one of the other types, describing the event that took place on the field. Example:

```
<EVENT eid="e4" class="OCCURENCE"
type="FINISH">skalla in</EVENT>
```

The links between time expressions and events are represented by TLINK elements (time links). Links between a time expression and an event have the attributes: time (tid of the TIMEX3), event (eid of the EVENT) and type (DURING in all cases, all events during the same minute of the game get this). Links between two events have the attributes: sevent (eid of source event), tevent (tid of target event) and type (so far always BEFORE). This means that the source event happens before the target event. Example:

```
<TLINK sevent="e5" tevent="e6"
type="BEFORE"/>
```

The root node is <TimeML> and the first child is <Text>. This element has one <s> child for each sentence in the report. Every <s> element contains text nodes and possibly <TIMEX3> and <EVENT> elements. The <TLINK> elements, if present, follow after the <Text> element. The appendix shows an example of a complete annotated match report.

3 Absolute time expressions and events

The test application has been implemented in Java and heavily uses the package for regular expressions provided, which showed to be good enough for finding the expressions needed. In total, the regular expressions take less than 50 lines of code and there are probably many possible optimizations left to be done. The usage of regular expressions makes the program fast, the analysis of the text can be made without noticeable delay.

3.1 Finding time expressions

The program is only implemented to find absolute time expressions (for instance ”in the 16th minute”) in order to put those on a timeline. Relative time expressions (for example ”5 minutes later”) are not considered at all. Apparently, there seems to be few ways of expressing absolute time in Swedish football reports. Two lines of code were required to get a very good recall. An example of a regular expression follows:

```
(I|i) [0-9]+:e (match)?minuten
```

3.2 Finding events

Events on the football field are described in numerous ways to make the text interesting to the reader.

Every reporter has one personal style of writing and since the texts in Hattrick have developed during many years and different people have been involved, the finding of the events proved to be more of a challenge. The diversity of the football language used demanded about 45 lines of regular expressions and about 10 hours work. Three examples follow:

```
(reducera|kvittera) till
[0-9]+ - [0-9]+
(komma|tagit sig) igenom
drygade [\\w]+ ut (sin ledning|
ledningen till [0-9]+ - [0-9]+)
```

4 Time links

The insertion of time links is of course critical to this application. If not an indisputable majority of the time links are correct, the resulting output file cannot be considered useful. Links between time expressions and events are always inserted in the same way, but I have tried two different approaches between events.

4.1 Connecting time expressions and events

Since all absolute time expressions in the football reports I have observed have expressed a certain minute of the game and since the reports are generated in real-time, the following strategy is used. If the sentence contains a time expression, all events within the same sentence are considered to occur during this time. I have not encountered any examples contradicting this so far. Therefore, one TLINK of the type ”DURING” is added for every event in the sentence. Cases of multiple absolute time expressions within the same sentence have not been encountered and are not treated in special ways.

4.2 Ordering events

The fact that the text is generated in real-time, means that the later the sentence was written, the later the contained events happened. Consequently, the task is to find the chain of events within the actual sentence being processed. The chronological order of a football report written after the final whistle is much harder to determine since this property seldom is the case.

In this project, it was assumed that the events involved with goal scoring opportunities always could be ordered in a linear fashion. If it is said that a striker scores a goal and the team got the equalizer, then the goal is considered to happen before the result changes. Other approaches could be used according to taste, but here those events are not thought as simultaneous and the time links are inserted accordingly. The first event is given a time link to the next one and so on until the last event has been reached.

The basic assumption used to implement the ordering was that the different events within a sentence appear in the text in the same order as they happened during the game. This very simple approach is used as the baseline in the evaluation in section 5.

The second strategy implemented, instead divides the events into six different types. The types are as follows:

#	*Type*	*Example*
0	RESULT-CHANGE	Team taking the lead, scored another goal etc.
1	SAVE	Keeper saves, defender blocks etc.
2	FINISH	Shots, touches etc. towards the goal
3	PREFINISH	Passes, crosses, rushes etc.
4	IDLEBALL	Set pieces, keeper throwing the ball etc.
5	OTHER	All events of the class ”STATE”

The types are then considered to always follow a certain order regarding each other, the types given high numbers happen before the lower ones. If multiple events of the same type are present, they get time links in the same order as they appear in the text.

5 Results

This section contains the results of an evaluation of the system, aimed at testing the recall and precision of the regular expressions used. The two different strategies of inserting time links were also tested. Since the size of the experiment is small, the results can only be taken as suggestive.

To make my application able to handle enough football related expressions, I used 25 different texts in the training set. The following composition of reports was used: 8 reports from league games in higher divisions, 7 reports from leagues in the middle, 5 reports from lower divisions and 5 reports from matches between national teams. This should ensure that reports from teams of various levels are covered by the system.

My test set contained 3 reports from different teams. I selected reports from matches with 4 goals, to be certain that enough goal scoring opportunities were described in the text. Then I annotated the texts by hand in what I consider to be the correct way, by finding all expressions and detecting the correct order of events. In the end I compared my results with the output from the system.

I started by looking at the absolute time expressions. This proved to be an easy task and the system found all of them. I suspected this early while working with the training set, absolute time appears to be expressed in limited ways in football reports.

As the next step I measured how many of the events were found. The reports in total contained 53 and my system reached a recall of 79.4% with a precision of 87.5%. The recall level could be increased simply by adding more texts to the training set. The precision found however, was lower than expected and further analysis showed that some mistakes were repeatedly made. Some key words the system is looking for are used in various situations. A good example of this is the word ”hörna” (Eng. ”corner”), which in Swedish is used both for the actual corner kick and when defenders or keepers save the ball by redirecting it and it passes the short line, resulting in a corner kick for the attacking team. With this in mind, getting an almost perfect precision would be possible. One way could be to first test if it appears after an event of the type ”FINISH” or not.

Apparently, without considering parts of speech or other language characteristics, it was possible to quickly get an acceptable recall with a system entirely based on regular expressions.

The final part of the evaluation was about testing if my ideas of dividing events into certain types gave a better ordering. The three reports contained 18 sentences with multiple events, in total 47 events, suggesting that they seldom are alone in a sentence in a football report. They were divided the following way: 12 sentences had 2 events, 1 sentence had 3 events and 5 sentences had 4 events. 5 of the events were wrongly detected, since the system treated some of the single events as two.

If the additionally found events were disregarded altogether, the baseline produced correct time links for 12 of the 18 sentences (66.7%). The strategy with the types giva a correct output between all the remaining events.

The additional events do not necessarily have to be disregarded however, since they can be assumed to happen after the core event they were derived from. With this assumption, the result is as follows: the baseline still produced the correct result for 12 sentences (66.7%). The more complex strategy produced correct time links for 15 sentences (83.3%).

The result of the baseline shows that the events in a football report cannot be considered to happen in the same way as they appear within a sentence. We can also conclude that dividing events into those different types and assuming that passes happen before shots etc., gives a better result. The failed time links are in this evaluation produced because of failed event detection. Since some additional set pieces were introduced, they were treated as the starts of the event chains. Examples of this were shots from the penalty area (treated like penalty kicks) and the issue of corners as previously described. Still, the more complex strategy gave a significant increase in producing correct time links.

6 Conclusions

This paper described a way to extract time information from football reports, generated in real-time by the game engine in Hattrick. The evaluation of the system showed that if a sentence contains events, there are usually more than one. Those events cannot be expected to have happened in the same order as they appeared within a sentence written in Swedish. Although the limited set of data prevents any firm conclusions, the work indicates that regular expressions together with type divided events can produce output well describing events on a football field. The methods should be possible to apply also on other domains with a somewhat limited vocabulary.

Further extensions could be to include also other types of events like injuries and substitutions, but I think that goals are more interesting to focus on at this stage. I also think it would make sense to add information about whether something actually happened or not, since this version of the system does not differentiate between ”had a chance to shoot but did not” and ”came through and shot”. Both shots would now be treated as the same ”FINISH” type.

The next step to make the system more robust could be to include a part of speech tagger. However, I consider the system already good enough to be tested for simple visualization purposes of Hattrick reports.

7 Acknowledgements

I would like to thank my supervisor Richard Johansson for his guidance during this project and Christian Henriksson, Language Administrator at Hattrick, for explaining internals of the report system and providing me with figures.

References

1 *Aftonbladet*
www.aftonbladet.se

2 *UEFA*
www.uefa.com

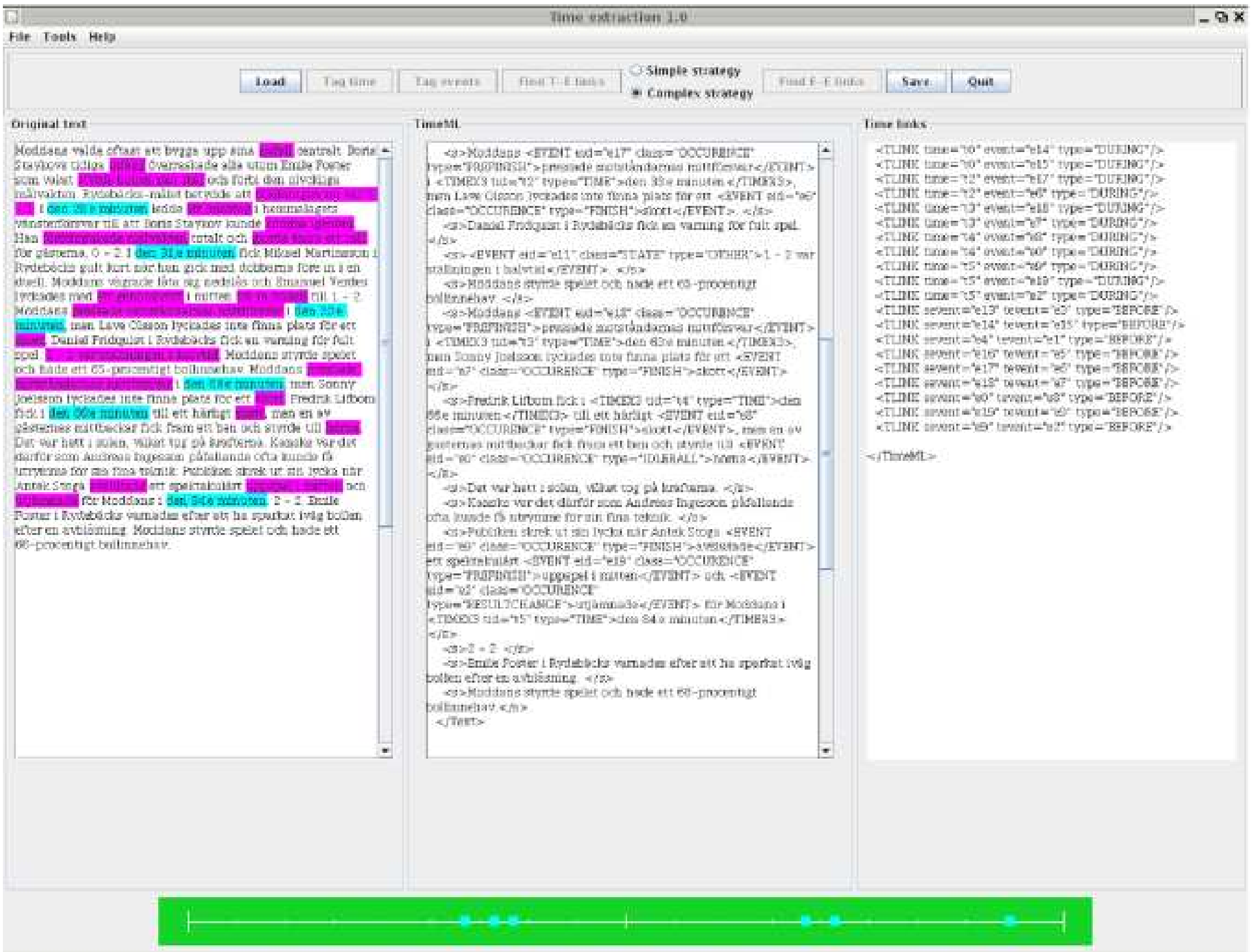
3 *Helsingborgs Dagblad*
www.hd.se

4 *The CarSim Project*
www.lucas.lth.se/lt/carsim.shtml

5 *Hattrick*
www.hattrick.org

6 *TimeML Specification Language*
www.timeml.org

A The user interface



Moddans valde oftast att bygga upp sina [illegible] centralt. Boris Stavkovs tidiga [illegible] överraskade alla utom Emile Foster som valde [illegible] och förde den olyckliga målvakten. Rydebäcks-målet berodde att [illegible] i den 31:a minuten ledde [illegible] i hemmalagets vänsterforward till att Boris Stavkov kunde [illegible]. Han [illegible] totalt och [illegible] för gästerna. 0 – 2. I den 31:a minuten fick Mikael Martinsson i Rydebäcks gult kort när han gick med dubbarna före in i en duell. Moddans vägrade låta sig nedslås och Emanuel Vendes lyckades med [illegible] i mitten [illegible] till 1 – 2. Moddans pressade motståndarnas mittförsvar i den 33:e minuten, men Lave Olsson lyckades inte finna plats för ett skott. Daniel Fridquist i Rydebäcks fick en varning för fult spel. 1 – 2 var ställningen i halvtid. Moddans styrde spelet och hade ett 65-procentigt bollinnehav. Moddans pressade motståndarnas mittförsvar i den 63:e minuten, men Sonny Joelsson lyckades inte finna plats för ett skott. Fredrik Lifbom fick i den 66:e minuten till ett härligt skott, men en av gästernas mittbackar fick fram ett ben och styrde till hörna. Det var hett i solen, vilket tog på krafterna. Kanske var det därför som Andreas Ingesson påfallande ofta kunde få utrymme för sin fina teknik. Publiken skrek ut sin lycka när Antek Stoga avslutade ett spektakulärt uppspel i mitten och utjämnade för Moddans i den 84:e minuten. 2 – 2. Emile Foster i Rydebäcks varnades efter att ha sparkat iväg bollen efter en avblåsning. Moddans styrde spelet och hade ett 68-procentigt bollinnehav.

```
    <s>Moddans <EVENT eid="e17" class="OCCURRENCE" type="PREFINISH">pressade motståndarnas mittförsvar</EVENT> i <TIMEX3 tid="t2" type="TIME">den 33:e minuten</TIMEX3>, men Lave Olsson lyckades inte finna plats för ett <EVENT eid="e6" class="OCCURRENCE" type="FINISH">skott</EVENT>. </s>
    <s>Daniel Fridquist i Rydebäcks fick en varning för fult spel. </s>
    <s><EVENT eid="e11" class="STATE" type="OTHER">1 – 2 var ställningen i halvtid</EVENT>. </s>
    <s>Moddans styrde spelet och hade ett 65-procentigt bollinnehav. </s>
    <s>Moddans <EVENT eid="e18" class="OCCURRENCE" type="PREFINISH">pressade motståndarnas mittförsvar</EVENT> i <TIMEX3 tid="t3" type="TIME">den 63:e minuten</TIMEX3>, men Sonny Joelsson lyckades inte finna plats för ett <EVENT eid="e7" class="OCCURRENCE" type="FINISH">skott</EVENT>. </s>
    <s>Fredrik Lifbom fick i <TIMEX3 tid="t4" type="TIME">den 66:e minuten</TIMEX3> till ett härligt <EVENT eid="e8" class="OCCURRENCE" type="FINISH">skott</EVENT>, men en av gästernas mittbackar fick fram ett ben och styrde till <EVENT eid="e0" class="OCCURRENCE" type="IDLEBALL">hörna</EVENT>. </s>
    <s>Det var hett i solen, vilket tog på krafterna. </s>
    <s>Kanske var det därför som Andreas Ingesson påfallande ofta kunde få utrymme för sin fina teknik. </s>
    <s>Publiken skrek ut sin lycka när Antek Stoga <EVENT eid="e9" class="OCCURRENCE" type="FINISH">avslutade</EVENT> ett spektakulärt <EVENT eid="e19" class="OCCURRENCE" type="PREFINISH">uppspel i mitten</EVENT> och <EVENT eid="e2" class="OCCURRENCE" type="RESULTCHANGE">utjämnade</EVENT> för Moddans i <TIMEX3 tid="t5" type="TIME">den 84:e minuten</TIMEX3>. </s>
    <s>2 – 2. </s>
    <s>Emile Foster i Rydebäcks varnades efter att ha sparkat iväg bollen efter en avblåsning. </s>
    <s>Moddans styrde spelet och hade ett 68-procentigt bollinnehav. </s>
  </Text>
```

```
<TLINK time="t0" event="e14" type="DURING"/>
<TLINK time="t0" event="e15" type="DURING"/>
<TLINK time="t2" event="e17" type="DURING"/>
<TLINK time="t2" event="e6" type="DURING"/>
<TLINK time="t3" event="e18" type="DURING"/>
<TLINK time="t3" event="e7" type="DURING"/>
<TLINK time="t4" event="e8" type="DURING"/>
<TLINK time="t4" event="e0" type="DURING"/>
<TLINK time="t5" event="e9" type="DURING"/>
<TLINK time="t5" event="e19" type="DURING"/>
<TLINK time="t5" event="e2" type="DURING"/>
<TLINK sevent="e13" tevent="e3" type="BEFORE"/>
<TLINK sevent="e14" tevent="e15" type="BEFORE"/>
<TLINK sevent="e4" tevent="e1" type="BEFORE"/>
<TLINK sevent="e16" tevent="e5" type="BEFORE"/>
<TLINK sevent="e17" tevent="e6" type="BEFORE"/>
<TLINK sevent="e18" tevent="e7" type="BEFORE"/>
<TLINK sevent="e0" tevent="e8" type="BEFORE"/>
<TLINK sevent="e19" tevent="e9" type="BEFORE"/>
<TLINK sevent="e9" tevent="e2" type="BEFORE"/>

</TimeML>
```

B Match report in Swedish

36171 åskådare hade kommit till Rydebäcks A-plan denna molniga matchdag. Rydebäcks hade valt att spela med en 3-5-2-uppställning. De ställde upp såhär påplanen: Lundman - Olenfeldt, Fery, van der Meijden - Rodrguez, Lystad, Martinsson, Jullien, Hörnsten - Fridquist, Evora.

Nynäshamns hade valt att spela med en 3-5-2-uppställning. Följande hade fått förtroendet: Jalajamani - Veselinovic, Kaltenbach, Konovalov - Ivarsson, Johansson, Gunkel, Remmel, Sörensson - Thestrup, Klöckner.

Efter 18 minuters spel bröt jublet lös dåNicolas Jullien kom igenom gästernas mittförsvar och dundrade in 1 - 0 för Rydebäcks. Daniel Fridquist i Rydebäcks tilldelades efter 20 minuter gult kort för osportsligt uppträdande. I den 22:e matchminuten fick gästernas mittförsvar se sig rundat av Mikael Martinsson som slog in 2 - 0 för Rydebäcks. I den 26:e minuten fick sten Sörensson i Nynäshamns gult kort när han gick med dobbarna före in i en duell. Rydebäcks tvingades samtidigt till ett byte eftersom John Hörnsten inte kunde fortsätta efter den omilda behandlingen. Alex Lunenburg fick kliva in i hans position. 2 - 0 var ställningen i halvtid. Halvleken dominerades av Rydebäcks som övertygade med ett 55-procentigt bollinnehav.

Storspel av målvakten Chittesh Jalajamani i den 61:e minuten räddade gästerna kvar i matchen, när Inge Olenfeldt fick påen riktig kanon efter ett vänsteranfall. Helge Lystad drygade ut ledningen för Rydebäcks till 3 - 0 genom att hålla sig framme påett skruvat högerinlägg. Nu ändrade matchen karaktär en smula, dåRydebäcks drog sig tillbaka för att bemöta motståndarna påegen planhalva. Dario van der Meijden betedde sig som en knattelagsspelare när han i den 68:e minuten sjabblade bort bollen till en motspelare, som dock missade friläget. Mest av spelet hade Rydebäcks med ett bollinnehav på 55 procent.

Den dominerande hos Rydebäcks var tveklöst Helge Lystad. Inge Olenfeldt hade däremot ingen lyckad dag. Den dominerande hos Nynäshamns var tveklöst Florian Remmel. Alexander Ivarsson hade däremot ingen lyckad dag. Matchen slutar 3 - 0.

C Match report in English

36171 spectators had come to Rydebäcks A-plan this cloudy day. Rydebäcks had chosen a strategic 3-5-2 formation. They fielded: Lundman - Olenfeldt, Fery, van der Meijden - Rodrguez, Lystad, Martinsson, Jullien, Hörnsten - Fridquist, Evora.

Nynäshamns had chosen a strategic 3-5-2 formation. The following players had been chosen: Jalajamani - Veselinovic, Kaltenbach, Konovalov - Ivarsson, Johansson, Gunkel, Remmel, Sörensson - Thestrup, Klöckner.

In the 18th minute cheers broke out as Nicolas Jullien found his way through the guests' central defence, clipping the 1 - 0 goal in for Rydebäcks. Daniel Fridquist of Rydebäcks received a yellow card in the 20th minute for unsportsmanlike behaviour. In the 22nd minute of the match, the visitors' central line of defence had to look on as Mikael Martinsson dashed through, knocking home 2 - 0 for Rydebäcks. In the 26th minute, Nynäshamns's sten Sörensson received a yellow card for going into a challenge studs first. Rydebäcks were forced to a substitution as John Hörnsten couldn't continue playing due to the rough treatment, forcing Alex Lunenburg to come in from the sidelines. 2 - 0 was the halftime score. The forty-five minutes were dominated by Rydebäcks, with an impressive 55 percent possession of the ball.

A great save by keeper Chittesh Jalajamani in the 61st minute kept the visitors in the game after Inge Olenfeldt struck from the left with a real cannonball. Helge Lystad increased Rydebäcks's lead to 3 - 0, putting a header away on a hooked ball from the right. The structure of the game started to change as Rydebäcks decided to pull back and meet their opponents in their own half. Dario van der Meijden behaved like an inexperienced youth player in the 68th minute as he gave the ball away to an opponent. Lucky for him though, there was no goal. Rydebäcks, bringing their ball possession to 55 percent, dominated the battle.

The most dominating Rydebäcks player was without a doubt Helge Lystad. Inge Olenfeldt on the other hand, had a terrible day. The most dominating Nynäshamns player was without a doubt Florian Remmel. Alexander Ivarsson on the other hand, had a terrible day. The match ends 3 - 0.

D Example of XML output

```
<TimeML>

  <Text>
    <s>Efter <TIMEX3 tid="t1" type="TIME">7 minuters spel</TIMEX3>
    blev publiken som galen efter att Mats Aronsson <EVENT eid="e1"
    class="OCCURENCE" type="PREFINISH">kom igenom bortalagets
    backlinje</EVENT> och <EVENT eid="e2" class="OCCURENCE"
    type="FINISH">dundrade in</EVENT> 1 - 0 fr Fortuna. </s>
    <s>Daniel Malmsten i Fortuna tilldelades <TIMEX3 tid="t1"
    type="TIME">efter 12 minuter</TIMEX3> gult kort efter farligt
    spel. </s>
    <s>I <TIMEX3 tid="t2" type="TIME">den 25:e matchminuten</TIMEX3>
    fick bortalagets mitt<EVENT eid="e4" class="OCCURENCE" type="PREFINISH">
    forsvar se sig rundat</EVENT> av Jonas Storm som <EVENT eid="e3"
    class="OCCURENCE" type="FINISH">slog in</EVENT> 2 - 0 fr Fortuna. </s>
    <s>I <TIMEX3 tid="t1" type="TIME">den 29:e minuten</TIMEX3> fick
    John Evans i Klippan gult kort efter en vansinnig tackling. </s>
    <s>Fortuna tvingades samtidigt till ett byte eftersom Stefan Blomdahl
    inte kunde spela vidare efter den omilda behandlingen. </s>
    <s>Dieter Fieback fick kliva in i hans position. </s>
    <s><EVENT eid="e5" class="STATE" type="OTHER">
    Det stod 2 - 0 i pausvilan</EVENT>. </s>
    <s>Halvleken dominerades av Fortuna som vertygade med ett
    55-procentigt bollinnehav.
  </Text>

  <TLINK time="t1" event="e2" type="DURING"/>
  <TLINK time="t1" event="e2" type="DURING"/>
  <TLINK time="t2" event="e4" type="DURING"/>
  <TLINK time="t2" event="e3" type="DURING"/>
  <TLINK sevent="e1" tevent="e2" type="BEFORE"/>
  <TLINK sevent="e4" tevent="e3" type="BEFORE"/>

</TimeML>
```