PARADIME

PARADIME

Parametrizable Domain-Adaptive Information and Message Extraction

Adapting the SMES System to a New Domain

Günter Neumann and Thierry Declerck

Goals of the PARADIME Project

Development of core technologies for Information Extraction (IE) allowing a fast and easy configuration for adapting the SMES system to new domains.

In order to support this task the project went for a systematic separation between the Natural Language Processing (NLP) components (dealing with the general linguistic knowledge) and the domain modeling components (handling the domain specific knowledge) and defined an interface between those two main modules:

The general linguistic processing is realized by a set of integrated NLP tools for chunk and shallow parsing.

The domain model is described in form of hierarchically organized abstract (uninstantiated) templates, declaratively defined within the Type Description Language (TDL), on the base of which inferences can be drawn.

The interface consists in a set of *linking types* defining a (partial) merging of the data types of the two main modules. A lookup in a domain lexicon helps selecting the type of templates to be filled by the particular IE task with the results of the NL analysis.

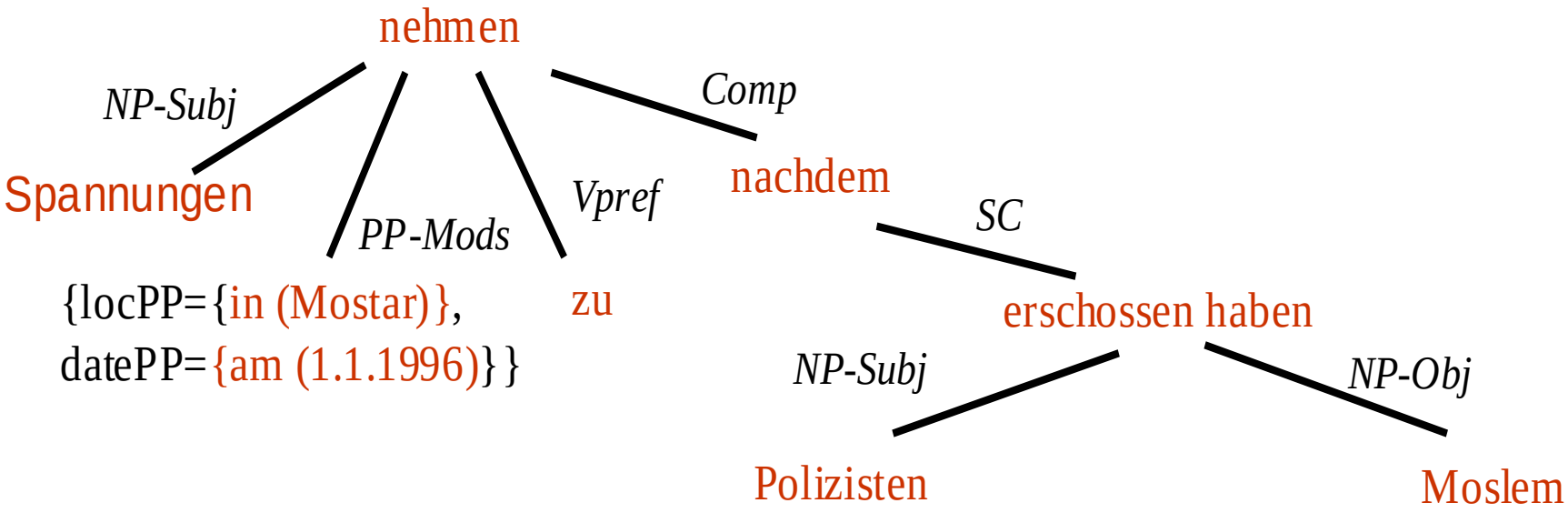
Source: TD & GN

The systematic separation of the NLP and the modeling components, dealing with two types of knowledge (1)

PARADIME

- The **linguistic analysis tools** comprise (1) a tokenizer, a morphological analyzer (incl. compound analysis) and a POS filter for the *lexical processing*, and (2) a *fragment recognizer* for Named Entities and generic phrases (NP, PP, Verbgroup). On the top of this (3) a *dependency based parser* computes a flat (partial) analysis of the text, enriched with information about grammatical functions.

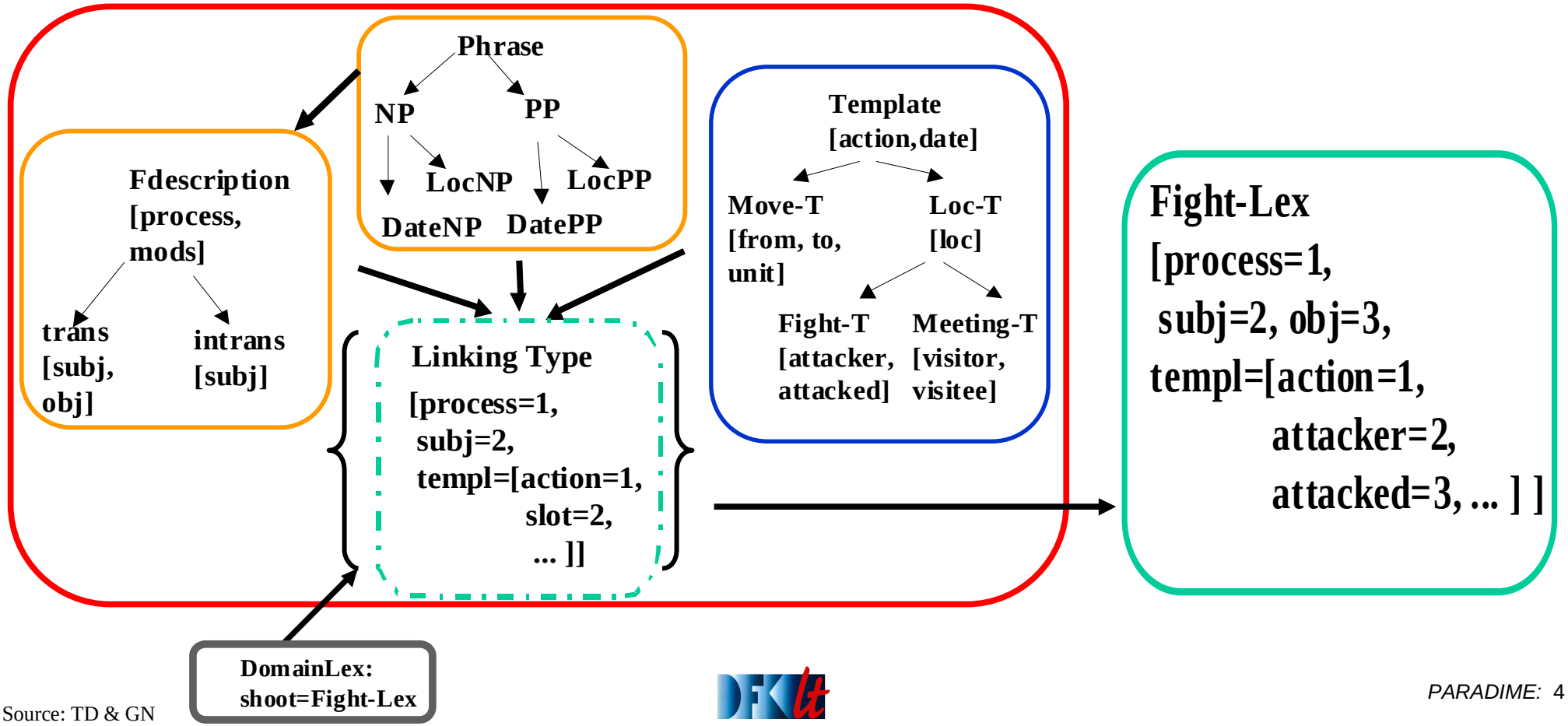
[$_{NP}$Die Spannungen] [$_{Loc\text{-}PP}$in Mostar] [$_{V}$nehmen] [$_{Date\text{-}PP}$am 1.Jan. 1996] [$_{Vpref}$zu] , [$_{Comp}$nachdem] [$_{NP}$kroatische Polizisten] [$_{NP}$einen 18jährigen Moslem] [$_{V}$erschossen haben], der ...



Source: TD & GN

The systematic separation of the NLP and the modeling components, dealing with two types of knowledge (2)

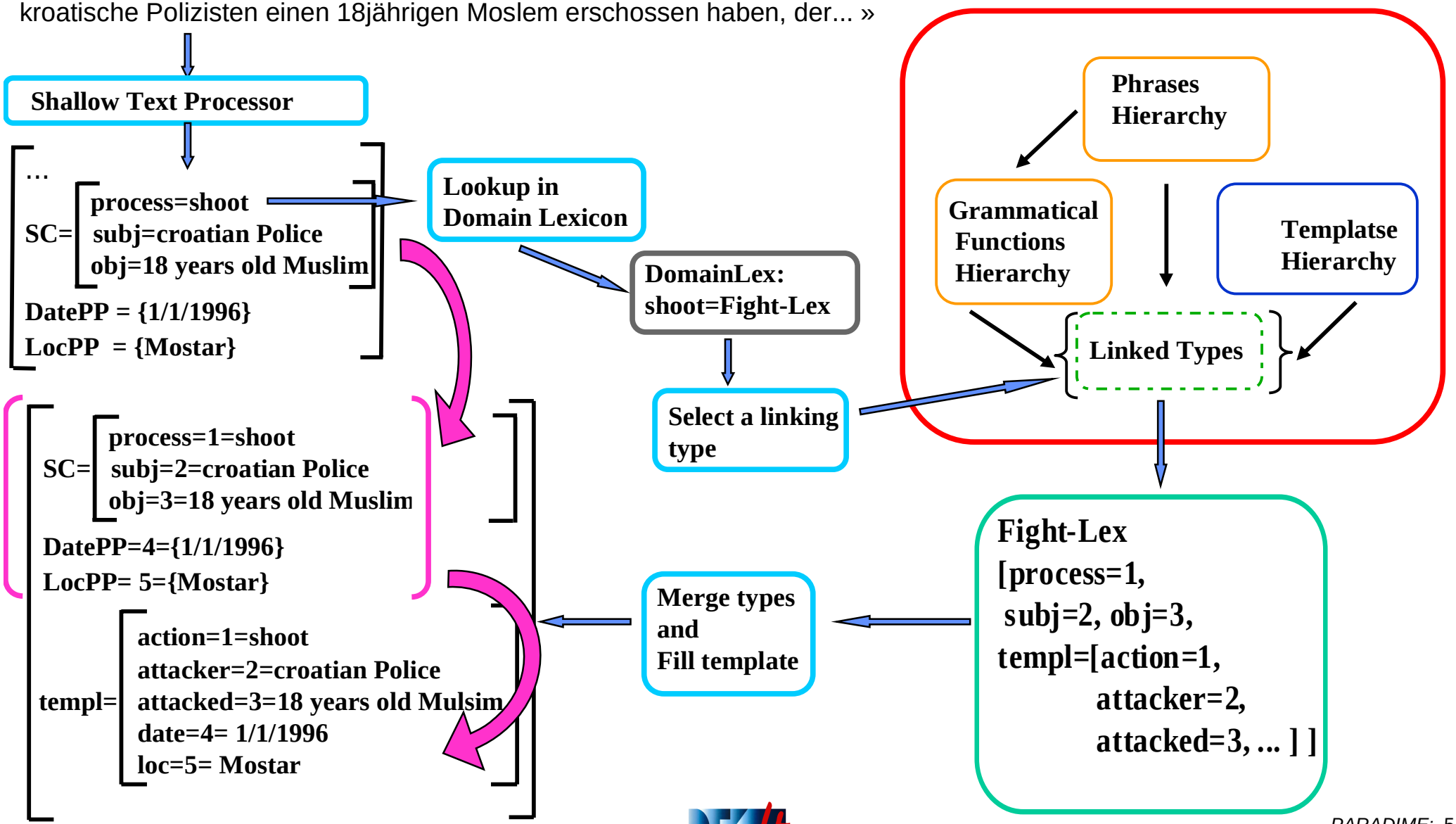
- The **domain modeling** is realized by hierarchically organized templates (blue box below), using the TDL formalism, in which also conceptual hierarchies abstracting over the results of the linguistic analysis are described and combined (yellow boxes).
- The **interface** between domain and linguistic knowledge is realized as a set of *linking types* (doted green box) describing merged abstract conceptual structures, out of which a domain-lexicon lookup (gray box) selects a task specific template (green box).



Source: TD & GN

Task Specific Template Filling, based on the TDL Model

« Die Spannungen in Mostar nehmen am 1.Jan. 1996 zu, nachdem kroatische Polizisten einen 18jährigen Moslem erschossen haben, der... »



Source: TD & GN

Adaptation of the SMES System to a New Domain (1)

- What are the steps involved in such an adaptation?
- Which modules are concerned by such an adaptation?
- How fast is such an adaptation?

⇨ The answer to those questions is among others dependent on the kind of Information Extraction subtask under consideration:

- Named Entity task (NE)
- Template Element task (TE)
- Template Relation task (TR)
- Scenario Template task (ST)
- Coreference task (CO)

Source: TD & GN

The Subtasks of IE (as defined in MUC-7)

- Named Entity task (NE): Mark into the text each string that represents, a person, organization, or location name, or a date or time, or a currency or percentage figure (this classification of NEs reflects the MUC-7 specific domain and task)
- Template Element task (TE): Extract basic information related to organization, person, and artifact entities, drawing evidence from everywhere in the text (TE consists in generic objects and slots for a given scenario, but is unconcerned with relevance for this scenario)
- Template Relation task (TR): Extract relational information on employee_of, manufacture_of, location_of relations etc. (TR expresses domain-independent relationships between entities identified by TE)
- Scenario Template task (ST): Extract prespecified event information and relate the event information to particular organization, person, or artifact entities (ST identifies domain and task specific entities and relations)
- Coreference task (CO): Capture information on corefering expressions, i.e. all mentions of a given entity, including those marked in NE and TE (not implemented in PARADIME yet).

Source: TD & GN

Adapting the SMES System to a New Domain (2)

- Data collection, corpus and domain analysis, identification of typical terms, relations and events, and description of the templates to be filled for the application.

  This task is a constant one for every adaptation to new domains (can be tackled by the user or by the developer, or a combination of both). The efficiency and accuracy of this task depends on the expertise of the persons and on the quality of the tools involved.

- Integration of the templates into a conceptual hierarchy (ontology) in order to describe the domain model and (partially) merge this conceptual structure into existing ontologies. This is the basis of the definition the *linking types* for template filling.

  The complexity of this task is varying with the domain and the application requirements.

- Selective adaptation of the modules of the NLP component of the IE system, if necessary, and description of the domain lexicon (containing at least the typical event words).

  Ideally this task should consist just in the identification of the key-words for NE and ST, and of some domain-specific patterns to be modularly integrated into the grammar.

Source: TD & GN

Adapting SMES to the Soccer Domain: Data Collection (1)

- Data Collection:
  - 323 texts about the Soccer World Championship 1998 have been collected from the Frankfurter Rundschau (on-line available German newspaper)
  - subclass of articles chosen for corpus analysis: game reports (74 texts), where only very rarely formal texts (tables etc.) are used (see next slide):

Source: TD & GN

Adapting SMES to the Soccer Domain: Data Collection (2)

PARADIME

Brasilien besiegt Schottland 2:1.

St. Denis (dpa) - Ein «Billard-Tor» hat Titelverteidiger Brasilien zu einem gelungenen Auftakt bei der 16. Fußball-Weltmeisterschaft verholfen.

Durch ein Eigentor von Boyd in der 73. Minute feierte der Top-Favorit am Mittwoch im WM-Eröffnungsspiel einen glücklichen 2:1 (1:1)-Sieg über den respektlosen Außenseiter Schottland. Vor 80000 Zuschauern im Stade de France von St. Denis hatte Cesar Sampaio den vierfachen Weltmeister in einer packenden und rasanten Partie bereits in der 4. Minute in Führung gebracht und damit das schnellste Tor in der Geschichte der WM-Eröffnungsspiele markiert. Collins gelang in der 38. Minute mit einem verwandelten Foulelfmeter der zwischenzeitliche Ausgleich für die Schotten. Mit ihren drei Treffern war die Auftaktpartie in Frankreich zugleich die torreichste in der WM-Geschichte.

Beflügelt durch ihr frühes Führungstor boten die Brasilianer über weite Strecken den erhofften Zauber-Fußball, scheiterten jedoch zu häufig im Abschluß. Superstar Ronaldo, der bei seinen Dribblings weder von Bewacher Calderwood noch von Abwehrchef Hendry wirkungsvoll gestört werden konnte, glänzte mit technischen Kabinettstückchen, tauchte aber im zweiten Durchgang völlig ab. Schwächen offenbarte der Titelverteidiger nicht ganz unerwartet im Deckungsverband, wo sich Cafu und Giovanni auf der rechten Abwehrseite als Achillesferse erwiesen. Den sich dort bietenden Raum nutzten Dailly und Collins allerdings nur selten gefahrbringend, obwohl Schottlands Trainer Brown mit dem Einsatz von Jackson als drittem Stürmer Offensive signalisiert hatte. «Wir können gegen Brasilien nicht 90 Minuten lang verteidigen. Das gibt ein Desaster», hatte Brown geäußert.

Besser hätte der Turnierstart für den Top-Favoriten nicht sein können. Nach gerade einmal 3:50 Minuten übersprang Cesar Sampaio nach dem ersten Eckball durch Bebeto zwei schottische Abwehrspieler und köpfte zum 1:0 ein. Der Rückstand verunsicherte die nervös beginnenden Schotten zusätzlich. In der 13. Minute hätte Kapitän Hendry seinen Keeper Leighton um ein Haar zum zweiten Mal bezwungen, als er eine weite Vorlage von Dunga per Kopf knapp am eigenen Tor vorbeisetzte. Anschließend lieferte Leighton, mit fast 40 Jahren der älteste Spieler des Turniers, zwei Kostproben seines Könnens. Eine Volley-Abnahme von Roberto Carlos (16.) entschärfte der Keeper des FC Aberdeen ebenso souverän wie den Schuß von Ronaldo (20.), der zuvor die halbe schottische Abwehr wie Slalomstangen umtanzt hatte.

Die Briten wirkten lange Zeit bieder, umständlich im Spielaufbau und harmlos in Strafraumnähe. Nach einer knappen halben Stunde machte der Außenseiter erstmals auf sich aufmerksam, als Junior Baiano in einer akrobatischen Aktion vor dem einschußbereiten Durie (29.) retten mußte. Acht Minten später wurde Torschütze Cesar Sampaio zur tragischen Figur des WM-Auftakts, als er den durchlaufenden Gallacher im Strafraum allzu energisch am Trikot zog. Collins (38.) verwandelte den vom spanischen Schiedsrichter Garcia Aranda verhängten Strafstoß zum nicht einmal unverdienten Ausgleich.

Das vor der Pause Versäumte versuchten die Brasilianer nach Wiederbeginn mit aller Macht nachzuholen und nahmen Leightons Tor unter Dauerbeschuß. Mit zwei Versuchen aus der Distanz scheiterte Rivaldo (49./52.) jeweils nur knapp. Aber auch die Schotten, die jeglichen Respekt vor den großen Namen des Gegners ablegten, suchten nun die Entscheidung, wobei sich auch vor Brasiliens Keeper Taffarel dramatische Szenen abspielten. Am Ende hatten jedoch die Brasilianer das Glück auf ihrer Seite. In der 73. Minute spielte Leighton «Billard» und lenkte einen Schuß von Cafu direkt auf den Körper von Boyd, von dem der Ball ins Tor prallte.

Source: TD & GN

Adapting SMES to the Soccer Domain: Identification of Terms, Relations and Events (1)

- Terms as descriptors for the NE task (more fine-grained as in MUC-7)
  - Team: *Titelverteidiger* Brasilien, den respektlosen *Außenseiter* Schottland
  - Player:*Superstar* Ronaldo, von *Bewacher* Calderwood noch von *Abwehrchef* Hendry, von Jackson als drittem *Stürmer, Torschütze* Cesar, von Roberto Carlos *(16.)*,
  - Referee: vom spanischen *Schiedsrichter* Garcia Aranda
  - Trainer: Schottlands *Trainer* Brown, *Kapitän* Hendry seinen *Keeper* Leighton
  - Location: im Stade de France *von* St. Denis (more fine-grained location detection would be: Stadion: *im* Stade de France and City: *von* St. Denis)
  - Attendance: Vor 80000 *Zuschauern*

- Terms for NE Task
  - Time: *in der 73. Minute*, *nach gerade einmal 3:50 Minuten*, von Roberto Carlos *(16.)*, *nach einer knappen halben Stunde*, scheiterte Rivaldo *(49./52.)* jeweils nur knapp, das *vor der Pause* Versäumte versuchten die Brasilianer *nach Wiederbeginn*, ...
  - Date: *am Mittwoch*, der Turnierstart (?), im WM-Eröffnungsspiel (?)
  - Score/Result: Brasilien besiegt Schottland *2:1*, einen *2:1 (1:1)*-Sieg, der zwischenzeitliche *Ausgleich*, in der 4. Minute *in Führung* gebracht, köpfte *zum 1:0* ein

Source: TD & GN

Adapting SMES to the Soccer Domain: Identification of Terms, Relations and Events (2)

PARADIME

- Relations for TR Task
  - Opponents: Brasilien *besiegt* Schottland, *feierte* der Top-Favorit ... einen glücklichen 2:1 (1:1)-Sieg über den respektlosen Außenseiter Schottland,
  - Player_of: hatte Cesar Sampaio den vierfachen Weltmeister ... *in Führung gebracht*, Collins gelang ... der zwischenzeitliche Ausgleich für die Schotten, der Keeper *des* FC Aberdeen, Brasilien*s* Keeper Taffarel
  - Trainer_of: Schottland*s* Trainer Brown
  - ...

- Events for ST task:
  - Goal: in der 4. Minute *in Führung gebracht*, das schnellste Tor ... *markiert*, Cesar Sampaio *köpfte* zum 1:0 *ein*, Collins (38.) *verwandelte* den Strafstoß, hätte Kapitän Hendry seinen Keeper Leighton um ein Haar zum zweiten Mal *bezwungen*, von dem der Ball ins Tor *prallte*
  - Foul: als er den durchlaufenden Gallacher im Strafraum allzu energisch am Trikot *zog*
  - Substitution: und mußе in der 59. Minute für Crespo *Platz machen*
  - ...

Source: TD & GN

DFKI lt

Adapting SMES to the Soccer Domain: Description of the Templates (1)

- On the basis of the corpus analysis and the design of the NE, TR and ST tasks for the soccer domain, the templates to be filled have been (manually) described, using for this the TDL formalism
  - The templates are realized in form of (typed) feature structures, where the values of some attributes can be an atom, a list or another template (all kind of values can be constrained)
  - All operations on feature structures are supported
  - (Multiple) inheritance can be described

    A declarative (and fine-grained) modeling of the domain specific knowledge is thus possible, without interaction with the NLP tools

- The relations are implicitely encoded in the templates and those are hierarchically organized on the base of a classification of the domain-specific objects and events (naming or *typing* the templates)

- Slots for (co-)referentiality resolution are foreseen, but the CO task has not been tackled yet

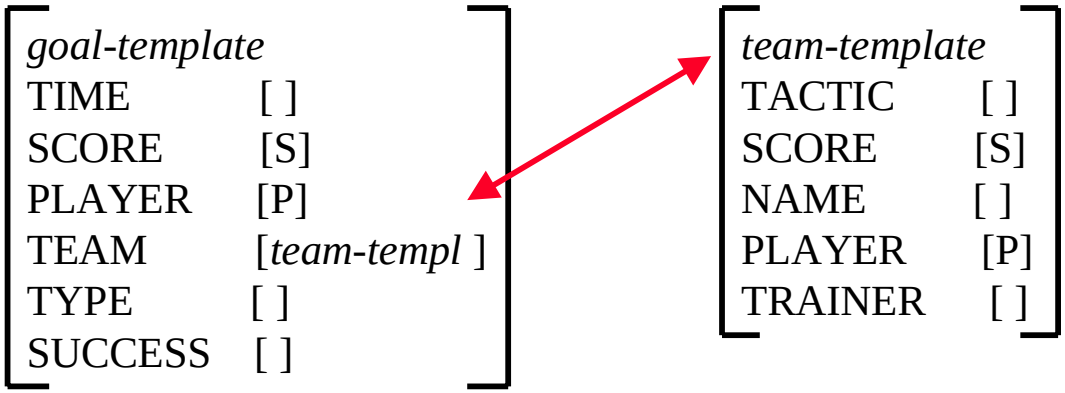
Source: TD & GN

Adapting SMES to the Soccer Domain: Description of the Templates (2)

- An Example: The Team-Template

team-template	
TACTIC	[ ]
SCORE	[ ]
NAME	[ ]
PLAYER	[ ]
TRAINER	[ ]

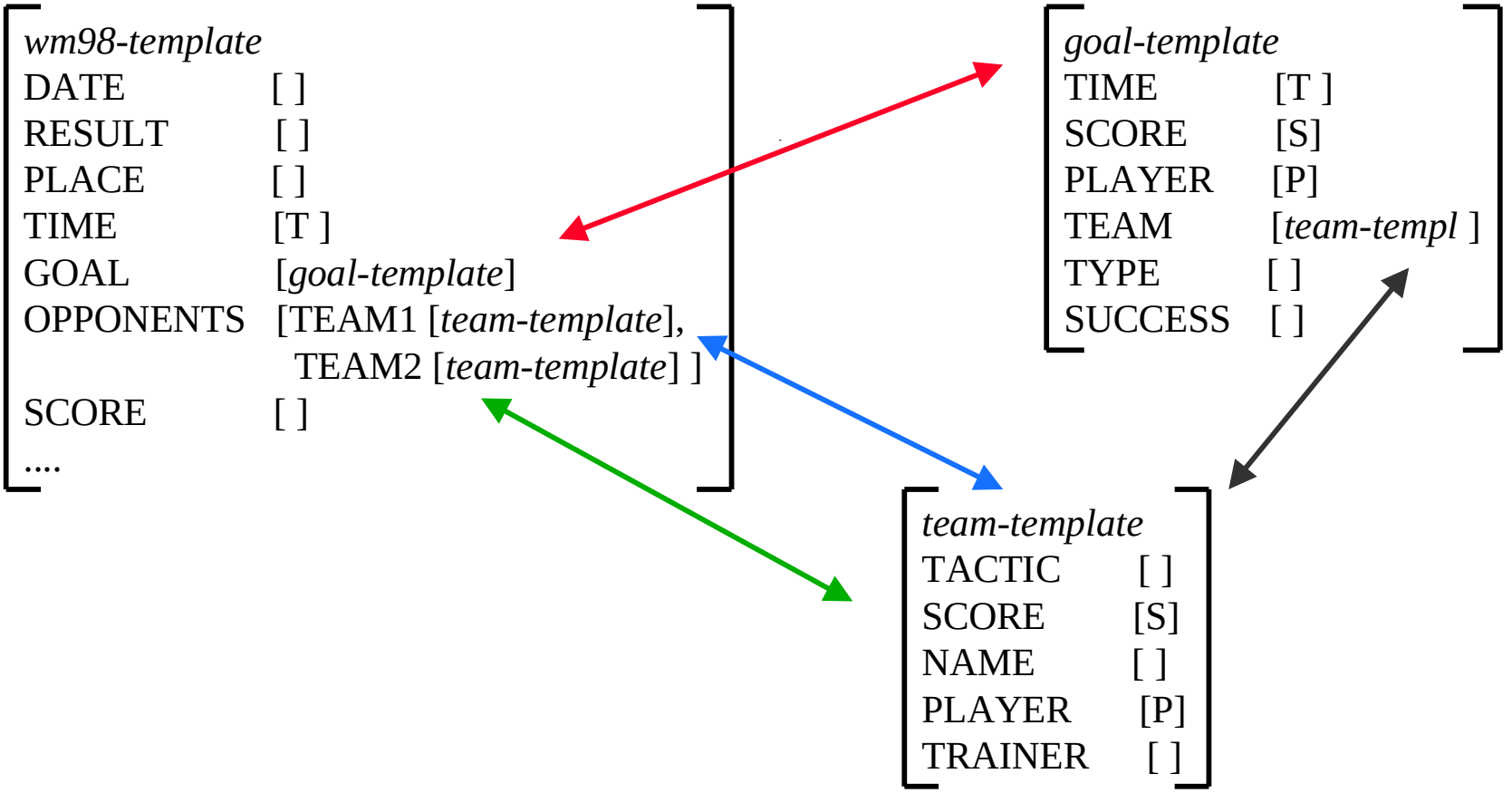
This kind of template (object-template) is typically embedded in an event-template as the value of a corresponding slot, where information-sharing is provided for the values of identical attributes at the distinct levels of template embedding:



Source: TD & GN

Adapting SMES to the Soccer Domain: Description of the Templates (3)

- The level of embedding templates is depending on the domain and the detail of the corpus analysis or of the user-requirement. In our case the top level is the one of a game of the World Championship 1998, being identified by the date and the opponents involved.



Source: TD & GN

Adapting SMES to the Soccer Domain: Definition of the Linking Types (1)

- The set of *linking types* is also defined within the TDL formalism and hierarchically organized on the basis of a classification of domain-specific verbs (we are concerned for the time being only with event words) taking into account the various verb frames, the polarity and the realization of certain modifiers within the clause under consideration.
- The top level linking type for the soccer-domain is called *soccer-domain* and associates at the abstract level a domain-specific verb with the general template *wm98-template*:
- On the basis of further properties of the verb and the associated linguistic material within the clause boundaries, a subtyping of the main linking type is designed. So for the verb refering to a “game result”, a linking type *soccer-domain-result* has been introduced, corresponding to the classification of such verbs in the domain lexicon “entry=besieg, cat=v, dom=soccer, type=goal-subj-obj” where also the subcat information is encoded.

Soccer-Lex
[process =1,
temp=[*wm98-template*
action=1]

Soccer-Result-Lex
[process =1,
syn [result-mods=2
temp=[*wm98-template*
action=1,
result=2]

Soccer-Result-Subj-Obj-Lex
[process =1,subj=2,obj=3
syn [result-mods=4
temp=[*wm98-template*
action=1,team1=2,
team2=3,result=4]

Source: TD & GN

Adapting SMES to the Soccer Domain: Definition of the Linking Types (2)

- Some linking types also take into consideration adjuncts in order to be able to describe certain relations. Example: “feierte der Top-Favorit ... einen glücklichen 2:1 (1:1)-Sieg *über den respektlosen Außenseiter Schottland*, where the prepositional modification (lexically constrained) is considered to function like an argument:

Soccer-Result-Subj-Pobj-Lex
[process =1,subj=2,
pobj(head=‘über’)=3,
syn [result-mods=4
temp=[*wm98-template*
action=1,team1=2,
team2=3,result=4]

Source: TD & GN

Adapting SMES to the Soccer Domain: Filling the Templates for the Soccer Domain

- Once the input text has been processed, a linking type for each clause will be selected, if the lookup in the domain lexicon (only verb entries for the time being, to be extended to other categories) is successful. So the template filling is done on the basis of the detection of some domain-specific event (maximally one template for each clause, but this restriction will be eliminated).
- As a consequence: the filling of object-template will take place during the process of the filling of an event-template within which the object-template is included as the value of a slot, even if the filling of specific attributes of the global template is in fact corresponding to distinct IE-subtasks:
  - Filling the NAME slot of the Team-Template is a TE Task
  - Filling the PLAYER and the TRAINER slots of the Team-Template is a TR Task based on the identification of the player and/or the trainer, and the team in the TE task (merging all Team-Templates filled after analysis of the text would give a list of players)
  - Filling the (actual) SCORE and the TACTIC slots is a ST task based on the definition of the specific event

Source: TD & GN

Adapting SMES to the Soccer Domain: Adapting the NLP Components

- For the NE Task some modification has been done, consisting in describing patterns for the recognition of domain-specific Named Entities, and adding those patterns to the NE recognition machinery. Also the output function had to be adapted for reflecting the new Named Entities to be detected. No modification at the level of the recognition of generic phrases was necessary.

- The information contained in the domain-lexicon has been modularly integrated into the general SMES lexicon (this step is optional but has the advantage of having one unique lexicon for lookup).

- The rest of the adaptation work for the new domain has been done in the context of the definition of the templates and the linking types, thus verifying our hypothesis that most of the job for porting SMES to a new domain can be done in the context of domain-modeling, whereas the maintenance of the NLP tools. This also offering efficiency benefits, since we could provide for a prototype for the new application twice as fast as with the former SMES architecture, with a similar coverage.

Source: TD & GN

Adapting SMES to a new Domain: Work to be done

- Implement the CO task (will be done within the NLP components).
- Investigate the automation of templates detection for the application to a new domain and their integration in a conceptual hierarchy. Use of learning mechanisms.
- Further investigate the integration of the domain-hierarchy with other conceptual structures (general purpose ontologies, thesaurus, WordNet etc...).
- Define the merging of instantiated (filled) templates for getting a better information extraction on the basis of a more complete discourse model. Allow also to get real “scenarios” for some events or objects.
- Use terminology extraction for building the domain-lexicon

Source: TD & GN