The Semantic Web: Research and Applications: 8th Extended Semantic Web Conference, ESWC 2011, Heraklion, Crete, Greece, May 29-June 2, 2011, Proceedings, Part I. Lecture Notes in Computer Science, Volumen 6643. Springer, 2011. 230-244.

DOI: http://dx.doi.org/10.1007/978-3-642-21034-1_16

FootbOWL: Using a generic ontology of football competition for planning match summaries

Nadjet Bouayad-Agha[1], Gerard Casamayor[1], Leo Wanner[1,2], Fernando Díez[3] and Sergio López Hernández[3]

[1] DTIC, University Pompeu Fabra, Barcelona, Spain
firstname.lastname@upf.edu
[2] Catalan Institute for Research and Advanced Studies (ICREA)
firstname.lastname@icrea.es
[3] DII, Universidad Autónoma de Madrid, Madrid, Spain
firstname.lastname@uam.es

Abstract. We present a two-layer OWL ontology-based Knowledge Base (KB) that allows for flexible content selection and discourse structuring in Natural Language text Generation (NLG) and discuss its use for these two tasks. The first layer of the ontology contains an application-independent base ontology. It models the domain and was not designed with NLG in mind. The second layer, which is added on top of the base ontology, models entities and events that can be inferred from the base ontology, including inferable logico-semantic relations between individuals. The nodes in the KB are weighted according to learnt models of content selection, such that a subset of them can be extracted. The extraction is done using templates that also consider semantic relations between the nodes and a simple user profile. The discourse structuring submodule maps the semantic relations to discourse relations and forms discourse units to then arrange them into a coherent discourse graph. The approach is illustrated and evaluated on a KB that models the First Spanish Football League.

1 Introduction

Natural language generators typically use as input external or purpose-built domain databases (DBs) or knowledge bases (KBs), extracting and/or transforming the relevant content during the text planning phase to instantiate schemas or other discourse representations, which are then verbalized during linguistic generation. See, for instance, [9]. More recent statistical, or heuristic-based, text planning tends to draw upon KBs crafted specifically for the task of Natural Language Generation (NLG) in order to assess relevance of its parts for inclusion into the text plan; see, among others, [4, 5]. Given the NLG-tuned nature of these KBs, the mapping from knowledge to linguistic representations is then quite straightforward.

In order to avoid linguistically-driven projection of relevant content onto discourse representations that intermingles conceptual information with linguistic

information or the creation of NLG-tuned KBs, we suggest a two-layer KB. The first layer consists of a base ontology modeled in OWL. This ontology is application-independent: it only models the domain and was not designed with NLG in mind. The second layer, which is added on top of the base ontology, models entities and events that can be inferred from the base KB, including logico-semantic relations that can be inferred between individuals. Evidence on the existence of the inferred individuals and relations between them is deduced from a reference text collection. The KB used in our experiments models Football Competitions, more specifically the First Spanish Football League.

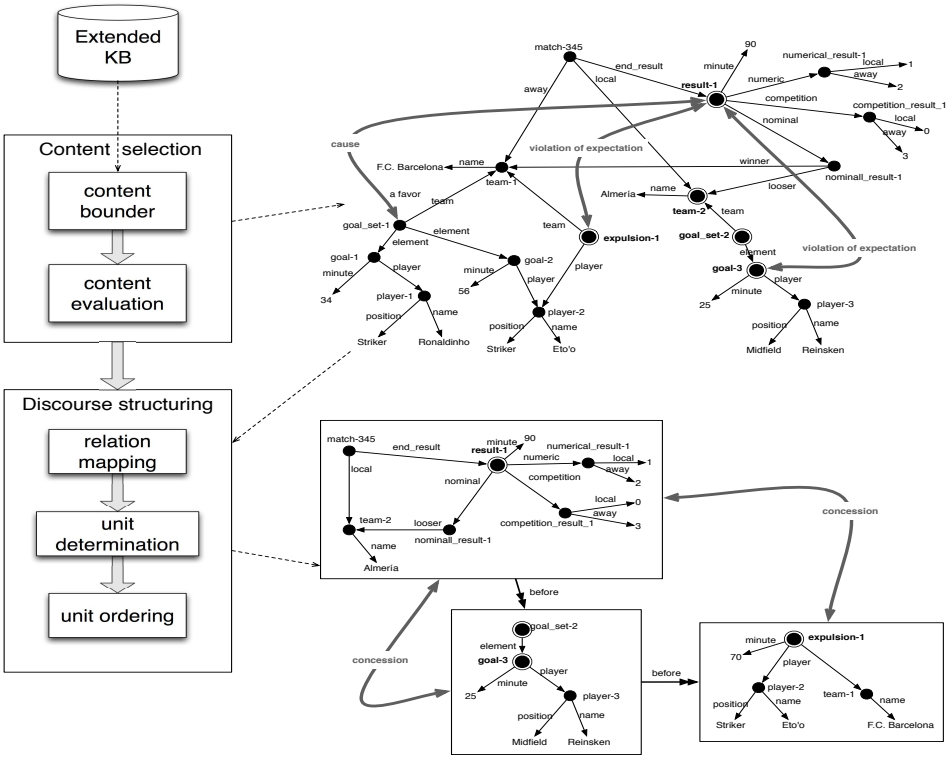

Fig. 1. Overview of text planning with associated content (top) and discourse (bottom) structures

In what follows, we describe how this KB is used for the two tasks of text planning, content selection and discourse structuring (Figure 1 illustrates the overall picture). Given their interconnection, these tasks are performed in an interplay between the content selection and discourse structuring modules. Thus, the relevance of the nodes in the KB is determined in the content selection module according to a simple user model, a set of relevance heuristics based on empirically determined weights, and the availability of logico-semantic rela-

tions that link the nodes to form a coherent content structure. The discourse representation module maps the logico-semantic relations to discourse relations, extracts the content marked as relevant during content selection using a set of templates to instantiate discourse units and to arrange them into a coherent discourse structure. The discourse structure is passed to the linguistic generator of Spanish, which produces short summaries of football matches of the kind found at the beginning of exhaustive articles about individual matches, only that they take the preferences of the targeted addressee for one of the teams involved into account. Consider Figure 2 that displays a generated summary that targets a fan of the team of Barcelona.[4]

> Victoria del F.C. Barcelona. El Barcelona ganó contra el Almería por 2-1 gracias a un gol de Ronaldinho en el minuto 34 y otro de Eto'o en el minuto 56. El Barcelona ganó aunque acabó el partido con 10 jugadores a causa de la expulsión de Eto'o. Gracias a esta victoria, permanece en la zona de champions. En la vigésimo quinta jornada, se enfrentará al Villarreal.

Fig. 2. A sample football match summary as produced by our generator

The remainder of the paper is structured as follows. In Section 2, we present the two-layer ontology used as input to the text planning. In Section 3, we detail our content selection module, and present how we empirically determine weights using supervised learning to assess the relevance of some of the content units found in football match summaries. Section 4 describes the discourse structuring module and Section 5 the evaluation of the content selection and discourse structuring modules. Section 6 presents related work on the use of knowledge bases, data assessment, text planning and empirical content selection in NLG, before in Section 7 some conclusions are given and plans for further work are outlined.

2 A two-layer OWL Ontology

2.1 Ontology design

There are some ontologies available that deal with sports and, more precisely, with (European) football (or soccer).[5] However, even the most detailed of them, let alone generic ontologies such as OpenCyc, did not contain specific football

[4] Translation: 'Victory of F.C. Barcelona. Barcelona won against Almería by 2-1 thanks to a goal by Ronaldinho in minute 34 and another goal by Eto'o in minute 56. Barcelona won despite ending the match with 10 players because of the sent off of Eto'o. Thanks to this victory, Barcelona remains in the Champions zone (of the classification). Gameweek 25 Barcelona will meet Villareal.'

[5] Among them `http://sw.deri.org/ knud/swan/ontologies/soccer` (the SWAN Soccer Ontology by DERI), the sports fragment of the OpenCyc Ontology [12] (`http://sw.opencyc.org/2009/04/07/concept/en/Soccer`), the sports fragments in the DAML repository [7] (`http://www.daml.org/ontologies/374`) and [18].

data we were interested in—among them, Approximation, Shot, Block or Header. Therefore, we developed our ontologies from scratch.

As already mentioned in the Introduction, our model foresees a two-layer ontology, the base ontology and the extended ontology.[6] The base ontology describes the football league domain. It is composed of two different ontologies: an object ontology which deals with the structural information of the competition (teams, competition phases, matches, players, etc.), and an event ontology which deals with information related to the events that happen in the match (penalties, goals, cards, etc.). To develop it, we followed the top-down strategy suggested by Uschold and King [19]: the more abstract concepts are identified first and subclassified then into more specific concepts. This is done to control the level of detail wanted. A known drawback of this strategy is that it can lead to an artificial excess of high-level classes. In our application, we achieved a sufficient level of detail for our application domain (i.e., the First Spanish Football League) with a moderate number of classes. More precisely, we model in the object ontology 24 classes and 42 properties, with 4041 instances in the corresponding KB (see Subsection 2.2 below). The top level classes of this ontology are Competition, Match, Period, Person, Result, Season, Team, TeamCompositionRelation and Title. In the event ontology, we model 23 classes and 8 properties, with 63623 instances in the corresponding KB (see Subsection 2.2 below). The top level classes of this ontology are ActionFault, Card, Corner, Fault, FaultKick, Goal, GoalKick, Interception, OffSide, Pass, Stop, Throw-in, Shot and Substitution.

The extended ontology adds an extra layer of meaning to the concepts modeled in the base ontology. Its concepts are deduced by the analysis of the target summaries, considering mainly what new knowledge can be inferred from the basic knowledge on the First Spanish Football League. We infer new knowledge about events and states of a match (goals and expulsions, results and classifications) typically found in summaries, excluding statistical information about matches within a season and across seasons (best scorer, consecutive wins, first victory in a given stadium, etc.). Some of the classes and properties were also added to make the navigation easier for the mapping to linguistic realization and for the inference of new knowledge. For example, 'for' and 'against' properties were added to the Goal class in order to know the team which scored respectively received the goal in case the information concerning team scored the goal was only available indirectly in the base ontology via the player. The inferred knowledge is divided into five categories, 1. result, 2. classification, 3. set, 4. match time, and 5. send-offs.

Result-related knowledge (nominal result and the points scored in the competition) is inferred from the numerical result of the match available in the base ontology (with winner/loser/drawing opponents specified), hence the classes NominalResult and CompetitionResult.

[6] The ontologies and corresponding knowledge bases are not available for free distribution. They are restricted to the i3media project consortium `https://i3media.barcelonamedia.org/`

Classification-related knowledge models information related to the position of each team in the competition, its accumulated points and relative zone. For the zone, in addition to the four official zones Champions, UEFA, neutral or relegation, we introduce two internal zones—Lead and BottomOfLeague. It is of interest to obtain after each gameweek a team's tendency (ascending, descending, stable) and distance with respect to its previous classification. Tendency represents the team's change of zone in the competition, whilst Distance represents a team getting closer (or further) to a higher (lower) zone. In addition to the real tendency, teams are assigned a virtual tendency which represents the team's change of zone taking a (virtual) result that may be different from the actual match result (for instance, if the team would have drawn instead of winning, what would be the tendency of its classification in the league table).

Set-related knowledge models sets of events or processes for a given team in a match or for a given match. It is needed to be able to talk about events or processes together in accordance with their chronological occurrence (first goal, team was winning then it drew, etc.), hence the classes Set and ConstituentSet. These classes also allow us to simply refer to the number of constituents within it (cf. *the team had two red cards*).

Match time-related knowledge models the state of the match along its duration, creating intermediate results after each goal, hence the class IntermediateResult. Thus, a team could be winning after a goal, even though the final result is a draw. It is also possible to refer to specific reference time points such as 'beginning of the match', and 'conclusion of the first period'.

Send-offs related knowledge includes the expulsion of a player after a red card, hence the Expulsion class and the number of players left after an expulsion, hence the PlayersInField class.

Each set of inferred knowledge triggers the inference of a number of logico-semantic relations, hence the class LogicoSemanticRelation with its subclasses such as Cause, Implication, ViolationOfExpectation, Meronymy, Precedence, Contrast. For instance:

- A cause relation is instantiated between the set of goals of a team and the final nominal result.
- A violation-of-expectation relation is instantiated between an instance of PlayersInField and a final winning/drawing result (e.g., *despite playing with 10, the team won*).
- A relation of precedence is instantiated between pairs of constituents in a set to show their immediate temporal precedence relation.
- A contrast relation is instantiated between the contrasting classification distances or tendencies of both teams of the match (e.g., *team A goes up in the classification whilst team B goes down*).

Figure 3 shows the representation of the set of four goals of a team in a match, including the precedence relation between the constituents; the figure also shows the division of concepts between the base and extended ontology.

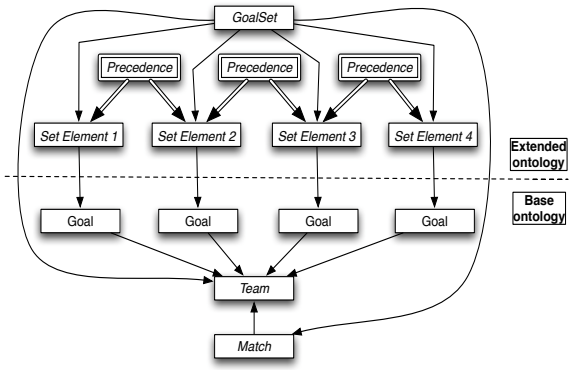

Fig. 3. Representation of an ordered set of goals of a team in a match

2.2 Ontology population

The base KB was automatically populated with data scraped from web pages about the Spanish League seasons to include general information about competitions, players, stadiums, etc, and specific information about matches. Currently, it contains three seasons: 2007/2008, 2008/2009 and 2009/2010. The scrapping was done by *ad hoc* programs that extract all the information required by the classes defined in the base ontologies.[7] The extended ontology population was carried out using the inference engine provided by Jena.[8] The engine works with a set of user-defined rules consisting of two parts: head (the set of clauses that must be accomplished to fire the rule) and body (the set of clauses that is added to the ontology when the rule is fired). We defined 93 rules, with an estimated average of 9,62 clauses per rule in the head part. Consider the following example of a rule for classifying the difference between the scores of the two teams as "important" if it is greater than or equal to three:

```
[rule2: (?rn rdf:type base:NumericResult)
(?rn base:localScore ?localScore) (?rn base:visitorScore ?visitorScore)
(?localScore base:result ?local) (?visitorScore base:result ?visitor)
differenceAbs(?local, ?visitor, ?r) ge(?r, 3)
-> (?rn inference:resultDifference "important")]
```

For the 38 gameweeks of the regular football season, the inference engine generates using the 93 rules from the data in the base ontologies a total of 55894 new instances. The inference rules are organized into five groups corresponding to the five categories of inferred knowledge described in Subsection 2.1.

[7] Object and event information were extracted from the Sportec (`http://futbol.sportec.es`) and AS (`http://www.as.com/futbol`) portals respectively.

[8] `http://jena.sourceforge.net/`

3 Content Selection

3.1 Approach

The content selection module consists of a content bounding submodule and a content evaluation submodule. The content bounding submodule selects from the KB, using a set of hand-written rules, individuals that are relevant to the match for which a text is to be generated, either because they can be related directly to the match (e.g., the players of the teams involved, the match events such as goals), or because of the more general context of the competition (e.g., the league's classification). It also includes the logico-semantic relations that link these individuals. Given the large size (by NLG standards) of the KB, the motivation for the content bounder is to filter out irrelevant information and to make thus the subsequent content selection task more manageable. The output of the content bounder is a subset of the KB which constitutes the maximal set of data available for generating any sort of summary for a given match. The content structure presented in Figure 1 is a simplified output of the content bounding submodule.

The content evaluation submodule is in charge of evaluating the relevance of the content according to 1) a simple user model, 2) a set of heuristics, and 3) the logico-semantic relations that link individuals of the KB. Both the user model and the heuristics are numeric functions that map instances of concepts in the KB to a numeric measure of their relevance. The user model consists of the specification of the user's team of interest for the requested match or of a "neutral" profile—if the user has no favourite team. The heuristics measure relevance according to empirical knowledge extracted from a corpus of texts.[9] The content evaluation currently gives a weight of '1' if the node is related to the user's team of interest (or if the user profile is "neutral") and '0' otherwise. This weight is multiplied by the node's relevance measure, which is set to '1' if the heuristic weight for selecting the instance outweighs the heuristic weight for not selecting it. Otherwise it is set to '0'. Finally, the nodes that represent the logico-semantic relations are marked as relevant if they link two nodes with a positive relevance weight. This ensures the coherence of the content being selected. In Figure 1, given the user interest for the local team, the content selection heuristics and the logico-semantic relations, the five double circled nodes in the content structure are marked as relevant by the content evaluation submodule.

3.2 Empirical Determination of Relevance Measures

The weights of the instances that are to be selected are obtained by supervised training on a corpus of aligned data and online articles. The corpus consists of eight seasons of the Spanish League, from 2002/2003 to 2009/2010 with a total of 3040 matches, downloaded from different web sources. The articles typically

[9] Relevance could also be measured according to other sources (e.g., past interaction with the user).

consist of a title, a summary and a body. The data for each match consist of the teams, stadium, referee, players, major actions like goals, substitutions, red and yellow cards, and some statistical information such as number of penalties. Table 1 shows the verbalization of some categories in each of the three article sections considered for a single season in any of the sources. As can be seen, the result of the match (whether nominal or numerical) is almost always included in all the sections, whilst the verbalization of other categories is more extensive in the article body than in the summary, and in the summary more extensive than in the title. In our work on the generation of summaries, we focused on learning weights for league classifications, goals and red cards.

	result	classification	goal	red card	stadium	referee	substitution
title	92.4%	16.3%	19.6%	9.3%	19.2%	2.9%	0%
summary	90.8%	22%	43.6%	32.2%	38.2%	3.7%	0.17%
body	97.6%	51.3%	95.2%	77.1%	82.4%	80%	18.1%

Table 1. Verbalization of some categories in title, summary and body of Spanish Football League articles (2007/2008 season) in all sources

The data-text alignment procedure implies as a first step a preprocessing phase that includes tokenization and number-to-digit conversion. Then, instances of the relevant categories (i.e., specific goals, specific red cards, etc.) are detected using data anchors in the text (such as player names and team names) and regular expressions patterns compiled from the most frequent N word sequences of the corpus (where 1<N<5). Data anchors are given priority over the use of regular expressions.

For the description of a goal or a red card, we used the same set of over 100 features. The features include deltas of minutes between the current event and the previous/next event of the same class, players and teams, information about individual players, a player's team and its classification. For modeling the classification, we used a more systematic approach to feature extraction by regarding a team's classification as event of a specific gameweek, comparing it to the events of the previous gameweek—that is, to the 20 classifications[10] of the previous gameweek and to the events of the same gameweek (also 20 classifications), such as the delta of category, points and team between classifications. In this way, we obtained a total of 760 features.

In order to classify the data, we used Boostexter [17], a boosting algorithm that uses decision stumps over several iterations and that has already been used in previous works on training content selection classifiers [1, 10].[11] For each of the three categories (goal, red card, classification), we experimented with 15 different classifiers. We considered a section dimension (title, summary and title+summary) and a source dimension (espn, marca, terra, any one of them

[10] The Spanish League competition involves 20 teams.
[11] After a number of experiments, the number of iterations was set to 300.

(any) and at least two of them), dividing the corpus each time into 90-10% of the matches for training and testing.

4 Discourse Structuring

The discourse structuring module receives as input a content plan which is a subset of the KB determined by the content bounding task, with some nodes marked as relevant by the content evaluation task (cf. Section 3). This subset of the KB in OWL-format is converted into the graph representation used by the Mate linguistic generation environment [2]. We use Mate for several reasons: 1) it comes with a handy API for graph manipulation, 2) it provides a straightforward representation of groups of nodes (i.e., "bubbles") necessary to represent discourse units, 3) the graph structures can be viewed in the Mate editor, and 4) the output of the discourse structuring is the input to the linguistic generation which uses these graph structures.

The discourse structuring works on the logico-semantic relations marked as relevant by the content selection (and their arguments, which are also relevant). It consists of three tasks which are: mapping logico-semantic to discourse relations, determining discourse units, and discourse units ordering. As already pointed out in [20] (for a different domain), a logico-semantic relation can be mapped onto different discourse relations depending on the user's previous knowledge, the content being communicated and the information structure. In the current prototype, the mapping between logico-semantic and discourse relations is one-to-one. The arguments of the logico-semantic relations are mapped onto nucleus–satellite arguments of the discourse relations following the Rhetorical Structure Theory [13]. For example, the CAUSE relation is mapped onto a *VolitionalCause* discourse relation, whilst the IMPLICATION relation is mapped onto a *NonVolitionalCause* discourse relation. As a consequence, the CAUSE relation between a set of goals and a victory can be verbalized during the linguistic generation by the discourse marker *gracias a* 'thanks to', whilst the IMPLICATION relation between a red card and an expulsion by *por* 'because of'.

The discourse unit determination is template-based; that is, we use our expertise of what can be said together in the same proposition in a football match summary. Currently, we have defined eleven discourse unit templates that cover the types of propositions that can be found in football summaries. Each core node, i.e., node that can be the argument of a discourse relation, can form a discourse unit. So, for each core node, a list of (possibly recursive) paths in the form *edge*> *Vertex* (where the edge is the object property and the vertex is the class range) is given to find in the graph the list of nodes that can be included in the discourse unit of that core node, starting from the core node. For example, the following is an excerpt of the template for expressing the result of a match:

```
partido>Partido,
periodo>PeriodoPartido,
resultNom>ResultNom,
resultNom>ResultNom>ganador>Equipo,
```

```
resultNom>ResultNom>perdedor>Equipo,
resultNom>ResultNom>protagonist>Equipo
```

This template includes the node *Partido* 'Match' when talking about the result, such that a sentence that introduces the match between the two teams and the final result (for example, nominal result with/without a numerical score) of the following kind can be produced: *The match between team A and team B ended with the victory of team A (2-1)*. Any node that stays outside the discourse units is not included in the discourse plan. In other words, the discourse unit determination is in charge of further – fine-grained – content selection.

The final discourse structuring task, namely discourse unit ordering, consists of a simple partial order on the discourse units that starts with 'ResultPuntual'. In Figure 1, the simplified discourse plan consists of three ordered discourse units, each of which includes (double-circled) node(s) marked as relevant by the content evaluation submodule and further content nodes added by the discourse unit determination templates.

5 Evaluation

5.1 Content Selection Evaluation

Our evaluation of the content selection consisted of three stages: (1) evaluation of the automatic data-article alignment procedure, (2) evaluation of the performance of the classifiers for the empirical relevance determination, and (3) the evaluation of the content selection proper.

The evaluation of the automatic alignment against 158 manually aligned summaries resulted in an F-score of 100% for red cards, 87% for goals and 51% for classification. The low performance of classification alignment is due to the low efficiency of its anchors: positions, zones and points are seldom mentioned explicitly and both team names often appear in the summary, leading to ambiguity. For this reason, classification alignment was edited manually.

Table 2 shows the performance of the classifiers for the determination of the relevance of the three categories (goal, red card and classification) with respect to their inclusion into the summary section, comparing it to the baseline, which is the majority class. For red cards, the classifier did not show any significant improvement over the baseline for summary section in any of the cases involving summary section only. However, when considering title and summary from a source together, the classifier accuracy for red cards is 85% and the baseline 53% with $t = 4.4869$ ($p<0.0001$, sample size=62). In all cases, the best performance is obtained by considering the content from any of the online sources.

The evaluation of the content selection proper includes the template-based content selection performed during the discourse unit determination. The evaluation is done by comparing the content of generated summaries with that of existing summaries (the gold standard).

Our test corpus consists of 36 randomly selected matches from the set of matches of the 2007–2008 season, each with three associated summaries from

category	source	sample size	classifier	baseline	paired t-test
goal	any	1123	64%	51%	t = 6.3360 (p<0.0001)
	terra	1121	65%	59%	t = 3.4769 (p=0.0005)
card	any	54	78.1%	65.4%	t = 1.6593 (p=0.1030)
classif	any	295	75%	61%	t = 4.4846 (p<0.0001)

Table 2. Performance of the best classifiers (vs majority baseline) on a test set for the summary section

three different web sources (namely espn, marca, terra). We compiled a list of all RDF-triples considered for inclusion in the content selection and discourse unit determination modules, including the logico-semantic relations. For each of the 108 (36×3) summaries, we manually annotated whether a triple was verbalized or not. We also annotated for each text the team of interest by checking whether the majority of content units was from one team or another; in case of equality, the user profile was considered neutral. This allowed us to compare the generated text of a given match for a given profile with the text(s) for the same profile. As baseline, we always select both teams and the final result regardless of profile since the result (and most likely the associated teams—as shown in Table 1) is almost always included in the summaries. This baseline is likely to have high precision and lower recall.

We performed three runs of generation: (1) a full run with relevance weights determined by the trained models ("estimated"), (2) a run in which the relevance of the instances is determined from the aligned texts, taking the profile into account ("real w., prof."), and (3) a run like (2), but without taking into account the user profile when determining relevance ("real w., no prof."). Table 3 shows the results of the evaluation for each of the three sources. Precision and recall are obtained by measuring the triples selected by the estimated or baseline model against the triples in the gold standard. The recall is predictably lower in the baseline than in the other runs. The F-measure in the source Marca is considerably lower for the three runs than the baseline. This is because the summaries in this source are very much like short titles (for marca, we had an average of 2 triples per summary vs. 4 for espn and 6 for terra). The runs without profile consideration have a somewhat lower F-measure than those with profile, especially for the two sources with the longest summaries. This shows that considering the profile of the user when selecting content is an important criterion. Finally, the performance of content selection with empirically estimated relevance is comparable to the performance of content selection with relevance taken from the target texts—which indicates that there are benefits in using supervised learning for estimating relevance.

Although a more formal error analysis would be needed, here are a few issues that we encountered during the (manual) counting of the triples for the evaluation:

1. errors in the automatic alignment for goals and red cards;
2. errors in the KB (we found at least a missing instance, and an error in the final score which meant that it was a draw instead of a victory);

source	#triples	baseline			estimated			real w., prof.			real w., no prof.		
		prec.	rec.	F1	prec.	rec.	F1	prec.	rec.	F1	prec.	rec.	F1
espn	157	83.3	57.3	67.9	43.2	77.1	55.4	42.5	79.6	55.4	35.1	85.4	49.7
marca	74	49.0	63.5	55.3	21.8	79.7	34.2	20.2	79.7	32.2	17.7	90.5	29.6
terra	223	98.1	47.5	64.0	54.2	64.1	58.7	56.1	65.9	60.6	44.8	75.8	56.3

Table 3. Content selection evaluation results

3. some inferred triples are missing, among them sets of goals for a given player or a given period of the match (e.g., first half) as well as some relations (e.g., violation of expectation between the fact that team A did not win and team B played with less than 11 players during a determined period of the game);
4. some of the considered triples are never included in the final content plan; for instance, the sets of goals without the listing of the individual goals (to say that a team marked 3 goals).

With respect to the second issue, we would like to point out that although we did not evaluate the correctness of the KB, we are aware that it is not error-free and that more testing and mending is needed. With respect to the third and fourth issues, the question comes up how to systematize the discovery of new inferred knowledge (including relations) and how to get relevance heuristics for content selection. Supervised learning can be unreliable and/or painstaking, especially if the data is scarce and/or requires manual annotation. Another promising avenue of research is to obtain those heuristics from the user using reinforcement learning.

5.2 Discourse Structuring Evaluation

To evaluate the coherence of the final texts, we relied on the evaluation of their *readability* done on 51 matches with three different outputs, one for each of the three user profiles (team A, team B and Neutral) performed by ten evaluators external to the project. As pointed out by [14]: "Fluency concerns the quality of generated text, rather than the extent to which in conveys the desired information. This is related to the notion of 'readability' and will include notions such a syntactic correctness, stylistic appropriateness, **organization and coherence**." *(the emphasis is ours)*

Figure 4 shows the questionnaire on readability passed to the evaluators. The questionnaire consists of a five point scale. We asked the evaluators not to judge the content of the texts as such, but rather their structure (and grammaticality). For each text, we obtained a total of three different judgements. These judgements were averaged to give the text its final score. We obtained an average performance for readability of 88%, which is indicative of the high degree of coherence of the texts.

6 Related Work

Natural Language Generation systems generally use hand-crafted toy knowledge bases (KBs) and/or external databases (DBs) as input. Sometimes, data

Please select one of the following:

5 The text is very easy to read; it seems perfectly natural.
4 The text is easy to read although there are some details that seem unnatural.
3 The text is not too difficult to read, but there are annoying repetitions or abusive agglutination of information in the same sentence.
2 The text is difficult to read, due to the reasons above, but it's still worth an effort.
1 The text is not readable.

Fig. 4. The readability questionnaire put to the subjects

is assessed or evaluated in order to produce new inferred knowledge that is more suitable for being communicated in natural language texts [16, 20]. From the DBs/KBs, it is extracted for inclusion in schemas or discourse plan operators [9]. A few systems reason directly on the input representation [5, 4]. In [5], a *content potential* is constructed based on the domain model (museum artifacts) that consists of predicates between entities, related by discourse relations. The content selection approach consists in weighting the relevant facts from a given node based on a number of manually set criteria and opportunistically navigating the best (and closest) facts. In [4], a content graph is hand-crafted that includes redundancy relations between facts. The nodes in the graph are weighted according to the PageRank centrality formula, with the best ranked nodes selected.

Some work has also been done on empirically estimating the relevance of content using supervised learning in the bibliographical domain [6] and the sports domain [1, 10].

In recent years, there has been also a surge of interest in using OWL knowledge bases for Natural Language Generation (NLG) [3, 8, 15, 21]. These works have mainly focused on verbalizing the taxonomic content of ontologies and/or annotating them with linguistic information for linguistic realization. None, to the best of our knowledge, is dedicated to text planning, be it content selection or discourse structure

7 Conclusions and future work

We have presented an application-independent two-layer OWL ontology and a text planning approach that exploits it. During our work, we have faced two typical problems faced by NLG practitioners when massaging content into text: the mapping between world knowledge and domain communication knowledge [11], and the mapping between world knowledge and linguistic knowledge. The problem related to the first type of mapping was resolved by adding to the basic ontology a second layer populated using inference rules. The problem related to the second type of mapping was resolved by grouping the content nodes into discourse units, which are then mapped onto a (near-standardized) conceptual structure that can be used by linguistic generation. Our content selection works

directly on the evaluation of the relevance of the nodes in the ontology based on empirically obtained relevance weights, a simple user profile and the conceptualized logico-semantic relations instantiated in the KB. The discourse structuring module consists of three well-defined albeit somewhat basic tasks. The evaluation shows that the user profile is an important criterion when selecting content for this domain, and that empirical determination of the relevance is a viable approach.

In the medium-term, we would like to make the tasks of our content selection and discourse structuring modules domain-independent, that is, parametrizable to a given domain but with clearly domain independent mechanisms. This is being addressed by applying the approach for ontology-based content selection to a completely different domain, namely environmental information. Thus, we want to be able to bound the content using a general algorithm that exploits domain-specific specifications of what content is to be bound. Similarly, the mapping operations we have developed for mapping the discourse structure to the conceptual structure should be general, although the actual mapped units are domain-dependent. We also need to develop a set of general purpose content extraction algorithms (such as PageRank [4]) that are applied once the content has been evaluated. We are furthermore working on the implementation of a constraint-based discourse structuring approach to replace the template-based discourse unit ordering task. Additional work is projected on the discourse unit determination, as it is still somewhat dependent on the ordering with respect to the information structure (what to say first where), and, finally on a mapping formalism between logico-semantic and discourse relations that shall include rules that take into account the user's previous knowledge, the content, etc.

The context for achieving our goal is optimal in that we can draw, on the one hand, upon the formalism and tools for inferencing provided by OWL (and thus count on standardized solutions for tasks prior to text planning) and, on the other hand, upon a theoretically motivated and mature linguistic generator, with clearly defined representations of the linguistic description (and thus count on standardized solutions for tasks coming after text planning).

References

1. Barzilay, R. and Lapata, M. 2005. Collective content selection for concept-to-text generation. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing.*
2. Bohnet, B. and Wanner, L. 2010. Open Source Graph Transducer Interpreter and Grammar Development Environment. *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10).*
3. Bontcheva, K. and Wilks, Y. 2004. Automatic Report Generation from Ontologies: the MIAKT approach. *Nineth International Conference on Applications of Natural Language to Information Systems (NLDB'2004).* Manchester, UK.
4. Demir, S., Carberry, S.,McCoy, K.F.. 2010. A Discourse-Aware Graph-Based Content-Selection Framework. *Proceedings of the International Natural Language Generation Conference.*

5. O'Donnell, M., Mellish, C., Oberlander, J., Knott, A. 2001 "ILEX: An architecture for a dynamic hypertext generation system". *Natural Language Engineering*, 7, 225–250.
6. Duboue, P.A., McKeown, K.R. 2003. Statistical Acquisition of Content Selection Rules for Natural Language Generation. *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, 121–128.
7. Dukle, K. 2003. A Prototype Query-Answering Engine Using Semantic Reasoning. *Master Thesis.* University of South Carolina.
8. Galanis, D., Androutsopoulos, I. 2007. Generating Multilingual Descriptions from Linguistically Annotated OWL Ontologies: the NaturalOWL System. *Proceedings of the 11th European Workshop on Natural Language Generation (ENLG 2007).* Schloss Dagstuhl. Germany.
9. Hovy, E.H. 1993. Automated Discourse Generation Using Discourse Structure Relations. *Artificial Intelligence*, 63(1-2):341–386.
10. Kelly, C., Copestake, A., Karamanis, N. 2009. Investigating content selection for language generation using machine learning. *Proceedings of the 12th European Workshop on Natural Language Generation*, 130–137.
11. Kittredge, R., Korelsky, T. and Rambow, O. 1991. On the need for domain communication knowledge. *Computational Intelligence*, 7(4):305–314.
12. Lenat, D. B. 1995. CYC: a large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38, 11 (November 1995), 33-38.
13. Mann, W. and Thompson, S. 1988. Rhetorical Structure Theory: Toward a functional thoery of text organization. *Text*, 8(3).
14. Mellish, C. and Dale, R. 1998. Evaluation in the Context of Natural Language Generation. *Computer Speech and Language*, 12, 349-373.
15. Mellish, C. and Pan, J. 2008. Natural Language Directed Inference from Ontologies. *Artificial Intelligence*, 172(10):1285-1315.
16. Reiter, E. 2007. An Architecture for Data-to-Text Systems. In Proceedings of ENLG-2007, 97–104.
17. Schapire, R.E and Singer, Y. 2000. BoosTexter: A boosting-based system for text categorization. Machine Learning, 39(2/3):135–168.
18. Tsinaraki, C., Polydoros, Kazasis, F., and Christodoulakis, S. 2005. Ontology-based semantic indexing for mpeg-7 and tv-anytime audiovisual content. *Multimedia Tools and Applications*, Vol. 26, Num .3, 2005, pp. 299-325.
19. Uschold, M., and King, M. 1995. Towards a Methodology for Building Ontologies. *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing*, held in conduction with IJCAI-95, pp. 6.1-6.10.
20. Wanner et al. 2010. MARQUIS: Generation of User-Tailored Multilingual Air Quality Bulletins. *Applied Artificial Intelligence*, 24(10):914–952.
21. Wilcock, G. 2003. Talking owls: Towards an ontology verbalizer. In Proceedings of the *Human Language Technology for the Semantic Web and Web Services, ISWC-2003*, 109–112, Sanibel Island, Florida.