Content selection from an ontology-based knowledge base for the generation of football summaries

Nadjet Bouayad-Agha
Gerard Casamayor
DTIC, University Pompeu Fabra
Barcelona, Spain
firstname@lastname.upf.edu

Leo Wanner
ICREA and
DTIC, University Pompeu Fabra
Barcelona, Spain
leo.wanner@icrea.es

Abstract

We present an approach to content selection that works on an ontology-based knowledge base developed independently from the task at hand, i.e., Natural Language Generation. Prior to content selection, a stage akin to *signal analysis* and *data assessment* used in the generation from numerical data is performed for identifying and abstracting patterns and trends, and identifying relations between individuals. This new information is modeled as an extended ontology on top of the domain ontology which is populated via inference rules. Content selection leverages the ontology-based description of the domain and is performed throughout the text planning at increasing levels of granularity. It includes a main topic selection phase that takes into account a simple user model, a set of heuristics, and semantic relations that link individuals of the KB. The heuristics are based on weights determined empirically by supervised learning on a corpus of summaries aligned with data. The generated texts are short football match summaries that take into account the user perspective.

1 Introduction

Content selection (or determination) forms one of the major tasks in Natural Language Generation (NLG). Traditionally, it has been done from purpose-built KBs intertwined with discourse structuring; see, e.g., (Hovy, 1993; Moore and Paris, 1993). In an attempt to systematize the structure of the used KBs and to build an intermediate knowledge-oriented layer between them and linguistic structures, language-oriented ontologies such as the Upper Models (Bateman et al., 1990; Henschel, 1992, 1993; Bateman et al, 1995) have been developed. However, in view of the rise of the semantic web and the rapidly increasing volumes of KBs codified in OWL/RDF, the question on content selection from large scale purpose-neutral ontologies becomes very essential—at least for practical applications of NLG— and has scarcely been addressed.

In what follows, we present a framework for content selection from large scale OWL/RDF ontology-based domain KBs that were developed independently from the task of NLG. The framework is novel in that it (i) foresees a separation of the *domain communication* ontology from the general purpose domain ontology, and (ii) implements mechanisms for selecting content from large scale (at least for NLG standards) ontology-based knowledge bases.

To identify and abstract regular patterns and trends and introduce semantic relations between the individuals of a generic domain ontology, which are critical for high quality generation, but absent from any general purpose ontology, prior to content selection a stage akin to *signal analysis* and *data assessment* used for the generation from numerical data (Reiter, 2007; Wanner et al., 2010) is performed. This new information is modeled as an additional layer on top of the domain ontology, which is populated via rule-based inferences. Content selection proper then takes place at a number of levels of increasing granularity. First, a content bounding task is in charge of selecting, based on the user query, a subset of the KB that includes the maximal set of information that might be communicated to the user. Next the main topics to be included in the content plan are selected, taking into account: 1) a user model, 2) a set of heuristics, and 3) the seman-

Proceedings of the 13th European Workshop on Natural Language Generation (ENLG), pages 72–81,
Nancy, France, September 2011. ©2011 Association for Computational Linguistics

tic relations that link individuals of the KB. Finally, discourse unit determination in the discourse structuring submodule is in charge of deciding which details to include (or not) in each message. The whole text planning procedure that includes both content selection and discourse structuring is presented in (Bouayad-Agha et al., 2011).

The framework has been implemented with a KB that models the First Spanish Football League competitions for the generation (in Spanish) of short user perspective-tailored summaries of the individual matches. The user model is a simple model that contains the preference of the user for one of the teams. The content bounding parameters include the time, location and protagonists of the match of interest. The heuristics are based on weights determined empirically by supervised learning on a corpus of summaries aligned with data, as in (Duboue and McKeown, 2003). The following is an example generated summary:[1]

> "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."

The first and the last sentences of the text are template-based. The content selection strategy is responsible for dynamically selecting the contents used to generate the text in between. For this example the system selected 30 RDF triples involving 17 individuals and 8 datatype values. For example, the fragment "a goal by Ronaldinho in minute 34 and another goal by Eto'o in minute 56" is generated from the following 6 triples: `minute(goal-1, 34)`, `player(goal-1, player-1)`, `name(player-1, Ronaldinho)`, `minute(goal-2, 56)`, `player(goal-2, player-2)`,`name(player-2, Eto'o)`.

In the next section, we outline the base and extended ontologies and their corresponding knowledge bases. In Section 3, we discuss the ontology-based content selection procedure. In Section 4, we present a corpus-based evaluation of the content selection procedure, before reviewing some related work in Section 5 and providing some conclusions and discussing future work in Section 6.

2 Creation of an ontology-based KB

In an ontology-based KB, the KB is an instantiation (or population) of the corresponding ontologies. In what follows, we thus first outline the (manual) design of the ontology underlying our framework and describe then their (automatic) instantiation (or population).

2.1 Design of the ontology

As mentioned in Section 1, our framework foresees a two-layer ontology, the base ontology and the extended ontology. The **base ontology** models the domain in question, namely a football league competition. It is composed of two different ontologies: an object ontology which deals with structural information of the domain and an event ontology. The object ontology contains the specification of the teams, competition phases, matches, players, etc. The event ontology covers the events that may happen in a match (penalties, goals, cards, etc.). The object base ontology consists of 24 classes and 42 properties, with 4041 instances in the corresponding KB; the top level classes of the object ontology are: Competition, Match, Period, Person, Result, Season, Team, TeamCompositionRelation and Title. The event ontology consists of 23 classes and 8 properties, with 63623 instances in the corresponding KB; the top level classes of the event ontology are: ActionFault, Card, Corner, Fault, FaultKick, Goal, GoalKick, Interception, OffSide, Pass, Stop, Throw-in, Shot and Substitution.

The **extended ontology** models types of knowledge that can be considered as inferred from the concepts of the base ontology. This knowledge and consequently the rules to infer it were obtained by manual analysis of a subset of the corpus of football match summaries described in Subsection 3.2 below. It includes (i) the most frequently verbalized

[1] 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.'

concepts that could be deduced from the events and states of a match specified in the base ontology,[2], and (ii) the semantic relations that implicitly hold between the individuals of the base and extended ontology concepts.[3]

The knowledge deduced from the events and states of a match is divided into five categories, each of them captured by several classes in the extended ontology: 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). *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. Furthermore, it is of relevance to obtain after each gameweek a team's tendency (ascending, descending, stable) and distance with respect to its previous classification. 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.). *Match time*-related knowledge models the state of the match along its duration, creating intermediate results after each goal. 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 and the number of players left after an expulsion.

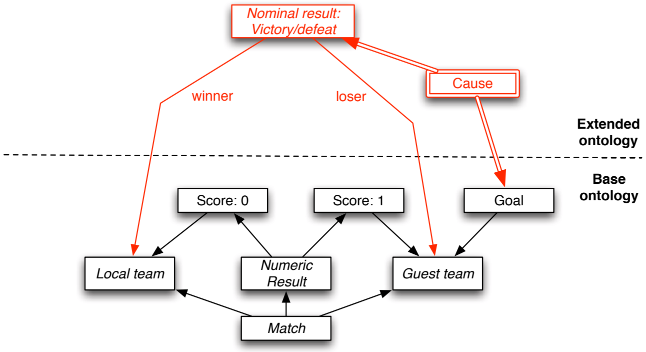

Figure 1: Fragment of the base and extended ontologies

In total, the five categories are modeled by 18 classes, among them: NominalResult, CompetitionResult, Tendency (a team's change of zone in the competition), Distance (to a higher/lower zone), Set, ConstituentSet,[4] Expulsion, PlayersInField, and IntermediateResult.

Consider Figure 1 for illustration.

Each class of deduced knowledge triggers the inference of a number of semantic relations; 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*

[2] Statistical information about matches within a season and across seasons (best scorer, consecutive wins, first victory in a given stadium, etc.), although mentioned in human produced summaries, has been excluded for now since it requires the assessment of a sequence of matches.

[3] More marginally, the extended ontology contains some information added to make the navigation easier for the mapping to linguistic realization and for the inference of new knowledge—for instance, 'for' and 'against' properties are added to the Goal class in order to know which team scored the goal and which team received it as this information was only available indirectly in the base ontology via the player who scored the goal.

[4] Set and ConstituentSet also allow us to simply refer to the number of constituents within it (cf. *the team had two red cards*).

goes up in the classification whilst team B goes down).

The semantic relations are modeled in terms of the class LogicoSemanticRelation and subclasses such as Cause, Implication, ViolationOfExpectation, Meronymy, Precedence, and Contrast.

2.2 Creation of the KB

The base KB has been 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.[5] The extended ontology population was carried out using the inference engine provided by Jena.[6] 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:NumResult)
(?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:resultDiff "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.

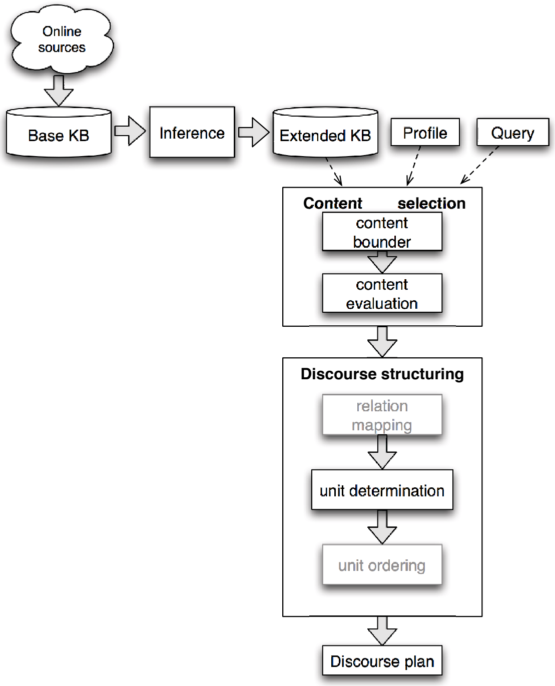

Figure 2: The view on text planning involving content selection *(the sub-modules that do not perform any content selection are grayed out)*

3 Ontology-based content selection

3.1 Approach to content selection

As mentioned in Section 1, content selection is performed at different stages of text planning, in increasing granularity. It includes content bounding and main topic selection performed within the content selection module proper, and fine-grained content selection performed during the discourse unit determination task of the discourse structuring module; see Figure 2 for the overall picture of text planning in which content selection is involved.

The content bounding sub-module selects from the ontology-based KB individuals that are relevant to the match for which a text is to be generated and the semantic relations that link these individuals. The selection works with a set of hand-written rules that draw upon relevance criteria concerning the direct involvement of the individuals (e.g., the players of the teams in question, goals during the match, etc.) and the general context of the competition (e.g., the league’s classification).

Given the large size (by NLG standards) of the

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

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

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 fragment of the KB which constitutes the maximal set of data available for generating any sort of summary for a given match.

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 semantic relations that link individuals in 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 favorite team. The heuristics measure relevance according to empirical knowledge extracted from a corpus of texts.[7] 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 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 Subsection 3.2 below, we describe how the relevance measures were empirically obtained.

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. The individuals that are not included in those discourse units are excluded from the final text. 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
```

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 consist of explicitly marked up title, summary and 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. These categories were automatically marked up using the alignment of text with data described below. 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.

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.

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

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

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

For the description of a goal or a red card, we used the same set of over 100 feature types since we considered them both as match events. The features include information about the current event (minute, event number in the match), the player involved (name,position, proportion of goals/cards in the match and in the season up to the match, proportion of games played in season up to the match, etc), the current game, gameweek, season and team (including classification and statistical information), and comparison of the current event with previous and next event of the same class (e.g., deltas of minute, player and team).

For modeling the classification, we used a more systematic approach to feature extraction by regarding a team's classification as the event of a specific gameweek, comparing it to the events of the previous gameweek—that is, to the 20 classifications[8] 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 feature types.

In order to classify the data, we used Boostexter (Schapire and Singer, 2000), a boosting algorithm that uses decision stumps over several iterations and that has already been used in previous works on training content selection classifiers (Barzilay and Lapata, 2005; Kelly et al., 2009).[9] For each of the three categories (goal, red card, classification), we experimented with 15 different classifiers by considering a section dimension (title, summary and title+summary) and a source dimension (espn, marca, terra, any one of them (any) and at least two of them). We divided the corpus each time into 90-10% of the matches for training and testing.

4 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) evaluation of the content selection as a whole.

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 results correspond to considering title and summary from a source together, given that the results are not significant when considering summary section only (accuracy is 78.1%, baseline accuracy is 65.4% and t = 4.4869 with p<0.0001). In all cases, the best performance is obtained by considering the content from any of the online sources.

The evaluation of the content selection as a whole is done by comparing the content of generated summaries with that of existing summaries (the gold standard). We say "as a whole" since this evaluation also considers the template-based content selection performed during discourse unit determination.[10]

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

[8] The Spanish League competition involves 20 teams.

[9] After a number of experiments, the number of iterations was set to 300.

[10] However, we do not evaluate discourse unit determination itself.

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	62	85%	53%	t = 4.4869 (p<0.0001)
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 (+title in case of red cards)

three different web sources (namely espn, marca, terra). We compiled a list of all individuals considered for inclusion in the content selection and discourse unit determination modules and for which explicitly references could be found in target texts, including instances of the semantic relations, which were modelled as classes in the KB. For each of the 108 (36×3) summaries, we manually annotated whether an individual 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.[11] 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. In the context of sports commentaries, readers usually tolerate better a certain excess of information than lack of (relevant) information. Therefore, recall can be considered of higher prominence than precision.

Precision and recall are obtained by measuring the individuals included in the content plan by the estimated or baseline model against the individuals mentioned 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 individuals mentioned per summary vs. 4 for espn and 6 for terra). The runs without profile have understandably a higher recall since content selection is less discriminative without a user profile (or rather with a *neutral* user profile). Nonetheless, they show a somewhat lower F-measure than those with a profile, especially for the two sources with the longest summaries. 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 individuals 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);
3. some inferred content is 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);

[11] Our observation is that sports commentaries (at least in web-based news media) are by far not always neutral and address thus readers with a specific (biased) profile.

source	#individuals	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

4. some of the considered individuals 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, 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 Related Work

The task of content selection in NLG can be characterized along three dimensions: 1) *what* is the source of the content, 2) *where* in the generation pipeline it is selected, and 3) *how* it is selected. The first dimension specifies, for instance, whether the content is structured or unstructured data in a relational database or hierarchical knowledge in a knowledge base, and whether the data / knowledge representation is built for the purposes of NLG or whether it is task-independent. The second dimension specifies whether content selection occurs before the actual generation (as an expert system task) or during it, and whether it is performed in a separate module or is integrated to a lesser or greater degree with other tasks. The third dimension reflects the strategy used: statistical or symbolic, top-down or bottom-up. Traditionally, content selection in NLG involves structured, purpose-built KBs processed using symbolic top-down approaches such as schemas or plan-based operators that perform content selection together with discourse structuring; see, e.g., (Hovy, 1993; Moore and Paris, 1993).

In a step towards more flexible content selection, (O'Donnell et al., 2001) put forward a proposal to select content by navigating a text potential. Also, in the recent past, determination of the relevant episodes in large time-series gained prominence (Yu et al., 2007; Portet et al., 2009). Although some of the data of a football league competition can also be expressed in terms of a time-series, in general, it goes beyond a numeric attribute-value pair sequence.

Statistical techniques on numerical data have also been investigated—among them (Duboue and McKeown, 2003; Barzilay and Lapata, 2005; Demir et al., 2010). Some of these techniques use classifiers trained with supervised learning methods to decide on the selection of individual units of data (e.g., a row of a table in a relational database, or entities in an RDF graph). Others construct a graph-based representation of the content and apply an optimisation algorithm for network analysis (i.e. a flow or a centrality algorithm) to find out the most relevant subset of content.

Ontologies have a long standing tradition in NLG, the most notable of which is the Upper Model (Bateman et al., 1990; Henschel, 1992, 1993; Bateman et al, 1995) which is a a linguistically motivated ontology. More directly related to our approach are ontology-oriented proposals in NLG whether to leverage linguistic generation (Bontcheva and Wilks, 2004), to verbalize ontologies (Wilcock, 2003; Power and Third, 2010) or to select content for the purpose of ontology verbalization (Mellish and Pan, 2008).

6 Conclusions and future work

We have presented an NLG content selection approach performed on a task-independent ontology-based knowledge base. The lack of domain communication knowledge (Kittredge et al., 1991) in the ontology was remedied by adding to the basic ontology a second layer populated using inference rules that includes the modelling of semantic relations between individuals. Ontological information, that is knowledge of classes and properties, was exploited at all stages of content selection, whether using schemas or empirically determined relevance measures for the main classes to include in the target text.[12] This latter task of selecting the main topics that are to be included in the final text takes into account coherence by exploiting the semantic relations between individuals, and the wanted perspective on the generated text by incorporating a simple user model and relevance measures empirically determined on a corpus of aligned text and data pairs. In the future, instead of using a heuristic-based content extraction approach for the main topic selection task, we plan to apply a set of general purpose content extraction algorithms such as PageRank (Demir et al., 2010).

In the medium-term, we also plan 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 goal is currently being addressed by applying the approach to ontology-based content selection to a completely different domain, namely environmental information. The environmental domain has been modeled in an ontology-based knowledge base which has been extended with domain communication knowledge. We want to be able to bound the content using a general algorithm that exploits domain-specific criteria.

We are also planning additional work on discourse unit determination, as it is still template-based and thus of restricted flexibility.

References

Regina Barzilay and Mirella Lapata. 2005. Collective Content Selection for Concept-to-Text Generation. *Proceedings of the Joint Human Language Technology and Empirical Methods in Natural Language Processing Conferences (HLT/EMNLP-2005)* Vancouver, Canada.

John A. Bateman, Robert T. Kasper, Johanna D. Moore, and Richard A. Whitney 1990 A General Organization of Knowledge for Natural Language Processing: the Penman Upper Model *Technical Report.* USC/Information Sciences Institute, Marina del Rey, California.

John A. Bateman, Renate Henschel, and Fabio Rinaldi 1995 Generalized Upper Model 2.0: documentation. *Technical Report.* GMD/Institut für Integrierte Publikations- und Informationssysteme, Darmstadt, Germany.

Kalina Bontcheva and Yorick Wilks. 2004. Automatic Report Generation from Ontologies: the MIAKT approach. *Proceedings of the Nineth International Conference on Applications of Natural Language to Information Systems (NLDB'2004).* Manchester, UK.

Nadjet Bouayad-Agha, Gerard Casamayor, Leo Wanner, Fernando Díez, and Sergio López Hernández. 2011. FootbOWL: Using a generic ontology of football competition for planning match summaries. *Proceedings of the 8th Extended Semantic Web Conference (ESWC2011).* Heraklion, Greece.

Seniz Demir, Sandra Carberry and Kathleen F. McCoy. 2010. A Discourse-Aware Graph-Based Content-Selection Framework. Proceedings of the International Language Generation Conference. Sweden.

Pablo A. Duboue and Kathleen R. McKeown. 2003. Statistical Acquisition of Content Selection Rules for Natural Language Generation. Proceedings of the 2003 conference on Empirical Methods in Natural Language Processing (EMNLP). Sapporo, Japan.

Renate Henschel 1992, 1993. Merging the English and German Upper Models. *Technical Report.* GMD/Institut für Integrierte Publikations- und Informationssysteme, Darmstadt, Germany.

Eduard Hovy. 1993 *Automated discourse generation using discourse relations.* Artificial Intelligence. 63 , 341 – 385.

Pamela W. Jordan and Marilyn A. Walker 2005 *Learning content selection rules for generating object descriptions in dialogue* Journal of Artificial Intelligence Research 24, 157–194.

[12] As pointed out by Referring Expression Generation researchers (Jordan and Walker, 2005), content selection occurs also further down the chain; for example, during the selection amongst the property for name, dorsal number, and role (e.g., attacker) to refer to a given player. In our generator, these properties are passed down to the linguistic generator for selection, although ad-hoc rules are used rather than strict ontological knowledge.

Colin Kelly, Ann Copestake, and Nikiforos Karamanis. 2009 Investigating content selection for language generation using machine learning. *Proceedings of the 12th European Workshop on Natural Language Generation.*. 130–137.

Richard Kittredge, Tanya Korelsky, and Owen Rambow. 1991. On the need for domain communication knowledge. *Computational Intelligence* 7(4):305–314.

Chris Mellish and Jeff Z. Pan. 2008 Language Directed Inference from Ontologies. *Artificial Intelligence*. 172(10):1285-1315.

Johanna D. Moore and Cécile L. Paris. 1993 Planning texts for advisory dialogs: capturing intentional and rhetorical information. *Computational Linguistics*. 19(4), 651-694.

Mick ODonnell, Chris Mellish, Jon Oberlander, and Alistair Knott. 2001. ILEX: an architecture for a dynamic hypertext generation system. *Natural Language Engineering*. 7(3):225–250.

François Portet, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer and Cindy Sykes. 2009. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence* 173(7-8): 789-816.

Richard Power and Allan Third. 2010. Expressing OWL axioms by English sentences: dubious in theory, feasible in practice. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*. Beijing, China.

Ehud Reiter 2007. An Architecture for Data-to-Text Systems. *Proceedings of the 11th European Natural Language Generation* Schloss Dagstuhl, Germany. page 97-104.

Robert E. Schapire and Yoram Singer 2000 BoosTexter: A boosting-based system for text categorization. *Machine Learning* 39(2/3):135–168.

Leo Wanner, Bernd Bohnet, Nadjet Bouayad-Agha, Francois Lareau, and Daniel Nicklaß. 2010 MARQUIS: Generation of User-Tailored Multilingual Air Quality Bulletins. *Applied Artificial Intelligence*. 24(10):914–952.

Graham Wilcock 2003 Talking owls: Towards an ontology verbalizer. *Proceedings of the Human Language Technology for the Semantic Web and Web Services, ISWC-2003*. 109–112. Sanibel Island, Florida.

Jin Yu, Ehud Reiter and Jim Hunter. 2007 Choosing the content of textual summaries of large time-series data sets. *Natural Language Engineering*. 13:25-49.