Discourse Generation from Formal Specifications Using the Grammatical Framework, GF

Dana Dannélls

NLP Research Unit, Department of Swedish Language,
University of Gothenburg, Sweden
dana.dannells@svenska.gu.se

Abstract. Semantic web ontologies contain structured information that do not have discourse structure embedded in them. Hence, it becomes increasingly hard to devise multilingual texts that humans comprehend. In this paper we show how to generate coherent multilingual texts from formal representations using discourse strategies. We demonstrate how discourse structures are mapped to GF's abstract grammar specifications from which multilingual descriptions of work of art objects are generated automatically.

Key words: MLG, Ontology, Semantic Web, CIDOC-CRM, Cohesion, Discourse strategies, Functional programming.

1 Introduction

During the past few years there has been a tremendous increase in promoting metadata standards to help different organizations and groups such as libraries, museums, biologists, and scientists to store and make their material available to a wide audience through the use of the metadata model RDF (Resource Description Framework) or the Web Ontology Language (OWL) [1, 2]. Web ontology standards offer users direct access to ontology objects; they also provide a good ground for information extraction, retrieval and language generation that can be exploited for producing textual descriptions tailored to museum visitors. These advantages have brought with them new challenges to the Natural Language Generation (NLG) community that is concerned with the process of mapping from some underlying representation of information to a presentation of that information in linguistic form, whether textual or spoken. Because the logical structure of ontologies becomes richer, it becomes increasingly hard to devise appropriate textual presentation in several languages that humans comprehend [3].

In this article we argue that discourse structures are necessary to generate natural language from semantically structured data. This argument is based on our investigations of text cohesive and syntactic phenomena across English, Swedish and Hebrew in comparable texts. The use of a discourse strategy implies that a text is generated by selecting and ordering information out of the underlying domain ontology, a process which provides a resulting text with

Special issue: Natural Language Processing and its Applications.
Research in Computing Science 46, 2010, pp. 167-178

Received 19/11/09
Accepted 16/01/10
Final version 09/03/10

fluency and cohesion. It is an approach that relies on the principles drawn from both linguistic and computer science to enable automatic translation of ontology specifications to natural language. We demonstrate how discourse structures are mapped to GF's abstract grammar specifications from which multilingual descriptions of work of art objects are generated automatically. GF is a grammar formalism with several advantages which makes it suitable for this task – we motivate the benefits GF offers for multilingual language generation. In this work, we focus on the cultural heritage domain, employing the ontology codified in the CIDOC Conceptual Reference Model (CRM).

The organization of this paper is as follows. We present some of the principles of cohesive text structure (section 2) and outline the difficulties of following these principles when generating from a domain ontology (section 3). We show how discourse strategies can bridge the gap between formal specifications and natural language and suggest a discourse schema that is characteristic to the cultural heritage domain (section 4). We demonstrate our grammar approach to generating multilingual object descriptions automatically (section 5). We conclude with a summary and provide pointers to future work (section 6).

2 Global and Local Text Structure

Early work on text and context [4] has shown that cultural content is reflected in language in terms of text as linguistic category of genre, or text type. A text type is defined as the concept of Generic Structure Potential (GSP) [5]. According to this definition, any text, either written or spoken, comprises a series of optional and obligatory macro (global) structural elements sequenced in a specific order and that the obligatory elements define the type to which a text belongs. The text type that is expressed here is written for the purpose of describing work of art objects in a museum.

To find the generic structure potential of written object descriptions, we examined a variety of object descriptions, written by four different authors, in varying styles. Our empirical evidence suggest there is a typical generic structure potential for work of art descriptions that has the following semantic groupings:

1. object's title, date of execution, creation place
2. name of the artist (creator), year of birth/death
3. inventory number when entered to the museum, collection name
4. medium, support and dimensions (height, width)
5. subject origin, dating, function, history, condition.

To produce a coherent text structure of an object description the author must follow this semantic specification sequences that convey the macro structure of the text. Apart from the macro structural elements, there is a micro (local) integration among semantic units of the text type that gives the text a unity. These types are reflected in terms of reference types that may serve in making a

text cohesive at the paragraph or embedded discourse level. Some examples of reference types are: conjunction, logical relationships between parts of an argument, consistency of grammatical subject, lexical repetition, consistency of temporal and spatial indicators. Thus local structure is expressed partly through the grammar and partially through the vocabulary.

3 The Realities of a Domain Specific Ontology

The ontology we utilize is the Erlangen CRM. It is an OWL-DL (Description Logic) implementation of The International Committee for Documentation Conceptual Reference Model (CIDOC-CRM) [6].[1] The CIDOC-CRM is an event-centric core domain ontology that is intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information and museum documentation.[2] One of the basic principles in the development of the CIDOC CRM has been to have empirical confirmation for the concepts in the model. That is, for each concept there must be evidence from actual data structures widely used. Even though the model was initially based on data structures in museum applications, most of the classes and relationships are surprisingly generic. In the following we use this model to illustrate the limitation imposed by a domain specific ontology on generation where concepts and relationships can not easily be mapped to natural language.

According to the CIDOC-CRM specifications, a museum object is represented as an instance of the concept *E22.Man_Made_Object*, which has several properties including:[3] *P55.has_current_location*, *P108B.has_dimension*, *P45F.consists_of*, *P101F.had_general_use*, *P108B.was_produced_by*. A concrete example of a formal specification (presented in turtle annotation) of the *RestOntheHunt_PE34604* object that was modeled according to the CIDOC Documentation Standards Working Group (DSWG) is given in Figure 1.

Taking the domain ontology structure as point of departure, the information in hand is an unordered set of statements that convey a piece of information about an object. The information the *RestOntheHunt_PE34604* statements convey spans at least four of the semantic sequences that we outline in section 2. To generate a coherent text, some ordering constraints must be imposed upon them. This is in particular important because a statement may map to an addition set of statements about an object, for example the relationship *P108B.was_produced_by* maps to an instance of the concept *E12.Production* that has the following properties: *P14F.carried_out_by*, *P7F.took_place_at*, *P4F.has_time_span*.

[1] The motivation behind the choice of DL is that it allows tractable reasoning and inference; it ensures decidability, i.e. a question about a concept in the ontology can always be answered; it supports the intuition that the model must be clear, unambiguous and machine-processable. These aspects are in particular important in computational setting, where we would like our logic to be processed automatically.

[2] The model was accepted by ISO in 2006 as ISO21127.

[3] *Property* is a synonym for *relationship* that maps between two instances. In this paper we use the term *statement* to refer to a relationship between instances.

```
1.:RestOntheHunt_PE34604
2.        a     :CIDOC5.0.1.rdfsE22.Man-Made_Object;
3.  :CIDOC5.0.1.rdfsP30B.custody_transferred_through       :RestOntheHunt_PE34604;
4.  :CIDOC5.0.1.rdfsP51F.has_former_or_current_owner     :Museum_Hallwylska;
5.  :CIDOC5.0.1.rdfsP51F.has_former_or_current_owner     :Galleria_Spada;
6.  :CIDOC5.0.1.rdfsP43F.has_dimension     :width_1.2_m;
7.  :CIDOC5.0.1.rdfsP43F.has_dimension     :length_1.54_m;
8.  :CIDOC5.0.1.rdfsP50F.has_current_keeper     :Museum_Hallwylska;
9.  :CIDOC5.0.1.rdfsP52F.has_current_owner     :Museum_Hallwylska;
10. :CIDOC5.0.1.rdfsP48F.has_preferred_identifier     :PE_34604;
11. :CIDOC5.0.1.rdfsP2F.has_type     :oil_cloth;
12. :CIDOC5.0.1.rdfsP103F.was_intended_for       :RestOntheHunt_function;
13. :CIDOC5.0.1.rdfsP101F.had_as_general_use     :RestOntheHunt_function;
14. :CIDOC5.0.1.rdfsP53F.has_former_or_current_location     :Stockholm;
15. :CIDOC5.0.1.rdfsP53F.has_former_or_current_location     :Rome;
16. :CIDOC5.0.1.rdfsP1F.is_identified_by     :PE_34604;
17. :CIDOC5.0.1.rdfsP1F.is_identified_by     :TA_959a;
18. :CIDOC5.0.1.rdfsP65F.shows_visual_item     :Inscription_of_RestOntheHunt_PE34604;
19. :CIDOC5.0.1.rdfsP45F.consists_of     :linseed_oil;
20. :CIDOC5.0.1.rdfsP45F.consists_of     :canvas_duck;
21. :CIDOC5.0.1.rdfsP62F.depicts         :Bambocciate;
22. :CIDOC5.0.1.rdfsP55F.has_current_location     :Stockholm;
23. :CIDOC5.0.1.rdfsP49F.has_former_or_current_keeper       :Museum_Hallwylska;
24. :CIDOC5.0.1.rdfsP54F.has_current_permanent_location     :Stockholm;
25. :CIDOC5.0.1.rdfsP70B.is_documented_in       :RestOntheHunt.jpg;
26. :CIDOC5.0.1.rdfsP108B.was_produced_by     :Creation_of_RestOntheHunt_PE34604 .
```

Fig. 1. Formal specification of a museum object modeled in the CIDOC-CRM.

4 From Formal Specifications to Coherent Representation

As we pointed out in the previous section, the structure of the ontology is not a good point of departure for producing coherent texts and therefore requires pre-processing. In broad terms this involves taking a set of information elements to be presented to a user and imposing upon this set of elements a structure which provides a resulting text with fluency and cohesion.

Some of the pre-processing steps that have been suggested by previous authors [7, 8] include removing repetitive statements that have the same property and arguments and grouping together similar statements to produce a coherent summary. Although there is a need to select statements that mirror linguistic complexity [9], most authors focus on the semantics of the ontology rather than on the syntactic form of the language. They assume that the ontology structure is appropriate for natural language generation, an assumption which in many cases only applies to English.

In this section we describe the approach we exploit to learn how the ontology statements are realized and combined in natural occurring texts. We perform a domain specific text analysis; texts are studied through text linguistics by which the critic seeks to understand the relationships between sections of the author's discourse.

4.1 Linking Statements to Lexical Units

When text generation proceeds from a formal representation to natural language output, the elements of the representation need to be somehow linked

to lexical items of the language. We examined around 100 object descriptions in English, Swedish and Hebrew and studied how statements are ordered, lexicalised and combined in the discourse. To capture the distribution of discourse entities across text sentences we perform a semantic and syntactic analysis, we assume that our unit of analysis is the traditional sentence, i.e. a main clause with accompanying subordinate and adjunct clauses. Below we exemplify how the ontology statements are mapped to lexical items in the studied texts.[4]
Statements:

1. *P55F. has_current_location* maps between instances of *E22.Man-Made-Object* and instances of *E53.Place* (see line 22, Figure 1)
2. *P52F. has_current_owner* maps between instances of *E22.Man-Made-Object* and instances of *E40. Legal Body* (see line 9, Figure 1)
3. *P82F.at_some_time_within* maps between instances of *E52. Time-Span* and *String* data values.

Text examples:
Eng> The subject made its first appearance [in 1880]$_{P82F}$. It is [now installed]$_{P52F}$ in the Wallace Collection[,]$_{P55F}$ London.

Swe> Först [på 1900 talet]$_{P82F}$ kom den till Sverige och [hänger nu på]$_{P55F}$ Gripsholms slott [i]$_{P52F}$ Statens porträttsamling.

Heb> ha-tmuwnah hegieh larisunah le-Aeretz yisraAel [be-snat 1960]$_{P82F}$. hyA [sayeket le]$_{P52F}$-quwleqitzyah sel Amir bachar [se-nimtzet]$_{P55F}$ be-muwzeyAuwn haAretz be-tel Aabiyb

These text examples exhibit a few local linguistic differences between the languages. In English and Hebrew, the order of the statements is: 3,2,1 while in the Swedish text it is: 3,1,2. It is interesting to note how the domain entities and properties are lexicalized in the different languages. In all three languages the property *P82F.at_some_time_within* is lexicalised with a preposition phrase. On the other hand, the lexicalisation of the property *P55F. has_current_location* differs significantly. Furthermore, in the Swedish text all statements are realized in one single sentence; the statements are combined with a simple syntactic aggregation using the conjunction *och* 'and'. Both in the English and the Hebrew examples, statements 3 and 2 are realized as two sentences which are combined with a referring pronoun, i.e. *it* and *hyA*. When generating natural occuring texts it is important to utilize a generation machinery that supports such syntactic variations. In section 5 we demonstrate how these variation are supported in the GF formalism.

Empirical representations of stereotypical clause structures such as presented above not only provide evidence on how to pair ontology statements with lexical units according to the language specific patterns, but also guide template constructions proceeding according to the organization of the domain semantics.

[4] The transliteration ISO-8859-8 ASCII characters of Hebrew are used to enhance readability.

Table 1. Template specification that governs text structures of a cultural object in a museum.

Name	Template slot
T1	(a) object's title | (b) object's creator | (c) creation date | (d) creation place
T2	(a) creator date of birth | (b) creator date of death
T3	(a) object id | (b) object material | (c) object size
T4	(a) current owner | (b) current location | (c) catalogue date | (d) collection
T5	(a) object's identifier | (b) identified place

4.2 Template Specifications

In section 2 we presented a five stage typical GSP for a work of art object description. To guarantee that the selected statements follow this structure, we defined a sequence of templates describing the discourse structure, this approach was first introduced by [10]. Each sequence in a template consists of slots that correspond to a set of statements in the domain knowledge.

The template specification as whole provides a set of ordering constraints over a pattern of statements in such a way that may yield a fluent and coherent output text. The templates and slots are specified in Table 1.

4.3 A Discourse Schema

A discourse schema is an approach to text structuring through which particular organizing principles for a text are defined. It straddles the border between a domain representation and well-defined structured specification of natural language that can be found through linguistic analysis. This idea is based on the observation that people follow certain standard patterns of discourse organization for different discourse goals in different domains.

Our text analysis has shown certain combinations of statements are more appropriate for the communicative goal of describing a museum object. Following our observations, we defined a discourse schema *Description schema* (see below) consisting of two rhetorical predicates (e.g. Identification–Property and Attributive–Property).[5] The schema encodes communicative goals and structural relations in the analyzed texts. Each rhetorical predicate in the schema is associated with a set of templates (specified in Table 1). The notation used to represent the schema: ',' indicates the mathematical relation *and*, '{}' indicates optionality, '/' indicates alternatives.

Description schema:

Describe–Object − >
  Identification–Property/
  Attributive–Property

Identification–Property − >

[5] The notion of *rhetorical predicates* goes back to Aristotle, who presented predicates as assertions which a speaker can use for persuasive argument.

T1 , {T2 / T3}

Attributive–Property − >
T4 / T5

An example taken from one of the studied texts:

> [T1b]Thomas Sully [T2](1783-1872) painted this half-length [T1a] Portrait of Queen Victoria [T1c] in 1838. The subject is now installed in the [T4d] Wallace Collection, [T4b] London.

The first sentence, corresponding to the rhetorical predicate *Identification–Property*, captures four statements (comprising the following relationships: *P82F.at_some_time_within*, *P14F.carried_out_by*, *P108B.was_produced_by* and *P102.has_title*) that are combined according to local and global text cohesion principles.

5 Domain Dependent Grammar-Based Generation

After the information from the ontology has been selected and organized according to the pre-defined schema, it is translated to abstract grammar specifications. The grammar formalism is the Grammatical Framework (GF) [11], a formalism suited for describing both the semantics and syntax of natural languages. The grammar is based on Martin-Löf's type theory [12] and is particularly oriented towards multilingual grammar development and generation. GF allows the separation of language-specific grammar rules that govern both morphology and syntax while unifying as many lexicalisation rules as possible across languages. With GF it is possible to specify one high-level description of a family of similar languages that can be mapped to several instances of these languages. The grammar has been exploited in many natural language processing applications such as spoken dialogue systems [13], controlled languages [14] and generation [15].

GF distinguishes between abstract syntax and concrete syntax. The abstract syntax is a set of functions (fun) and categories (cat) that can be defined as semantic specifications; the concrete syntax defines the linearization of functions (lin) and categories (lincat) into strings that can be expressed by calling functions in the resource grammar. [6] Each language in the resource grammar has its own module of inflection paradigms that defines the inflection tables of lexical units and a module for specifying the syntactic constructions of the language.

Below we present the abstract and concrete syntax of the rhetorical predicate *Identification–Property* presented in section 4.3.[7] Figure 2 illustrates the abstract syntax tree of our abstract grammar that reflects on the semantics of the domain and that is common for all languages.

[6] A resource grammar is a fairly complete linguistic description of a specific language. GF has a resource grammar library that supports 14 languages.

[7] The GF Resource Grammar API can be found at the following URL: `<http://www.grammaticalframework.org/lib/doc/synopsis.html>`.

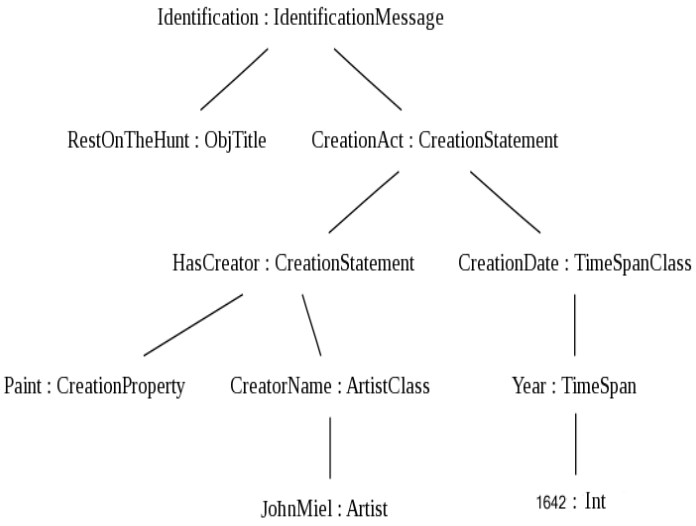

Fig. 2. Abstract syntax tree for *Rest on the Hunt was painted by John Miel in 1642.*

abstract syntax
cat
IdentificationMessage; ObjTitle; CreationProperty; Artist; TimeSpan; CreationStatement; ArtistClass; TimeSpanClass;
fun
Identification: ObjTitle → CreationStatement → IdentificationMessage;
CreationAct: CreationStatement → TimeSpanClass → CreationStatement;
HasCreator: CreationProperty → ArtistClass → CreationStatement;
CreatorName: Artist → ArtistClass;
CreationDate: TimeSpan → TimeSpanClass;
Year : Int → TimeSpan ;
RestOnTheHunt: ObjTitle;
JohnMiel: Artist;
Paint: CreationProperty;

The abstract specification expresses the semantics of the ontology and is language independent. What makes the abstract syntax in particular appealing in this context is the ability to expand the grammar by simply adding new constants that share both common semantics and syntactic alternations. For example, Beth Levin's [16] English *Performance Verbs* class contains a number of verbs that can be added as constants of type *CreationProperty*, such as *draw* and *produce*, as follows: Paint, Draw, Produce : CreationProperty.

GF offers a way to share similar structures in different languages in one parametrized module called *functor* [17]. In our implementation the common structure of the concrete syntax for English and Swedish is shared in a functor. Since the function *CreationDate* is linearized differently, it is defined separately for each language. This is illustrated below.

incomplete concrete syntax[8]
lincat
IdentificationMessage = S ; TimeSpanClass, ArtistClass = Adv ; TimeSpan = NP ; CreationStatement = VP ; CreationProperty = V2 ; ObjTitle, Artist = PN ;
lin
Identification np vp = mkS pastTense (mkCl (mkNP np) vp);
CreationAct vp compl = mkVP vp compl;
HasCreator v np = (mkVP (passiveVP v) np) ;
CreatorName obj = (mkAdv by8agent_Prep (mkNP obj));
Year y = mkNP (SymbPN y) ;

concrete English syntax
lin CreationDate obj = (mkAdv in_Prep obj);

concrete Swedish syntax
lin CreationDate obj = mkAdv noPrep (mkCN year_N (mkNP obj));

The lexicon is implemented as an interface module which contains *oper* names that are the labels of the record types. It is used by the functor and by each of the language specific lexicons.

interface lexicon
oper
year_N : N;
restOnTheHunt_PN : PN ;
johnMiel_PN : PN ;
paint_V2 : V2 ;

instance English lexicon
oper restOnTheHunt_PN = mkPN [“Rest on the Hunt”]; johnMiel_PN = mkPN “John Miel”; year_N = regN “year”; paint_V2 = mkV2 “paint” ;

instance Swedish lexicon
oper restOnTheHunt_PN = mkPN [“Rastande jägare”]; johnMiel_PN = mkPN “John Miel”; year_N = regN “år”; paint_V2 = mkV2 “måla” ;

[8] The word *incomplete* suggests that the functor is not a complete concrete syntax by itself.

In GF it is possible to built a regular grammar for new languages by using simple record types. In our case we implemented a small application grammar for Hebrew, i.e. *concrete Hebrew* that uses the same abstract syntax as for English and Swedish. In this module functions are linearized as strings where records {*s : Str*} are used as the simplest type.[9] We introduce the parameter type *Gender* with two values: Masc and Fem, these are used in table types to formalize inflection tables. In Hebrew, verb phrases are parameterized over the gender and are therefore stored as an inflection table {s : Gender => Str}; noun phrases have an inherent gender that is stored in a record together with the linearized string {s : Str ; g : Gender}.[10]

concrete Hebrew syntax
lincat
IdentificationMessage, TimeSpan, ArtistClass, TimeSpanClass = {s : Str}; Artist, ObjTitle = {s : Str ; g : Gender}; CreationProperty, CreationStatement = { s : Gender => Str};
lin
Identification np vp = {s = np.s ++ vp.s ! np.g };
CreationAct vp compl = { s = \\g => vp.s ! g ++ compl.s };
HasCreator v obj = { s = \\g => v ! g ++ obj.s};
CreatorName obj = { s = [“al yedey”] ++ obj.s };
CreationDate obj = { s = [“be”] ++ obj.s };
ObjTitle = {s = [“menuhat tzayydym”] ; g = Fem};
JohnMiel = {s = [“guwn miyAe”] ; g = Masc};
Paint = { s = table {Masc => “tzuwyr”; Fem => “tzuwyrah”}};
Param
Gender = Fem | Masc ;

The complete grammar specifications yield the following text, in English, Swedish and Hebrew:

> Eng> Rest on the Hunt was painted by John Miel in 1642. The painting is located in the Hallwyska museum in Stockholm.
> Swe> Rastande jägare blev målad av John Miel år 1642. Tavlan hänger på Hallwyska museet i Stockholm.
> Heb> menuhat tzayydym tzuwyrah ’al yedey guwn miyAel be-1642. htmwnh memukemet be-muwzeyAuwn hallwiska be-stukholm.

This kind of multi-level grammar specification maps non-linguistic information to linguistic representation in a way that supports local and global text variations. For example, in the English and the Hebrew concrete syntax, the sentence complement is realized as a prepositional phrase (signalled by the prepositions *in* and *be*), but in the Swedish sentence, the complement is realized as a noun phrase (signalled by the noun *år*). In the above example this is

[9] The resource grammar for Hebrew is currently under development.
[10] Hebrew has a more complex morphology as the one described here. However, in this implementation we changed the grammar so that it takes only care of gender agreement.

illustrated in the linearization of *CreationDate*. In the Swedish concrete syntax no preposition is used (*noPrep*), and a different NP rule is applied to generate the noun phrase *år 1642*, i.e. CN→ NP → CN. Lexical variations are supported by the grammar as well, for instance, the verb *located* is not a direct translation of the Swedish verb *hänger* 'hang' but the interpretation of the verb in this context implies the same meaning, namely, the painting exists in the Hallwyska museum. The choice of the lexical unit are governed by the semantic structure of the ontology that is reflected in the abstract syntax.

While the functional orientation of isolated sentences of language is supported by GF concrete representations, there are cross-linguistic textual differences that we touched upon in section 4.1 and that are not yet covered in the grammar specifications, i.e. patterns with which cohesive and coherent texts are created. In English, cohesive means comprise conjunction, substitution and ellipsis that can frequently be used to realize a logical relation. In Swedish, cohesive means is often realized as elliptical item, preposition phrase, and/or punctuation. Whereas in Hebrew means of cohesion are realized through the verbal form, usage of ellipsis and conjunctive elements are not common.

6 Conclusion

In this paper we have presented a grammar driven approach for generating object descriptions from formal representations of a domain specific ontology. We illustrated how the lexicons of individual languages pair ontology statements with lexical units which form the backbone of the discourse structure. We demonstrated how schema based discourse structure is mapped to an abstract grammar specification using the domain specific ontology concepts and properties.

We are now in the process of the development of schemata that are being continually modified and evaluated; each rhetorical predicate should capture as many sentence structure variations as possible. A limitation of discourse schemata development is that it requires a lot of human efforts, however once a discourse schema is defined it can automatically be translated to abstract grammar specifications. This method of assembling coherent discourses from basic semantic building blocks will allow any generation system to assemble its texts dynamically, i.e. re-plan portion of its text and communicate successfully.

In the nearest future we intend to extend the grammar to support grouping of rhetorical predicates which requires a certain coverage of linguistic phenomena such as ellipsis, focus, discourse and lexical semantics. The long challenge of this work is in capturing linguistic properties of a language already during the schema development process to guide further development of language independent grammar specifications.

Acknowledgements

The author would like to express her appreciation to Robert Dale for helpful discussions and acknowledge three anonymous readers for commenting on the paper. The GF summer school 2009 and the Centre for Language Technology (CLT) for sponsoring it.

References

1. Schreiber, G., Amin, A., van Assem, M., de Boer, V., Hardman, L., Hildebrand, M., Hollink, L., Huang, Z., Kersen, J., Niet, M., Omelayenko, B., Ossenbruggen, J., Siebes, R., Taekema, J., Wielemaker, J., Wielinga, B.: Multimedian e-culture demonstrator. In Cruz, I.F., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L., eds.: International Semantic Web Conference. Volume 4273., Springer (2006) 951–958 E-culture-demonstrator-2006.
2. Bryne, K.: Having triplets – holding cultural data as rdf. In: Proceedings of IACH workshop at ECDL2008 (European Conference on Digital Libraries), Aarhus (2009)
3. Hielkema, F., Mellish, C., Edwards, P.: Evaluating an ontology-driven wysiwym interface. In: Proc. of the Fifth International NLG Conference. (2008)
4. Hasan, R.: Linguistics, language and verbal art. Geelong: Deakin University. (1985)
5. Halliday, M.A., Hasan, R.: Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective. Oxford: Oxford University Press (1989)
6. Crofts, N., Doerr, M., Gill, T., Stead, S., Stiff, M.: Definition of the CIDOC Conceptual Reference Model. (2005)
7. O'Donnell, M.J., Mellish, C., Oberlander, J., Knott, A.: Ilex: An architecture for a dynamic hypertext generation system. NL Engineering **7** (2001) 225–250
8. Bontcheva, K.: Generating tailored textual summaries from ontologies. In: Second European Semantic Web Conference (ESWC). (2005) 531–545
9. Mellish, C., Pan, J.Z.: Natural language directed inference from ontologies. Artifical Intelligence **172** (2008) 1285–1315
10. McKeown, K.R.: Text generation : using discourse strategies and focus constraints to generate natural language text. Cambridge University Press (1985)
11. Ranta, A.: Grammatical framework, a type-theoretical grammar formalism. Journal of Functional Programming **14** (2004) 145–189
12. Martin-Löf, P.: Intuitionistic type theory. Bibliopolis, Napoli (1984)
13. Ljunglöf, P., Larsson, S.: A grammar formalism for specifying isu-based dialogue systems. In: Advances in Natural Language Processing, 6th International Conference, GoTAL 2008, Gothenburg, Sweden. Volume 5221 of Lecture Notes in Computer Science., Springer (2008) 303–314
14. Khegai, J., Nordström, B., Ranta, A.: Multilingual syntax editing in gf. In Processing, I.T., (CICLing-2003), C.L., eds.: LNCS 2588, Mexico, Springer (2003) 453–464
15. Johannisson, K.: Formal and Informal Software Specifications. PhD thesis, Chalmers University of Technology (2005)
16. Levin, B.: English Verb Classes and Alternations: A Preliminary Investigation. University of Chicago Press, Chicago (1993)
17. Ranta, A.: The gf resource grammar library. Linguistic Issues in Language Technology (LiLT) (2009)

An Improved Indonesian Grapheme-to-Phoneme Conversion Using Statistic and Linguistic Information

Agus Hartoyo, Suyanto

Faculty of Informatics - IT Telkom, Jl. Telekomunikasi No. 1 Terusan Buah Batu
Bandung, West Java, Indonesia
truegushar@yahoo.co.id, suy@ittelkom.ac.id

Abstract. This paper focuses on IG-tree + best-guess strategy as a model to develop Indonesian grapheme-to-phoneme conversion (IndoG2P). The model is basically a decision-tree structure built based on a training set. It is constructed using a concept of information gain (IG) in weighing the relative importance of attributes, and equipped with the best-guess strategy in classifying the new instances. It is also leveraged with two new features added to its pre-existing structure for improvement. The first feature is a pruning mechanism to minimize the IG-tree dimension and to improve its generalization ability. The second one is a homograph handler using a text-categorization method to handle its special case of a few sets of words which are exactly the same in spelling representations but different each other in phonetic representations. Computer simulation showed that the complete model performs well. The two additional features gave expected benefits.

Keywords: Indonesian grapheme-to-phoneme conversion, IG-tree, best-guess strategy, pruning mechanism, homograph handler.

1 Introduction

Many methods of data driven approach was proposed to solve grapheme-to-phoneme (G2P) conversion problem, such as instance-based learning, artificial neural networks, and decision-tree. In [7], it was stated that an IG-tree + best-guess strategy has high performance. It compresses a given training set into an interpretable model. In this research, the method is adopted to develop a new model for Indonesian G2P (IndoG2P). In the new model, two new features for improvement are added: a pruning mechanism using statistic information and a homograph handler based on some linguistic information provided by a linguist.

According to the fact that the model is a lossless compression structure, which means that it stores all data including those of outliers into rules, a pruning mechanism is proposed to prune some rules accommodating outliers. Hence, the model is expected to increase its generalization, but decrease its size.

Furthermore, the model does not handle homograph problems. The letter-based inspection mechanism performed letter by letter internally in a word cannot handle a few sets of words which are exactly the same in spelling representations but different

Special issue: Natural Language Processing and its Applications.
Research in Computing Science 46, 2010, pp. 179-190

Received 23/11/09
Accepted 16/01/10
Final version 12/03/10