Generation of Informative Texts with Style

Stephan M. Kerpedjiev
Institute of Mathematics
Acad. G. Bonchev St., Bl.8
1113 Sofia, Bulgaria

Abstract: An approach to the computational treatment of style is presented in the case of generation of informative texts. We regard the style mostly as a means of controlled selection of alternatives faced at each level of text generation. The generation technique, as well as the style specification, are considered at four levels — content production, discourse generation, surface structure development, and lexical choice. A style is specified by the frequency of occurrence of certain features examined through observation of particular texts. The algorithm for text generation ensures efficient treatment of the style requirements.

1 Introduction

The manner of presentation, i.e. the repeating pattern of expression produced by a given subject or in a certain community, is known as style. Thus each newspaper renders the weather information in a specific style by adopting a particular scheme of presentation. We assume that a text generator, like humans, should has its own style. Furthermore, we regard the style as a means to control the selection of particular constructs among the great variety provided by the language. In this paper, we study what defines a style and how a system could produce texts with style.

The problem is tackled in the framework of text generation created by some of the pioneering works in the field [9,10]. According to this framework, the process of generation is considered at four levels: content production, discourse planning, surface structure generation and lexical choice. We trace out the features that characterize the texts generated from a given content portion in the case of informative texts (introduced in section 2) and show how they can be treated computationally. For illustration of our considerations we make use of the class of weather reports — informative texts about which a good deal of material has been collected, mostly through METEOVIS — an experimental system for handling multimodal weather reports.

The development of the METEOVIS project began with the transformation of weather forecasts from text to map [5]. Then we studied the conversion of weather forecast texts into texts with another discourse structure or in another language [6]. This year, the system was redesigned so that multimodal weather products could be generated from datasets [7]. The domain specific knowledge was isolated in knowledge bases (terminological, rhetorical and grammatical), and the processing modules were made independent from the subject domain as much as possible. At this point we became aware that additional information was necessary to produce high quality texts. Thus we came to the notion of style which was later on generalized to the case of informative texts.

2 Informative texts

The considerations in this paper refer to a particular category of texts which I call *informative texts*. Examples, in addition to the weather reports, are war communiqués, summaries on the ecological situation over a given region, etc. An informative text describes a phenomenon or a situation, either observed or predicted. It consists of assertions, each one relating an event, observation or prediction to a given location and time period.

Informative texts differ from descriptive texts (studied in [10]) in that they are not intended to create permanent long-term memory traces about certain conceptual structures; instead, they draw a mental picture of a particular situation. Informative texts differ also from instructional (operative) texts in that they are not associated with particular actions on the part of the reader (a lot of studies on instructional texts have been carried out, consider e.g. [8]). Informative texts are a type of objective narrative texts well-classified by their subject domains (e.g. weather, ecology) and inheriting from those domains properly devised models.

The source information for the generation of an informative text is a dataset produced by an application program or collected by humans. The dataset encodes the situation comprehensively according to a certain model created to support the research and practical work in the corresponding field. Usually, that model defines the parameters, both quantitative and qualitative, that characterize the phenomena concerned, as well as some relations between parameters.

Since each assertion specifies the value of a parameter referred to a particular location and time period, territory and time models are employed as well. They define the granularity of the territory and the time axis, and certain relations between time periods or regions (e.g. inclusion, partial order, neighbourhood, paths of regions). Depending on the size of grain, either temporal or spatial, the assertions are characterized with a certain degree of imprecision which, if greater than a given threshold, has to be explicitly stated in order to prevent the readers from getting mislead.

The predictive character of some informative texts requires that the assertions are marked with the probability of their occurrence. Similarly to imprecision, this information, called certainty, is necessary for the creation of a proper picture of the situation being presented.

3 Style of informative texts

The concept of style is fundamental in this work. In the light of NL generation systems, the style is a means of adapting the system to a particular manner of text formulation, thus making possible the expression of the same content portion as different texts, according to the available styles.

In [2] an approach to the computational treatment of style is suggested in the case of machine translation. The internal stylistics of the source language is used to determine the specific goals of style such as clarity or concreteness; then the comparative stylistics for the two languages is employed to infer the corresponding style goals of the target text; and finally, the internal stylistics of the target language says how to construct the target text so that the inferred goals be achieved. The relationship between stylistic and syntactic features is expressed through stylistic grammars.

Our approach to using style features in text generation is similar to the approach in [9]. Both allow adopting one or another generation alternative on the basis of certain stylistic rules. Unlike our approach, however, their rules define the preferences explicitly (we use features distributions) and concern only the surface structure development (we cover all levels of text generation).

To provide an evidence for the existence of styles in informative texts, we observed a number of weather forecasts published in different newspapers in three languages — Bulgarian, English and Russian. Samples of such texts are given below (the translations from Bulgarian and Russian into English preserve the features of the original texts as much as possible):

> Today it will be cloudy. In many portions it will drizzle, turning into snow over the regions higher than 500 m above the sea level...
> Outlook for Friday: The rain will stop and it will clear gradually.
> *Trud* (translated from Bulgarian)

> Rain in south-east England will soon clear and with the rest of southern and central England and Wales the day will be mainly dry. However, further rain is likely in southernmost counties by midnight ... It will feel cool everywhere in the strong winds which will reach gale force in the south-east.
> *The Times*

> Much of Britain will be dry with sunny spells but south-west England, the Channel Islands and north and west Scotland will be mostly cloudy with showery rain ... Reasonably warm in sunnier parts of the West, but cool, especially on the east coast.
> *Observer*

> In Moscow, warm weather will remain with occasional rain. Temperatures in the night from 0 to -5 Centigrade, in the day about 0.
> In Leningrad, occasional rain, temperatures in the night from -3 to +2 Centigrade, in the day 0 - 4.
> In Irkutsk region, snow, snowstorm, temperatures in the day from -8 to -13 Centigrade. Towards the weekend the temperatures will fall by 4 - 6 degrees.
> *Izvestia* (translated from Russian)

The Bulgarian weather forecasts are usually organized in two paragraphs corresponding to the first and the second day of the forecast. The sentences most often are simple. Complex and compound sentences occur rarely but in various types: complex sentences with a main and a relative clause connected by the adverbial phrases 'where' and 'when'; compound sentences with co-ordinating conjunctions of addition 'and', co-occurrence 'with', or contrast 'but'. The use of impersonal verbs ('it will be', 'it will rain') is typical whereas verbless sentences are rather an exclusion than a norm.

In English forecasts, impersonal verb phrases are rarely used; instead, the formal subject of the sentences most often is the region or the weather element, and less frequently — the time period. Compound sentences are used intensively for assertions with opposite weather values connected by the co-ordinating conjunction for contrast 'but' (cf. the forecast from *Observer*).

In *Izvestia*, because of the large area of this country, the text is almost always structured by regions and all the weather information about a given region is rendered in one long compound sentence the constituents of which are laconic, verbless clauses divided by commas. Complex sentences rarely occur.

The features of the observed weather forecast texts allow us to summarize the basic properties that characterize a style:

- the extent to which details are provided;
- text organization (by regions, time periods, etc.);
- the prevalent types of sentences according to [1] (simple, complex, compound);
- the prevalent length of sentences (short, medium, long);
- the most typical patterns of surface structures;
- the lexical entries preferred in the expression of the assertions elements.

Since style features are regarded as typical, prevalent, preferred, they should be defined through the frequencies of their occurrence rather than as obligatory characteristics.

4 Text generation

In this section we concisely introduce the principles and techniques of text generation employed in METEOVIS and, as I believe, relevant to other kinds of informative texts. Along with this, we show how one or another alternative is selected on the basis of certain stylistic rules.

4.1 Content production

The content production (CP) component generates the set of assertions from a dataset using domain-specific techniques. In METEOVIS, we employed weather verification techniques that match the generated set of assertions with the dataset and evaluate the precision of the set as a whole.

Although CP is not responsible for the logical consistency of the set of assertions, it is guaranteed that there are no serious contradictions. An example of a weakly inconsistent set of two assertions is given below:

```
<clouds=broken, region=Bul, time=today>
<clouds=clear, region=NE_Bul, time=noon>
```

This type of inconsistency, easily resolved by the readers, is inevitable because of the roughness of the territory and time models. If we required that the generated set of assertions be absolutely uncontradictory, we might loose completeness (some territories or time periods remain uncovered) or conciseness.

A style feature at the CP level is the extent to which details are provided — from summary information only to the finest detail. It is specified by any of the terms *summary, normal* or *detailed*. In the case of *summary* information one assertion is extracted for each weather attribute. A *detailed* style requires that the set of assertions giving the highest precision rate is extracted, without any restrictions on the number of assertions. The extraction of *normal* information is limited to no more than $(1+d)/2$ assertions giving the highest precision rate, where d is the number of assertions that would be extracted if the style were *detailed*.

4.2 Discourse generation

The assertions generated can immediately be transformed into simple NL sentences, but the text obtained most probably will be awkward, unorganized and inefficient. In order to be coherent, a text has to be organized according to rhetorical schemas that take into account semantic relations between entities presented in the text [3,10]. Thus the user will perceive the information with minimal cognitive effort.

For the generation of discourse structures, we employ seven rhetorical schemas based on certain semantic relations [7]:

Parameter progression. An assertion about a given parameter cannot interpose a sequence of assertions concerning another parameter.

From a summary to details. An assertion with a region and a time period containing the region and the time period of another assertion is conveyed before the second assertion.

Temporal progression. The assertions are ordered by the successive time intervals they pertain to.

Spatial progression. The assertions are arranged in such a way that their regions, if taken in this order, make a path defined in the territory model.

Coupling related values. Assertions with co-occurring values are rendered in a group.

Contrast. Two assertions with opposite values are conveyed together to contrast with each other.

Value progression. The assertions about a given parameter with an ordered domain are conveyed in successive groups relating to the particular values.

For each rhetorical schema there is a rule which decides whether the schema is applicable to a given set of assertions, and if it is, structures the set accordingly. This is a hierarchical top-down process starting from the original set of assertions and resulting in a complete discourse structure of the text represented as a tree. The terminals correspond to the assertions and each node represents the discourse relation existing between its successors; hence the root represents the discourse structure at the highest level. We also regard the nodes as chunks of assertions that have to be rendered in a group.

The following properties say how the discourse structure influences the surface structure:

Property 1. Each sentence presents all assertions of a given chunk.

Property 2. The order of the sentences follows the left-to-right order of the chunks of the discourse tree.

Property 3. For each type of discourse structure (temporal progression, related values, etc.), there are sentence grammars each of which can convert the corresponding set of assertions into a sentence surface structure.

A style at the discourse level specifies the rhetorical schemas applicable at each level of discourse generation. For example, the following specification

$$\left[\begin{array}{l} 1: \left\{\begin{array}{l} spat\_progr(W\_Bul, C\_Bul, E\_Bul) \\ spat\_progr(N\_Bul, S\_Bul) \\ temp\_progr(day\_1, day\_2) \end{array}\right\} \\ 2: relate \\ 3: par\_progr(clouds, precip, wind, temp) \\ 4: any \end{array}\right]$$

implies that at the highest level, one of the two spatial progressions (by the paths West, Central and East Bulgaria or North and South Bulgaria) or the temporal progression by the two days of the forecast, should be applied, depending on the set of assertions. Thus if it is better stratified by the time periods *day_1* and *day_2* than by the two paths of regions, then the temporal progression will be applied, else — one of the two spatial progressions. At the second level, all assertions with related values will be coupled into indivisible chunks. At the next level, parameter progression should be employed to further break down the chunks obtained as a result of the previous divisions. Finally, for each terminal chunk the schema that best applies to it will be used to complete the corresponding subtree.

4.3 Surface structure development

One of the major problems in the creation of informative texts is how to avoid text monotony. Perhaps it is the poorly designed surface structure that most of all contributes to the monotony of a text. The ever repeating same sentence pattern makes the text artificial, awkward and boring for the reader. Adversely, a text with diverse surface structure, expressive function words, alternating short and long sentences helps the reader perceive the important elements quickly, extract and memorize the facts, and enjoy the proper pace of reading.

Partially the surface structure of a sentence is predetermined by the current discourse unit through the correspondence *discourse structure* → *possible grammars* introduced in Property 3, section 4.2. The main vehicle for the selection of one or another syntactic structure from the great variety offered by the grammar is the focussing mechanism. The idea is that a sentence should

begin with some concepts or objects already introduced (topic) and end with new information about them (focus) [3,4,10].

Here we put forward a treatment of the focussing mechanism applicable to the generation of informative texts. According to the particular discourse structure, one of the assertion elements — parameter, time period or region — should be the topic of the current sentence. For example, in a spatial progression, it is the path of regions that is the common element of the assertions unified in the chunk and this path is represented in the separate assertions by their regions. Therefore, it is natural to construct the corresponding sentences with the regions being their topics. This decision puts additional constraints on the possible grammars converting the chunk contents into a text surface structure.

Even though the discourse structure and the focussing mechanism restrict to a large extent the possible surface structures provided by the grammar, still more than one alternatives may exist. At this point the style decides which alternative is most suitable as a surface structure of the current chunk. For example, a discourse structure of type *contrast* is a very appealing pre-condition for the creation of a compound sentence in which the constituent clauses (corresponding to the assertions linked by the contrast relation) are connected by the conjunction 'but'. However, the creation of two simple sentences without any function words is acceptable as well. It is just a question of style to make one or another decision — whether a simple or a compound sentence is preferred at this point, if some of the potential surface structures have priority over the others, which function word is preferred to lead a sentence or to connect two clauses, etc.

The style features at surface level supported by the system are sentence type, sentence length and syntactic roles of the assertions elements. These features characterize the style with different levels of detail. Thus a specification of the sentence type or length provides less detail than a specification of the syntactic roles.

Sentence type is specified by the frequencies of the simple, compound and complex sentences. For example, the statement:

$$sentence\_type = \begin{bmatrix} simple & : & 0.5 \\ compound & : & 0.3 \\ complex & : & 0.2 \end{bmatrix}$$

is understood as an instruction for minimizing the function:

$$r = \sqrt{(x-0.5)^2 + (y-0.3)^2 + (z-0.2)^2}$$

where x, y, and z are the portions of the simple, compound and complex sentences, respectively, in the actually generated text. As a result, about half of the sentences in the final form should be simple, 3/10 — compound, and 1/5 — complex.

Sentence length is treated in a similar manner by specifying the frequencies of the short, medium and long sentences. A sentence is considered short, if it contains at most 4 entities (parameter values, regions or time periods); medium — between 5 and 8 entities; and long — more than 8 entities.

Syntactic roles are specified by enumerating the allowed sentence patterns together with their relative frequencies as follows:

$$syntactic\_roles = \begin{bmatrix} g_1 : f_1 \\ g_2 : f_2 \\ \dots \\ g_n : f_n \end{bmatrix}$$

where $f_1, f_2, ..., f_n$ are the relative frequencies for the grammars $g_1, g_2, ..., g_n$, respectively. This specification makes the system minimize the function:

$$r = \sqrt{(x_1 - f_1)^2 + (x_2 - f_2)^2 + ... + (x_n - f_n)^2}$$

where $x_1, x_2, ..., x_n$ are the portions of sentences actually generated by means of grammars $g_1, g_2, ..., g_n$.

Only one of the features *sentence_type*, *sentence_length* and *syntactic_roles* should be specified, for there are certain co-relations between them and the specifications of two features may contradict each other.

The following algorithm for surface structure generation makes use of Properties 1, 2 and 3 of the discourse tree (cf. section 4.2), the focussing mechanism and the style requirements.

The process begins with counting all grammars that implement the chunks of the discourse tree as sentences, using the correspondence *discourse structure → possible grammars*. Those grammars form the current stock of candidates which in the process of generation of the surface structure is updated as specified in steps 5 and 6 below. The generation proceeds in a loop as follows.

1. For each chunk on the path from the root to the left-most terminal, the grammars candidates to implement the chunk are considered.
2. Those grammars that do not satisfy the focussing condition are left out of consideration.
3. The final selection is performed taking into account the style specification. Suppose that the style specifies n sentence types with frequency rates $f_1, f_2, ..., f_n$ and the portions of these sentence types in the current stock are $s_1, s_2, ..., s_n$, resp. Then the system selects from the remaining candidates the grammar of type k satisfying the conditions:

   $$f_k - s_k = max(f_1 - s_1, ..., f_n - s_n),$$
   $$f_k > 0, s_k > 0.$$

   The heuristics behind this rule is "select the grammar that best compromises the frequency rate specified by the style and the deficiency rate in the current stock".
4. The set of assertions constituting the corresponding chunk is converted into the surface structure of a sentence through the selected grammar.
5. The discourse tree is pruned by removing the subtree rooting at the chunk that was converted into a surface structure and the grammars corresponding to this subtree are deducted from the current stock.
6. The portion of grammar candidates deducted from the current stock is subtracted from the frequency rate f_i of the selected sentence grammar, and the portions $s_1, s_2, ..., s_n$ are re-calculated.

7. Steps 1–6 are repeated until the discourse tree is exhausted.

The selection of a surface structure as described above avoids the combinatorial explosion expected during the examination of the minimizing conditions. This efficiency is achieved at the expense of a looser treatment of those conditions. Thus the technique ensures an actual distribution of the surface features that is sufficiently close but not necessary the closest to the distribution specified. The only drawback of the algorithm is observed when short texts are generated. Then the surface structures with low frequency rates either tend to appear at the end of the text or are not generated at all.

4.4 Lexical choice

The last step in text generation is the linearization of the surface structure into a string. METEOVIS makes use of a phrasal lexicon to replace the terminals of the surface structure tree with entries from the lexicon using the terminal's type and value as a key. The freedom given to the generator at this level of processing allows it to choose from two or more synonyms for the same entity. For example, the following structure

adv_region →
prep('in') reg_mod(much) prep('of') noun(Bul)

can be linearized as *'in many portions of the country'*, *'in much of Bulgaria'*, etc. The style may give preference to one of these expressions specifying the frequency of each member of the synonymous groups *reg_mod(much)* and *noun(Bul)*.

Similarly to the selection of sentence grammars, the lexical choice between synonyms is made on the basis of a distribution specified by the style. Thus the statement

$$reg\_mod(much) = \begin{bmatrix} much & : 0.5 \\ many\ portions & : 0.25 \\ many\ parts & : 0.25 \end{bmatrix}$$

specifies a distribution of the elements of the synonymous group representing the entity *region modifier* according to which the modifier 'much' will occur twice as frequently as any of the other two modifiers. Such kind of style specification can be made for each synonymous group. The default is even distribution.

5 Conclusion

The problem of text generation with style has been described in the case of informative texts. We stepped on the platform of the experimental METEOVIS system designed to handle multimedia weather information. In order to get efficient control over the generated texts, we employed the concept of style, examined the features that make up a style, and adapted the technique of text generation to take into account those features. This new opportunity makes possible the controlled generation of various texts from the same dataset.

Style specification is feasible at the four levels of text generation — content production, discourse generation, surface structure development, and lexical choice. It drives the system to select from the many alternatives offered by the rhetorical knowledge, grammar and lexicon those providing text features sufficiently close to the specified ones. The algorithm for text generation provides efficient treatment of the style requirements.

Acknowkedgement: This work was supported by the Ministry of Science and Education under grant I23/91 and by the Bulgarian Academy of Sciences under grant 1001003.

References

[1] L. G. Alexander. *Longman English Grammar.* Longman, 1988.

[2] C. DiMarco and G. Hirst. Stylistic Grammars in Language Translation. In: *Proc. COLING 88*, Vol.1, Budapest, 1988, 148-153.

[3] N. E. Enkvist. Introduction: Coherence, Composition and Text linguistics. In: *Coherence and Composition: A Symposium*, ed. N.E.Enkvist, Abo Academy, 1985, 11-26.

[4] E. Hajičová. Focussing - a Meeting Point of Linguistics and Artificial Intelligence. In: *Artificial Intelligence II: Methodology, Systems, Applications*, eds. Ph.Jorrand and V.Sgurev, North-Holland, 1987, 311-321.

[5] S. Kerpedjiev. Transformation of Weather Forecasts from Textual to Cartographic Form. *Computer Physics Communications.*, 61(1990), 246-256.

[6] S.Kerpedjiev, V.Noncheva. Intelligent Handling of Weather Forecasts. In: *Proc. COLING 90*, vol. 3, Helsinki, 1990, 379-381.

[7] S. Kerpedjiev. Automatic Generation of Multimodal Weather Reports from Datasets. In: *Proc. 3rd Conf. on Applied Natural Language Processing*, Trento, 1992, 48-55.

[8] K. Linden et al. Using Systems Networks to Build Rhetorical Structures. In: *Lecture Notes in Artificial Intelligence*, 587, 1992, 183-198.

[9] D. McDonald and J. Pustejovsky. A Computational Theory of Prose Style for Natural Language Generation. In: *Proc. 2nd Conf. of the European Chapter of ACL*, Geneva, 1985, 185-193.

[10] K. R. McKeown. *Text Generation.* Cambridge University Press, 1985.