A Linguistic Discovery Program that Verbalizes its Discoveries

Vladimir Pericliev*
Max-Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany
pericliev@eva.mpg.de

Abstract

We describe a discovery program, called UNIVAUTO (UNIVersals AUthoringTOol), whose domain of application is the study of language universals, a classic trend in contemporary linguistics. Accepting as input information about languages, presented in terms of feature-values, the discoveries of another human agent arising from the same data, as well as some additional data, the program discovers the universals in the data, compares them with the discoveries of the human agent and, if appropriate, generates a report in English on its discoveries. Running UNIVAUTO on the data from the seminal paper of Greenberg (1966) on word order universals, the system has produced several linguistically valuable texts, two of which are published in a refereed linguistic journal.

1 Introduction

Previous works in machine scientific discovery have mostly focussed on historical reconstruction (work culminating in the book by Langley et. al. 1987), but more recent efforts are directed towards designing programs that discover new scientific knowledge. Such systems operate in disciplines as diverse as mathematics, chemistry, astronomy, medicine or linguistics. The field is currently very active (for recent developments, cf. e.g. the special issues on discovery of the journals *Artificial Intelligence*, April 1997 or *Foundations of Science* 1999; the ECAI-98 Workshop on Discovery, The International Conferences on Discovery Science, Japan, 1998 and 1999).

In this paper, we present UNIVAUTO (UNIVersals AUthoringTOol), a system whose domain of application is linguistics, and in particular, the study of language universals, a classic trend in contemporary linguistics. This trend was initiated by the pioneering paper of Joseph Greenberg (1966), investigating word order in a database of 30 languages of wide genetic and areal coverage, described in terms of 15 ordering features. Greenberg discovered a number of universals relating diverse ordering properties of languages, and his example was followed by attempts at similar generalizations at other linguistic levels or across levels (for a review of the state-of-the-art, cf. e.g. Croft 1990).

UNIVAUTO was run on various data sets (word order, phonology, morpho-syntax), with very promising linguistic results. The published outcomes of UNIVAUTO so far include : two whole journal articles Pericliev (1999, 2000) based on data from Greenberg (1966) (with no post-editing, the first one with no disclosure of articles' "machine origin"); around 50 statistically significant phonological universals based on Maddieson's UPSID-451 database, published without post-editing at the Universals Archive at the University of Konstanz; the substance discovery (rather than verbalization) part of Pericliev (2002). To the best of our knowledge, this is the first computer program to generate a whole scientific article.

2 Overview of UNIVAUTO

Below is a brief description of UNIVAUTO (UNIVersals AUthoringTOol), drawing for illustration on data from Greenberg (1966).

2.1 The input

UNIVAUTO accepts as input the following, manually prepared, information:

(1) A *database* (=a table), usually comprising a sizable number of languages, described in terms of some properties (feature-value pairs), as well as a list of the abbreviations

* Currently on leave from the Institute of Mathematics & Informatics, 1113 Sofia, Bulgaria (peri@math.bas.bg).

used in the database. Below is a (simplified) description of the language Berber in terms of just 4 features: v-order (=the position of verb, subject and object), na/an (=the position of noun and adjective), cn/pn (=the position of common noun and proper noun), and pref/suf (=the presence of prefix or suffix):

data(berber,
[v-order=vso,na/an=na,cnpn/pncn=,pref/suf=both]*
).

The value "*" is special, and is used to designate that either the feature cnpn/pncn is inapplicable for Berber or that the value for that feature is unknown.

(2) a *human agent's discoveries* (represented as simple logical propositions, if originally formulated as complex ones); e.g.:

discovery(agent=greenberg,no=3,nonstatistical,
implication(v- order=vso,pr/po=pr)).

This record states that a human agent, Greenberg, has found the implicational universal, relating two variables, to the effect that for all languages, if a language has a Verb-Subject-Object order then this language has *pr*epositions (rather than *po*stpositions), that this universals is non-statistical (holds without exceptions in the studied database), and that it is stated as Universal No. 3 in the original publication of the human agent.

Aside from these basic sources of information, the input includes information on: the origin of database, if any (the full citation of work where the database is given); reference name(s) of database, if any; the kinds of objects rows and columns represent; etc..

2.2 The task

The *task* UNIVAUTO addresses can be formulated as follows: Given the input information (as described in 2.1), find the language universals valid in the data, compare them with those discovered by some human agent, and write a report, if appropriate.

E.g. a query to the system may look like:

?-discover(implication(A,B),non_statistical,
positive_examples=4,compare_with=greenberg).

It amounts to requesting that non-statistical implicational universals holding between two variables and supported in at least 4 positive examples be found, the results be compared with the findings of Greenberg, and, if judged as interesting enough, a report of these discoveries be written. Other queries may also be formulated (cf. 3.1), but currently only such involving one type of universal and one database at a time.

2.3 The output

Below we list some excerpts from Pericliev (1999) as an illustration of the system output. The program was run on the data from Greenberg (1966), with the query in the preceding section. It discovered some problems in his analyses (which forms the bulk of the text below) as well as 59 novel universals of type "If A then B, non-statistical" (as against 12 found by Greenberg, one of which further turned out to be wrong!). The paragraphs have bold face numeration to be used for later reference.

Figure 1. Illustrative output of *UNIVAUTO*

---

[1] *Further implicational universals in Greenberg's data*

[2] *The goal of this article is to study the implicational universals in the 30 languages sample of Greenberg 1966 and compare the results of the two studies.* <…>*

[3] *We confirmed the validity of universals [12,13,15-a,15-b,21-a,22-a,27-a].*

[4] *Universal [27-b] is also true, however it violates our restriction pertaining to the occurrence of at least 4 positive examples in the dataset. [27-b] is supported in 1 language (Thai).*

[5] *Universals [16-a,16-b,16-c] are uncertain, rather than indisputably valid in the database investigated, since they assume properties in languages, which are actually marked in the database as "unknown or inapplicable" (notated with "*" in Table 1). Universal [16-a] would hold only if the feature AuxV/VAux is applicable for Berber, Hebrew, and Maori and in these languages the inflected auxiliary precedes the verb. Universal [16-b] would hold only if the feature AuxV/VAux is applicable for Burmese and Japanese and in these languages the verb precedes the inflected auxiliary. Universal [16-c] would hold only if the feature AuxV/VAux is applicable for Loritja and in this language the verb precedes the inflected auxiliary.*

[6] *Universal [23-a] is false. It is falsified in Basque, Burmese, Burushaski, Finnish, Japanese, Norwegian, Nubian, and Turkish, in which the proper noun precedes the common noun but in which the noun does not precede the genitive.*

[7] *We found the following previously undiscovered universals in the data.*
[8] *Universal 1. If in a language the adjective precedes the adverb then the main verb precedes the subordinate verb.*
[9] *Examples of this universal are 8 languages: Fulani, Guarani, Hebrew, Malay, Swahili, Thai, Yoruba, and Zapotec. <…>*
[10] *Universal 59. If a language has an initial yes-no question particle then this language has the question word or phrase placed first in an interrogative word question.*** *<…>*

*The generated text continues with description of what an implicational universal is, a table of Greenberg's 30 language sample, accompanied by the abbreviations used, and a listing of the universals he found. His universals, verbalized by our program, are listed with their numeration in the original publication. An alpha-numeric numeration means that an originally complex universal has been split into elementary ones of the form "If A then B".
**There follows a conclusion which is a summary of the results.

---

3 The UNIVAUTO System

UNIVAUTO comprises two basic modules: one in charge of the discoveries of the program, called UNIV(ersals), and the other in charge of the verbalization of these discoveries, called AU(thoring)TO(ol).

3.1 The discovery module UNIV

UNIV discovers logical patterns (=universals), including (but not limited to):

- *A (*absolute, non-implicational universal)
- *If A1 and A2 and A3 and...An, then B* (implicational universal)

UNIV can compute "non-statistical" universals (holding without exceptions) or "statistical" universals (holding with some user-specified percentage of exceptions).

Also, UNIV can compute (implicational) universals valid in (at least) a user-specified number of positive examples (=languages), as well as compute the statistical significance of universals (based on the χ^2 statistic). A minimal set-cover subroutine may guarantee the discovery of the smallest set(s) of universals, generating a typology (Pericliev 2002).

Importantly, given the discoveries of another, human agent, UNIV employs a diagnostic program to find (eventual) errors in the humanly proposed universals. Currently, we identify as PROBLEMS the following categories:
(1) *Restriction Problem*: Universals found by human analyst that are "under-supported", i.e. are below a user-selected threshold of positive evidence and/or percentage of validity (the latter applying to statistical universals).
(2) *Uncertainty Problem*: Universals found by human analyst that tacitly assume a value for some linguistic property which is actually unknown or inapplicable (marked by '*' in the database).
(3) *Falsity Problem*: Universals found by human analyst that are false or are logically implied by simpler universals.

The DISCOVERIES of UNIV are two lists, falling into one of the *types*: (1) new universals (absolute or implicational, and statistical or non-statistical), and (2) problems (sub-categorized as above).

3.2 The authoring module AUTO

AUTO accepts as input the discoveries made by UNIV, but also has access to the input data (cf. 2.1) to make further computations, as necessary.

AUTO can generally be characterized as a *practical* text generation system, of opportunistic type, intended to meet the needs of our particular task, rather than as a system intended to handle, in a general and principled way, scientific articles' composition or surface generation of a wide range of linguistic phenomena (reminiscent of earlier work on generation from formatted data of metereological bulletins (Kittredge et.al.'s RAREAS) or stock market reports (Kukich's Ana)). For applied NLG, cf. e.g. Reiter et. al. (1995); also *Computational Linguistics* 1998 4(23), and elsewhere. Xuang & Fielder (1996) and later work verbalize machine-found mathematical proofs.

First, AUTO needs to know whether the discoveries of UNIV are interesting enough for generating a report, and to this end, it uses a natural and simple numeric method: UNIV's discoveries (new universals+problems) are judged worthy of generating a report if they are at least as many in number as the number of the

published discoveries of the human agent studying the same database.

Having decided upon report generation, AUTO follows a fixed scenario for DISCOURSE COMPOSITION (scientific papers are known to follow such fixed structure in "genre analysis"). The details of this scenario, however, will vary in accordance with a number of parameters, related with the query to the system, the discoveries made in response to this query, as well as other considerations. The basic *components* of the scenario (alongside with some minor elaboration) are given below. Each component is structured as a separate text paragraph (possibly with sub-(sub)-paragraphs) .

1. Statement of title. Title is selected from one of the following foci : (i) new_universals, (ii) problems, (iii) new_universals+problems. (Focus (i) selected in Fig. 1, **[1]** .)

2. Introduction of goal. Choice among same foci. (Focus (iii) selected in Fig. 1, **[2]** .)

3. Elaboration of goal. Logical definition of type of universal investigated, constructed by our system, plus message on user-specified constraints (supporting evidence, etc.).

4. Description of the investigated data and the human discoveries. Based on data available from input.

5. Explaining the problems in the human discoveries. UNIV's diagnostic subroutine feeds to AUTO problems classed in one of three sub-categories (cf. 3.1) for AUTO to decide how to explain them.

6. Statement of machine discoveries. Input from the discoveries of UNIV.

7. Conclusion. Summary of findings (new_universals and/or problems).

8. References. Based on data avialable from input.

Below we briefly outline component (5). This paragraph comprises 4 sub-paragraphs, in this order: one conveying information on the confirmed humanly found universals (Fig. 1, **[3]**), and the remaining on problems of restrictions (=under-support), uncertainty and falsity (Fig. 1, **[4,5,6]**). Each sub-paragraph starts with an *intro_part*, making a statement about a collection of discoveries (e.g. "Universals [1,2,..] are under-supported/uncertain/false.."). All but the first sub-paragraph (referring to confirmed discoveries) also have a *body_part*, justifying why these predications hold for each individual discovery in the collection.

The body_parts appeal either solely to *examples* (as in Fig. 1, **[4]**, where mentioning an example of less support, appearing immediately after mentioning of the required one, suffices for an explanation) or to both *examples and explanation* of why these are indeed examples. The latter situation is illustrated by (Fig. 1, **[5,6]**). Thus, for instance, the examples justifying that a universal is false are actually its counter-examples and AUTO will find these counterexamples as well as the reason for that (in the case of implication, antecedent true, but consequent false).

AUTO also has a limited SENTENCE-PLANNING FACILITY to decide how to split up a paragraph's content into sentences and clauses. Assume, for the sake of illustration, that we need to verbalize an under-support body_part, like that on (Fig. 1, par. **[4]**), but, say, requiring at least 8 supporting languages. The input to the sentence planning facility of AUTO would look like this (the last constituents indicating the number of supporting languages):

[3]--is_supported--Berber,Hebrew,Maori,Masai,
Welsh,Zapotec--6
[12]--is_supported--Berber,Hebrew,Maori,Masai,
Welsh,Zapotec--6
[15-a]--is_supported--Berber,Hebrew,Maori,Masai,
Welsh,Zapotec--6
[27-b]--is_supported--Thai--1
[13]--is_supported--Burmese,Burushaski,Hindi,
Japanese,Kannada,Turkish—6

AUTO will form separate sentences from the propositions having an equal number of supporting evidence. Within the framework of each such sentence, the system will group together the propositions supported by the same languages, taking care that the universals with smaller numeration appear first. After some further transformations, the system outputs this:

[27-b] is supported in 1 language (Thai). [13] is supported in 6 languages (Burmese, Burushaski, Hindi, Japanese, Kannada, and Turkish), and so are [3,12,15-a] (Berber, Hebrew, Maori, Masai, Welsh, and Zapotec).

For SURFACE GENERATION we use a hybrid approach, employing both templates and grammar rules, as required by the needs at the specific portions of text we are producing.

The *templates* consist of canned text, interspersed with variables whose values are to be computed. The variables may stand either for individual words or for more abstract entities than words whose values are computed by grammar rules. To ensure agreement e.g. AUTO employs rules for agreement between subject and predicate, noun and determiner, demonstrative, relative-marker, apposition; between noun and pronoun (for pronominal reference); external sandhi, etc. If e.g. a variable stands for a list of languages, it will be handled by a grammar rule for *and*-coordinated NP to get e.g. "Masai, Welsh, and Zapotec". Also, the templates are often randomly chosen among a set of "synonymous" alternatives in order to increase the variability of the produced texts.

We have *grammar rules* to handle a variety of syntactic constructions, but the most important of them are those responsible for the verbalization of universals (forming by far the largest bulk of the produced texts). The dictionary part of that grammar is supplied from input (cf. 2.1). There are diverse ways of expressing implications in English (and we do not confine only to implications), and the grammar tries to attend to this fact. The grammar is a *random generator*, ensuring the avoidance of intra-textual repetitions in the statement of the many universals UNIV usually finds.

Finally, AUTO also supports formatting facilities, e.g. for capitalization, correct spacing around punctuation marks, etc.

4 Conclusion

We have shown how a simple text generator can be linked to a linguistic discovery program in order to verbalize its discoveries. Despite the seemingly bizarre nature of the task of article generation, this work was actually inspired by the practical need to verbalize the great number of universals UNIV has systematically found in the various databases we have explored, as well as by the need to compare these with the findings of previous researchers. Presumably, such problems have not confronted previous discovery programs because they searched non-conventional spaces (necessitating additional human interpretation of results), because their solution objects (e.g. numerical laws in physics/mathematics, reaction path-ways in chemistry, etc.) are not amenable to verbal expression or simply because the set of solution objects has been too small to require automated verbalization.

In sum: UNIVAUTO models scientific domains in which a machine is likely to find numerous and verbalizable solution objects (conceivably, low-level generalisations), and the scientific discourses in these domains are basically limited to description of these findings. We believe that such domains are not exceptional in empirical sciences generally, and hence systems like ours are not unlikely to emerge to aid scientists in these domains.

Acknowledgment. *The writing of this paper was supported through an EC Marie Curie Fellowship MCFI-2001-00689. The author is solely responsible for information communicated and the European Commission is not responsible for any views or results expressed.*

References

Croft, W. (1990). *Typology and Universals*. Cambridge University Press, UK.

Greenberg, J. (1966). Some universals of grammar with particular reference to the order of meaningful elements. In J. Greenberg, ed., *Universals of Language*, MIT Press, pp. 73-113.

Xuang, X. & A. Fielder (1996). Presenting machine-found proofs. *CADE13, LNCS 1104*: 221-225.

Langley, P., Simon, H., Bradshaw, G., and Zytkow, J. (1987) *Scientific Discovery: Computational Explorations of the Creative Processes*. MIT Press.

Pericliev, V. (1999). Further implicational universals in Greenberg's data (a computer-generated article). *Contrastive Linguistics 24*: 40-51. (Sofia)

Pericliev, V. (2000). More statistical implicational universals in Greenberg's data (another computer-generated article). *Contrastive Linguistics 25*: 115-125. (Sofia)

Pericliev, V. (2002). Economy in formulating typological generalizations. *Linguistic Typology 6*: 49-68.

Reiter, E., C. Mellish & J. Levine (1995). Automatic generation of technical documentation. *Applied Artificial Intelligence 9*: 259-287.