Coll. Antropol. **29** (2005) 2: 735–738
UDC 808.62-3:519.6
Original scientific paper

Probability Distribution on the Parse Trees

Božidar Tepeš[1], Dubravko Hunjet[2] and Slobodan Elezović[3]

[1] Department of Information Sciences, Faculty of Philosophy, University of Zagreb, Zagreb, Croatia
[2] University Computing Centre »SRCE«, University of Zagreb, Zagreb, Croatia
[3] Institute for Anthropological Research, Zagreb, Croatia

A B S T R A C T

Paper describes probability distribution on the parse trees of natural language by using Bayesian networks. First parts of the paper describes probabilistic context-free grammar and parse trees. In the second part of the paper, Bayesian network was modelled and joint probability distribution on their vertex. On these theoretical ideas, in the third part, we describe our model tested on Database of grammatical sentences of Croatian language (http://infoz.ffzg.hr/tepes/). At the end was presented a backward procedure and evaluation of our results.

Key words: *parse tree, Bayesian network, sentences, Croatian language, cross entropy*

Introduction

An anthropological research of population structure of the East Adriatic rural populations regarding a number of Adriatic islands, estimates basic geographical, historical, economic, demographic and linguistic factors. Linguistic factors influence the formation of the island population structure[1–8]. The method of hidden Markov model was aimed at identification of internal and external impulse of change or continuity of rural population within socio-cultural contex[9–11]. On this studies now been extended to investigation on morphological and syntactical structure of Croatian language. First step was construction Database of grammatical sentences of Croatian language (http://infoz.ffzg.hr/tepes/)[12]. The theoretical background is theory of formal and natural language[13].

Grammars and parse trees

Context-free grammar[14] CFG is a grammar CFG(V, T, S, P). It consist of a set nonterminal symbols V, a set of terminal symbols T, a start symbol $S \in V$ and a set of rule productions P of the form: $A \to \alpha$, were $A \in V$ and $\alpha \in (V \cup T)^*$. These rules can be interpreted as saying that nonterminal symbol A expands into string α of nonterminal and terminal symbols. String $w_1 = \beta A \gamma$ directly derive string $w_2 = \beta\alpha\gamma$, if rule of production $A \to \alpha$ exists in set P or as relation $w_1 \Rightarrow w_2$. Reflexive and transitive closure of this relation is derive relation $\Rightarrow^*$. Context-free language CFL is a set of strings of terminal symbols $w \in T^*$ derived with rules productions from start symbol S. Those strings are sentences. Context-free language CFL is:

$$CFL = \{w \mid (w \in T^*) \wedge (S \Rightarrow^* w)\} \tag{1}$$

A sequence of rules which directly derive a word of a context-free language $w \in$ CFL from start symbol can be represented in a parse tree. A root of parse tree is start symbol S and leaves are terminal symbols t_1, t_2,..., t_n of sentence $w = t_1 t_2 ... t_n$.

Probabilistic context-free grammar[15–16] PCFG is a context-free grammar CFG with a set of probability function F of form: $p_A(A \to \alpha)$ assigned to each left-hand side A of rule $A \to \alpha$. Definition for a probability function p_A is:

$$\sum_{\alpha} p_A(A \to \alpha) \tag{2}$$

Probability of parse tree $pt(w)$ of a word w is product of probability functions $p_A(A \to \alpha)$ in a parse tree. PCFG also defines probability function of the words $w \in$ CFL as sum of probabilities of parse trees $pt(w)$:

$$p(w) = \sum_{pt(w)} \prod_{(A \to \alpha) \in pt(w)} p_A(A \to \alpha) \tag{3}$$

The maximum-likelihood estimator of probability function is:

Received for publication February 11, 2005

$$\hat{p}_A(A \to \alpha) = \frac{\sum_w f_A(A \to \alpha; w)}{\sum_{(A \to \alpha)} \sum_w f_A(A \to \alpha; w)} \tag{4}$$

Bayesian network for parse trees

Bayesian network[17] is a directed acyclic graph $G(V, E)$ which consists of a set vertexes V and a relation on vertexes $E \subseteq V^2$. The vertices in graph G will correspond to random variables $V = \{X_1, X_2,\ldots,X_n\}$. Each edge in graph is directed $(X_i, X_j) = X_i \to X_j$. Joint probability distribution of vertexes on Bayesian network is:

$$\begin{aligned} &P(x_1, x_2,\ldots, x_n) = \prod_{i=1}^{n} P(x_i | pa_i) \\ &PA_i = \{PA_{i1}, PA_{i2},\ldots PA_{ik}\} \subset V \\ &PA_{i1} \to X_i, PA_{i2} \to X_i,\ldots, PA_{ik} \to X_i \end{aligned} \tag{5}$$

Variables PA_{ij} are parents of variable X_i, x_i is value of random variable X_i and pa_{ij} is value of parent variable PA_{ij}.

Every parse tree of sentence is graph named tree with root start symbol S and the leaves are words or terminal symbols t_1, t_2,..., t_n in the sentence. Natural language sentences have phrase structure. A parse tree of sentence has root category or phrase S and leaves categories of words $t_1 = C_1$, $t_2 = C_2$,..., $t_n = C_n$. What interests us in this analysis is category structure of sentence. In parse tree of a sentence, there are phrases of categories PC of head categories C with structure in Figure 1.

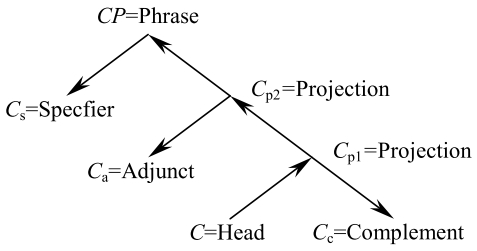

Fig. 1. Phrase structure.

In a phrase structure specifier, adjunct, complement and projections can be categories of words or new phrases with similar structure. Head-driven phrase grammar increased importance of head category. This is the reason for directions from head to projections and phrase, direction from phrase to specifier and direction from projections to adjunct and complement. We can annotate the phrase structure (Figure 1.) in the Chomsky form[18]:

$$CP\lfloor C_s C_{p2} \lfloor C_a C_{p1} [CC_c] \rfloor \rfloor \tag{6}$$

In front of every bracket is the parent category and in the bracket are children of the branch in parse tree of phrase structure.

Results

For our experiment we chose $n = 41$ simple sentences in Database of grammatical sentences of Croatian language (http://infoz.ffzg.hr/tepes). Parse tree of this simple sentences are:

$$\begin{aligned} S_1 &= VP[N_1 V] \\ S_2 &= VP[V N_2] \\ S_3 &= VP[N_1 V_p [V N_2]] \end{aligned} \tag{7}$$

In our database we find $n_1 = 21$ sentences of form S_1, $n_2 = 15$ sentences of form S_2 and $n_3 = 5$ sentences of form S_3. It means that we have frequencies of these sentences:

$$\begin{aligned} f(S_1) &= 21{:}41 = 0.51 \\ f(S_2) &= 15{:}41 = 0.37 \\ f(S_3) &= 5{:}41 = 0.12 \end{aligned} \tag{8}$$

If we put together these three sentences, joint probability distribution of vertexes on Bayesian network is:

$$\begin{aligned} P(V, V_c, VP, N_1, N_2) = P(V) \cdot P(VP|V) \cdot P(V_c|V) \\ \cdot P(N_1|VP) \cdot P(N_2|VP, V_c) \end{aligned} \tag{9}$$

A set of random variables is $\{V, V_c, VP, N_1, N_2\}$. Every variable has two values 0 or 1. Zero means that we have no category in a sentence and one means that we have a category in a sentence. Probabilities of sentences from our model are:

$$\begin{aligned} P(S_1) &= \frac{P(1,0,1,1,0)}{P(1,0,1,1,0) + P(1,0,1,0,1) + P(1,1,1,1,1)} \\ P(S_2) &= \frac{P(1,0,1,0,1)}{P(1,0,1,1,0) + P(1,0,1,0,1) + P(1,1,1,1,1)} \\ P(S_3) &= \frac{P(1,1,1,1,1)}{P(1,0,1,1,0) + P(1,0,1,0,1) + P(1,1,1,1,1)} \end{aligned} \tag{10}$$

Probabilities from (7) and (8) in our model (9) are:

$$\begin{aligned} &P(V) = V \\ &P(VP|V) = (1-V) \cdot (1 - VP) + \\ &+ V \cdot ((1-P(VP = 1|V = 1)) \cdot (1-VP) + P(VP = 1|V = 1) \cdot VP) \\ &P(V_c|V) = (1-V) \cdot (1-V_c) + \\ &+ V \cdot ((1-P(V_c = 1|V = 1)) \cdot (1-V_c) + P(V_c = 1|V = 1) \cdot V_c) \\ &P(N_1|VP) = (1-VP) \cdot (1-N_1) + \\ &+ VP \cdot ((1-P(N_1 = 1|VP = 1)) \cdot (1-N_1) + P(N_1 = 1|VP = 1) \cdot N_1 \\ &P(N_2|VP,V_c) = (1-VP) \cdot (1-V_c) \cdot (1-N_2) + \\ &+ VP \cdot (1-V_c) \cdot ((1-P(N_2 = 1|VP = 1)) \cdot (1-N_2) + \\ &+ P(N_2 = 1|VP = 1) \cdot N_2) + \\ &+ (1-VP) \cdot V_c \cdot (1-N_2) + VP \cdot V_c \cdot N_2 \end{aligned} \tag{11}$$

with parameters:

$$\begin{aligned} P(VP = 1|V = 1) &= 0.88 \\ P(V_c = 1|V = 1) &= 0.12 \end{aligned}$$

$$P(N_1 = 1|VP = 1) = 0.63$$
$$P(N_2 = 1|VP = 1) = 0.37 \quad (12)$$

From our model we have:

$$P(S_1) = 0.138628$$
$$P(S_2) = 0.640462$$
$$P(S_3) = 0.220910 \quad (13)$$

Evaluation and backward procedure

For evaluation of our results, we are using the cross entropy[19] and the χ^2-statistical test[20].

The cross entropy for probability function of our model with (13) and (8) is:

$$H(P,f) = -\sum_{i=1}^{3} P(S_i)\cdot \ln f(S_i) = 1.02305 \quad (14)$$

The χ^2-statistical test for probability function of our model with (13) and numbers of sentences is:

$$\chi^2 = \sum_{i=1}^{3} \frac{(n_i - n\cdot P(S_i))^2}{n\cdot P(S_i)} = 12.5543 \quad (15)$$

For better results of our model we use backward procedure. The idea of backward procedure is to go back from nouns N_1 and N_2 in our Bayesian network (9) and change parameters of our models (12). New parameters are:

$$P(VP = 1|V = 1) = \sum_{V_c=0}^{1}\sum_{N_1=0}^{1}\sum_{N_2=0}^{1} P(1,V_c,1,N_1,N_2):$$
$$\sum_{V_c=0}^{1}\sum_{VP=0}^{1}\sum_{N_1=0}^{1}\sum_{N_2=0}^{1} P(1,V_c,VP,N_1,N_2)$$

$$P(V_c = 1|V = 1) = \sum_{VP=0}^{1}\sum_{N_1=0}^{1}\sum_{N_2=0}^{1} P(1,1,VP,N_1,N_2):$$
$$\sum_{V_c=0}^{1}\sum_{VP=0}^{1}\sum_{N_1=0}^{1}\sum_{N_2=0}^{1} P(1,V_c,VP,N_1,N_2)$$

$$P(N_1 = 1|VP = 1) = \sum_{V=0}^{1}\sum_{V_c=0}^{1}\sum_{N_2=0}^{1} P(V,V_c,1,1,N_2):$$
$$\sum_{V=0}^{1}\sum_{V_c=0}^{1}\sum_{N_1=0}^{1}\sum_{N_2=0}^{1} P(V,V_c,1,N_1,N_2)$$

$$P(N_2 = 1|VP = 1) = \sum_{V=0}^{1}\sum_{V_c=0}^{1}\sum_{N_1=0}^{1} P(V,V_c,1,N_1,1):$$
$$\sum_{V=0}^{1}\sum_{V_c=0}^{1}\sum_{N_1=0}^{1}\sum_{N_2=0}^{1} P(V,V_c,1,N_1,N_2)$$

After five steps of model and backward procedure, the best found solution is:

$$P(S_1) = 0.115006$$
$$P(S_2) = 0.576938$$
$$P(S_3) = 0.308056 \quad (16)$$

New cross entropy is $H(P, f) = 0.975708$ and χ^2-test is $\chi^2 = 1.61715$.

Acknowledgements

The research is funded by the Ministry of Science and Technology of the Republic of Croatia under grant 0130441 for project »Annotated database and syntactic structure of Croatian sentences«.

REFERENCES

1. RUDAN, P., D. F. ROGERTS, A. SUJOLDŽIĆ, B. MACAROL, ŽUŠKIN, A. KAŠTELAN, Coll. Antropol., 6 (1982) 39. — 2. RUDAN, P., D. ŠIMIĆ, N. SMOLEJ-NARANČIĆ, L. A. BENNETT, B. JANIČIJEVIĆ, V. JOVANOVIĆ, M. F. LETHBRIDGE, J. MILIČIĆ, D. F. ROBERTS, A. SUJOLDŽIĆ, L. SZIROVICZA, Am. J. Phys. Antropol., 74 (1987) 417. — 3. RUDAN, P., A. CHAVENTRE, Coll. Antropol., 13 (1989) 177. — 4. SUJOLDŽIĆ, A., Coll. Antropol., 13 (1989) 189. — 5. SUJOLDŽIĆ, A. Coll. Antropol., 17 (1993) 17. — 6. SUJOLDŽIĆ, A., P. RUDAN, V. JOVANOVIĆ, B. JANIČIJEVIĆ, A. CHAVENTRE, Coll. Antropol., 11 (1987) 181. — 7. SUJOLDŽIĆ, A., P. ŠIMUNOVIĆ, B. FINKA, L. A. BENNETT, J. L. ANGEL, P. RUDAN, Antropol. Linguistics, 28 (1987) 405. — 8. ŠKREBLIN, L., L. ŠIMIČIĆ, A. SUJOLDŽIĆ, Coll. Antropol. 26 (2002) 333. — 9. TEPEŠ, B., T. ŽUBRINIĆ, L. SZIROVICZA, I. HUNJET, Int. Conf. Information Technology Interface ITI'96, (1996) 91 — 10. SZIROVITCZA, L., A. SUJOLDŽIĆ, B. TEPEŠ, Coll. Antropol. 21 (1997) 609 — 11. TEPEŠ, B., L. SZIROVICZA, A. SUJOLDŽIĆ, M. PRIMORAC, Journal of Computing and Information Technology, 5 (1997) 265 — 12. TEPEŠ B., V. MATELJAN, Coll. Antropol., 27. Suppl. 1. (2003) 195 — 13. TEPEŠ, B.: Računarska ligvistika. (University of Zagreb, Zagreb, 2001). — 14.. HOPCROFT, J. E., J. D. ULLMAN: Introduction to automata theory, languages and computation. (Addison-Weseley Publ. Comp., Reading, 1979). — 15. CHI, Z., S. GERMAN, Computational Linguistics, 24 (1998) 299. — 16. CHI, Z., Computational Linguistics, 25 (1999) 131. — 17. PEARL, J.: Causality, models, resoning and inference. (Cambridge University Press, Cambridge, 2000). — 18. CHOMSKY, N.: The minimalist program, (The MIT Press, Cambridge, 1996). — 19. ROARK, B., Computational Linguistics, 27 (2001) 249. — 20. MOOD, M. A., F. A. GRAYBILL, BOES D. C.: Introduction to the theory of statistics. (McGraw-Hill Book Company, N. Y. 2001)

B. Tepeš

Department of Information Sciences, Faculty of Philosophy, University of Zagreb, I. Lučića 4, 10000 Zagreb, Croatia
e-mail: btepes@ffzg.hr

RAZDIOBA VJEROJATNOSTI SASTAVNICA STRUKTURE

S A Ž E T A K

Ovaj rad opisuje razdiobu vjerojatnosti sastavnica strukture (granaljaka dijelova rečenica) prirodnoga jezika uporabom kauzalnih mreža. Prvi dio rada opisuje bezokolinsku gramatiku s vjerojatnostima i sastavnice strukture. U drugom dijelu opisali smo kauzalnu mrežu i razdiobu vjerojatnosti na njenim vrhovima. Na tim teorijskim temeljima u trećem je dijelu testiran model na Bazi gramatičkih rečenica hrvatskoga jezika(http://infoz.ffzg.hr/tepes/). Na kraju je pokazan povratni postupak i vrednovanje rezultata.