Coll. Antropol. **29** (2005) 2: 731–733
UDC 808.62-3:519.6
Original scientific paper

Causal Bayesian Network for Tagging Syntactical Structure of Croatian Sentences

Božidar Tepeš[1], Lajos Szirovicza[2] and Slobodan Elezović[2]

[1] Department of Information Sciences, Faculty of Philosophy, University of Zagreb, Zagreb, Croatia
[2] Institute for Anthropological Research, Zagreb, Croatia

A B S T R A C T

Paper describes tagging syntactical structure of Croatian language sentences using causal Bayesian network. In the first part of the paper we describe Bayesian model for tagging sentences. Base on this idea, we will test our model on Croatian language sentences on Database of grammatical sentences of Croatian language (http://infoz.ffzg.hr/tepes/). This paper is result of our new research connected with the paper hidden Markov model for tagging of Croatian language texts for project Linguistic Analysis of The European languages and the paper Probability distribution on the parse trees for the project Annotated database and syntactic structure of Croatian languages.

Key words: *causal Bayesian network, tagging syntactic structure, sentences of Croatian language*

Introduction

Anthropological research of the East Adriatic rural population regarding a number of Adriatic islands, estimates basic geographical, historical, economic, demographic and linguistic factors[1–8]. The method of hidden Markov model was aimed to develop the automatic tagging grammatical categories of Croatian language corpora[9–11]. In research syntactic structure of Croatian language we made Database of grammatical sentences of Croatian language[12–13] with grammatical categories and syntactic structure of thousand sentences. With causal Bayesian network we develop automatic tagging of syntactic structure of sentences of Croatian language.

Causal Bayesian network for tagging of syntactic structure

Causal Bayesian network[14] tell us how probabilities in Bayesian network would change as a result of external interventions. Organization of causal Bayesian network is similar to one's knowledge in such modular configurations that permits one to predict the effect of external interventions. In our model modules for syntactic structure of language are phrases. Structure of phrases are defined in X-bar theory in the minimalist program[15]. Syntactic structure of phrase *XP* is constituent structure tree (Figure 1.). Main elements of constituent structure tree are phrase head *X*, spec of head the phrase *ZP*, complement of head the phrase *YP* and adjunct of head the phrase *WP*. We can annotate this phrase structure in Chomsky form[15].

$$\left[{}_{XP} ZP \left[{}_{X'} WP \left[{}_{X'} XYP \right] \right] \right] \qquad (1)$$

For example an verb phrase *VP* can be annotated:

$$\left[{}_{VP} NP \left[{}_{V'} V\,AP \right] \right] \qquad (2)$$

In the verb phrase (2) head is verb *V*, spec of head *V* is the noun phrase *NP* and complement of head *V* is adjective phrase *AP*. This verb phrase has no adjunct of the head *V*. Sentences in natural language have similar phrase structure *IP* with head tense *I* or phrase structure *CP* with head complement *C*.

For causal Bayesian model, it is important to annotated directed paths representing mutual causations. The head of constituent structure three is the most important category for phrase and we directed our paths from head (Figure 2.).

It is very important that causal Bayesian network has directed and acyclic graph. In our model we use

Received for publication February 11, 2005

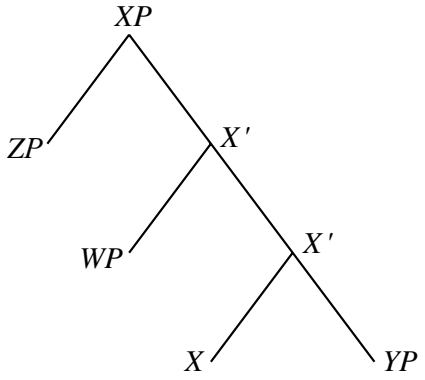

Fig. 1. Phrase structure.

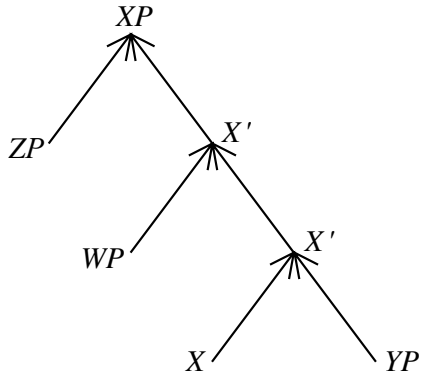

Fig. 2. Causal network of phrase.

graphs where every category has two parents category. For example in causal Bayesian network of phrase (Figure 2.) the category XP has two parent category X' and ZP and the category X' has two parent category X and XP. The head X has no parent category. The phrase categories YP, ZP and WP have parents categories like phrase XP.

This idea is from merging operation in minimalist program[15], where by merging the two categories we get a new category that is only one of these two categories. For example, merging category X with category Y in phrase YP new category is X annotated as projections X' (Figure 2.).

Joint probability of causal Bayesian network (Figure 2.) is a product of probabilities of parents and conditional probabilities:

$$\begin{aligned} &P(XP) = P(X)P(X')P(X')P(YP)P(ZP)P(WP) \\ &P(X'|X,YP)P(X'|WP,X')P(XP|ZP,X') \end{aligned} \quad (3)$$

If we have words in one sentence, we know that we can have $N(n)$ possible constituent structure trees:

$$N(n) = \frac{1}{n}\binom{2(n-1)}{n-1} \quad (4)$$

For sentences with $n = 4$ words we have $N(4) = 5$ possible constituent structure trees (Figure 3.).

Fig. 3. Five possible structures for four words.

We are interested in categories of that structure with two conditions:

1. Merging only possible categories
2. Merging with maximal probability of phrase or sentence

First condition means that we cannot merge, for example projection with projection or phrase with phrase. Second condition means optimization of joint probability of phrase or sentence (3).

Tagging of Croatian language sentences

In our Database of grammatical sentences of Croatian language (http://infoz.ffzg.hr/tepes/) we have syntactic structure of thousand sentences. From this database we can calculate probabilities for joint probability of possible phrases and sentences. In the first step, we have to form possible phrase structures and in the second step we have to evaluate these structures. We can use dynamic programming algorithm in these two steps. For calculation of probabilities and conditional probabilities we can use maximum likelihood algorithm. Modal structure of causal Bayesian network helps us in this calculations. Elementary structure is merging two categories X and Y in phrase XP or projection X'. Joint probability of phrase XP is:

$$P(XP) = P(X)P(Y)P(XP|X)P(XP|Y) \quad (5)$$

For projection X' joint probability is:

$$P(X') = P(X)P(Y)P(X'|X)P(X'|Y) \quad (6)$$

Dynamic programming algorithm means that using (5) and (6) in a modal structure of causal Bayesian network we can calculate joint probability of sentence from joint probability of their phrases.

Conclusion

We can use causal Bayesian network with hidden structure for tagging the corpora of Croatian sentences using structure of sentences from database of grammatical sentences of Croatian language (http://infoz.ffzg.hr/tepes/).

Acknowledgements

This work was supported by the Ministry of Science, Education and Sports of the Republic of Croatia grant to B.T. (0130441) and to L. Sz. (001960003).

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.. PEARL, J.: Causality models, reasoning, and inference. (Cambridge University Press, Cambridge, 2001). — 15. CHOMSKY, N.: The minimalist program. (The MIT Press, Cambridge, Mass.1996).

B. Tepeš

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

KAUZALNE BAYESOVE MREŽE ZA OZNAČAVANJE SINTAKTIČKE STRUKTURE HRVATSKIH REČENICA

SAŽETAK

Rad opisuje označavanje sintaktičke strukture rečenica hrvatskoga jezika. uporabom kauzalnih Bayesovih mreža. U prvom dijelu opisujemo kauzalni Bayesov model za označavanje rečenica. Na temelju toga ćemo provjeriti model na rečenicama hrvatskoga jezika iz baze gramatičkih rečenica hrvatskoga jezika (http://infoz.ffzg.hr/tepes/). Ovaj rad je rezultat naših novih istraživanja vezana za rad Prikriveni Markovljev model za označavanje teksta hrvatskoga jezika za projekt Lingvistička analiza europskih jezika i rad Razdioba vjerojatnosti sastavnica strukture za projekt Označena baza i sintaktička struktura hrvatskih rečenica.