Exploring Word Order Universals: a Probabilistic Graphical Model Approach

1 Introduction

Previous statistical methods in the research of word order universals have yielded interesting results but they have to make strong assumptions and do considerable amount of data preprocessing to make the data fit the statistical model (Greenberg, 1963; Hawkins, 1982; Dryer, 1989; Nichols, 1986; Justeson & Stephens, 1990). Recent studies using probabilistic models are much more flexible can handle noise and uncertainty better (Daume & Campbell, 2007; Dunn et al., 2011). However these models still rely on strong theoretic assumptions and heavy data treatment, such as using only two values of word order pairs while discarding other values, purposefully selecting a subset of the languages to study, or selecting partial data with complete values. In this paper we introduce a novel approach to use a probabilistic graphical model to study word order universals. Using this model we can have a graphic representation of the structure of language as a complex system composed of linguistic features. Then the relationship among these features can be quantified as probabilities.

2 Method

Probabilistic graphical models combine a graphical representation with a complex distribution over a high-dimensional space (Koller & Friedman, 2009).There are two advantages of using this model to study word order universals. First the graphical structure can reveal much finer structure of language as a complex system. We assume there's a meta-language that has the universal properties of all languages in the world. We want a model that can represent this meta-language and make inferences about linguistic properties of new languages. This system is composed of multiple sub-systems such as phonology, morphology, syntax, etc. which correspond to the subfields in linguistics. In this paper we focus on the sub-system of word order only. The other advantage of PGM is that it enables us to quantify the relationships among word order features. A PGM model for word order subsystem encodes a joint probabilistic distribution of all word order feature pairs. Using probability we can describe the degree of confidence about the uncertain nature of word order correlations.

The WALS data has posed a difficulty for applying statistical methods because the languages are not independent and identically distributed due to relatedness in genealogy or geography. To solve the problem of limited data we use model averaging by using bootstrap replicates. To solve the dependence problem among the languages we select each subset randomly and learn a DAG (directed acyclic graph) structure for this subset. First we use bootstrap to create a resample from the original dataset. Then we divide the samples into four groups of equal number of languages randomly and learn the DAG structure and conditional probabilities for each subset; then using the graph fusion algorithm (Matzkevich & Abramson, 1993) we combine all the graphs into the final consensus DAG structure, and use the original data to learn the parameters.

The final "consensus" DAG structure is shown in Figure 1. From this graph we can see word order features are on different tiers in the hierarchy. The root S_O_V "dominates" all the other features; O_V is an important node since it directly "dominates" three other branches of nodes; noun modifiers and noun are in the middle tier while Neg_V, Ad-Sub_Cl, IntPhr and Num_N are the leaf nodes which are the least important features in terms of their contribution to the word order properties of a language.

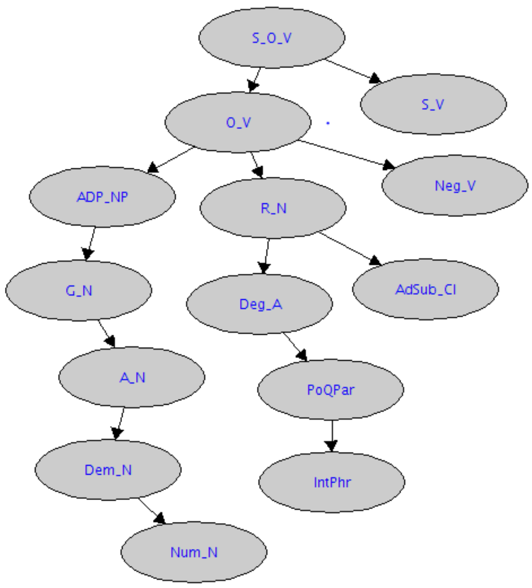

Figure 1. DAG for our PGM model

3 Results

We use SamIam[1] to do probabilistic inference queries since it has an easy-to-use interface for. Figure 2 gives an example: when we know the language is SV and NegV, we can get the probabilities for all values of other features of this language.

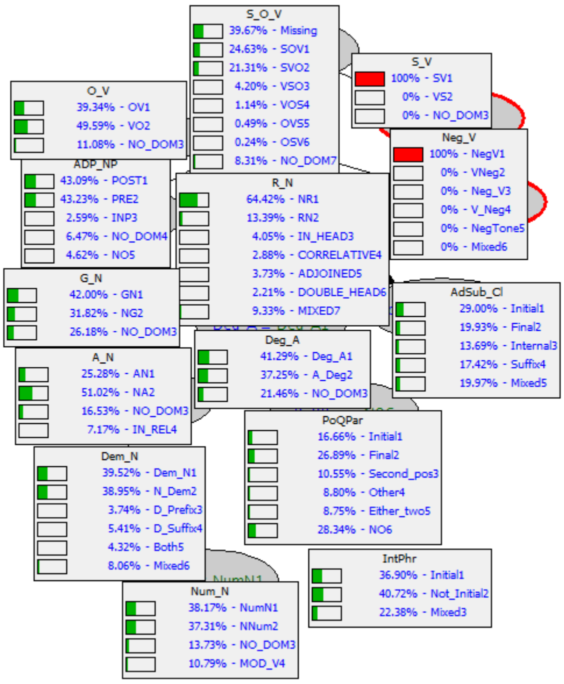

Figure 2. Query example in SamIam

The other type of query is MAP which aims to find the most likely assignments to all of the unobserved variables. For example, when we only know that language is VO, we can use MAP query to find the combination of values which has the highest probability (0.0032 as shown in Table 1).

[1] http://reasoning.cs.ucla.edu/samiam/

P(MAP,e)=0.0015052949102098631
P(MAP|e)=0.003213814742532023

Variable	Value
A_N	NA2
ADP_NP	PRE2
AdSub_Cl	Initial1
Deg_A	Deg_A1
Dem_N	N_Dem2
G_N	NG2
IntPhr	Not_Initial2
Neg_V	NegV1
Num_N	NNum2
O_Obl_V	VOX1
PoQPar	Final2
R_N	NR1
S_O_V	SVO2
S_V	SV1

Table 1: MAP query example

One more useful function is to calculate the likelihood of a language in terms of word order properties. If all values of 13 features of a language are known, then the probability (likelihood) of having such a language can be calculated. We calculated the likelihood of eight languages and got the results as shown in Figure 3.

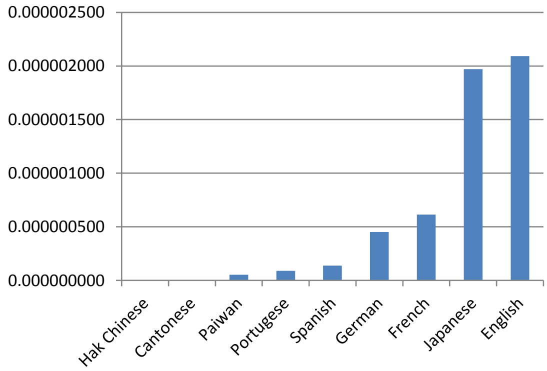

Figure 3. Likelihood of eight languages in terms of word order properties

As we can see, English has the highest likelihood to be a language while Hak Chinese has the lowest. German and French have similar likelihood; Portuguese and Spanish are similar but are less than German and French. In other words English is a typical language regarding word order properties while Hak Chinese is an untypical one.

4 Evaluation and Implications

We did qualitative evaluation through comparison with the well-known findings in word order correlation studies: those of Greenberg's, Dryer's, and Daume and Campbell's.

Compare with Greenberg's and Dryer's work

Universals	**Dependencies**	**UNIV**
U2: ADP_NP<=>N_G	POST->GN PRE->NG GN->POST NG->PRE	88.51 74.63 80.11 86.08
U3: VSO->PRE	VSO->PRE	83.61
U4: SOV->POST	SOV->POST	90.88
U5: SOV&NG->NA	SOV&NG->NA	69.36

U9: PoQPar<=>ADP_NP	Initial->PRE Final->POST PRE->Initial POST->Final	43.12 50.81 15.07 13.99
U10: PoQPar<=> VSO	all values of PoQPar: VSO below 10%	below 10%
U11: IntPhr->VS	Initial->VS	20.88
U12: VSO->IntPhr	VSO->Initial SOV->Initial SOV->not_initial	47.95 25.00 69.06
U17: VSO->A_N	VSO->A_N	24.98
U18&19: A_N<=>Num_N<=>Dem_N	AN->NumN AN->DemN NA->Nnum NA->NDem	84.82 63.67 69.53 54.18
U24: RN->POST (or AN)	RN->POST RN->AN	87.63 30.20

Table 2. Comparison with Greenberg's work

OV	**UNIV**	**VO**	**UNIV**
Correlated pairs			
postposition	90.88	preposition	83.66
GenN	87.92	NGen	67.83
RelN	39.11	NRel	94.36
SQ(final Q)	35.69	QS	15.13
S-AdSub	30.82	AdSub-S	84.69
"wh" phrase in situ	67.86	initial "wh" phrase	32.02
Non-correlated pairs			
A_N	30.11	N_A	66.53
DEM_N	54.41	N_DEM	55.66
NUM_N	48.81	N_NUM	55.10
DEG_A	45.77	A_DEG	37.47
NEG_V	31.05	V_NEG	16.93

Table 3. Comparison with Dryer's work

Compare with Daumé and Campbell's work

We compared the probabilities of single value pairs of the top ten universals with Daume and Campbell's results, which are shown in the following graphs (p(true) is the probability of having the particular implication; prob is the probability calculated in a different way which is not specified[2]):

[2] http://www.umiacs.umd.edu/~hal/WALS/

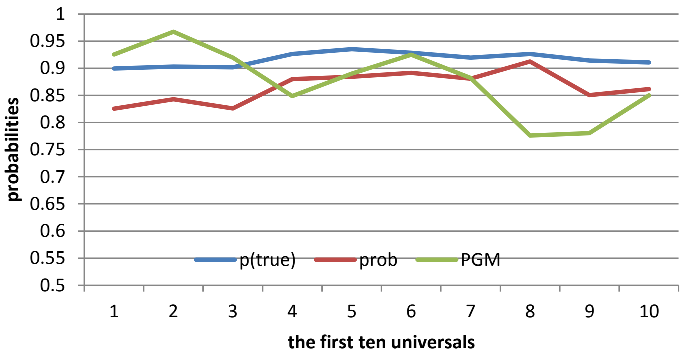

Figure 4: Compare with Daume and Campbell's DIST model

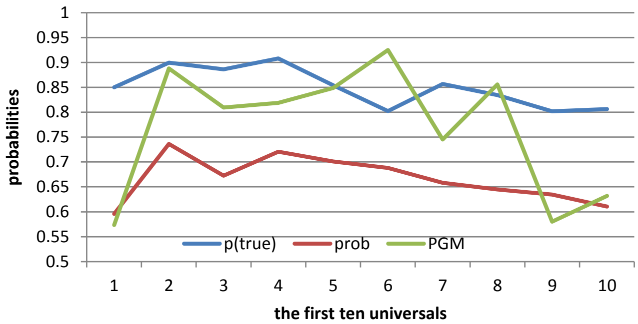

Figure 5: Compare with Daume and Campbell's HIER model

It can be seen that our model provides moderate numbers which fall between the two probabilities in Daume and Campbell's results. In Figure 4 the four universals that have biggest gaps are: 1) VS->VO, 2) OV->SV, 8) Noun-Genitive->Initial subordinator word, 9)Noun-Genitive->Prepositions and in Figure 5 the two universals that have the biggest gaps are: 6)Prepositions ->VO and 7) Genitive-Noun->Postpositions. Our model shows that the word order pair S_V and O_V has higher dependency than the DIST model; and the pair ADP_NP and G_N has lower dependency in both models.

Probabilistic graphic modeling provides solutions to the problems we noticed in the current study of word order universals, which are summarized in the following table:

Problem	**Solution**
only deal with individual features	take language as a complex system
hard to quantify strength of relationships	probabilities can measure the strength of dependencies
interaction between features not clear	nodes are connected to each other in different ways
direction and flow of influence	arrows in the graph
preprocessing of data	no
remove values of features	very little
Null Hypothesis Significance Testing	probability theory

Table 5: Summary of advantages of PGM for word order universal study

References

Bickel, B. 2010a. *Absolute and statistical universals*. In Hogan, P. C. (ed.) The Cambridge Encyclopedia of the Language Sciences, 77-79. Cambridge: Cambridge University Press.

Bickel, B. 2010b. *Capturing particulars and universals in clause linkage: a multivariate analysis*. In Bril, I. (ed.) Clause-hierarchy and clause-linking: the syntax and pragmatics interface, pp. 51 - 101. Amsterdam: Benjamins.

Croft, William. 2003. *Typology and universals*. 2nd edn. Cambridge: Cambridge University Press.

D.M. Chickering, D. Heckerman, and C. Meek. 1997. A Bayesian approach to learning Bayesian networks with local structure. *Proceeding UAI'97 Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*.

Daphne Koller and Nir Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques (Adaptive Computation and Machine Learning series)*. MIT Press, Aug 31, 2009

Dryer, M. S. 1989. Large linguistic areas and language sampling. *Studies in Language 13*, 257 – 292.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2011. *The world atlas of language structures online*. München: Max Planck Digital Library.

Dryer, Matthew S. 2011. The evidence for word order correlations. *Linguistic Typology 15*. 335–380.

Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson & Russell D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature 473*. 79–82.

E. T. Jaynes. 2003. *Probability Theory: The Logic of Science*. Cambridge University Press, Apr 10, 2003.

Greenberg, J. H. 1963. *Some universals of grammar with particular reference to the order of meaningful elements*. In Universals of Language, J. H. Greenberg, Ed. MIT Press, Cambridge, MA, 73-113.

Greenberg, Joseph H. 1966. Synchronic and diachronic universals in phonology. *Language 42*. 508–517.

Greenberg, J. H. (1969). Some methods of dynamic comparison in linguistics.*Substance and structure of language*, 147-203.

Daumé, H., & Campbell, L. (2007). A Bayesian model for discovering typological implications. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS* (Vol. 45, No. 1, p. 65).

Hawkins, John A. 1983. *Word Order Universals.* Academic Press, 1983.

Justeson, J. S., & Stephens, L. D. (1990). Explanations for word order universals: a log-linear analysis. In *Proceedings of the XIV International Congress of Linguists* (Vol. 3, pp. 2372-76).

Leray, P., & Francois, O. (2004). BNT structure learning package: Documentation and experiments.

Maslova, Elena & Tatiana Nikitina. 2010. Language universals and stochastic regularity of language change: Evidence from cross-linguistic distributions of case marking patterns. Manuscript.

Maslova, Elena. 2000. A dynamic approach to the verification of distributional universals. *Linguistic Typology 4*. 307–333.

Matzkevich, I., & Abramson, B. (1992, July). The topological fusion of Bayes nets. In *Proceedings of the Eighth international conference on Uncertainty in artificial intelligence* (pp. 191-198). Morgan Kaufmann Publishers Inc..

Moninder Singh. 1997. Learning Bayesian networks from incomplete data. *In Proceedings of the 14th National Conference on Artificial Intelligence (AAAI)*. AAAI Press, Menlo Park.

Murphy, K. (2001). The bayes net toolbox for matlab. *Computing science and statistics, 33*(2), 1024-1034.

Perkins, Revere D. 1989. Statistical techniques for determining language sample size. *Studies in Language 13*. 293–315.

William Croft, Tanmoy Bhattacharya, Dave Kleinschmidt, D. Eric Smith and T. Florian Jaeger. 2011. Greenbergian universals, diachrony and statistical analyses [commentary on Dunn et al., Evolved structure of language shows lineage-specific trends in word order universals*]. Linguistic Typology 15*.433-53.