Modeling the Relationship among Linguistic Typological Features with Hierarchical Dirichlet Process

Chu-Cheng Lin[a], Yu-Chun Wang[a], and Richard Tzong-Han Tsai[b]

[a]Department of Computer Science and Information Engineering, National Taiwan University,
No. 1, Sec. 4, Roosevelt Rd., Taipei, 10617 Taiwan
r97060@csie.ntu.edu.tw, albyu35@gmail.com
[b]Department of Computer Science and Engineering, Yuan Ze University,
No. 135, Far-East Rd., Chung-Li, 320 Taiwan
thtsai@saturn.yzu.edu.tw

Abstract. We propose that topic models can be used to represent the relationship among linguistic typological features. Typological features are typically analyzed in terms of universal implications. We argue that topic models can better capture some phenomena, such as universal tendencies, which are hard to be explained by implications. We conduct experiments to evaluate the predictive accuracy of our Hierarchical Dirichlet Process (HDP) model on the WALS dataset. We discover some interesting findings. Topics regarding word order types are recognized. We also find a topic that regards areal tendency.

Keywords: topic model, typological feature, Hierarchical Dirichlet Process

1 Introduction

Languages do not exhibit syntactic features, such as position of verbs, affixes, and nouns, at random. For instance, languages that have prepositions tend to have verbs precede nouns. *Implicational universals*, proposed by Greenberg (1966), model these phenomena with logical implications. These implications describe co-occurrence conditions between two features in the form "if a language has feature *x*, then it has feature *y*." For example, one of the universals regarding constituent order found by Greenberg states "If a language has VSO order feature, then it is prepositional."

However, using logical implications to model the relationship among structural features has some problems. One problem is with the interpretation of implication: while implications take the form "feature *x* implies feature *y*", these are hypothesized implications induced from empirical observations, not logically deduced propositions. And thus using implications — which are empirically discovered under the implicational universals model — to explain the relationship among features is tautological (Cysouw, 2003). Additional knowledge is required to interpret them.

One another problem is that implications are in nature *asymmetric*. But it is questionable that the relationship between features are uni-directional, as there do exist bi-directional implications, in which it's impossible to determine which feature of the two is more marked (Croft, 2002).

Also there are implications that have exceptions — called universal tendencies (Rolf and Halvor, 2005). An example is "if a language has OV (object-verb) word order, then it has GenN (genitive-noun) word order," which has numerous counter-example languages. Both tendency and implication can be derived from a feature table (Croft, 2002). Even though many implications have been found, the existence of implications cannot explain why tendencies — which can be identified by the same method — also exist, with varying degrees of exception. This suggests

that implications is a special case where there are no exceptions, and a more general model is demanded.

Last but not least, as Cysouw argues (Cysouw, 2003), implications found in a feature table may have little significance under statistical tests. On the other hand, meaningless feature relations, which cannot be appropriately modeled with implications, may be statistically significant. Cysouw further suggests that using statistical tests to examine the relationship between features is better than indicating the relationship with implications.

As a result, we take a different approach here: we propose a plausible *computational generative model*: unlike the implications, under this model the structural features are generated by some high-level concepts. The major work then is to infer the high-level concepts. And we can then judge the model's sanity by examining the inferred concepts.

Topic models seem adequate for this purpose. They receives more and more attractions in document modeling (Steyvers and Griffiths, 2007), and are widely employed in various NLP applications. In general, a topic model first generates several *topics* of a corpus. Each topic is a distribution over all possible terms in the corpus. To generate a document, one first draws a weighted mixture of topics from some prior, say, Dirichlet distribution. With the common bag-of-words assumption, to generate each word in this document one first draws a topic from the previously drawn mixture; then one draws a word from this topic. Each word's topic assignment, each document's topic mixture, and each topic's distribution over terms can easily be inferred or estimated using Monte Carlo method, which we shall go into more details in Section 3.

Under the assumption of independence between words, it is clear that terms that co-occur more frequently are more likely to belong to a same topic. Empirical results show that for natural language corpora, the inferred topics consist of semantically related terms (Blei et al., 2003). In other words, semantically related words co-occur frequently, which largely adheres to the intuition.

When applying the topic model to interpret the relations between language features, consider that (1) how implications are interpreted (2) what a topic represents. In the implication $f_x \Rightarrow f_y$, f_x should be more significant than f_y. Natually under a topic model we have

$$\sum_m P(f_y|m)P(m|f_x) \gg \sum_{y' \in Y, y' \neq y} \sum_m P(f'_y|m)P(m|f_x),$$

where f_x is a value of feature X and f_y is a value of feature Y. For instance, the feature "word order" may have two values "VO" and "OV". The language cannot have multiple values of a same feature at the same time. m is the latent topic to which both f_x and f_y belong. We do not explicitly model the relation between two features; instead, each feature is an observation of the underlying topic. The less marked a feature is, the more probability it has in a topic. The asymmetric implications then correspond to the case where $P(f_y|m) \gg P(f_x|m)$. Therefore a continuum between uni-directional and bi-directional implications, and continuum between tendency — when there are mutually incompatible features in the same topic — and implication is allowed.

Topic model also handles situations where multiple features are involved at the same time quite neatly. For example, the two object-verb orders OV and VO, which belong to the well-known *word order types*, often co-occur with several features. The OV feature often co-occurs with features such as postpositions, GenN order, SV order, sentence-final question particles, and so on. On the other hand, the VO order co-occurs with prepositions, NGen order, VS order, and sentence-initial question particles. However, it might have some same features co-occuring with the two features such as the two features OV and VO. For example, NAdj order was thought to co-occur with the VO word order. But later studies argue that the ordering of adjective and noun may not be related to verb-object ordering (Dryer, 1988), but to other features. With the introduction of latent topics,

we can deal with the relationship of the features which co-occur with NAdj and AdjN. The detailed discussion is in Section 5.1.

To our knowledge, there is no prior work which takes this view in the literature of linguistic typology. This is understandable since, until the advent of a computational approach (Daumé III and Campbell, 2007), implicational universals have been carried out by painstaking manual analysis; and it is hard to derive the underlying topics with a plain two-dimension feature table.

We propose that a widely-used topic model, Hierarchical Dirichlet Process (HDP), can be used to model the latent topics, which represent relationship among typological features. We evaluate HDP on the WALS dataset (Haspelmath et al., 2005). Then we discuss about implications of the topics we have discovered.

2 Related Work

Daumé III and Campbell are the first to take computational approach to the implicational universal problem (Daumé III and Campbell, 2007). They proposed a Bayesian model that discovers implicational universals. But as we have stated in Section 1, implicational universal has some urgent problems that are to be solved. And later Daumé III proposed a Bayesian model that captures the areal influence effect of neighboring languages (Daumé III, 2009). He considered two factors in his model: areal influence and ancestral influence. While it is widely accepted that a language's structural features are influenced by its ancestors, it remains relatively questionable whether such features can be directly used to build a phylogeny tree (Donohue et al., 2008). The phylogeny trees of Indo-European languages Daumé had built shows significant improvement of the accuracy over those without areal knowledge. However, the phylogeny trees built by features are still less accurate than those built with cognate lists (Gray and Atkinson, 2003). Moreover, while Daumé's model explicitly considered areal influence, the phylogeny trees still show obvious taints of areal diffusion. For example, the Romansch language, which is mainly spoken in Switzerland nowadays, is wrongly put in a Germanic clade in one of his reconstructed phylogeny tree.

In contrast, we do not assume knowledge of how features are transmitted from time to time and place to place. The only assumption of the topic model is that co-occurring features are likely to be affected by the same factor. The factors can be word order system, areal influence, or others. Therefore, we consider topic model to be more objective since it requires less assumptions. In fact, in our experiments we also discover the areal influence in structural features.

3 Methodology

Topic mixture models are widely used to discover the semantic structure of a corpora. With an unlabeled corpora as input, a topic model outputs every word's topic; and hence each document's distribution over topics, and each topic's distribution over features can be determined. Nowadays a very popular topic model is Latent Dirichlet Allocation (LDA) (Blei et al., 2003), a generative Bayesian hierarchical model. Many models have been developed based on LDA (Steyvers and Griffiths, 2007).

While LDA is very useful, it requires the number of topics as parameters. Non-parametric topic models are topic models that do not require pre-specified number of topics. They are suitable because we do not know the actual number of topics a priori. We use Hierarchical Dirichlet Process (HDP), which is the non-parametric extension of LDA, for our purpose.

Here we briefly review HDP from the perspective of the Chinese restaurant franchise metaphor: in this metaphor, each language l_j is a Chinese restaurant with infinite number of infinitely-large tables; and each feature of a language is a customer entering that restaurant. Upon entering the restaurant, this customer $x_{j,i}$, which represents a specific value of some feature, may join an occupied table and eat the dish on the table with probability proportional to number of that table's customers — it's worth noting that order of seating is irrelevant. The dishes correspond to topics;

and customers sitting at the same table — and therefore eating the same dish — corresponds to features coming from the same topic. Or she may sit at an unoccupied table, and order a dish from the menu, with probability proportional to α. When the table t orders from the menu, the probability that a dish $d_{j,t}$ is chosen is proportional to number of tables in *all restaurants* which have chosen $d_{j,t}$ — less metaphorically, that is number of topic t in all languages. Analogously, the probability of choosing a never-chosen dish is proportional to β.

From this metaphor we can see that features of one language share a same topic by clustering around the same table; and languages share the topics by clustering the tables around a dish. Since the number of tables and dishes are not fixed a priori, we can inspect the appropriate number of topics (number of dishes in the metaphor) from posterior sampling.

Inference is carried out by Gibbs sampling in the Chinese restaurant franchise. The Chinese restaurant franchise metaphor yields that the conditional probability of $x_{j,i}$ sitting at table t:

$$p(t_{j,i} = t|\mathbf{t}^{-j,i}, \mathbf{d}, \mathbf{x}) \propto$$
$$\begin{cases} n_t^{-j,i} f(x_{j,i}|\theta_{d_t}) & \text{if } t \text{ is not new,} \\ \alpha f(x_{j,i}|\theta_{d_t}) & \text{if } t = t_{new}. \end{cases}$$

And the probability of a table t of restaurant j choosing dish d:

$$p(d_{j,t} = d|\mathbf{t}, \mathbf{d}^{-j,t}, \mathbf{x}) \propto$$
$$\begin{cases} n_d^{-j,t} \prod_{i:t_{j,i}=t} f(x_{j,i}|\theta_d) & \text{if } d \text{ is not new,} \\ \beta \prod_{i:t_{j,i}=t} f(x_{j,i}|\theta_d) & \text{if } d = d_{new}, \end{cases}$$

where $n_t^{-j,i}$ is the number of customers sitting at table t excluding customer $x_{j,i}$, $n_d^{-j,t}$ is the number of tables in all restaurants that have chosen dish d, excluding table t in restaurant j, and θ_d is the multinomial distribution over features d attaches to. The base Dirichlet distribution (from which θ_d is drawn) is integrated out because it's a conjugate distribution. And likewise we don't really sample each θ_d; rather we integrate it out, too. Therefore only membership of customers and tables get sampled. Detailed implementation of HDP model and Dirichlet Process are described in (Teh et al., 2006; Teh, 2007).

The features are "flattened" to make each topic a plain multinomial distribution over feature values. Posterior inference is done as described in (Teh et al., 2006). The hyperparameters for concentration parameters α and β is $(0.0001, 0.0001)$ for every DP. We record a sample, which is the state of features' topic assignments, after 60,000 burn-in iterations.

4 Experiments

4.1 Data

We use the *World Atlas of Language Structure* (WALS) dataset (Haspelmath et al., 2005), on which we conduct experiments. WALS contains typological features of 2,650 languages. As of the time of writing, 142 features are recorded. Every language can have a value for each feature; however, the dataset is very sparse. Only about 18% of all features have values assigned. WALS divides its features into several, including phonology, morphology, simple clause, and others. For simplicity, we use syntax-related features. They are 102 features in count, with 478 possible values. Since some languages have no recorded values of such structural features, 2,060 languages are used in our experiments.

4.2 Predictive Accuracy & Number of Topics

We evaluate HDP's predictive accuracy on the WALS dataset. First we evenly divide languages that have more than 2 features into 60 sets. For each set, we randomly remove a feature from

each language; the removed features are used as held out. We use the rest data — whole dataset minus the removed features — for topic inference. Predictive probability of all possible values are calculated as follows. If 83_1 (the first value of feature 83 in WALS' numbering) is removed and feature 83 has three possible values in the WALS dataset, then the predictive probability of 83_1, 83_2, and 83_3 are calculated. The one with highest probability is regarded as predicted value. The predictive probability is calculated as

$$p(f) \propto \sum_t n_t \cdot n_{f|t} + \alpha \sum_{t'} \frac{n_{t'}}{\sum_{t''} n_{t''} + \beta} \cdot n_{f|t'},$$

where α and β are concentration parameters, n_t is the number of features in the language with topic t, and $n_{f|t}$ is the occurrence of f in topic t in all languages.

HDP's mean predictive accuracy is 66.01% with standard deviation 0.086.

5 Discussion

Using all features described in Section 4.1, we have obtained 26 topics. We have examined these topics and have discovered some meaningful linguistic phenomena.

5.1 Topics that Resemble Word Order Types

Two topics (#4, #9) in our results regards SVO word order, two (#10, #13) are SOV word order, and one (#8) is VSO word order. The most probable features of these topics are listed in Table 1.

It is interesting to see that although both Topic 4 and 9 exhibits SVO word order type features, they do not describe the same type. Topic 4 suggests an SVO language that has no case marking nor Tense-Aspect-Mood inflection, like Mandarin; while Topic 9 suggests an SVO language that has prefixes, like Zulu. Topics 10 and 13 clearly discuss about SOV languages. AdjN is highly probable in Topic 10, while NAdj in Topic 13. And Topic 8 resembles VSO word order.

In spite of minor difference, Topic 10 and 13 largely agree on features of the OV word order type, while Topic 4, 8, 9 agree on that of VO type. This provides a new perspective on a longly disputed problem: it was originally proposed that the AdjN order co-occurs with OV, and NAdj with VO. However Dryer (1988) argues the order of Noun and Adjective may not be related to the order of Verb and Object: he has shown that OV languages having NAdj and AdjN have different geographic distribution. Later in Justeson and Stephens' analysis they find AdjN correlates with RelN (Justeson and Stephens, 1990). Our analysis does not yield pairwise correlation between features; but shows the context of the features in which they appear. In the topics we have found, we find OV co-occurs with both NAdj (in Topic 13) and AdjN (in Topic 10). VO co-occurs with NAdj (Topic 4 and 9); but VSO co-occurs with AdjN (Topic 8). We can also find AdjN co-occurs with RelN in Topic 10, while NAdj co-occurs with NRel in Topic 4, 9, and 13. Topic 8 is an exception, however. Support of the former arguments can be seen in individual topics: Dryer's in Topic 10 and 13; Justeson and Stephens' in Topic 4, 9, 10, and 13. However, under the topic model, the presence of one feature does not depend on one another, but rather depends on the underlying topic — which is manifested in features it contains.

5.2 Areal Influence

At first glance, features in Topic 1 (Table 1) do not constitute a meaningful topic. After furthering examination, we observed that languages with over 50% features in Topic 1 come from several language families, which are majorly spoken in Europe and India. Figure 1 depicts the geographic location of the 45 languages with top topic-1 feature ratios. Beside the two skeptical languages listed in Table 2, all other languages belong to Indo-European, Afro-Asiatic, Dravidian, Turkic, and Finno-Ugric families. To our knowledge, a linguistic area of this size has not been reported. However two smaller linguistic areas in this region have been proposed (Emeneau, 1956; Haspelmath, 2001). The India linguistic area covers Dravidian and Indo-European languages, while the

Table 1: Highly probable features: in descending order of probability.

Topic	Features
1	sex-based gender system, situational possibility with verbal constructions, morphologically dedicated second singular and plural imperatives, identical encoding of nominal and locational predication, predicative adjectives have nonverbal encoding, semantic and formal assignment of gender, verbal constructions of epistemic possibility, 'and' and 'with' are not identical, semantic assignment of gender, number of Genders: 2
4	NDem, No case affixes, NAdj, SV, VO, polar questions use question particles, SVO, no tense-aspect inflection, NRel, NNum, little affixation
8	VO, prepositions, NumN, NRel, VS, NGen, initial subordinator word, initial interrogative phrase, DemN, AdjN, VSO, tense-aspect suffixes
9	VO, SVO, NAdj, SV, prepositions, NDem, NGen, NNum, NRel, interrogative phrases not obligatorily initial, subject affixes on verb, plural prefix, tense-aspect prefixes
10	AdjN, OV, SV, GenN, postpositions, DemN, tense-aspect suffixes, SOV, plural suffix, NumN, case suffixes, interrogative phrases not obligatorily initial, RelN
13	OV, SV, NAdj, SOV, postpositions, GenN, NNum, tense-aspect suffixes, interrogative phrases not obligatorily initial, NDem, NRel

European linguistic area covers Indo-European, Turkic, Afro-Asiatic, and Finno-Ugric languages. The joint area covers a large proportion of the plotted region in Figure 1. Although the results have not been thoroughly analyzed yet, we think this Übersprachbund-like topic is a remarkable finding. Note that we do not use any spatial information a priori.

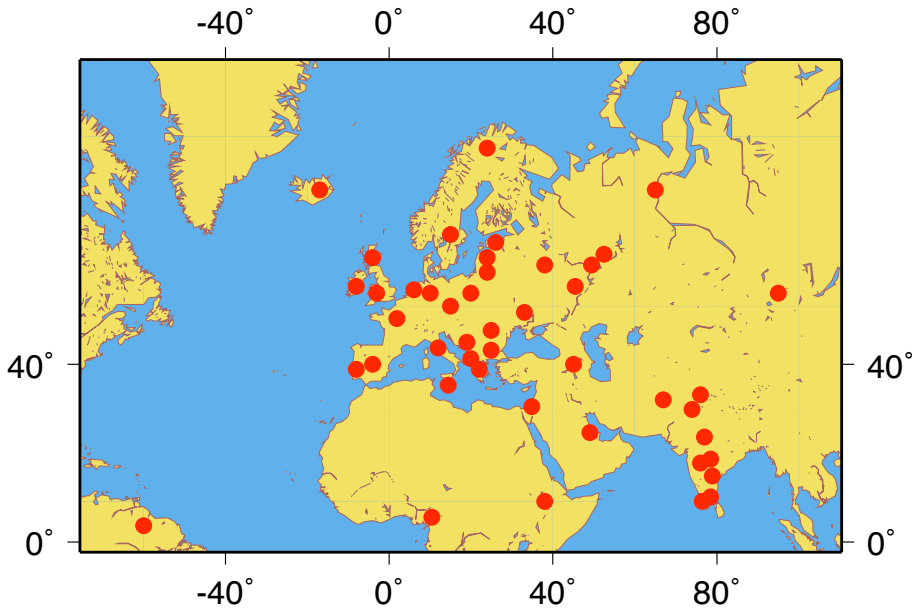

Figure 1: Geographic distribution of Topic 1's 45 most prominent languages.

6 Conclusion

We have described our motivation to use topic models to capture the relationship among typological features, which is a novel approach to typological feature modeling. We have conducted experiments on the WALS data with the Hierarchical Dirichlet Process model, and have evaluated its predictive accuracy. Initial findings show that topics can capture important phenomena of word order types and areal influence.

Table 2: Skeptical languages in Figure 1

Language	Family	Longitude	Latitude
Macushi	Cariban	4°N	60°E
Babungo	Benue-Congo	6.12°N	10.42°E

7 Future Work

As we have discussed in Section 5, we have found topics that are universal (the word order ones) and areal (Topic 1). The areal topic suggests that neighboring languages may share common features. Our future work will focus on the incorporating areal factor into the model.

References

Blei, D. M., A. Y. Ng and M. I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:2003.

Croft, W. 2002. *Typology and universals*. Cambridge University Press.

Cysouw, M. 2003. Against implicational universals. *Linguistic Typology*, 7(1):89–101.

Daumé III, H. and L. Campbell. 2007. A Bayesian model for discovering typological implications. In *Conference of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic.

Daumé III, H. 2009. Non-parametric Bayesian model areal linguistics. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, Boulder, CO.

Donohue, M., S. Wichmann and M. Albu. 2008. Typology, areality, and diffusion. *Oceanic Linguistics*, 47(1):223–232.

Dryer, M. S. 1988. Object-verb order and adjective-noun order: dispelling a myth. *Lingua*, 74:185–217.

Emeneau, M. B. 1956. India as a lingustic area. *Language*, 32(1):3–16.

Gray, R. D. and Q. D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426(6965):435–439, November.

Greenberg, J. H. 1966. Some universals of grammar with particular reference to the order of meaningful elements.

Haspelmath, M., M. Dryer, D. Gil and B. Comrie. editors. 2005. *The world atlas of language structures*. Oxford University Press.

Haspelmath, M. 2001. The European linguistic area: Standard average European. *Language typology and language universals*, 20:1492–1510.

Justeson, J. S. and L. D. Stephens. 1990. Explanations for word order universals: a log-linear analysis. In *Proceedings of the XIV International Congress of Linguists*, volume 3, pages 2372–76. Mouton de Gruyter.

Rolf, T. and E. Halvor. 2005. Linguistics for students of Asian and African languages.

Steyvers, M. and T. Griffiths. 2007. *Probabilistic Topic Models*, chapter 21. Lawrence Erlbaum Associates.

Teh, Y. W., M. I. Jordan, M. J. Beal and D. M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Teh, Y. W. 2007. Dirichlet processes. Submitted to Encyclopedia of Machine Learning.