Parameter Estimation for LDA-Frames

Jiří Materna
Centre for Natural Language Processing
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00, Brno, Czech Republic
xmaterna@fi.muni.cz

Abstract

LDA-frames is an unsupervised approach for identifying semantic frames from semantically unlabeled text corpora, and seems to be a useful competitor for manually created databases of selectional preferences. The most limiting property of the algorithm is such that the number of frames and roles must be predefined. In this paper we present a modification of the LDA-frames algorithm allowing the number of frames and roles to be determined automatically, based on the character and size of training data.

1 Introduction

Semantic frames and valency lexicons are useful lexical sources capturing semantic roles valid for a set of lexical units. The structures of linked semantic roles are called semantic frames. Linguists are using them for their ability to describe an interface between syntax and semantics. In practical natural language processing applications, they can be used, for instance, for the word sense disambiguation task or in order to resolve ambiguities in syntactic analysis of natural languages.

The lexicons of semantic frames or verb valencies are mainly created manually or semi-automatically by highly trained linguists. Manually created lexicons involve, for instance, a well-known lexicon of semantic frames FrameNet (Ruppenhofer et al., 2006) or a lexicon of verb valencies VerbNet (Schuler, 2006). These and other similar lexical resources have many promising applications, but suffer from several disadvantages:

- Creation of them requires manual work of trained linguists which is very time-consuming and expensive.
- Coverage of the resources is usually small or limited to some specific domain.
- Most of the resources do not provide any information about relative frequency of usage in corpora. For instance, both patterns `[Person]` *acquire* `[Physical_object]` and `[Person]` *acquire* `[Disease]` reflect correct usage of verb *acquire*, but the former is much more frequent in English.
- Notion of semantic classes and frames is subjectively biased when the frames are created manually without corpus evidence.

In order to avoid those problems we proposed a method for creating probabilistic semantic frames called LDA-frames (Materna, 2012). The main idea of LDA-frames is to generate the set of semantic frames and roles automatically by maximizing posterior probability of a probabilistic model on a syntactically annotated training corpus. A semantic role is represented as probability distribution over all its realizations in the corpus, a semantic frame as a tuple of semantic roles, each of them connected with some grammatical relation. For every lexical unit (a verb in case of computing verb valencies), a probability distribution over all semantic frames is generated, where the probability of a frame corresponds to the relative frequency of usage in the corpus for a given lexical unit. An example of LDA-frames

Proceedings of NAACL-HLT 2013, pages 482–486,
Atlanta, Georgia, 9–14 June 2013. ©2013 Association for Computational Linguistics

computed on the British National Corpus is available at the LDA-frames website[1].

The original LDA-frames algorithm has two parameters that must be predefined – number of frames and number of roles – which is the most limiting property of the algorithm. A simple cross-validation approach can be used in case of very small data. However, real data is much bigger and it is not recommended to use such techniques. For example, the inference on the British National Corpus using a single core 2.4 GHz CPU takes several days to compute one reasonable combination of parameters.

In this paper we present a non-parametric modification of the LDA-frames algorithm allowing to determine the parameters automatically, based on the character and size of training data.

2 LDA-Frames

LDA-frames (Materna, 2012) is an unsupervised approach for identifying semantic frames from semantically unlabeled text corpora. In the LDA-frames, a frame is represented as a tuple of semantic roles, each of them connected with a grammatical relation i.e. subject, object, modifier, etc. These frames are related to a lexical unit via probability distribution. Every semantic role is represented as probability distribution over its realizations.

The method of automatic identification of semantic frames is based on probabilistic generative process. Training data for the algorithm consists of tuples of grammatical relation realizations acquired using a dependency parser from the training corpus for every lexical unit. For example, suppose that the goal is to generate semantic frames of verbs from a corpus for grammatical relations *subject* and *object*. The training data for lexical unit *eat* may look like {`(peter, cake)`, `(man, breakfast)`, `(dog, meat)`, `...`}, where the first component of the tuples corresponds to *subject* and the second to *object*.

In the generative process, each grammatical relation realization is treated as being generated from a given semantic frame according to the realization distribution of the corresponding semantic role. Supposing the number of frames is given by parameter F, the number of semantic roles by R, the number of slots (grammatical relations) by S and the size of vocabulary is V. The realizations are generated as follows.

For each lexical unit $u \in \{1, 2, \dots, U\}$:

1. Choose a frame distribution φ_u from Dir(α).
2. For each lexical unit realization $t \in \{1, 2, \dots, T_u\}$ choose a frame $f_{u,t}$ from Multinomial(φ_u), where $f_{u,t} \in \{1, 2, \dots, F\}$.
3. For each slot $s \in \{1, 2, \dots, S\}$ of frame $f_{u,t}$, generate a grammatical relation realization $w_{u,t,s}$ from Multinomial$(\theta_{r_{f_{u,t},s}})$, where $r_{f,s}$ is a projection $(f, s) \mapsto r$, which assigns a semantic role for each slot s in frame f. The multinomial distribution of realizations, symbolized by θ_r, for semantic role r is generated from Dir(β).

The graphical model for LDA-Frames is shown in figure 1. It is parametrized by hyperparameters of prior distributions α and β, usually set by hand to a value between 0.01 – 0.1.

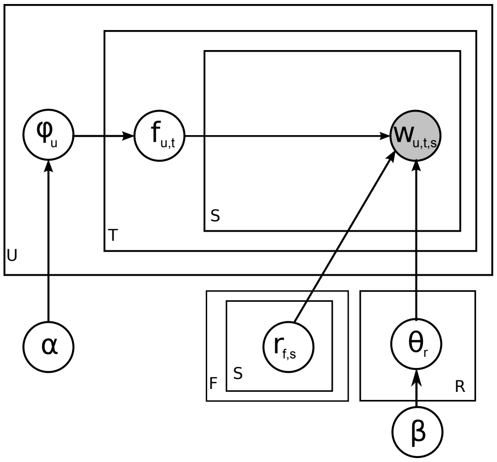

Figure 1: Graphical model for LDA-frames.

The inference is performed using the Collapsed Gibbs sampling (Neal, 2000), where the θ and φ distributions are marginalized out of the equations. In each iteration, latent variables $f_{u,t}$ and $r_{f,s}$ are sampled as follows

$$P(f_{u,t}|\mathbf{f}^{-(u,t)}, \mathbf{r}, \mathbf{w}, \alpha, \beta) \propto$$

$$(fc^{-(u,t)}_{f_{u,t},u} + \alpha) \prod_{s=1}^{S} \frac{wc^{-(u,t,s)}_{w_{u,t,s},r_{f_{u,t},s}} + \beta}{wc^{-(u,t,s)}_{*,r_{f_{u,t},s}} + V\beta} \qquad (1)$$

[1] http://nlp.fi.muni.cz/projekty/lda-frames/

$$P(r_{f,s}|\mathbf{f}, \mathbf{r}^{-(f,s)}, \mathbf{w}, \alpha, \beta) \propto \prod_{v=1}^{V} \left(\frac{wc_{v,r_{f,s}}^{-(f,s)} + \beta}{wc_{*,r_{f,s}}^{-(f,s)} + V\beta} \right)^{wc_{f,s,v}}, \quad (2)$$

where $fc_{f,u}^{-(u,t)}$ is the number of times frame f is assigned to lexical unit u excluding (u,t), $wc_{v,r}^{-(u,t,s)}$ is the number of times word v is assigned to role r excluding (u,t,s), and $wc_{f,s,v}$ is the number of times word v is assigned to slot s in frame f. The asterisk sign * stands for any value in its position.

After having all latent variables **f** and **r** inferred, one can proceed to compute the lexical unit–frame distribution and the semantic role–word distribution using the following formulas:

$$\varphi_u = \frac{fc_{f,u} + \alpha}{\sum_f fc_{f,u} + F\alpha} \quad (3)$$

$$\theta_r = \frac{wc_{v,r} + \beta}{\sum_v wc_{v,r} + V\beta}. \quad (4)$$

3 Parameter Estimation

As one can see from the LDA-frames model, the requirement is to define the number of frames and roles in advance. It is not clear, however, how to select the best values that depend on several factors. First of all, the number of frames and roles usually increase with the growing size of training corpus. If the training data is small and covers just a small proportion of lexical unit usage patterns, the number of semantic frames should be small as well. The parameters are also affected by the granularity of roles and frames. One way to estimate the parameters automatically is to select those that maximize posterior probability of the model given training data.

LDA-frames algorithm generates frames from the Dirichlet distribution (DD) which requires a fixed number of components. Similarly, the latent variables $r_{f,s}$ are chosen from a fixed set of semantic roles. In order to be able to update the number of frames and roles during the inference process, we propose to add the Chinese restaurant process (CRP) (Aldous, 1985) prior for the $r_{f,s}$ variables, and to replace the Dirichlet distribution the semantic frames are generated from with the Dirichlet process (Ferguson, 1973).

3.1 Number of Semantic Roles

In the original version of the LDA-frames model, the latent variables $r_{f,s}$, representing semantic role assignment for slot s in frame f, are chosen from a fixed set of semantic roles without any prior distribution. We propose to generate $r_{f,s}$ from the CRP, which is a single parameter distribution over partitions of integers. The generative process can be described by using an analogy with a Chinese restaurant. Consider a restaurant with an infinite number of tables, each of them associated with some dish, and N customers choosing a table. The first customer sits at the first table. The n^{th} customer sits at table t drawn from the following distribution

$$\begin{aligned} P(t = \text{occupied table } i) &= \frac{n_i}{\gamma + n - 1} \\ P(t = \text{next unoccupied table}) &= \frac{\gamma}{\gamma + n - 1}, \end{aligned} \quad (5)$$

where n_i is the number of customers sitting at the table i and $\gamma > 0$ is a concentration parameter which controls how often a customer chooses a new table. The seating plan makes a partition of the customers (Aldous, 1985).

In the proposed modification of the LDA-frames model, the dishes are replaced with the semantic role numbers and customers with slots of frames. In the model we use prior distribution ω corresponding to the CRP with concentration parameter γ. The latent variables $r_{f,s}$ are then sampled as follows

$$P(r_{f,s}|\mathbf{f}, \mathbf{r}^{-(f,s)}, \mathbf{w}, \alpha, \beta, \gamma) \propto (rc_{r_{f,s}}^{-(f,s)} + \gamma) \prod_{v=1}^{V} \left(\frac{wc_{v,r_{f,s}}^{-(f,s)} + \beta}{wc_{*,r_{f,s}}^{-(f,s)} + V\beta} \right)^{wc_{f,s,v}}, \quad (6)$$

where $rc_r^{-(f,s)}$ is the number of times role r is used in any frame and slot excluding slot s in frame f. Notice that the sampling space has $R+1$ dimensions with the probability of the last unseen component proportional to

$$\gamma \prod_{v=1}^{V} \frac{1}{V^{wc_{f,s,v}}}. \quad (7)$$

3.2 Number of Semantic Frames

Estimating the number of frames is a little bit more complicated than the case of semantic roles. The idea is to replace DD φ_u with the Dirichlet process.

The Dirichlet process $DP(\alpha_0, G_0)$ is a stochastic process that generates discrete probability distributions. It has two parameters, a base distribution G_0 and a concentration parameter $\alpha_0 > 0$. A sample from the Dirichlet process (DP) is then

$$G = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}, \tag{8}$$

where ϕ_k are independent random variables distributed according to G_0, δ_{ϕ_k} is an atom at ϕ_k, and weights β_k are also random and dependent on the parameter α_0 (Teh et al., 2006). Simply, DP is a distribution over some infinite and discrete distributions. It is the reason why DP is often used instead of DD in order to avoid using a fixed number of components.

The question, however, is how to make the sampled frames shared between different lexical units. We propose to generate base distributions of the DPs from GEM distribution (Pitman, 2002) τ with concentration parameter δ. The idea is inspired by the Hierarchical Dirichlet Process (Teh et al., 2006) used for topic modeling. The graphical model of the non-parametric LDA-frames is shown in figure 2.

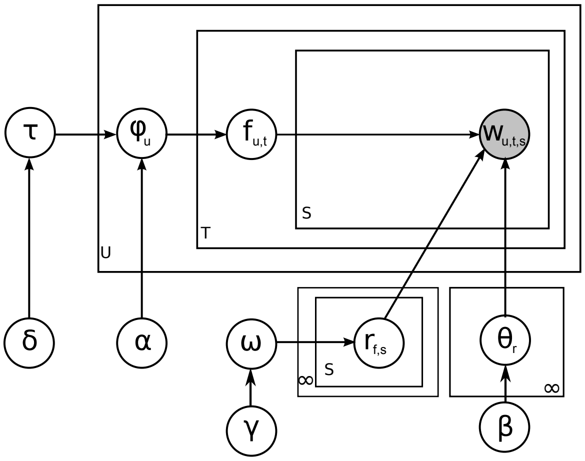

Figure 2: Graphical model for non-parametric LDA-frames.

Since it is hard to integrate out the DP with base distribution generated from GEM in this model, we proceeded to sample τ separately (Porteous, 2010). The base distribution proportions can be sampled by simulating how new components are created for $fc_{f,u}$ draws from DP with the concentration parameter $\alpha\tau_f$, which is a sequence of Bernoulli trials for each u and f (Heinrich, 2011):

$$\begin{gathered} P(u_{f,u,r} = 1) = \frac{\alpha\tau_f}{\alpha\tau_f + r - 1} \forall r \in [1, fc_{f,u}] \\ \tau \sim \text{Dir}(\{u_f\}_f, \delta) \text{ with } u_f = \sum_u \sum_r u_{f,u,r}. \end{gathered} \tag{9}$$

Finally, the latent variables $f_{u,t}$ are sampled as follows

$$\begin{gathered} P(f_{u,t} | \mathbf{f}^{-(u,t)}, \mathbf{r}, \mathbf{w}, \alpha, \beta, \tau) \propto \\ (fc^{-(u,t)}_{f_{u,t},u} + \alpha\tau_f) \prod_{s=1}^{S} \frac{wc^{-(u,t,s)}_{w_{u,t,s}, r_{f_{u,t},s}} + \beta}{wc^{-(u,t,s)}_{*, r_{f_{u,t},s}} + V\beta}. \end{gathered} \tag{10}$$

4 Evaluation

The non-parametric algorithm was evaluated by an experiment on a synthetic data set consisting of 155 *subject-object* tuples. The training data was generated randomly from a predefined set of 7 frames and 4 roles for 16 verbs using the following algorithm. For every lexical unit u:

1. Choose a number of corpus realizations $N_u \in \{5, \ldots, 15\}$ from the uniform distribution.

2. For each realization $n_u \in \{1, \ldots, N_u\}$, among all permitted frames for lexical unit u, choose a semantic frame f_{n_u} from the uniform distribution.

3. For each frame f_{n_u}, generate a realization of all its roles from the uniform distribution.

Each semantic role had 6 possible realizations on average, some of them assigned to more than one semantic role to reflect the character of real languages. Since the data was generated artificially, we knew the number of frames and roles, how the frames were defined, and which frame and which role was responsible for generating each realization in the data.

We ran the non-parametric algorithm with hyperparameters $\alpha = 5, \beta = \gamma = 0.1, \delta = 1.5$. It has been shown that the selection of hyperparameters has little impact on the resulting frames when they are in some reasonable range, thus, the hyperparameters were chosen empirically by hand. The experiment led to correct assignments of $f_{u,t}$ and $r_{f,s}$ after 56 iterations on average (based on 10 independent runs of the algorithm).

In order to compare the non-parametric algorithm with the original, we ran the original algorithm with the same data that had the number of frames and roles set to $R \in \{1 \dots 10\}$, $F \in \{1 \dots 20\}$, and measured the perplexity of the data given to the model after convergence. The perplexities for all settings are shown in figure 3. The lowest perplexity was reached with $F = 7$, $R = 4$ and had the same value as the case of the non-parametric algorithm. The $f_{u,t}$ and $r_{f,s}$ assignments were correct as well.

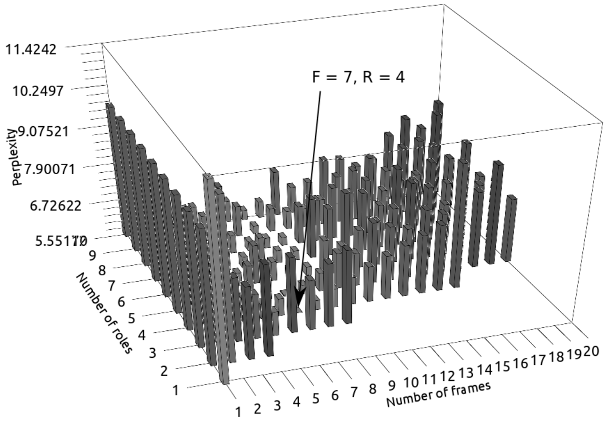

Figure 3: Perplexities for different values of F and R.

We also ran the non-parametric algorithm with the same hyperparameters on real data (1.4 millions of *subject-object* tuples) acquired from the British National Corpus[2] using the Stanford Parser (de Marneffe et al., 2006). The algorithm reached the optimal perplexity with 427 frames and 144 roles. This experiment has been performed only for illustrating the algorithm on real data. Because of long running time of the algorithm on such huge data set, we did not perform the same experiments as with the case of the small synthetic data.

5 Conclusion

In this paper we presented a method for estimating the number of frames and roles for the LDA-frames model. The idea is based on using the Chinese Restaurant Process and the Dirichlet Process instead of the Dirichlet Distributions and selecting such parameters that maximize the posterior probability of the model for given training data. An experiment showed that the non-parametric algorithm infers correct values of both the number of frames and roles on a synthetic data set.

Acknowledgments

This work has been partly supported by the Ministry of Education of the Czech Republic under the project LINDAT-Clarin LM2010013.

References

Aldous, D. J. (1985). Exchangeability and Related Topics. *École d'Été de Probabilités de Saint-Flour XIII – 1983*, 1117:1 – 198.

de Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating Typed Dependency Parses from Phrase Structure Parses. In *The International Conference on Language Resources and Evaluation (LREC) 2006*.

Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1:209 – 230.

Heinrich, G. (2011). "Infinite LDA" – Implementing the HDP with Minimum Code complexity. Technical report.

Materna, J. (2012). LDA-Frames: An Unsupervised Approach to Generating Semantic Frames. In Gelbukh, A., editor, *Proceedings of the 13th International Conference CICLing 2012, Part I*, pages 376–387, New Delhi, India. Springer Berlin / Heidelberg.

Neal, R. M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of computational and graphical statistics*, 9(2):249–265.

Pitman, J. (2002). Combinatorial Stochastic Processes. *Lecture Notes for St. Flour Summer School*.

Porteous, I. (2010). *Networks of mixture blocks for non parametric bayesian models with applications*. PhD thesis, University of California.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., and Scheffczyk, J. (2006). FrameNet II: Extended Theory and Practice. http://www.icsi.berkeley.edu/framenet.

Schuler, K. K. (2006). *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. PhD thesis, University of Pennsylvania.

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

[2] http://www.natcorp.ox.ac.uk