EFFICIENT IMPLEMENTATION OF MORPHOLOGICAL FINITE-STATE TRANSITION NETWORKS EMPLOYING THEIR STATISTICAL PROPERTIES

N.Glushnev[†], B.O'Donovan[‡], A.Troussov[‡]

[†] 415-204 Seigniory, Pte-Claire, QC, H9R 1K2,Canada
nickgl@sympatico.ca
[‡] IBM Dublin Software Lab, Airways Ind. Est., Cloghran, Dublin 17, Ireland
{Brian_ODonovan, atrousso}@ie.ibm.com

ABSTRACT

Here we study numerically the structure of directed state transition graphs for several types of finite-state devices representing morphology of 16 languages. In all numerical experiments we have found that the distribution of incoming and outcoming links is highly skewed and is modeled well by the power law, not by the Poisson distribution typical for classical random graphs. Studied for three languages, distribution of nodes according to the traffic they experience during corpora processing obeys the power law as well. Traffic and out-degree are the parameters, which affect performance of finite-state devices. We discuss how specific properties of power law, like distribution of these parameters (coexistence of small number of "hubs" with large number of "small events"), can be exploited for efficient computer implementation of finite-state devices used in morphology.

Keywords: Finite-state processing, Power law

1. INTRODUCTION

Finite-state devices, including finite-state automata and transducers, are widely used in natural language processing to produce morphological information. Constructed as applications of formal finite-state techniques, they can be considered as networks where nodes represent states and arcs represent the transitions.

In computational linguistics semantic and co-occurrence networks were already studied. In these networks nodes correspond to words. In semantic networks the links show semantic relations between words. In co-occurrence networks links represent the fact that words occur beside each other in a corpus. In [8] the average out-degree of random non-deterministic automata is shown to be a good predictor for the expected number of states in the determinized automaton, the same technique is used in [13].

In the Introduction we present the basics of finite-state processing in morphological applications. In Section 2 – Random Networks – we briefly outline methods and results of this relatively new theory to identify which of them are related to the study of finite-state devices. In Section 3 we describe the morphological data used in our experiments, and in Section 4 - our cross-linguistic experimental study of the degree distribution, which we have found well approximated by the power-law. In Section 5 we describe applications to per-node optimization of finite-state processing.

1.1 Finite-state devices used in morphology

In our experiments we analyzed two major types of finite-state devices, used in natural language processing for word verification and producing morphological information.

Finite-state automata. The input list of surface forms is compiled into a letter tree, which is then minimized to reuse common postfixes. Each word can be loaded with additional information (its part-of-speech categories, etc.), which can be attached to the leaves (the terminals)

of the letter tree. In this case two postfixes can be merged only if they lead to exactly the same information. Finite-state automata (FSAs) constructed this way, are acyclic and deterministic.

In our experiments we also analyzed IBM lexical transducers that implement two-level morphology rules. Some of them have cycles and are non-deterministic.

1.2 Transition networks of finite-state devices

All finite-state devices considered here are finite-state transition networks and can be viewed as the compact description of morphology in the form of algorithm. If information about conditions and/or intentional descriptions of the transitions is ignored, we are left with a graph, which in case of deterministic devices exactly represents control flow of the algorithm. These transition networks range in size from 31,219 to 429,577 nodes.

2. RANDOM NETWORKS

2.1 Graph theory and random graphs

In the theory of random graphs, the simplest and the most intensively studied one-vertex characteristic is degree. The in-degree, k_i, is the number of incoming arcs of a vertex in an directed graph. The out-degree, k_o, is the number of its outgoing arcs. Total distributions of vertex degrees of an entire network, $P_i(k_i)$ – the in-degree distribution, and $P_o(k_o)$ - the out-degree distribution - are basic statistical characteristics of random networks.

Random graphs were first studied in the late 1950s by Erdös and Rényi. In general terms, a random graph is a graph in which properties such as the number of nodes, edges, and connections between them are determined in a random way. In Erdös and Rényi simplest classical model the graph has a fixed number of vertices, which are connected, at random, by edges. The degree in classical random graphs follows binomial distribution which can be approximated by a Poisson distribution $P(k) = e^{-\lambda} \lambda^k / k!$.

2.2 Random massive networks

During the last decade random networks became an interdisciplinary area of research with a strong influence from statistical physics. Empirical and theoretical studies were applied to numerous real world networks in communications, biology, social sciences and economics. Standard indicators or measurements that characterize the structure of a graph are:

- The statistical distribution of links (characterizing homogeneity and scaling properties of the graph);
- The mean or maximum intervertex distance, giving an idea of its size, or diameter;
- The clustering index (a measure of independence of neighboring links).

The following less frequently used characteristic is important for finite-state transition networks (as we consider them as control flow networks):

- the traffic (the number of trajectories passing through each vertex or arc, and so identifying the most active hubs).

The following phenomena were found in many real networks (see [2]):

- Small path length (small-word concept);
- Large degree of clustering;
- Power-law tail degree distribution (scale-free concept).

2.3 Applications to computational linguistics

Methods of random networks theory were already successfully applied to the study of lexical-semantic resources like WordNet - a database of word meanings with basic semantic relations between them, such as synonymy, hyponymy etc. See, for example, ([9], [10] and [6]). The major focus was on the small-world concept. Degree distribution was found to follow the power-law.

3. EXPERIMENTAL DATA

3.1 Morphological data

Finite-state transition networks of IBM morphological dictionaries were used for experiments. For the purpose of this paper, it is necessarily to describe the types of glosses in the dictionaries, because this directly affects topology of the network through minimization stage which eliminates some nodes.

3.2 Description of the dictionaries

Germanic languages: English, German, Dutch, Swedish, Norwegian, Danish. Dictionaries contain word formation elements used for compounding.

Romance languages: French, Italian, Spanish, Portuguese. Clitics are present in the dictionaries.

Ideographic languages: Chinese traditional and simplified. Chinese FSAs are compiled from the lists of words provided with glosses: part-of-speech and "frequencies". The frequencies are used for statistical word segmentation because Chinese language has no orthographic word boundaries. Implementation is complicated by the fact that Chinese is an ideographic language with a repertoire of thousands of characters. In [7] binary search was suggested for implementation of finite-state devices for ideographic languages. We use the polymorphic node structure suggested in [12] .

Lexical transducers. Languages: Finnish, Turkish, Czech, Polish, Thai. Thai dictionary contains words and collocations and is used for word segmentation. Other languages provide inflectional and derivational morphology based on two-level morphology rules.

4. EXPERIMENTAL STUDY OF STATISTICS

4.1 Degree distribution

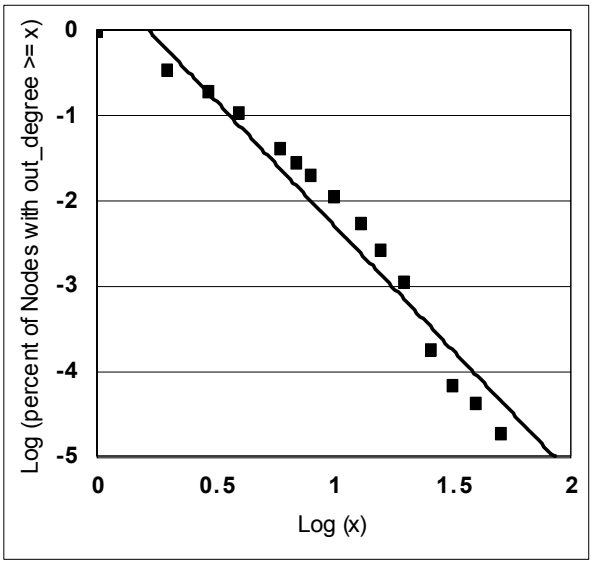

Fig. 1. The log-log (base 10) plot of power-law approximation of the out-degree distribution for English with data binned into exponentially expanding bins so that they will appear evenly spaced on a log scale (the method is discussed in [1] and [4]).

We studied separately distribution of in-, out- and full-degree. In all our experiments, performed for two types of deterministic finite-state devices representing morphology of 16 languages, we have found that degree distribution in all state transition graphs that were considered is well approximated by the power-law. The random variable x is said to be distributed according to the power-law with the exponent γ if its probability density function satisfies $P(x) \sim x^{-\gamma}$.

Language	FS device	In-Degree exponent	Out-Degree exponent
Chinese simplified	FSA	2.47	2.53
Chinese traditional	FSA	2.55	2.56
Czech	FST	2.44	2.69
Danish	FSA	2.62	3.19
Dutch	FSA	2.71	3.35
English	FSA	2.52	3.18
Finnish	FST	2.37	2.35
French	FSA	2.50	3.38
German	FSA	2.67	3.32
Italian	FSA	2.43	3.14
Norwegian	FSA	2.56	3.17
Polish	FST	2.56	2.96
Portuguese	FSA	2.42	3.20
Spanish	FSA	2.45	3.20
Swedish	FSA	2.58	3.09
Thai	FST	2.82	2.95

Table 1. The exponents of power-law approximation of the distribution of in- and out-degree found for our experimental data. Quantification was done with the data binned into exponentially expanding bins (the method is discussed in [1] and [4]).

The usual way to fit power-law distribution is to perform a linear regression on a log-log plot of the cumulative distribution function.

4.2 Traffic

Nodes of a finite-state device experience traffic when the device is used to process corpora. The traffic is defined as a the number of visits to a node while processing a corpus. Traffic was measured for English, French and German dictionaries based on several

corpora with the sizes varying from several hundreds of kilobates to a dozen of megabytes. In all cases traffic demonstrated highly skewed Zipf-like distribution. Fig.2 shows traffic distribution for English dictionary.

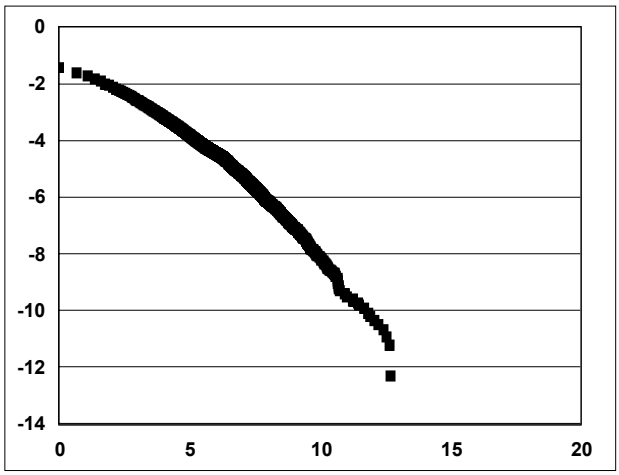

Fig. 2. The log-log (natural log) plot of cumulative distribution function for nodes traffic in English dictionary.

In our experiments top 1% of the most visited nodes covered on average about 90% of the whole traffic. In all our experiments for three languages top 5% of the nodes covered more than 91.1% of the traffic.

4.3 Fitness of the degree-distribution to power-law

Our experiments demonstrate, that the in-degree and full-degree distribution of finite-state transition networks are well approximated by the power-law. In-degree in all our experiments cleanly fit a power-law over about three decades. Empirical distribution is considered to exhibit a regime of power-law decay if log-log plot fits visually to a strait line at least for two decades preferably on both axes (see e.g. [4]).

Deterministic finite-state transition nets, considered in this paper, have finite node's capacity of the out-flowing links, because these links are uniquely labeled by characters from the dictionary alphabet. For example, English dictionary contain 96 unique characters (including case variations). However, out-degree degree distribution also demonstrate limited power-law regime, which is quite different from the Poisson distribution typical for classical random graphs.

5. APPLICATIONS OF STATISTICS FOR EFFICIENT IMPLEMENTATION

5.1 Consequences of power-law behavior

Distributions with peaks characterizing random networks display a strong tendency to cluster around particular values. Thus it may be useful to characterize the distribution by several quantities related to its moments, including the mean (or alternatively – by median and mode). By contrast, in systems exhibiting power law distribution, the mean and median are typically not very useful.

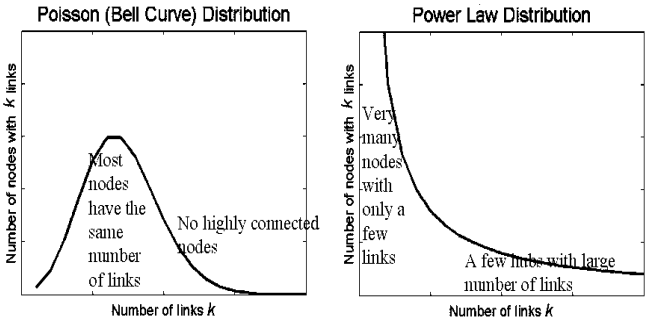

Fig. 3. Consequences of power-law behavior of the out-degree distribution vs. Poisson distribution. In Poisson distribution most of the nodes have number of links close the average. In power-law distribution small number of nodes have significant number of links, coexisting with the many nodes which have only a few links.

Consider applications of a power-law behavior of a distribution of a parameter related to per-node optimization of finite-state processing, such as out-degree distribution or traffic. If this distribution is Poissonean (Bell curve), than most nodes are similar, and optimization based on this parameter is likely to be inefficient. If the distribution is more like a power-law distribution, than

- "rare events" or hubs emerge
- numerous "small" events coexist with a few large ones (spanning for several decades of the parameter values)

Fig. 3 illustrates this, the parameter being the out-degree distribution.

5.2 Applications to the polymorphic node format

Program structures, used to navigate through a finite-state net, have different performance and different memory requirements. For example, deterministic finite-state devices are most efficiently implemented with transition tables that enable rapid selection of links between states, where these links are stored in an array indexed by characters from the input language. However, efficiency comes at the expense of considerable memory overheads. Thus, efficiency of application of transition tables depends primarily on the distribution of nodes with respect to the traffic they experience during corpora processing. Transition tables can be efficiently used if this distribution is a power-law-like fat-tailed distribution, so that a small number of nodes are responsible for a significant portion of traffic.

The paper [12] provided classification of nodes according to their traffic-related role in an FSA and a formal procedure for assigning a polymorphic format to each node based on this role and out-degree of the node.

However, the optimization model [12] was based on heuristic assumptions about the structure of dictionaries. Experimental results of this paper support most of these assumptions; in particularly, in our experiments top 1% of the nodes with the biggest traffic ("hubs") are responsible for about 90% of the whole traffic. This explains high efficiency of the per-node optimization reported in [12].

Another specific property of a power-law-like distribution is a large number of "small events", which for out-degree is related to emergence of filament-like "letter chains", where the out-degree of several consecutive nodes is one. A dedicated node format can be assigned to the head node of each letter chain, with the responsibility of this node/format to perform the necessary test to allow direct transition to the end of the chain. This gives an opportunity to eliminate without any loss some nodes, which is known in the construction of word graphs as compaction or path compression method. The known efficiency of the methods hinges upon power-law-like distribution of out-degree.

5.3 Applications to the storage of sparse transition tables

Transition tables are memory expensive, which suggests that methods for storing sparse tables, like those described in [11], might be useful. One such method can be described as follows. Instead of occupying different memory segments, transition tables, assigned for different nodes are all mapped into one large array, so that transitions associated with different nodes do not overlap. Each node is assigned a displacement, which determines position of the transition table associated with this node in the common array. A procedure of mapping is presentred in [11]. It starts from the transition table with the largest out-degree, i.e. number of actual transition, then displacement for the second large transition table is calculated, etc., finishing with the smallest transition table.

Efficiency of this method can be measured as a ratio of the volume of the resultant common table to the sum of volumes of the original transition tables. Efficiency of the method is proved analytically [11] for the particular case of the out-degree distribution, satisfying the condition of the so-called harmonic decay.

According to the results of this paper, out-degree distribution can be approximated by the power law $p(x > k) \cong ak^{-\gamma}$ where $\gamma \sim 3$. The cumulative probability $p(x > k)$ can be expressed in terms of the probability density function $f(z)$:

$$p(x > k) \cong \int_{k}^{\infty} f(z)dz \cdot$$

Thus, the probability density function $f(z)$ (which is the probability that a given node contains exactly z non-empty transitions) takes the form

$$f(z) \cong \frac{ax^{-(\gamma+1)}}{\gamma} \quad .$$

A decay is called harmonic [11] if the following inequality holds:

$$\frac{n(l)}{n} < \frac{1}{l+1} \quad ,$$

where $n(l)$ is the number of non-empty transitions in all nodes with out-degree greater than l, and n is the total number of non-empty transitions.

The following inequalities

$$n(l) \cong \sum_{k=l+1}^{N} kf(k) = \frac{a}{\gamma} \sum_{k=l+1}^{N} k\left(\frac{1}{k}\right)^{\gamma+1} = \frac{a}{\gamma} \sum_{k=l+1}^{N} \left(\frac{1}{k}\right)^{\gamma} \le \frac{a}{\gamma} \int_{l}^{N} x^{-\gamma} dx =$$

$$= \frac{ax^{1-\gamma}}{\gamma(1-\gamma)}\Big|_{l}^{N} = = \left(\frac{a}{\gamma(\gamma-1)}\right)\left(\left(\frac{1}{l}\right)^{\gamma-1} - \left(\frac{1}{N}\right)^{\gamma-1}\right) ,$$

where N is the size of transition tables, demonstrate that $n(l)$ has asymptotically a power law distribution with the exponent $1-\gamma \sim -2$ (-1.35 for the Finnish in the worst case, -2.38 for French in the best case), which is more favorable for compression of sparse transition arrays than the respective condition for the harmonic decay in [11] (the latter can be expressed as $1-\gamma \sim -1$).

Our analytical results for the out-degree distribution studied here suggest that the method of storing transition tables in one array is efficient for dictionary's implementation, especially for ideographic languages.

6. Conclusions

Simulations for morphology of 16 languages, represented by several types of deterministic finite-state devices, show that an important structural property of finite-state transition graphs – degree distribution – is well approximated by the power law. Traffic, measured for English, French and German dictionaries, exhibited a highly skewed power law distribution. Top 1% of the most visited nodes covered on average about 90% of the whole traffic.

Traffic and out-degree are the parameters affecting performance of finite-state devices. Specific properties of power-law distribution of these parameters (coexistence of small number of "hubs" with large number of "small events") can be exploited for efficient computer implementation of finite-state devices used in morphology.

References

[1] L. Adamic, "Zipf, power-laws, and Pareto–a ranking tutorial." http://www.parc.xerox.com/istl/groups/iea/papers/ranking/ranking.html

[2] R. Albert, and A.-L. Barabasi, "Statistical Mechanics of Complex Networks," cond-mat/0106096 (June 2001).

[3] S.N. Dorogovtsev, and J.F.F. Mendes, "Language as an evolving word web," Proceedings of The Royal Society of London. Series B, Biological Sciences, 268(1485):2603–2606, 2001.

[4] S.N. Dorogovtsev, and J.F.F. Mendes, Evolution of networks. Oxford U. Press, New York, 2003.

[5] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin, "Modern architecture of random graphs: constructions and correlations," arXiv:cond-mat/0206467, v1, 24, Jun 2002.

[6] R. Ferrer Cancho, and R.V. Solé, "The small world of human language," Proceedings of The Royal Society of London. Series B, Biological Sciences, 268(1485):2661–2665, 2001.

[7] T. Goetz, H. Wunsch, "An Abstract Machine Approach to Finite State Transduction over Large Character Sets," Finite State Methods in Natural Language Processing 2001. ESSLLI Workshop, August 20-24, Helsinki.

[8] Ted Leslie, Efficient approaches to subset construction, Master's thesis, Computer Science, University of Waterloo. 1995.

[9] A. E. Motter, A. P. S. de Moura, Y.-C. Lai, and P. Dasgupta, "Topology of the conceptual network of language," Physical Review E, 65, 065102, (2002)

[10] M. Sigman, and G.A. Cecchi, "Global organization of the lexicon," cond-mat/0106509, (2002)

[11] R.E. Tarjan, and A. Yao, "Storing a sparse table," Communications of the ACM, 22(11):606-611, November 1979.

[12] A. Troussov, B. O'Donovan, S. Koskenniemi, and N. Glushnev, "Per-Node Optimization of Finite-State Mechanisms for Natural Language Processing," Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science N 2588, Springer-Verlag, pp.223–227, (2003).

[13] G. van Noord, "Treatment of Epsilon Moves in Subset Construction," Computational Linguistics 26(1): 61-76 (2000)