An Empirical Study of the Occurrence and Co-Occurrence of Named Entities in Natural Language Corpora

K Saravanan, Monojit Choudhury, Raghavendra Udupa, A Kumaran
Microsoft Research India
{v-sarak, monojitc, raghavu, a.kumaran}@microsoft.com

Abstract

Named Entities (NEs) that occur in natural language text are important especially due to the advent of social media, and they play a critical role in the development of many natural language technologies. In this paper, we systematically analyze the patterns of occurrence and co-occurrence of NEs in standard large English news corpora - providing valuable insight for the understanding of the corpus, and subsequently paving way for the development of technologies that rely critically on handling NEs. We use two distinctive approaches: normal statistical analysis that measure and report the occurrence patterns of NEs in terms of frequency, growth, etc., and a complex networks based analysis that measures the co-occurrence pattern in terms of connectivity, degree-distribution, small-world phenomenon, etc. Our analysis indicates that: *(i)* NEs form an open-set in corpora and grow linearly, *(ii)* presence of a *kernel* and *peripheral* NE's, with the large periphery occurring rarely, and *(iii)* a strong evidence of *small-world* phenomenon. Our findings may suggest effective ways for construction of NE lexicons to aid efficient development of several natural language technologies.

Keywords: Named entities, occurrence and co-occurrence profiles, language corpora

1. Introduction

Named Entities (NEs) that occur in natural languages text play a significant role in development of many Natural Language Processing (NLP) technologies used for information access and extraction, machine transliteration and translation, cross-language search, etc. In addition, the exponential rise in the content generated in social media that necessarily anchors on NEs only underscores the criticality of NEs in many of the new generation information and entertainment systems. For the first time in the research literature – to the best of our knowledge – we systematically analyze the patterns of occurrence and co-occurrence of NEs in natural language corpora, through two distinct approaches: First, a conventional analysis of profiling the volume, frequency and growth of occurrence of NEs in the natural language text, and, second, complex networks based analysis to profile the co-occurrence patterns of NEs in the same corpora, comparing and contrasting with the normal words. The analysis and our findings could provide valuable insight in the development of NE lexicons and technologies that rely critically on handling NEs in many languages.

1.1 Corpora used in this Study

For all our experiments, we considered large and diverse sets of standard English news corpora of different genre, time-period and publication demographics (details in Table 1). Specifically, we used parts of the Gigaword-AFE corpus[1] published by the Linguistic Data Consortium (LDC) that included English news published in several demographics around the world (referred as AFE1994), the LA Times corpora[2] consisting of news articles published by The Los Angeles Times during 1994 and 2002 (referred as LAT1994 and LAT2002, respectively), and The Telegraph of India corpora[3] consisting of news articles published during 2004-07 (referred as TTI2004).

2. Related Work

Starting with the Sixth Message Understanding Conference (MUC-6) (Grishman & Sundheim, 1996), Named Entities were the focus of many conference tracks, workshops and shared tasks, spanning many aspects such as, recognition, classification, categorization, relationship extraction, etc. In addition to English, NEs in other languages were also exclusively focused on in specific forums, such as, CONLL-2003 in Spanish, German and Dutch NEs, MUC-6 in Japanese NEs, etc. The profiles of NEs in six natural languages text were profiled in (Palmer et al., 1997); showing that the occurrence of NEs in all the languages followed Zipfian distribution. Co-occurrence profiles of NEs in text corpora have been explored in the past (Hasegawa et al., 2004; Jinxiu, 2007), but specifically in relationship extraction tasks using semi-supervised and unsupervised learning techniques. Even though their approaches depended on co-occurrence of NEs, they didn't explicitly measure and report the co-occurrence profiles of NEs. In this work, our objective is to analyze and report occurrence and co-occurrence patterns of NEs in natural language corpora.

3. Occurrence Profile of NEs

Manual annotation of NE's is not viable for the large news corpora that we used (consisting of several hundreds of thousands of articles). Hence, we used the standard

[1] Published by the Linguistic Data Consortium (LDC), Catalog number: LDC2003T05.

[2] Published as a part of the Cross-Language Evaluation Form (CLEF) corpora in 2006 & 2007.

[3] Published as a part of the Forum for Information Retrieval Evaluation (FIRE) corpora in 2008.

Stanford Named Entity Recognizer (Finkel et al., 2005) for identifying the named entities.

This Named Entity Recognition (NER) tool version 1.2.1 was trained with the three class model files corresponding to the three core NE types '*person*', '*location*' and '*organization*', and used for annotation of NE's, with one of the three tags. Note that a set of continuous words that are annotated with the same tag are considered as single NE. Hence, we consider both single and multi word NEs for our analysis. In addition, typographically different strings identified by the NER are considered as unique NEs. The below example shows the way NEs are extracted from news articles.

Sample input: `John, Jane Albert and Joseph Shaw went to opera in London.`

NER output: `John/PERSON ,/O Jane/PERSON Albert/PERSON and/O Joseph/PERSON Shaw/PERSON went/O to/O opera/O in/O London/LOCATION ./O`

NEs list after grouping the multi word NEs:

Person: {`John, Jane Albert, Joseph Shaw`}

Location: {`London`}

The published accuracy of the NER is 91.88% for the '*Person*' tags, 82.91% for the '*Organization*' tags and 88.21% for the '*Location*' tags. We confirmed the accuracy by manually annotating 10 randomly selected news articles from our corpora with one of the above tags. The accuracy of the NER in this sample data is 87.5% for the '*Person*' tags, 87% for the '*Organization*' tags and 96% for the '*Location*' tags. We assume similar NER performance for the entire corpus for subsequent analysis.

In addition to all type NEs, person type NEs also considered as they constitute the largest and arguably, the most interesting type. Table 1 outlines the occurrence figures for all type NEs and person type NEs identified in various corpora, both in terms of total NEs and unique NEs (corpora-wide). For example, in the AFE1994 corpus, we find totally about 9.2M NEs, occurring at the rate of about 8.5K NEs per day, and about 20 NEs per article. Considering only unique NEs, we find about 491K NEs, occurring at the rate of 451 NEs per day, and approximately 1 NE per article. The first significant observation from these figures is that the NEs occur in large volumes in news corpora. Second, we observe that there are a large number of new NEs occurring on a daily basis in every corpus, as high as two thousand new NE's in the LAT1994 corpus. And the plot of daily growth of unique NE's (Figure 1) indicates a trend that is nowhere near saturation. The growth of vocabulary in standard natural language corpora is known to follow Heap's law (Heaps, 1978), according to which vocabulary grows sub-linearly with corpus size (typically as N^{β}, with β varying from 0.4 to 0.6 for English). However, from Figure 2 we observe that the number of unique NEs grows almost linearly (β varying from 0.73 to 0.82) with corpus size, i.e., much faster than vocabulary growth in all the four corpora. Further, we observe that such a trend seems to be universal, irrespective of size, time period or publication geography of the news articles from other corpora, confirming the empirical fact that NEs in a language are an open set. This finding points to the futility of manually constructing exhaustive NE lexicons; as no matter how big such a constructed lexicon is, new NEs will always be encountered.

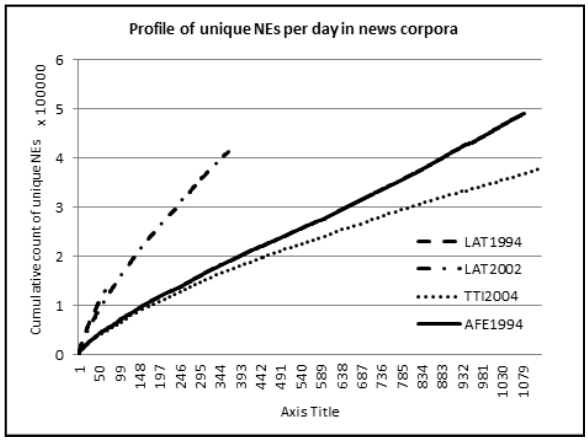

Figure 1: Growth of unique NEs per day in all corpora

Figure 3 shows the cumulative frequency distribution of NEs – in log-log plot – for all corpora. The plot indicates that majority of the NEs occur very infrequently. For example, about 90% of the NEs occur 10 times or less in the AFE1994 corpus and such trend seems universal across all corpora types. Also, we observe that this distribution of occurrence of NEs conforms to power-law with two regimes (Cancho & Solé, 2001), and such two-regime power-law distributions have been attributed to the presence of a distinct kernel and peripheral lexicons. Kernel lexicon refers to common shared vocabulary and the peripheral lexicon refers to domain specific jargons used infrequently. In the context of News corpus NEs, the kernel is comprised of only 0.01% of all the NEs indicating that most of the NEs occur only in the news pertaining to specific domains. We observe similar patterns in our four corpora (Figure 4). This indicates strongly that any technology developed based on NEs should be fairly robust to the occurrence frequency, in order not to miss out majority of them.

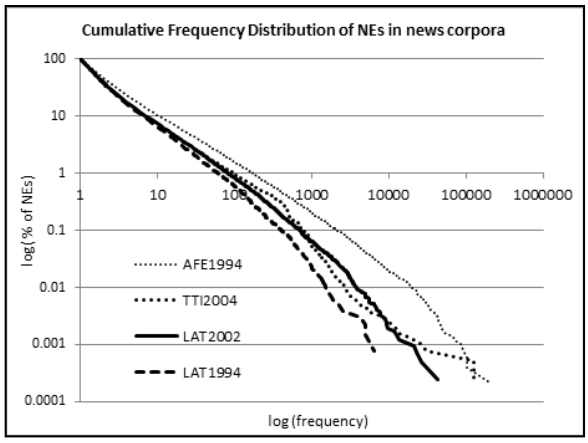

Figure 3: Cumulative Frequency Distribution of NEs (in all News Corpora)

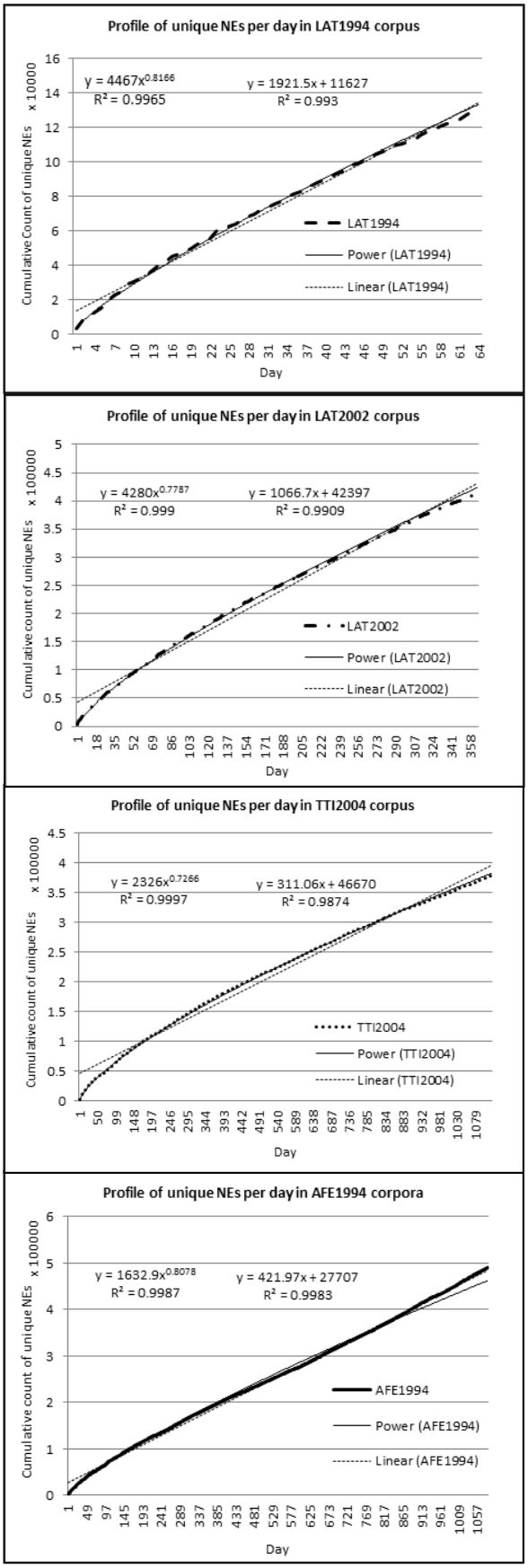

Figure 2: Growth of unique NEs per day (in each corpus)

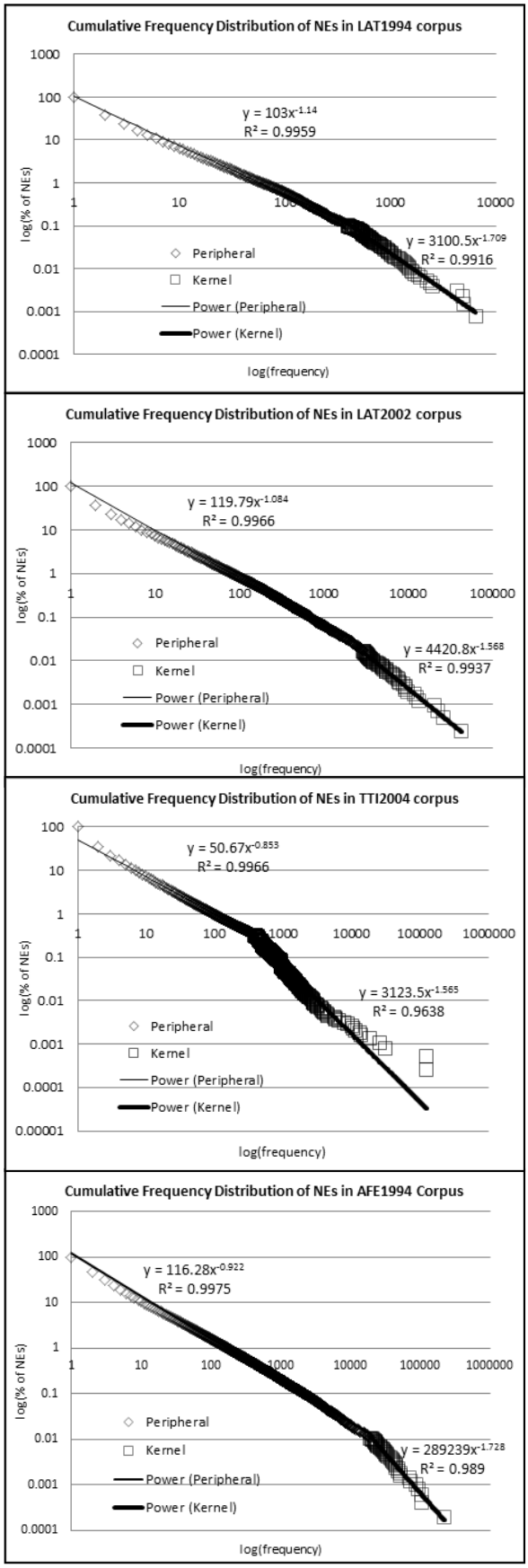

Figure 4: Cumulative Frequency Distribution of NEs (in each corpus)

3.1 Occurrence Profile of NEs in Wikipedia

As a comparative profile, we also analyzed the occurrence of NEs in Wikipedia corpora – both in English and German Wikipedias, and present our findings here.

English and German Wikipedias are among the largest Wikipedias with about 3.5 Million and 1.5 Million articles each. Since our objective is to profile the occurrence of NEs in Wikipedia articles, we randomly selected about 1 Million articles from English and German Wikipedia primarily from the general articles, discarding special pages, such as user, talk, admin, etc. We extracted only the content of an article, and the NEs in the body are identified using the NER tools in English (Finkel et al., 2005) and German (Faruqui & Pado, 2010) respectively. Table 2 provides the NEs that occur in the articles, and Figure 5 shows the growth of unique NEs with size of Wikipedia (in number of articles).

NE Type in Corpora	English		German	
	Total NEs count	NEs per Article	Total NEs count	NEs per Article
All	13,776k	13.8	13,686k	13.69
Unique	2,784k	2.8	2,538k	2.54

Table 2: NEs in English & German Wikipedia.

We observe that these two profiles of occurrence of NEs (in English and German Wikipedia) are very similar to each other, as well as, to what we find in the news corpora, corroborating our findings on occurrence and frequency of NEs in natural language text.

4. Co-occurrence Profile of NEs

Recently, complex networks theory (CNT) (Albert & Barabási, 2002) and (Newman, 2003).has been used as a tool to model and study the statistical properties of a corpus (see, e.g., Cancho & Solé, 2001). Complex network refers to a system of nodes denoting physical or abstract entities and a set of edges connecting them, which usually represent certain interactions between the entities, for example, the cell, the Internet, or a network of computers. In linguistics, network models have been used to study linguistic entities and their interactions (Choudhury & Mukherjee, 2009). Of special interest to us is the concept of Word Collocation Networks (Cancho & Solé, 2001), where the nodes represent unique words and edges connect nodes that co-occur in a sentence close to each other. Since then word collocation networks have been used to model corpora in various languages (see, e.g., Choudhury et al., 2010) and various other kinds of datasets, such as query logs (Roy et al., 2011) and Indus valley symbols (Sinha et al., 2009) to establish linguistic nature of apparently non-linguistic datasets.

Through empirical studies, authors showed that these networks exhibit small-world property similar to social network of people, and feature two-regime power-law which, the authors argued, is due to the existence of distinct kernel and peripheral lexicon in a language.

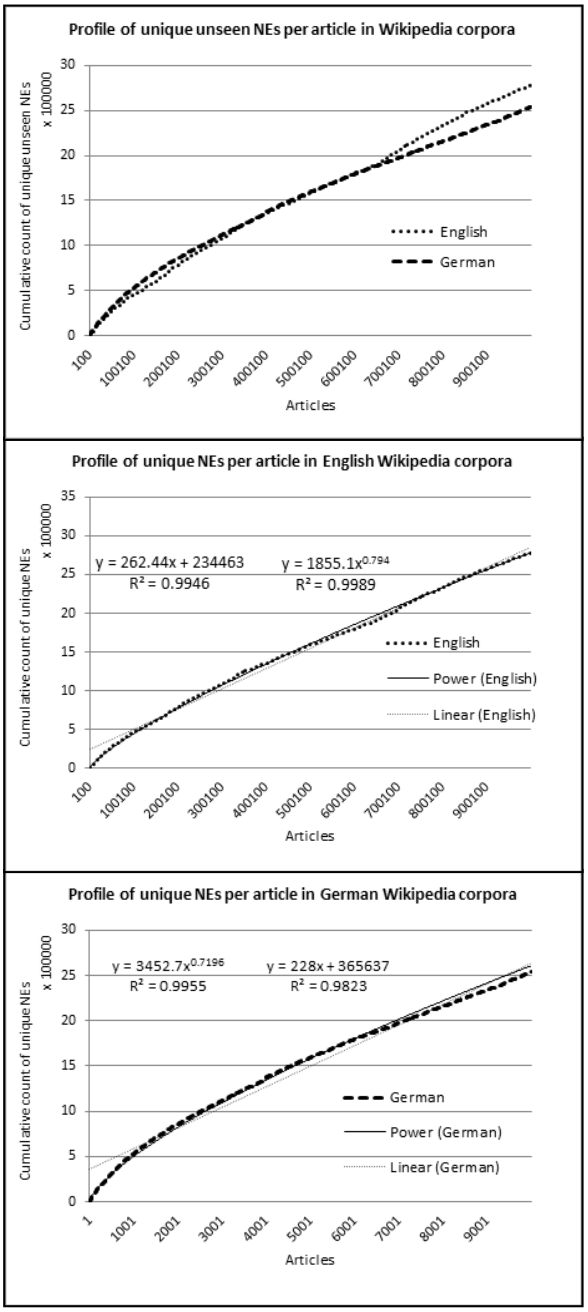

Figure 5: Growth of NEs in Wikipedia

The present study takes inspirations and ideas from this rich and fascinating body of research work and models the co-occurrence network of NEs. The objective of the study is to discover the principles of co-occurrence for NEs, and compare and contrast those principles with that of words.

4.1 Network Definition

We define the co-occurrence network NEs as a graph $G = \langle V, E \rangle$, where V is the set of nodes each labelled by a distinct NE. Two nodes u and $v \in V$ are connected by an edge $e_{uv} \in E$ if and only if the two NEs, NE_u and NE_v represented respectively by the nodes u and v, co-occur in at least one news article in the corpus. We define G_{type} as projection of G onto a particular type of NE, say person, location or organization, as the subgraph of G induced by nodes which represent NEs only of that particular type and the corresponding edges. In this study, we only report

our findings on G_{person} as person NEs constitute the largest and arguably, the most interesting type.

An important distinction between the word collocation network defined in (Cancho & Solé, 2001) and the co-occurrence network of NEs that has been defined here, is that in the former there is an edge between two words only if they co-occur in close vicinity of each other (within 3 words) in a sentence, while in the latter two NEs are considered connected if they co-occur anywhere in the same article. This we believe is a meaningful and natural way of defining co-occurrence between NEs because usually in a news article there are only a few distinct NEs per article (12.7 to 24.1 on average in the studied corpora). Furthermore, co-occurrence in an article would usually imply some causal relation between the NEs (e.g., NE_u initiated an event that happened in location NE_v and affected NE_w). Other reason is in several IR applications, document level co-occurrence is more important than sentence level co-occurrence.

4.2 Observations and Discussions

Table 1 summarizes the basic statistics of the networks – G and G^{person} constructed from the four collections of news articles.

Connectivity: We observe that the networks so constructed have one large connected component that constitutes more than 99% of the nodes for G and 91 to 97% for G_{person}. That is, most of the nodes are connected to each other, and a few nodes form small communities of their own. Usually these are NEs that have occurred in only one or very few articles, in which there is no other common NE. Moreover, slightly smaller sizes of the largest component for the corresponding G_{person} suggests that other types of NEs, especially locations, play a vital role in connecting different NEs.

Degree Distribution: Figures 6 and 7 shows the cumulative degree distributions of all type NEs and only person type NEs of all the networks, which seem to be a two-regime power-law in most of the cases. All the networks show very similar degree distributions. Exceptionally, corpus LAT1994 with all type NEs doesn't follow the two-regime power-law in its cumulative degree distributions. The power-law fits are also shown in the Figures 6 and 7 for better visualization. With all type NEs the power-law exponents, γ_1 varies from -2.35 to -1.86 and γ_2 varies from -3.13 to -2.55. With person NEs γ_1 varies from -2.04 to -1.91 and γ_2 varies from -3.95 to -3.12. These values are much lower than those reported in (Cancho & Solé, 2001) for words, which are 1.5 and 2.7 respectively, indicating a steeper or more skewed degree distribution for NEs. Here also, the two regime power-law distributions show that the existence of a kernel and peripheral lexicons. In natural language, the kernel and peripheral lexicons are defined as domain-independent and domain-specific vocabularies. We observe that for NEs, the kernel units are generic and often domain independent; i.e., these units are present in news articles from almost all or at least a very wide range of domains. In contrast, the peripheral units are present in news articles only from specific domains. In the TTI2004 corpus, examples of kernel units include Mumbai, Delhi, Bengal, London, Pakistan and Bollywood. Examples of peripheral unites in the same corpus include Irugu, Nation Academy Press, ROI and Sachin Varma.

Small World Property: All the networks have high clustering coefficient and very small average shortest path length implying that they are all *small worlds*. While in the context of word collocation networks, (Cancho & Solé, 2001) argues that this property might arise as a part of an optimization in language to facilitate faster access to the mental lexicon, such an argument does not naturally extends to NE. However, the NE networks define certain social relationship between the NEs, especially persons; two person type NEs co-occurring in an article definitely imply that certain news or event connects the two. Therefore, most probably the people know each other personally. Thus, the networks represent "acquaintances" or "social relations" between the NEs. It is a well-known fact that social network of people exhibit the small world property (Watts & Strogatz, 1998). Hence, we believe that our networks reflect the small world nature and other properties of the social networks.

5. Conclusions

In this paper we presented a study of occurrence and co-occurrence patterns of NEs in several large standard English corpora. Our analysis shows that, first, NEs occur in large numbers, indicating that they will play a significant role in the development of any NLP technologies. Second, the unique NEs grow almost linearly, and indefinitely with corpora size, indicating that any hand crafted names lexicon will become obsolete quickly. However, the two-regime pattern of cumulative frequency distribution and the degree distribution suggest that there is a small fraction of kernel NEs that occurs quite frequently, as well as, uniformly distributed in time. Hence, it leaves a possibility of construction of robust lexicons for the kernel, and handling the peripheral NEs based on their co-occurrence with the kernel NEs, for specific technologies. Third, the large fraction of rare NEs suggests that any technology designed to handle NEs should be robust for low frequency NEs. Fourth, the complex network based analysis indicates that the pattern of co-occurrence of NEs is very similar to that of the words and exhibits the small-world phenomenon. We believe that this is a reflection of the underlying social and causal relations between NEs. These networks can potentially be mined for inferring interesting connections between persons (and other types), a potential we plan to explore in our future research. The analysis presented here helps characterize the richness and coverage of any language corpus with respect to any type of NEs, and points to a systematic approach for construction of NE lexicons, especially for many resource poor languages.

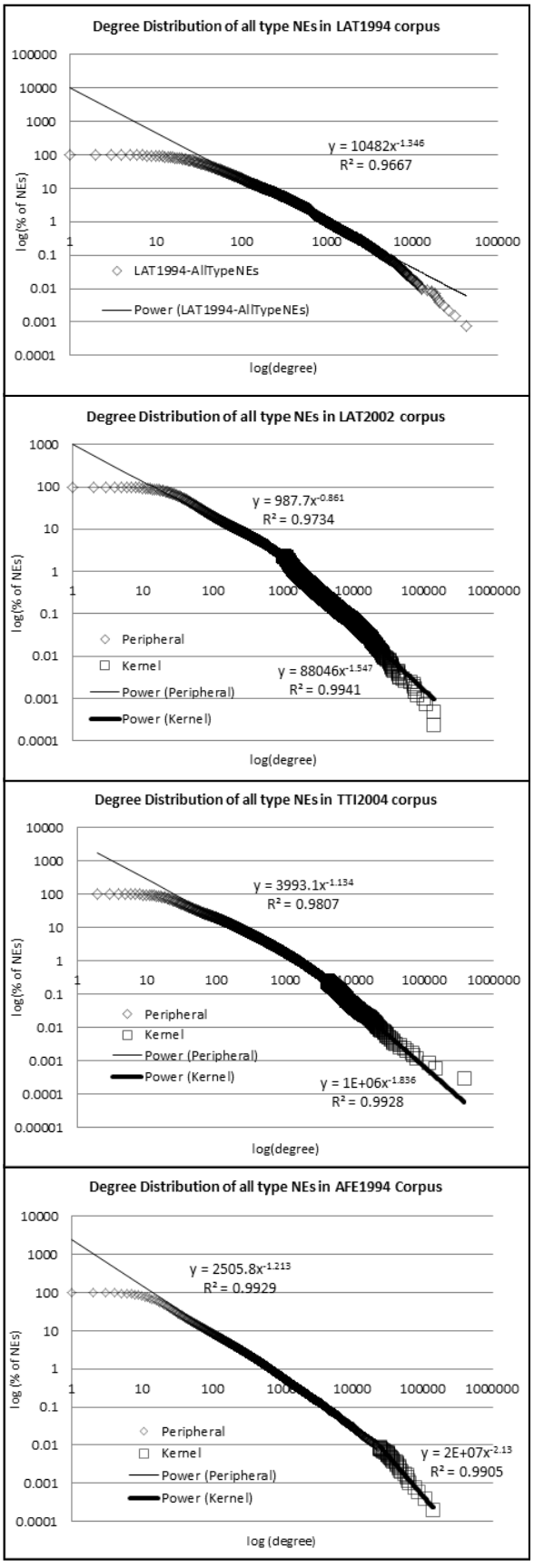

Figure 6: Degree distribution of all type NEs (in each corpus)

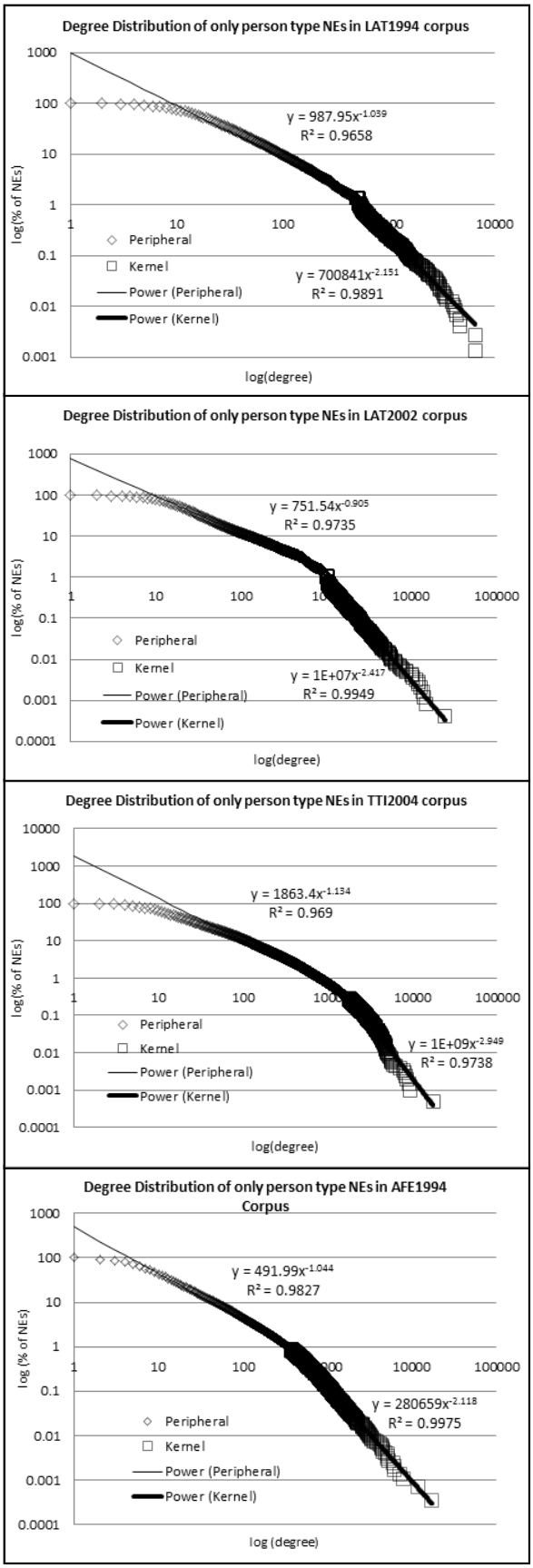

Figure 7: Degree distribution of person type NEs (in each corpus)

6. References

Albert, R., Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics* 74, 47–97.

Cancho, R.F.I., Solé, R.V. (2001). The small world of human language. In *Proceedings of the Royal Society of London* B, 268(1482):2261-2265.

Chen Jinxiu (2007). Automatic relation extraction among named entities from contents. Doctoral Thesis. University of Singapore.

Choudhury, M., Mukherjee, A. (2009). The structure and dynamics of linguistic networks. In *N. Bellomo, N. Ganguly, A. Deutsch, and A. Mukherjee, editors, Dynamics On and Of Complex Networks, Modeling and Simulation in Science, Engineering and Technology*, pages 145–166. Birkhäuser Boston.

David, D. Palmer, David, S. Day (1997). A Statistical Profile of the Named Entity Task. In *Proceedings of Fifth ACL Conference for Applied Natural Language Processing (ANLP-97)*, Washington D.C.

English Giga-word Corpus (LDC 2003T05), Linguistic Data Consortium, http://www.ldc.upenn.edu/.

Faruqui, M., Pado, S. (2010). Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. In *Proceedings of Konvens*, Saarbrucken, Germany.

Finkel, J., Trond Grenager, Christopher Manning (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43nd Annual Meeting of the ACL.*

Grishman Ralph, Sundheim, B. (1996). Message Understanding Conference - 6: A Brief History. *In Proceedings of International Conference on Computational Linguistics.*

Heaps, H.S. (1978). Information Retrieval: Computational and Theoretical Aspects. New York: Academic Press.

Heli Leena Marjukka Teitto (2010). HUMAN REFERENTS IN SUBTITLES - A Study on Personal Pronouns and Proper Nouns in Translated and Original Finnish. Doctoral Thesis. University of Eastern Finland.

Nadeau David, Satoshi Sekine (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes* 30(1):3–26.

Newman, M.E.J. (2001). Scientific collaboration networks. Phys. Rev. E 64

Monojit Choudhury, Diptesh Chatterjee, Animesh Mukherjee (2010). Global topology of word co-occurrence networks: Beyond the two-regime power-law. In *Proceedings of Coling 2010.*

Rishiraj Saha Roy, Niloy Ganguly, Choudhury, M., Mukherjee, A. (2011). Complex Network Analysis Reveals Kernel-Periphery Structure in Web Search Queries. In *Proceedings of SIGIR Workshop on Query Understanding and Representation.*

Sitabhra Sinha, Raj Kumar Pan, Nisha Yadav, Mayank Vahia, Iravatham Mahadevan (2009). Network analysis reveals structure indicative of syntax in the corpus of undeciphered Indus civilization inscriptions. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, Suntec, Singapore.

Stanford Named Entity Recognizer. http://nlp.stanford.edu/software/CRF-NER.shtml

Takaaki Hasegawa, Satoshi Sekine, Ralph Grishman (2004). Discovering relations among named entities from large corpora. In *Proceedings of the 42nd Annu-al Meeting on Association for Computational Linguistics*, p.415-es, Barcelona, Spain

The New Indian Express. http://expressbuzz.com

The Cross-Language Evaluation Forum (CLEF). http://clef-campaign.org

Watts, D.J., Strogatz, S.H., (1998). Collective dynamics of 'small-world' networks. *Nature* 393(6684):440-442.

Corpora	LA Times 1994 (LAT1994) (CLEF 2006[4])		LA Times 2002 (LAT2002) (CLEF 2007[2])		The Telegraph of India (TTI2004) (FIRE 2008)		Giga Word AFE (AFE1994)	
Property	**All NEs**	**Person NEs**	**All NEs**	**Person NEs**	**All NEs**	**Person NEs**	**All NEs**	**Person NEs**
Pub. date	1994		2002		2004-07		1994-97	
No of days covered	64		365		1125		1027	
No. of articles	19,561		87,767		125,526		475,092	
Word count	12M		50M		54M		127M	
No. of unique NEs	130k	74k	413k	251k	379k	210k	491k	298k
No. of all NEs	653k	289k	2973k	1327k	4093k	1228k	9243k	2814k
Avg. no. unique NEs per article	6.65	3.78	4.71	2.86	3.02	1.67	1.03	0.63
Avg. no. NEs per article	33.4	14.77	33.87	15.12	32.61	9.78	19.46	5.92
Avg. no. of unique NEs per day	2032.1	1156.3	1131.9	687.7	336.9	186.7	451.7	274.2
Avg. no. of NEs per day	10207	4515.6	8145.2	3635	3638.8	1091.6	8503.2	2588.8
Average degree	105	48	135	68	111	57	62	24
γ_1	-2.35	-2.04	-1.86	-1.91	-2.13	-2.13	-2.21	-2.04
γ_2	*	-3.15	-2.55	-3.42	-2.84	-3.95	-3.13	-3.12
Average clustering coefficient (CC)	0.781	0.84	0.855	0.823	0.801	0.802	0.863	0.749
CC(rand)	8e-4	6e-4	3e-4	2.7e-4	3e-4	2.7e-4	1.2e-4	8e-05
No. of connected components	139	1067	477	3675	1	3016	542	15286
% of NEs in the largest component	99.77	97.09	99.83	97.4	100	97.33	99.82	91.06
Average shortest path (d)	2.6	3.0	2.7	3.1	2.0	3.2	2.7	3.5
d(rand)	2.5	2.9	2.6	2.9	2.7	3.0	3.2	4.0
max(maxSPL)	6	8	6	9	2	9	6	13
2*min(maxSPL)	8	10	8	10	4	10	8	14

Table 1: Statistics for the news corpora (*only a single regime was observed)

[1] This corpus occurs as a part of CLEF 2006 Corpora, released by (CLEF).

[2] This corpus occurs as a part of CLEF 2007 Corpora, released by (CLEF).