Knowledge-Based Systems 279 (2023) 110942

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Marshall–Olkin power-law distributions in length-frequency of entities

Xiaoshi Zhong [a,b], Xiang Yu [a], Erik Cambria [c,*], Jagath C. Rajapakse [c]

[a] School of Computer Science and Technology, Beijing Institute of Technology, China
[b] State Key Laboratory of Software Development Environment, Beihang University, China
[c] School of Computer Science and Engineering, Nanyang Technological University, Singapore

A R T I C L E I N F O

Article history:
Received 26 April 2023
Received in revised form 9 August 2023
Accepted 26 August 2023
Available online 1 September 2023

Keywords:
Entities
Length-frequency of entities
Power-law distributions
Marshall–Olkin power-law (MOPL) model

A B S T R A C T

Entities involve important concepts with concrete meanings and play important roles in numerous linguistic tasks. Entities have different forms in different linguistic tasks and researchers treat those different forms as different concepts. In this paper, we are curious to know whether there are some common characteristics that connect those different forms of entities. Specifically, we investigate the underlying distributions of entities from different types and different languages, trying to figure out some common characteristics behind those diverse entities. After analyzing twelve datasets about different types of entities and eighteen datasets about entities in different languages, we find that while these entities are dramatically diverse from each other in many aspects, their length-frequencies can be well characterized by a family of Marshall–Olkin power-law (MOPL) distributions. We conduct experiments on those thirty datasets about entities in different types and different languages, and experimental results demonstrate that MOPL models characterize the length-frequencies of entities much better than two state-of-the-art power-law models and an alternative log-normal model. Experimental results also demonstrate that MOPL models are scalable to the length-frequency of entities in large-scale real-world datasets.

1. Introduction

Estoup [1] and Zipf [2,3] found a very long time ago that the rank-frequency of words in natural languages follows a family of power-law distributions. During his exploration, Zipf also found that the meaning-frequency of words follows power-law distributions as well. The rank-frequency distribution of words is later credited as Zipf's law and provides a direction to understand the use of languages in our communicative system. Zipf's law has been observed in many languages [3,4] and has attracted tremendous attention of researchers from diverse areas for more than eighty years [5].

The Zipf distribution has a linear behavior in the log–log scale and is widely used to model phenomena such as word frequencies, city sizes, income distribution, and network structures. However, the Zipf distribution may not fit well the probabilities of the first positive integer numbers, which are often observed to be higher or lower than expected by the linear model. Besides the rank-frequency and meaning-frequency of words, Zipf also analyzed word length, sentence length, and phonemes [3].

Although Zipf explained the use of these three language units under the same principle of least effort as he explained word frequency and word meaning in a qualitative way, unfortunately, extensive studies have demonstrated that the frequencies of these three language units do not follow a power-law distribution, but follow variants of Poisson distributions, lognormal distributions, or gamma distributions [6–15].

In the last two decades, the field of natural language processing and related areas have constructed numerous datasets for diverse linguistic tasks [16–18]. Those datasets provide us opportunities to analyze some other forms of languages, among which entity is an important one. An entity is a real-world object, such as persons, locations, and organizations [19,20]. Entities generally involve important concepts with concrete meanings and usually act as (part of) the subject or the object or even both in a sentence.

For example, in the sentence "Michael Jordan could be an NBA player, or a professor of University of California, Berkeley", the entity "Michael Jordan" acts as the subject while other two entities "NBA" and "University of California, Berkeley" are parts of the object. Because of its importance in language, entities have been extensively studied and are involved in diverse linguistic tasks, such as named entity recognition [19,20] and entity linking [21,22].

* Corresponding author.
E-mail addresses: xszhong@bit.edu.cn (X. Zhong), yuxiang@bit.edu.cn (X. Yu), cambria@ntu.edu.sg (E. Cambria), asjagath@ntu.edu.sg (J.C. Rajapakse).

https://doi.org/10.1016/j.knosys.2023.110942
0950-7051/

Table 1
Some examples of entities in English and their corresponding entity lengths (l). Symbols and punctuations in entities are taken into account during the calculation.

Entity	Entity length (l)
NBA	1
Michael Jordan	2
United Arab Emirates	3
University of California, Berkeley	5
10:00 p.m. on August 20, 1940	7
Human cytomegalovirus (HCMV) major immediate	7

To the best of our knowledge, however, there is no existing literature that investigates the underlying distribution(s) of entities which may provide a better understanding on language use and provide insights into designing effective and efficient algorithms for entity-related linguistic tasks. In this paper, we fill in this gap and conduct a thorough investigation on the length-frequency distributions of entities in different types and different languages. We aim to fit the length-frequency of entities with a uniform model or a family of models. Entity length is defined by the number of words in an entity. Entity length is an important feature of natural language processing that reflects the complexity and structure of texts. Table 1 presents some examples of entities and their corresponding lengths. After a careful exploration, we find that the length-frequency of entities cannot be well characterized by pure power-law models, but can be well characterized by the Marshall–Olkin power-law (MOPL) models that are developed by Pérez-Casany and Casellas [23]. MOPL models are a family of generalized models of power-law models. Compared with pure power-law models, MOFL models have more flexibility to adjust the probabilities of the first few data points while keeping the linearity of the remaining probabilities.

Specifically, we collect twelve datasets about different types of entities (e.g., named entities and time expressions) and eighteen datasets about entities in different languages (e.g., English and French). Those datasets are dramatically diverse from each other in terms of their sources, domains, text genres, generated time, corpus sizes, and entity types, and those languages have significant differences in terms of their phonetic systems and spelling systems (see Section 4.1 for details). However, we find that the length of these diverse entities demonstrates some similar characteristics, and the length-frequency distributions of these diverse entities can be well characterized by a family of MOPL models. To evaluate the quality of MOPL models fitting to the length-frequency of diverse entities, we use the Kolmogorov–Smirnov (KS) test [24,25] and define an average-error metric to evaluate the goodness-of-fit of the MOFL models and compare the fitting results with two state-of-the-art power-law models, namely CSN2009 [26] and LS_{avg} [27], and an alternative log-normal model. We conduct experiments on thirty datasets about entities in different types and different languages, and experimental results demonstrate that MOPL models well characterize the length-frequency distributions of diverse entities, and the fitting results of MOPL are much better than the ones of the three compared models. Specifically, MOPL achieves much better results in the KS test and average-error metric than the three compared models. Experimental results also demonstrate that MOPL models fit the length-frequency of entities in an individual dataset less than one minute, which is comparable with the most efficient model LS_{avg} and much better than the CSN2009 model. This indicates that MOPL models are more suitable to characterize the length-frequency of diverse entities than the three compared models and that MOPL models are scalable to entities in large-scale real-world datasets.[1]

[1] Source codes and data are available at https://github.com/xszhong/MOPL.

To summarize, we mainly make in this paper the following contributions.

- We investigate the underlying distributions of diverse entities, finding that the length-frequency of entities in different types and languages can be characterized by MOPL models. Our finding adds a piece of stable knowledge to the field of language and provides insights for entity-related linguistic tasks.
- We demonstrate the superiority of MOPL models against two state-of-the-art power-law models and a log-normal model in terms of fitting to the length-frequency of diverse entities in different types and languages.
- Experiments demonstrate that MOPL is scalable to large-scale real-world datasets without linearly nor exponentially increasing the runtime when the number of entities increases.

The remaining of this paper is organized as follows. Section 2 reviews the literature about power-law distributions in languages. Section 3 introduces the MOPL models that we use to characterize the length-frequency of divers entities. Section 4 reports experimental results and computational efficiency of MOPL models and compared models fitting to the length-frequency distributions of entities in different types and different languages. Section 5 discusses possible implications and limitations of this paper while Section 6 draws the conclusion.

2. Related works

While power-law distributions have been observed to appear in numerous natural systems and societal systems [26,28], in this paper, we are concerned with power-law distributions in languages. Following we review related works about the power-law distributions in languages and about the length-frequency distributions of words and sentences, and discuss the connection and differences between these related works and our work.

2.1. Power-law distributions in languages

The most famous power-law distribution in languages is the one in the rank-frequency of words. This linguistic phenomenon was originally discovered by Estoup [1] and then further explored by Zipf [2,3]; such linguistic phenomenon is later credited as Zipf's law. Zipf's law reveals that the rth most frequently occurring word in a corpus has the frequency defined by $f(r) \propto r^{-z}$, where r denotes the frequency rank of a word in the corpus and $f(r)$ denotes its frequency. The Zipf's law has been observed in many languages [3–5,29], and the scaling exponent z is observed to be close to 1. During his exploration, Zipf found as well that the meaning-frequency of words in a corpus also follows a family of power-law distributions.

Besides real languages, researchers have also explored randomly generated texts and genetic regulatory networks [30–32]. Miller [33,34] and Li [35] found that the rank-frequency of random texts also follows power-law distributions. Malone and Maher [36] and Wang et al. [37] found that the rank-frequency of user passwords from different websites can be characterized by power-law distributions.

We now discover another form of human languages, namely entities, whose length-frequency distributions can be characterized by the Marshall–Olkin extended power-law distributions. There are significant differences between power-law distributions in the length-frequency of entities and in the rank-frequency of words. Firstly, the meanings and functions of words and of entities in a sentence are different. In the rank-frequency of words, those most frequent words are always auxiliary words without concrete meanings (random texts and user passwords

Table 2
Statistics of datasets about entities in different types. Entity length l is defined by the number of words in an entity.

Dataset	Entity type	Num of entities	Max l	Average l	StdDev. l
ABSA	Aspect terms	9,979	21	1.45	0.89
ACE04	Named entities	29,949	57	2.43	9.29
BBN	Named entities	98,427	15	1.26	0.36
BioMed	Biomedical entities	450,729	86	1.80	4.05
CoNLL03	Named entities	35,087	14	1.45	0.48
COVID19	Pandemic entities	10,260,797	117	1.27	0.63
LitBank	Literary entities	13,912	129	2.93	19.66
OntoNotes5	Named entities	155,413	28	1.85	1.58
Re3d	Defense entities	3,394	20	2.32	3.20
TimeExp	Time expressions	18,484	22	1.80	1.31
Twitter	Informal entities	20,515	14	1.39	0.71
WikiAnchor	Anchor text	2,690,849	49	2.10	3.09

have no concrete meanings as well), while entities generally involve important concepts with concrete meanings and play important roles in a sentence, such as the subject and the object.

Secondly, the numbers of their data points are different. In the rank-frequency of words, an r-rank word appears as a data point, while in the length-frequency of entities, all the l-length entities composite a data point. So the number of data points in the rank-frequency of words is as large as the size of vocabulary in a corpus, while the number of data points in the length-frequency of entities is generally less than 100, and our analysis shows that, in about 93.3% of datasets (28 out of 30), the longest entity contains no more than 100 words (see Tables 2 and 3).

Thirdly, the scaling exponents of these two kinds of power-law distributions are different. The scaling exponents in the rank-frequency of words are observed to approximate to 1, indicating that these power-law distributions do not have theoretical means nor finite variances. By contrast, the exponents in the length-frequency of entities are greater than 2, theoretically indicating well-defined means in all these power-law distributions; and in real-world datasets, these power-law distributions have finite means and variances.

2.2. Length-frequency distributions of words and sentences

A line of researches that is somewhat related to our work is about the length distributions of words and sentences. According to a review article by Grotjahn and Altmann [12] and Fucks [7,8] first theoretically and experimentally demonstrated that the length-frequency of words in a corpus follows a family of Poisson distributions. This linguistic phenomenon has been observed in more than 32 languages [14]. On the other hand, Williams [6] and Wake [9] observed that the length-frequency of sentences in different languages can be characterized by a family of log-normal distributions. Sigurd et al. [15] observed that the length-frequencies of words and sentences from English, Swedish, and German corpora can be characterized by variants of log-normal distributions or gamma distributions.

Unlike the length-frequency of words and sentences that can be characterized by variants of Poisson distributions, log-normal distributions, or gamma distributions, we find from experiments on datasets about entities in different types and different languages that the length-frequency of entities cannot be characterized by Poisson distributions nor log-normal distributions but are well characterized by a family of Marshall–Olkin power-law (MOPL) distributions. Moreover, our extensive experiments demonstrate that MOPL models characterize the length-frequency of entities much better than two state-of-the-art power-law models and one alternative log-normal model and that MOPL models are scalable to the length-frequency of entities in large-scale real-world datasets.

3. Methodology

We first briefly introduce the discrete power-law distributions and then detail the Marshall–Olkin power-law (MOPL) models that we use to characterize the length-frequency distributions of entities in different types and different languages. After that we introduce the Kolmogorov–Smirnov (KS) test [24, 25] and the average-error metric that are used to evaluate the goodness-of-fit.

3.1. Discrete power-law distribution

The discrete power-law distribution is a special case of the power-law distributions with discrete values and is defined by Eq. (1):

$$P(X=x)=\frac{x^{-\alpha}}{\zeta(\alpha)} \tag{1}$$

where $x \in N^{+}$, $\alpha > 0$ is the scaling exponent, and $\zeta(\alpha) = \Sigma_{k=1}^{\infty} k^{-\alpha}$ is the Riemann Zeta function.

Eq. (1) can be written as Eq. (2), which demonstrates the linear behavior in the log–log scale:

$$\log P(X=k) = -\alpha \log x - \log \zeta(\alpha) \tag{2}$$

The survival function (SF) of the power-law distribution is given by Eq. (3):

$$\overline{F}(X) = P(X > x) = \frac{\zeta(\alpha, x+1)}{\zeta(\alpha)} \tag{3}$$

where $\zeta(\alpha, x) = \Sigma_{k=x}^{\infty} k^{-\alpha}$ is the Hurwitz zeta function.

3.2. Marshall–olkin power-law distribution

Pérez-Casany and Casellas [23] explore a new form of power-law distributions by extending the original power-law function through the Marshall–Olkin transformation. They extend the original power-law function to a more general function called Marshall–Olkin power-law distribution. This function have two parameters, α and β, and its survival function (SF) is given as below:

$$P(X > x) = \overline{G}(x; \alpha, \beta) = \frac{\beta \overline{F}(X)}{1 - \overline{\beta} \overline{F}(X)} = \frac{\beta \zeta(\alpha, x+1)}{\zeta(\alpha) - \overline{\beta} \zeta(\alpha+1)} \tag{4}$$

where $\beta > 0$, $\alpha > 1$ and $\overline{\beta} = 1 - \beta$.

The probability mass function (PMF) can be computed through Eq. (5):

$$\begin{aligned} P(X=x) &= \overline{G}(x-1; \alpha, \beta) - \overline{G}(x; \alpha, \beta) \\ &= \frac{x^{-\alpha} \beta \zeta(\alpha)}{[\zeta(\alpha) - \overline{\beta} \zeta(\alpha, x)][\zeta(\alpha) - (\overline{\beta}) \zeta(\alpha, x+1)]} \end{aligned} \tag{5}$$

Table 3
Statistics of entities in different languages.

Language	Entity type	Num of entities	Max l	Average l	StdDev. l
Afrikaans	Named entities	13,947	27	1.86	1.87
Arabic	Named entities	44,284	41	2.15	6.06
Basque	Named entities	4,748	20	1.47	0.62
Bokmal	Named entities	13,950	15	1.10	0.19
Croatian	Named entities	21,105,675	11	1.95	2.37
Czech	Named entities	62,867	9	1.53	0.79
France	Named entities	9,836	17	1.41	0.75
German	Named entities	12,778	34	1.53	0.91
Italian	Named entities	1,071,045	41	2.35	2.37
Netherland	Named entities	7,102	9	1.42	0.99
Nynorsk	Named entities	12,726	10	1.13	0.25
Polish	Named entities	12,038,419	13	1.86	1.16
Romanian	Named entities	153,226	30	1.77	1.94
Russian	Named entities	3,152,930	12	1.70	1.16
Samnorsk	Named entities	29,407	15	1.11	0.22
Slovak	Named entities	136 435	11	1.72	1.44
Slovene	Named entities	13,055,756	8	2.07	2.03
Ukrainian	Named entities	18,347,492	14	2.23	2.31

where $x \in N^+$ and $\zeta(\alpha, x) = \Sigma_{k=x+1}^{\infty} k^{-\alpha}$ stands for the Hurwitz Zeta function.

The Marshall–Olkin power-law (MOPL) distributions are a generalization of power-law distributions and overcome some limitations of pure power-law distributions by introducing a parameter. Such parameter allows for more flexibility in adjusting the probabilities of small values while keeping the linearity in tails. The MOPL models are capable of fitting the concave and convex issues encountered in realistic situations, and have been applied to characterize various data such as music compositions and web page visits [23].

In this paper, we use the MOPL models to characterize the length-frequency distributions of entities in different types and different languages.

3.3. Kolmogorov–Smirnov test

Like many previous researches [26,27,37–41], we employ the Kolmogorov–Smirnov (KS) test [24,25] to examine the goodness-of-fit. The KS statistic (D_n) quantifies the distance between the cumulative distribution function (CDF) of a set of data points ($F_n(l)$) and the CDF of a theoretic distribution ($F(l)$), as defined by Eq. (6):

$$D_n = \sup_l |F_n(l) - F(l)| \tag{6}$$

where $\sup_l$ is the supremum of the set of distances. The KS statistic $D_n \in [0, 1]$ is the maximal distance between the two CDF curves $F_n(l)$ and $F(l)$. The smaller the D_n value is, the better the theoretic distribution fits the data points.

The KS test can also be used to examine whether two underlying distributions are significantly different. In such case, the two-sample KS statistic ($D_{n,m}$) is defined by Eq. (7):

$$D_{n,m} = \sup_l |F_n(l) - F_m(l)| \tag{7}$$

where $F_n(l)$ and $F_m(l)$ are the CDF curves of two sets of data points.

In the KS test, the null hypothesis (H_0) is that the data points are drawn from a theoretic distribution, where the theoretic distribution can be any parametric distribution, such as Zipf distribution, normal distribution, power law distribution, and lognormal distribution; the alternative (H_1) is that the data points are not drawn from the theoretic distribution. A larger p-value suggests that it is safer to draw a conclusion that these data points are not significantly different from the hypothesized distribution. In two-sample KS test, the null hypothesis (H_0') is that the two sets of data points are drawn from the same underlying distribution, while the alternative (H_1') is that they are not from the same distribution. Similarly, a larger p-value suggests that it is safer to draw a conclusion that the two sets of data points are drawn from the same underlying distribution.

3.4. Average error

Besides the KS test, we also define a metric called average error to examine the goodness-of-fit. The average error is defined by Eq. (8):

$$E_{avg} = \frac{1}{N} \sum_{x_i} \frac{|p_N(x_i) - p(x_i)|}{\sqrt{p_N(x_i) \cdot p(x_i)}} \tag{8}$$

where $p_N(x)$ and $p(x)$ are the probability density functions (PDF) of the raw data and the hypothesized data. $N = |\{(x_i, p_N(x_i)\}|$ stands for the number of data points. Defining the average-error metric by Eq. (8) is to remove the impact of different sample sizes. For different models fitting to the same dataset, the smaller the model achieves the E_{avg}, the better the model fits the dataset.

4. Experiments

We fit Marshall–Olkin power-law (MOPL) models to twelve datasets about different types of entities and eighteen datasets about entities in different languages and compare the fitting results of MOPL with two state-of-the-art models, namely CSN2009 [26] and LS_{avg} [27], and an alternative log-normal model.

4.1. Datasets

The datasets we use in this paper mainly involve two kinds: (1) entities in different types and (2) entities in different languages. Most of these datasets contain annotated entities while some contain automatically annotated entities. We collect from both their training and test sets of these datasets for their entities.

4.1.1. Entities in different types

This kind of datasets contains twelve datasets regarding different types of entities collected from dramatically diverse sources, including general named entities [19,20,42], entity mentions [43,44], time expressions [45–47], aspect terms [48,49], literary entities [50], defense entities, informal entities [51,52], and domain-specific entities [53,54] that are well studied in the field of natural language processing and related areas. In this paper, we use the term of "*entity*" to broadly represent these diverse concepts, and these specific concepts are treated as *different types of entities*.

In a specific type of entities, researchers may also assign some pre-defined labels (e.g., PERSON, LOCATION, and ORGANIZATION) to these entities. We use "different types of entities" or "entity types" to represent the above general named entities, time expressions, aspect terms, etc., while use "different categories of entities" or "entity categories" to represent these pre-defined labels. In our analysis, we are concerned with "different types of entities" and do not care much about "different categories of entities". Because each type of entities may also contain different categories/labels and can reveal general habits of our humans in using language, while a certain category of entities reveal only our specific/narrow habit(s). In this paper, we care more about those general habits and principles than specific/narrow one(s). Since English is the most studied language in natural language processing and related areas, we analyze these different types of entities in English.

The twelve datasets are (1) ABSA [49,55], (2) ACE04 [56], (3) BBN [57], (4) BioMed [58], (5) CoNLL03 [20], (6) COVID19 [59], (7) LitBank [50], (8) OntoNotes5 [44], (9) Re3d, (10) TimeExp [46, 60–63], (11) Twitter [52,64], (12) WikiAnchor [43]. They are briefly described below in alphabetical order.

ABSA contains two corpora that are used in SemEval-2014 [49] and SemEval-2015 [55] for aspect-based sentiment analysis. While the two corpora have several language units for different tasks, we are concerned with aspect terms and collect these aspect terms for the analysis of their length-frequency distribution.

ACE04 is a benchmark dataset used for the 2004 Automatic Content Extraction (ACE) technology evaluation [56]. It consists of various types of data collected from different sources (e.g., newswire and broadcast news) for the analysis of entities and relations in three languages: Arabic, Chinese, and English. We use its English entities for the analysis of different types of entities, while use its Arabic entities for the analysis of entities in different languages.

BBN consists of Wall Street Journal articles for pronoun coreference and entity analysis [57]. It includes 28 entity categories in total. We collect all of its entities for analysis, without considering its entity categories.

BioMed contains fourteen corpora that are developed for the analysis of biomedical entities. Crichton et al. [58] collect the fourteen corpora and we can get these corpora from their paper for the biomedical entities.

CoNLL03 is a benchmark dataset with 1393 news articles derived from the Reuters RCV1 Corpus, which is collected between the period of August 1996 and August 1997 [20]. We collect its entities without entity categories for the analysis of the length-frequency distribution.

COVID19 is a newly constructed dataset for the analysis of entities related to the recent COVID-19 pandemic [59]. We collect and analyze its entities for the length-frequency analysis.

LitBank is a dataset collected from 100 different English-language literary articles across over a long period of time and it is developed for the analysis of literary entities [50].

OntoNotes5 is a large-scale dataset collected from different sources (e.g., news articles, newswire and web data) over a long period of time for the comprehensive analyses of syntax, coreference, proposition, word sense, and named entities in three languages (i.e., English, Chinese, and Arabic) [44]. In this paper we are concerned with its entities in English for analysis.

Re3d[2] is a dataset with various documents relevant to the conflict in Syria and Iraq. The dataset is constructed for the analysis of entity and relation extraction in the domain of defense and security. We collect its entities for analysis.

TimeExp consists of three corpora that are developed for the analysis of time expressions [62,63,65]. These corpora include TempEval-3 (including TimeBank [46], TE3-Silver, AQUAINT, and the Platinum corpus) [61], WikiWars [60], and Tweets [62].

Twitter consists of two corpora whose text is collected from Twitter: WNUT16 [64] and Broad Twitter Corpus [52]. These two corpora are developed for the analysis of entities in informal text.

WikiAnchor treats the anchor text (i.e., the text in the hyperlinks) from Wikipedia (the 20 110 513 version) as entity mentions [43]. We collect these entity mentions (i.e., anchor text) for length-frequency analysis.

For each of these datasets that contain two or more corpora (i.e., ABSA, BioMed, TimeExp, and Twitter), we simply merge all the entities from the whole corpora. Note again that we collect from these datasets only their entities for the analysis of length-frequency distribution; we do not care about their entity categories (or pre-defined labels).

Table 2 reports the entity types and statistics of the twelve datasets. As mentioned in Section 3.2, the entity length l is defined by the number of words in an entity. Table 2 shows that the numbers of entities in the twelve datasets are diverse dramatically, ranging from 3394 (Re3d) to 10,260,797 (COVID19); and the maximal lengths and standard deviations of these entities are also diverse: the maximal lengths are varied from 14 to 129 and the standard deviations are varied from 0.36 to 19.66, respectively. However, the average lengths of these entities are comparable and range around 2 (only from 1.26 to 2.93). This indicates that the average length is a common characteristic among these diverse entities.

4.1.2. Entities in different languages

This kind of datasets contains named entities in eighteen different languages. These datasets are collected from 2004 Automatic Content Extraction (ACE) evaluation [56], European Newspapers,[3] NCHLT Afrikaans Named Entity Annotated Corpus,[4] Basque EIEC (version 1.0),[5] BSNLP 2017,[6] Italian KIND [66], Norwegian Navnkjenner [67], and RONEC [68].

The eighteen languages include (1) Afrikaans, (2) Arabic, (3) Basque, (4) Bokmal, (5) Croatian, (6) Czech, (7) France, (8) German, (9) Italian, (10) Netherland, (11) Nynorsk, (12) Polish, (13) Romanian, (14) Russian, (15) Samnorsk, (16) Slovak, (17) Slovene, and (18) Ukrainian. We do not include English in this kind of datasets because different types of entities are analyzed in English. Table 3 summarizes the statistics of entities in the eighteen languages. It shows that the numbers of these entities are significantly diverse, ranging from 4748 (Basque) to 21,105,675 (Croatian). The maximal lengths and standard deviations of these entities in different languages are somewhat diverse but not that dramatical; while the average lengths of these entities are comparable, ranging around 2 (specifically, from 1.10 to 2.35). These statistics are consistent with corresponding ones of different types of entities reported in Table 2. This indicates that entities across different types and different languages share some similar characteristics.

4.2. Compared methods

We evaluate the quality of MOPL models in fitting the length-frequency distributions of entities against two state-of-the-art models, namely CSN2009 [26] and LS_{avg} [27], and an alternative log-normal model.

CSN2009: Clauset et al. [26] propose a maximum-likelihood fitting method, which is denoted by CSN2009, that combines with

2 https://github.com/dstl/re3d.

3 https://github.com/EuropeanaNewspapers/ner-corpora.

4 https://repo.sadilar.org/handle/20.500.12185/299.

5 http://www.ixa.eus/node/4486?language=en.

6 http://bsnlp-2017.cs.helsinki.fi/shared_task.html.

Table 4
Fitting results of MOPL and compared models fitting to the length-frequency distributions of entities in different types. C indicates the coverage which is defined by the percentage of data covered by a model. M_{log} denotes logarithmic mean while V_{log} denotes logarithmic variance.

Dataset	MOPL			LS_{avg}		CSN2009			LogNormal		
	$\hat{\alpha}$	$\hat{\beta}$	C (%)	$\hat{\alpha}$	C (%)	$\hat{\alpha}$	$\hat{x}_{min}$	C (%)	M_{log}	V_{log}	C (%)
ABSA	4.07	5.44	99.82	2.34	99.95	3.68	2	28.79	0.26	0.19	100.00
ACE04	2.69	2.50	99.54	1.61	99.97	2.73	4	15.38	0.55	0.51	100.00
BBN	4.74	5.43	99.97	3.03	100.00	6.77	4	1.23	0.16	0.11	100.00
BioMed	2.84	2.17	99.92	2.02	99.99	3.36	4	8.53	0.36	0.33	100.00
CoNLL03	5.83	29.48	99.97	2.51	100.00	5.09	2	36.78	0.28	0.15	100.00
COVID19	3.68	1.94	99.00	2.42	99.99	4.96	4	2.10	0.15	0.13	100.00
LitBank	3.44	14.98	99.47	2.94	99.68	2.61	2	70.99	0.62	0.41	100.00
OntoNotes5	3.71	3.12	99.90	0.73	99.99	5.31	5	1.28	0.22	0.17	100.00
Re3d	3.26	8.79	98.70	1.12	99.82	4.67	6	5.10	0.69	0.55	100.00
TimeExp	4.19	14.15	99.91	1.46	100.00	5.34	4	8.09	0.45	0.26	100.00
Twitter	4.20	5.21	99.91	2.54	99.99	3.86	2	26.19	0.23	0.16	100.00
WikiAnchor	4.21	23.02	100.00	2.55	100.00	3.81	3	24.69	0.58	0.30	100.00

Table 5
Goodness-of-fit testing results of MOPL and compared models fitting to the length-frequency distributions of entities in different types. D_n indicates the KS statistic defined by Eq. (6). E_{avg} indicates the average error defined by Eq. (8). *DEC* indicates the decision to accept or reject the hypothesis H_0 that a model well fits the data, based on the p-value of the KS test. For each of D_n and E_{avg}, the best result on each dataset is highlighted in bold.

Dataset	MOPL			LS_{avg}			CSN2009			LogNormal		
	D_n	E_{avg}	*DEC*	D_n	E_{avg}	*DEC*	D_n	E_{avg}	*DEC*	D_n	E_{avg}	*DEC*
ABSA	**1.67E−03**	**0.18**	**Accept**	4.17E−01	1.48	Reject	2.63E−02	0.35	Reject	3.97E−02	1.28	Reject
ACE04	**6.15E−03**	**0.18**	**Accept**	5.28E−01	1.60	Reject	4.29E−02	0.32	Reject	1.21E−01	1.51	Reject
BBN	**6.51E−04**	0.43	**Accept**	2.73E−01	1.88	Reject	1.24E−02	**0.25**	**Accept**	5.69E−02	4.61	Reject
BioMed	**1.58E−03**	0.62	**Accept**	6.27E−01	2.61	Reject	9.71E−03	**0.34**	Reject	1.15E−01	3.28	Reject
CoNLL03	**3.36E−04**	**0.32**	**Accept**	3.33E−01	2.34	Reject	4.46E−03	0.36	**Accept**	6.68E−02	1.11	Reject
COVID19	**7.88E−05**	1.40	**Accept**	6.25E−01	3.96	Reject	8.69E−03	**0.66**	Reject	4.97E−02	11.27	Reject
LitBank	**1.73E−03**	0.87	**Accept**	8.00E−01	3.39	Reject	2.00E−02	**0.32**	Reject	6.50E−02	0.92	Reject
OntoNotes5	**2.04E−03**	0.51	**Accept**	3.85E−01	1.60	Reject	1.83E−02	**0.30**	**Accept**	5.40E−02	2.66	Reject
Re3d	**1.22E−02**	**0.28**	**Accept**	4.62E−01	1.53	Reject	6.02E−02	0.39	**Accept**	5.64E−02	0.36	Reject
TimeExp	**1.22E−03**	0.37	**Accept**	5.88E−01	4.57	Reject	1.00E−02	**0.36**	**Accept**	3.14E−02	0.72	Reject
Twitter	**1.24E−03**	**0.21**	**Accept**	3.33E−01	1.22	Reject	1.92E−02	0.36	Reject	4.02E−02	2.21	Reject
WikiAnchor	**1.63E−04**	0.92	**Accept**	2.92E−01	1.12	Reject	1.20E−02	**0.59**	Reject	1.76E−02	4.46	Reject

goodness-of-fit tests based on the Kolmogorov–Smirnov statistic to fit power-law distributions to empirical data. CSN2009 estimates the exponent of a power-law model and the minimal value from which the power-law distribution starts. Besides data fitting, CSN2009 also adopts the KS test with likelihood ratios to evaluate the goodness-of-fit of how well a model fits to data. CSN2009 has been the most popular method in the last decade in fitting power-law distributions.

LS_{avg}: Zhong et al. [27] demonstrate through extensive experiments that least-squares methods can accurately fit to power-law distributions. They propose a least-squares method to fit power-law distributions to empirical data and use an average strategy to reduce the impact of noisy data that deviate from the fitted line.

LogNormal: Log-normal distributions are alternative distributions that researchers usually use to fit data when considering power-law distributions. Therefore, besides CSN2009 and LS_{avg}, we also compare MOPL models with the log-normal model in terms of fitting the length-frequency of entities.

4.3. Implementation details

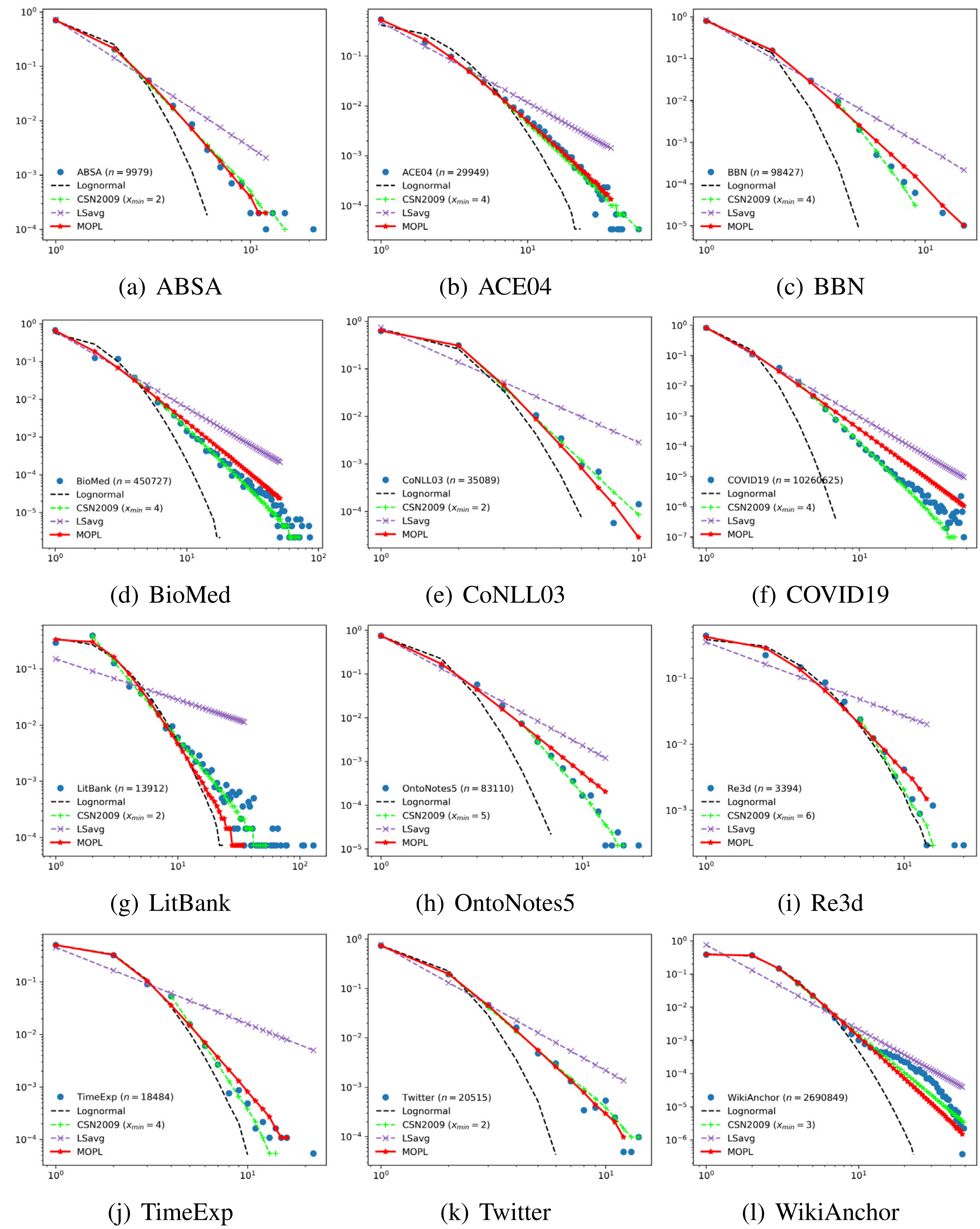
For the experiments of data fitting, we use the zipfextR package [23] in the R programming language to implement our method and apply the codes of CSN2009[7] and LS_{avg}[8] to the datasets. For the KS test, we use the *dgof*[9] [69] and *KSgeneral*[10] [70] packages in the R programming language for MOPL, LS_{avg}, and the log-normal model, while use CSN2009's KS-test module for CSN2009. In experiments, we find that for the same model on the same dataset, *dgof* and *KSgeneral* achieve the same D_n value (i.e., the KS statistic) but different p-values. This suggests that the D_n values are accurate while the p-values may not be accurate. In this paper, we use the *dgof* package to report the D_n values and make the final Accept/Reject decisions. All our experiments are conducted on a Dell PowerEdge R740 server with a 96-CPUs processor, 256 GB memory, and the CentOS-7 system.

4.4. Experimental results

Tables 4 and 5 report the fitting and goodness-of-fit testing results of MOPL and the three compared models on the length-frequency distributions of entities in different types. Specifically, Table 4 reports the estimated parameters of the models and the coverages (i.e., percentages of data that models cover) while Table 5 reports the goodness-of-fit testing results of the models on the datasets, including D_n, E_{avg}, and *DEC* where *DEC* indicates the decision to accept or reject the hypothesis H_0. Fig. 1 visualizes the results of MOPL and the three compared models fitting to the length-frequency distributions of entities in different types. Table 6 reports the fitting results while Table 7 reports the goodness-of-fit testing results of MOPL and the three compared models fitting to the length-frequency of entities in different languages. Figs. 2 and 3 visualize those fittings to the length-frequency of entities in different languages.

What follows are separate discussions on model fitting and testing results on the length-frequency of entities in different types and different languages.

7 https://aaronclauset.github.io/powerlaws/.
8 https://github.com/xszhong/LSavg.
9 https://cran.r-project.org/web/packages/dgof/index.html.
10 https://github.com/raymondtsr/ksgeneral.



(a) ABSA (b) ACE04 (c) BBN

(d) BioMed (e) CoNLL03 (f) COVID19

(g) LitBank (h) OntoNotes5 (i) Re3d

(j) TimeExp (k) Twitter (l) WikiAnchor

Fig. 1. Plots of MOPL and the three compared models fitting to the length-frequency distributions of entities in different types in the twelve datasets. The horizontal axis indicates the entity length (l) while the vertical axis indicates the percentage ($p(l)$).

4.4.1. Results on the length-frequency of entities in different types

Let us first look at the three measures that examine the goodness-of-fit in Table 5: D_n, E_{avg}, and DEC. Table 5 shows that MOPL achieves the best results in all the three measures on all the twelve datasets, in comparison with the three compared models. Specifically, MOPL achieves the performance of D_n in the range from 7.88E−05 to 1.22E−02 and the E_{avg} value from 0.18 to 1.40 as well as all the "**Accept**" across the twelve datasets. By contrast, LS_{avg} achieves the performance of D_n in the range from 2.73E−01 to 8.00E−01 and the E_{avg} value from 1.12 to 4.57 as well as all the "Reject" across the datasets. The three measures that CSN2009 achieves are 4.46E−03∼6.02E−02 for D_n, 0.25∼0.66 for E_{avg}, and 5 "**Accept**" and 7 "Reject" for DEC.

The three measures of LogNormal are 1.76E−02∼1.21E−01 for D_n, 0.36∼11.27 for E_{avg}, and all 12 "Reject" for DEC. This indicates that MOPL fits the length-frequency distributions of entities in different types much better than LS_{avg} and CSN2009, which are developed to fit power-law distributions, and LogNormal, which is often used as an alternative model for power-law models to fit empirical data.

Fig. 1 intuitively visualizes the difference between MOPL and the three compared models in fitting the length-frequency distributions of entities on the twelve datasets. From Fig. 1 we can see that the fittings of MOPL are much better than the ones of the three compared models.

Table 6
Results of MOPL and compared models fitting to the length-frequency distributions of entities in different languages. C indicates the coverage which is defined by the percentage of data covered by a model. M_{log} denotes logarithmic mean while V_{log} denotes logarithmic variance.

Dataset	MOPL			LS_{avg}		CSN2009			LogNormal		
	$\hat{\alpha}$	$\hat{\beta}$	C (%)	$\hat{\alpha}$	C (%)	$\hat{\alpha}$	$\hat{x}_{min}$	C (%)	M_{log}	V_{log}	C (%)
Afrikaans	3.42	6.01	99.63	1.59	99.99	4.90	5	4.92	0.44	0.31	100.00
Arabic	2.66	3.02	99.57	2.25	99.96	4.72	14	0.80	0.47	0.45	100.00
Basque	4.91	13.74	99.77	4.25	99.96	5.60	3	8.34	0.29	0.17	100.00
Bokmal	4.69	1.66	99.71	1.58	99.99	4.12	1	99.71	0.06	0.05	100.00
Croatian	3.67	8.78	99.40	2.37	100.00	3.12	2	49.58	0.48	0.32	100.00
Czech	5.08	18.68	99.70	1.98	100.00	4.41	2	39.92	0.32	0.18	100.00
France	3.83	3.73	99.69	2.12	99.95	5.30	4	3.29	0.23	0.18	100.00
German	4.74	13.38	99.82	1.09	99.91	4.53	3	9.38	0.31	0.19	100.00
Italian	3.91	23.10	99.95	0.71	100.00	7.35	9	0.60	0.68	0.33	100.00
Netherland	3.06	1.49	99.34	3.89	100.00	2.74	1	98.47	0.22	0.20	100.00
Nynorsk	4.49	1.95	99.94	1.30	100.00	3.77	1	88.37	0.08	0.06	100.00
Polish	4.79	29.87	99.79	1.82	100.00	3.76	2	56.15	0.49	0.23	100.00
Romanian	3.21	3.81	99.80	2.14	100.00	5.94	8	0.85	0.39	0.30	100.00
Russian	5.12	28.91	99.62	4.06	100.00	4.19	2	49.85	0.41	0.21	100.00
Samnorsk	4.53	1.70	99.98	2.25	100.00	3.95	1	99.63	0.07	0.05	100.00
Slovak	4.24	12.01	99.77	1.24	100.00	3.62	2	45.30	0.40	0.25	100.00
Slovene	3.68	11.37	98.77	0.86	100.00	4.38	4	13.11	0.54	0.33	100.00
Ukrainian	3.98	21.16	99.47	1.83	100.00	4.77	5	7.60	0.63	0.32	100.00

Table 7
Goodness-of-fit testing results of MOPL and compared models fitting to the length-frequency distributions of entities in different languages. D_n indicates the KS statistic defined by Eq. (6). E_{avg} indicates the average error defined by Eq. (8). DEC indicates the decision to accept or reject the hypothesis H_0 that a model well fits the data, based on the p-value of the KS test. For each of D_n and E_{avg}, the best result on each dataset is highlighted in bold.

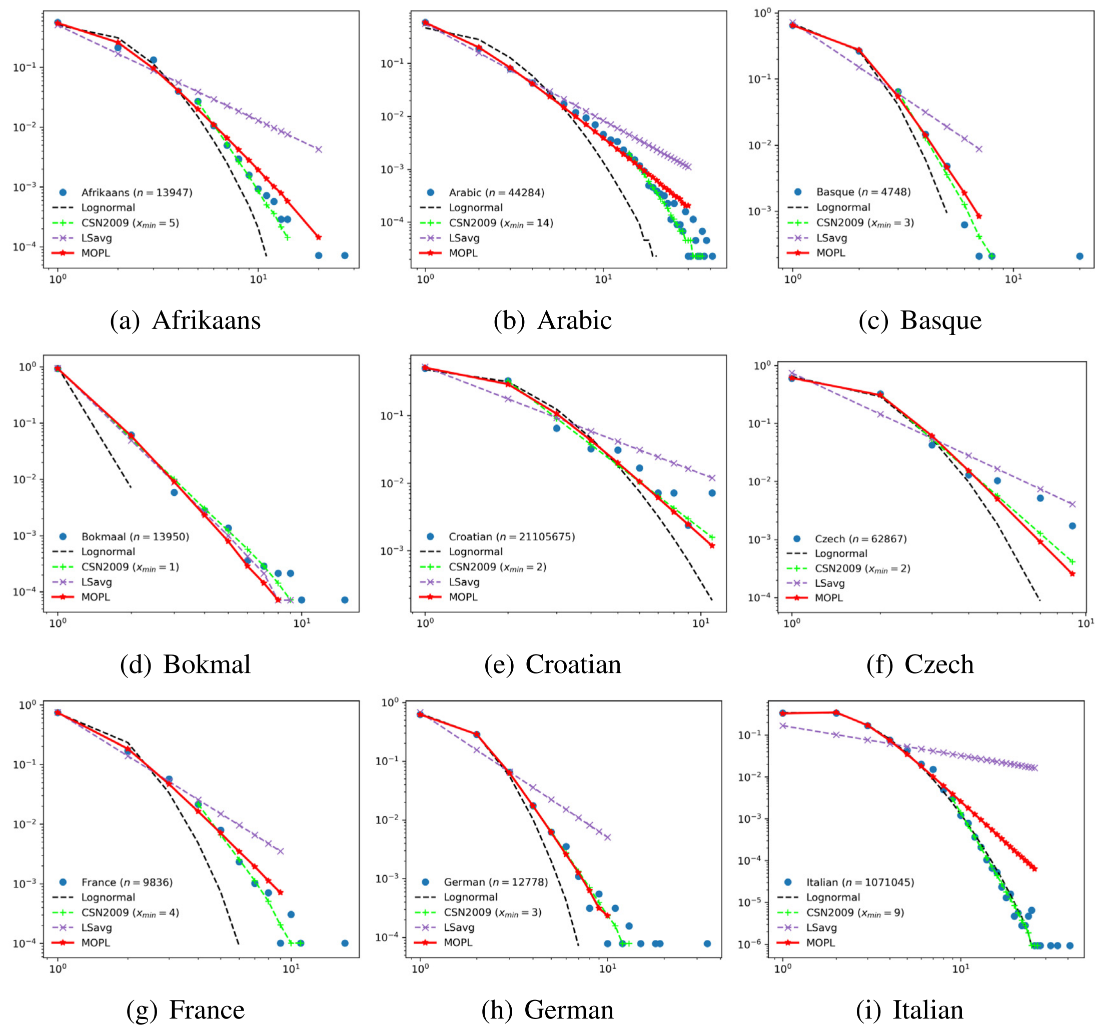
Dataset	MOPL			LS_{avg}			CSN2009			LogNormal		
	D_n	E_{avg}	DEC	D_n	E_{avg}	DEC	D_n	E_{avg}	DEC	D_n	E_{avg}	DEC
Afrikaans	**1.72E−03**	0.42	**Accept**	4.67E−01	2.16	Reject	2.24E−02	**0.22**	**Accept**	6.53E−02	0.86	Reject
Arabic	**6.07E−03**	**0.37**	**Accept**	4.33E−01	1.41	Reject	5.66E−02	0.39	**Accept**	1.24E−01	1.80	Reject
Basque	1.50E−02	**0.24**	**Accept**	2.86E−01	1.21	Reject	**7.06E−03**	0.31	**Accept**	8.63E−02	0.65	Reject
Bokmal	**1.34E−02**	0.41	Reject	2.00E−01	0.43	Reject	5.41E−02	**0.32**	Reject	4.69E−02	1.34	Reject
Croatian	**1.53E−02**	0.30	Reject	3.00E−01	0.80	Reject	2.08E−02	**0.29**	Reject	5.88E−02	0.70	Reject
Czech	**4.01E−02**	0.55	Reject	1.43E−01	**0.49**	Reject	5.69E−02	1.89	Reject	4.60E−02	1.70	Reject
France	**2.13E−03**	**0.27**	**Accept**	3.33E−01	0.87	Reject	4.92E−03	0.51	**Accept**	4.49E−02	1.73	Reject
German	**2.42E−03**	**0.20**	**Accept**	4.00E−01	1.69	Reject	2.18E−02	0.32	**Accept**	6.73E−02	1.16	Reject
Italian	**1.16E−02**	2.18	Reject	7.69E−01	23.99	Reject	3.47E−02	0.38	Reject	6.89E−02	**0.34**	Reject
Netherland	**8.98E−03**	0.32	**Accept**	2.22E−01	0.34	Reject	1.67E−02	**0.29**	Reject	7.06E−02	1.86	Reject
Nynorsk	**8.90E−03**	0.50	**Accept**	2.00E−01	**0.33**	Reject	2.17E−02	0.34	Reject	4.03E−02	4.81	Reject
Polish	2.04E−02	2.47	Reject	3.33E−01	8.78	Reject	5.21E−03	**0.35**	Reject	4.00E−02	2.12	Reject
Romanian	2.74E−02	**1.18**	Reject	5.45E−01	4.31	Reject	7.06E−03	3.18	**Accept**	3.72E−02	1.77	Reject
Russian	**5.51E−03**	0.49	Reject	1.25E−01	0.71	Reject	1.77E−02	**0.30**	Reject	4.03E−02	1.17	Reject
Samnorsk	**2.08E−03**	0.57	**Accept**	1.82E−01	0.36	Reject	1.52E−02	**0.25**	Reject	2.47E−02	6.81	Reject
Slovak	**9.13E−03**	0.40	Reject	1.00E−01	0.45	Reject	2.49E−02	**0.28**	Reject	5.55E−02	1.60	Reject
Slovene	3.63E−02	**0.24**	Reject	3.75E−01	0.56	Reject	8.79E−03	0.25	Reject	1.70E−02	0.37	Reject
Ukrainian	**2.26E−02**	0.17	Reject	4.55E−01	1.61	Reject	3.06E−02	**0.15**	Reject	7.39E−02	0.34	Reject

More importantly, MOPL achieving all the "**Accept**" on the twelve datasets indicates that MOPL is a suitable model to characterize the length-frequency of entities in different types. The fact that MOPL achieves the best goodness-of-fit testing results indicates that MOPL achieves the best estimated parameters. As shown in Table 4, therefore, the $\hat{\alpha}$ of MOPL should be considered as the relatively accurate estimated exponents fitting to the power-law segments of the length-frequency distributions of entities in different types. All the $\hat{\alpha}$ of MOPL fitting to these different types of entities range from 2.69 to 5.83, and most of these $\hat{\alpha}$ range from 2.69 to 4.74. This indicates that the length-frequency of entities in different types have stable scaling property.

Let us now look at the fittings of the two state-of-the-art compared models, LS_{avg} and CSN2009. The $\hat{\alpha}$ of LS_{avg} are deviated relatively far away from the $\hat{\alpha}$ of MOPL. The reason is that LS_{avg} assumes that a power-law starts from the very beginning of an empirical dataset, but Fig. 1 shows that such assumption is not applicable to the length-frequency of entities. This indicates that a pure power-law model is unsuitable to characterize the length-frequency of entities in different types. On the other hand, the $\hat{\alpha}$ of CSN2009 are deviated slightly from the $\hat{\alpha}$ of MOPL. The reason is that CSN2009 adopts a minimum-KS-statistic strategy to choose larger lower bound (i.e., $\hat{x}_{min}$) and fits only the long tails. Consequently, CSN2009 discards the majority of data and achieves low coverages, which are only from 1.23% to 70.99%. By contrast, other models cover more than 98.70% of data. This result that CSN2009 achieves low coverage in fitting to empirical data is consistent with the observation reported in Zhong et al. [27].

4.4.2. Results on the length-frequency of entities in different languages

Let us first look at the three goodness-of-fit testing measures in Table 7 as well: D_n, E_{avg}, and DEC. Table 7 shows that none of the four models (i.e., MOPL, LS_{avg}, CSN2009, and LogNormal) can perfectly characterize the length-frequency distributions of entities in the eighteen languages. The fittings to the length-frequency of entities in different languages are much worse than the fittings to the length-frequency of entities in different types. A possible reason is that some of these datasets in the non-English languages contain a large number of noises. As we mentioned above, English is the most studied language in the field of natural language processing and related areas; other languages are also studied, but their annotated datasets may not be as accurate as the datasets in English.



(a) Afrikaans (b) Arabic (c) Basque

(d) Bokmal (e) Croatian (f) Czech

(g) France (h) German (i) Italian

Fig. 2. Plots of MOPL and the three compared models fitting to the length-frequency distributions of entities in different languages in the first nine datasets. The horizontal axis indicates the entity length (l) while the vertical axis indicates the percentage ($p(l)$).

Another possible reason is that none of our authors are familiar with those languages and cannot guarantee the accuracy of the annotations for these datasets. Let us now look at the comparison among the four models fitting to the length-frequency of entities. While MOPL does not well characterize the length-frequency distributions of entities in all the eighteen languages, MOPL outperforms the three compared models.

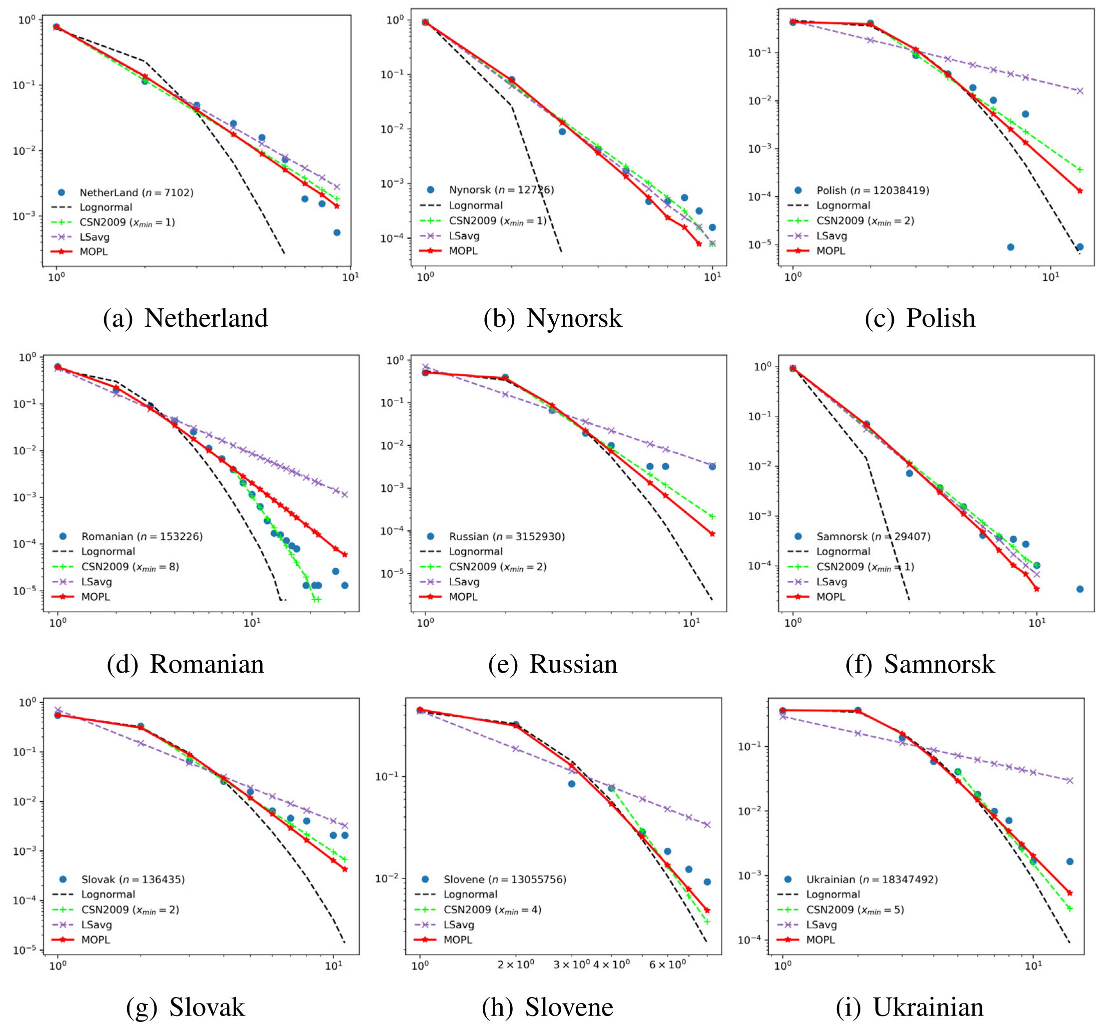
Specifically, MOPL achieves the D_n value in the range from 1.72E−03 to 4.01E−02, achieves the E_{avg} value in the range from 0.17 to 2.47, and achieves 8 "**Accept**" and 10 "Reject" for *DEC* across all the eighteen languages. By contrast, LS_{avg} achieves the D_n value from 1.00E−01 to 7.69E−01, achieves the E_{avg} value from 0.33 to 23.99, and achieves all 18 "Reject" for *DEC* across the eighteen languages. CSN2009 achieves the D_n value from 4.92E−03 to 5.69E−02, achieves the E_{avg} value from 0.15 to 3.18, and achieves 6 "**Accept**" and 12 "Reject" for *DEC*.

LogNormal achieves the D_n value from 1.70E−02 to 1.24E−01, achieves the E_{avg} value from 0.34 to 6.81, and achieves all 18 "Reject" for *DEC*. The comparison among the four models fitting to the length-frequency of entities is intuitively visualized in Figs. 2 and 3. The fitting and testing results indicate that MOPL is more suitable to characterize the length-frequency distributions of entities in different languages than LS_{avg}, CSN2009, and LogNormal. Table 6 shows that the $\hat{\alpha}$ of MOPL fitting to the length-frequency distributions of entities in different languages range only from 2.66 to 5.12, which is consistent with the $\hat{\alpha}$ of MOPL fitting to different types of entities, as shown in Table 4. This indicates that the length-frequency distributions of entities in different languages also have stable scaling property. In terms of data coverage, MOPL, LS_{avg}, and LogNormal cover almost all the data (i.e., from 99.91% to 100%), while CSN2009 achieves relatively low coverages (i.e., lower to 0.60%). Specifically, CSN2009 discards at least 50% of data in 13 out of 18 languages, and discards at least 90% of data in 8 out of 18 languages. The low coverage of CSN2009 on the length-frequency of entities in different languages is consistent with the one of CSN2009 on the length-frequency of entities in different types reported in Table 4 as well as the observation reported in [27].

4.5. Computational efficiency

Table 8 reports the runtimes of MOPL, LS_{avg}, CSN2009 and LogNormal fitting to the length-frequency distributions of entities in different types and different languages.[11] Table 8 shows that while the runtimes of MOPL fitting to length-frequency of entities in both different types and different languages are less efficient than ones of LS_{avg} and LogNormal, they are significantly more

11 Note that the reported runtimes only include the time of the four models fitting to the length-frequency distributions; they do not include the time of the KS testing.



(a) Netherland (b) Nynorsk (c) Polish (d) Romanian (e) Russian (f) Samnorsk (g) Slovak (h) Slovene (i) Ukrainian

Fig. 3. Plots of MOPL and the three compared models fitting to the length-frequency distributions of entities in different languages in the remaining nine datasets. The horizontal axis indicates the entity length (l) while the vertical axis indicates the percentage ($p(l)$).

efficient than the ones of CSN2009. Moreover, while the number of entities in individual dataset ranges from 3394 to 10,260,797 in different types (see Table 2) and from 4748 to 21,105,675 in different languages (see Table 3), the runtime of MOPL performing on individual dataset ranges only from 41.71 to 409.67 ms, all of which are less than one second. That means the runtime of MOPL neither increases linearly nor exponentially as the number of entities increases. This suggests that MOPL can be easily applied on large-scale datasets with high efficiency.

5. Discussion

5.1. Some implications on entity-related linguistic tasks

We here briefly discuss some implications of this linguistic phenomenon (i.e., the length-frequency of entities in different types and different languages can be characterized by Marshall–Olkin power-law distributions) on entity-related linguistic tasks. This linguistic phenomenon may be able to explain why many statistical models and deep-learning models, such as conditional random fields [71], long short-term memory networks [72], and transformer [73], can be applied for recognizing all these different types of entities from unstructured text [20,48,49,51–54,74–78]. This linguistic phenomenon may also be able to provide insights into analyzing those languages with low-resources.

Since entities in different types and different languages share many common characteristics (e.g., their length-frequency distributions, average lengths, and scaling property), we could transfer knowledge and resource available in those well-studied languages to those low-resource languages. We could also apply those statistical modes and deep-learning models that have demonstrated to be effective and efficient in well-studied languages to those low-resource languages. Distilling this knowledge about the length-frequency distributions of entities can also drive us to design effective and efficient algorithms for specific linguistic tasks. For example, Zhong et al. [62] found that an average time expression contains only about two words of which one is time token and the other is modifier or numeral, and then they designed proper rules to recognize time expressions from unstructured text. To apply this linguistic knowledge and achieve more progress in linguistic tasks, however, we still need to explore into deeper understanding of this linguistic phenomenon.

5.2. Limitations

While we find that the length-frequency distributions of entities in different types can be well characterized by Marshall–Olkin power-law (MOPL) models, and the ones in different languages can also be roughly characterized by MOPL models, we should note that our analysis on these datasets about different languages may be inaccurate because many of these languages are not well

Table 8
Runtime of MOPL, LS_{avg}, CSN2009, and LogNormal fitting to the length-frequency distributions of entities in different types and different languages. The unit of the runtime is millisecond, denoted by ms.

Dataset	MOPL	LS_{avg}	CSN2009	LogNormal
ABSA	188.93 ms	5.89 ms	29.51 ms	6.20 ms
ACE04	293.97 ms	6.40 ms	308.19 ms	7.14 ms
BBN	69.83 ms	6.81 ms	134.39 ms	6.32 ms
BioMed	360.48 ms	7.03 ms	4368.31 ms	7.43 ms
CoNLL03	360.48 ms	5.71 ms	42.93 ms	6.92 ms
COVID19	261.38 ms	7.52 ms	39 544.32 ms	27.45 ms
LitBank	409.67 ms	6.78 ms	474.60 ms	6.57 ms
OntoNotes5	96.58 ms	5.60 ms	183.25 ms	8.53 ms
Re3d	111.97 ms	6.20 ms	19.79 ms	6.90 ms
TimeExp	137.48 ms	6.54 ms	59.12 ms	6.66 ms
Twitter	89.37 ms	152.74 ms	53.19 ms	1371.74 ms
WikiAnchor	357.21 ms	7.05 ms	17 060.66 ms	12.55 ms
Total	**2737.35** ms	**224.27** ms	**62 278.26** ms	**1474.41** ms
Afrikaans	312.27 ms	6.34 ms	53.83 ms	6.58 ms
Arabic	224.97 ms	7.13 ms	284.04 ms	6.68 ms
Basque	64.78 ms	6.44 ms	13.29 ms	6.30 ms
Bokmal	92.05 ms	6.13 ms	22.85 ms	6.03 ms
Croatian	73.45 ms	6.09 ms	31 483.92 ms	88.09 ms
Czech	69.13 ms	6.50 ms	80.67 ms	6.09 ms
France	79.26 ms	6.48 ms	23.68 ms	7.02 ms
German	168.32 ms	227.47 ms	88.78 ms	783.02 ms
Italian	295.43 ms	6.26 ms	6335.01 ms	9.42 ms
Netherland	41.71 ms	6.84 ms	11.21 ms	6.37 ms
Nynorsk	69.92 ms	6.28 ms	21.86 ms	6.61 ms
Polish	67.35 ms	5.47 ms	20 347.38 ms	99.88 ms
Romanian	132.39 ms	6.20 ms	527.88 ms	6.26 ms
Russian	82.65 ms	6.06 ms	4555.56 ms	12.21 ms
Samnorsk	89.67 ms	5.80 ms	41.98 ms	6.03 ms
Slovak	114.66 ms	6.12 ms	185.98 ms	6.17 ms
Slovene	60.35 ms	6.30 ms	15 422.35 ms	39.23 ms
Ukrainian	94.12 ms	7.39 ms	37 443.65 ms	50.21 ms
Total	**2132.46** ms	**335.30** ms	**116 943.92** ms	**1152.21** ms

studied in the field of natural language processing and related areas and we authors do not have sufficient expertise knowledge to cover our analysis on these different languages.

6. Conclusion

In this paper, we discover that the length-frequency distributions of entities in different types and different languages can be characterized by a family of Marshall–Olkin power-law (MOPL) models. Our discovery adds a stable knowledge to the field of language and provides some insights into conducting entity-related linguistic tasks and may also provide a new perspective for future potential research in understanding the language use. Experimental results on the length-frequency of entities in both different types and different languages demonstrate the superiority of MOPL models against a log-normal model and two state-of-the-art power-law models, namely LS_{avg} that is developed by Zhong et al. [27] and CSN2009 that is developed by Clauset et al. [26]. Experimental results also demonstrate that MOPL models are scalable to the length-frequency of entities in large-scale real-world datasets.

CRediT authorship contribution statement

Xiaoshi Zhong: Writing – original draft, Methodology, Investigation. **Xiang Yu:** Software, Resources, Conceptualization. **Erik Cambria:** Writing – review & editing, Supervision. **Jagath C. Rajapakse:** Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article

Acknowledgments

This research is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project #A18A2b0046).

References

[1] J.B. Estoup, Gammes Stenographiques, Institut Stenographique de France, Paris, 1916.
[2] G. Zipf, The Psychobiology of Language, Routledge, London, 1936.
[3] G. Zipf, Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology, Addison-Wesley Press, Inc., 1949.
[4] B. Corominas-Murtra, R.V. Solé, Universality of Zipf's law, Phys. Rev. E 82 (1) (2010) 011102.
[5] S.T. Piantadosi, Zipf's word frequency law in natural language: A critical review and future directions, Psychon. Bull. Rev. 21 (5) (2014) 1112–1130.
[6] C.B. Williams, A note on the statistical analysis of sentence-length as a criterion of literary style, Biometrika 31 (3/4) (1940) 356–361.
[7] W. Fucks, Theorie der wortbildung, Math.-Phys. Semesterber. 4 (1955) 195–212.
[8] W. Fucks, Die mathematischen gesetze der bildung von sprachelementen aus ihren bestandteilen, Nachr.tech. Fachber. 3 (1956) 7–21.
[9] W.C. Wake, Sentence-length distributions of Greek authors, J. R. Stat. Soc. Ser. A (Gen.) 120 (3) (1957) 331–346.
[10] G.A. Miller, E.B. Newman, E.A. Friedman, Length-frequency statistics for written English, Inf. Control 1 (1958) 370–389.
[11] C.B. Williams, Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon, Biometrika 62 (1) (1975) 207–212.
[12] R. Grotjahn, G. Altmann, Modelling the distribution of word length: Some methodological problems, in: Contributions to Quatitative Linguistics, 1993, pp. 141–153.
[13] G. Wimmer, R. Kohler, R. Grotjahn, G. Altmann, Towards a theory of word length distribution, J. Quant. Linguist. 1 (1) (1994) 98–106.
[14] K.-H. Best, Word length in old icelandic songs and prose texts, J. Quant. Linguist. 3 (2) (1996) 97–105.
[15] B. Sigurd, M. Eeg-Olofsson, J. van de Weijer, Word length, sentence length and frequency - Zipf revisited, Stud. Linguist. 58 (1) (2004) 37–52.
[16] C. Manning, H. Schutze, Foundations of Statistical Natural Language Processing, MIT Press, Cambride, 1999.
[17] D. Jurafsky, J. Martin, Speech and Language Processing, second ed., Prentice Hall, 2008.
[18] D. Jurafsky, J. Martin, Speech and Language Processing, third ed. Draft ed., 2020.
[19] N.A. Chinchor, MUC-7 named entity task definition, in: Proceedings of the 7th Message Understanding Conference, 1997.
[20] E.F.T.K. Sang, F.D. Meulder, Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition, in: Proceedings of the 7th Conference on Natural Language Learning, 2003, pp. 142–147.
[21] H. Ji, R. Grishman, Knowledge base population: Successful approaches and challenges, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, 2011, pp. 1148–1158.
[22] X. Ling, S. Singh, D.S. Weld, Design challenges for entity linking, Trans. Assoc. Comput. Linguist. 3 (2015) 315–328.
[23] M. Pérez-Casany, A. Casellas, Marshall-olkin extended zipf distribution, 2013, arXiv preprint arXiv:1304.4540.
[24] N. Smirnov, Table for estimating the goodness of fit of empirical distributions, Ann. Math. Stat. 19 (2) (1948) 279–281.
[25] M.A. Stephens, EDF statistics for goodness of fit and some comparisons, J. Amer. Statist. Assoc. 69 (347) (1974) 730–737.
[26] A. Clauset, C.R. Shalizi, M.E.J. Newman, Power-law distributions in empirical data, SIAM Rev. 51 (4) (2009) 661–703.
[27] X. Zhong, M. Wang, H. Zhang, Is least-squares inaccurate in fitting power-law distributions? The criticism is complete nonsense, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 2748–2758.
[28] M.E. Newman, Power laws, Pareto distributions and Zipf's law, Contemp. Phys. 46 (5) (2005) 323–351.
[29] W. Li, Zipf's law everywhere, Glottometrics 5 (2002) 14–21.
[30] A. Pratap, R. Raja, J. Cao, G. Rajchakit, H.M. Fardoun, Stability and synchronization criteria for fractional order competitive neural networks with time delays: An asymptotic expansion of Mittag Leffler function, J. Franklin Inst. B 356 (4) (2019) 2212–2239.

[31] P. Anbalagan, E. Hincal, R. Ramachandran, D. Baleanu, J. Cao, M. Niezabitowski, A Razumikhin approach to stability and synchronization criteria for fractional order time delayed gene regulatory networks, 6 (5) (2021) 4526–4555.
[32] A. Pratap, R. Raja, R.P. Agarwal, J. Alzabut, M. Niezabitowski, E. Hincal, Further results on asymptotic and finite-time stability analysis of fractional-order time-delayed genetic regulatory networks, Neurocomputing 475 (2022) 26–37.
[33] G. Miller, Some effects of intermittent silence, Am. J. Psychol. 70 (1957) 311–314.
[34] G. Miller, "Introduction" in The Psycho-Biology of Language: An Introduction to Dynamic Philology (1935), MIT Press, 1965.
[35] W. Li, Random texts exhibit Zipf's-law-like word frequency, IEEE Trans. Inform. Theory 38 (6) (1992) 1842–1845.
[36] D. Malone, K. Maher, Investigating the distribution of password choices, in: Proceedings of the 21th International Conference on World Wide Web, 2012, pp. 301–310.
[37] D. Wang, H. Cheng, P. Wang, X. Huang, G. Jian, Zipf's law in passwords, IEEE Trans. Inf. Forensics Secur. 12 (11) (2017) 2776–2791.
[38] R. Hanel, B. Corominas-Murtra, B. Liu, S. Thurner, Fitting power-laws in empirical data with estimators that work for all exponents, PLoS ONE 12 (2) (2017) 1–15.
[39] M. Gerlach, E.G. Altmann, Testing statistical laws in complex systems, Phys. Rev. Lett. 122 (16) (2019) 168301.
[40] I. Artico, I. Smolyarenko, V. Vinciotti, E.C. Wit, How rare are power-law networks really? in: Proceedings of the Royal Society A, 2020, 20190742.
[41] B. Nettasinghe, V. Krishnamurthy, Maximum likelihood estimation of power-law degree distributions via friendship paradox-based sampling, ACM Trans. Knowl. Discov. Data 15 (6) (2021) 1–28.
[42] R. Grishman, B. Sundheim, Message understanding conference - 6: A brief history, in: Proceedings of the 16th International Conference on Computational Linguistics, 1996.
[43] X. Ling, D.S. Weld, Fine-grained entity recognition, in: Proceedings of the Twenty-Sixth Conference on Artificial Intelligence, 2012.
[44] S. Pradhan, A. Moschitti, N. Xue, H.T. Ng, A. Bjorkelund, O. Uryupina, Y. Zhang, Z. Zhong, Towards robust linguistic analysis using ontonotes, in: Proceedings of the 7th Conference on Computational Natural Language Learning, 2013, pp. 143–152.
[45] J. Pustejovsky, J. Castano, R. Ingria, R. Sauri, R. Gaizauskas, A. Setzer, G. Katz, D. Radev, TimeML: Robust specification of event and temporal expressions in text, in: New Directions in Question Answering, Vol. 3, 2003, pp. 28–34.
[46] J. Pustejovsky, P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, B. Sundheim, D. Radev, D. Day, L. Ferro, M. Lazo, The TIMEBANK corpus, Corpus Linguist. 2003 (2003) 647–656.
[47] X. Zhong, E. Cambria, Time expression recognition and normalization: A survey, Artif. Intell. Rev. 56 (9) (2023) 9115–9140.
[48] B. Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, 2012.
[49] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, SemEval-2014 task 4: Aspect based sentiment analysis, in: Proceedings of the 8th International Workshop on Semantic Evaluation, 2014, pp. 27–35.
[50] D. Bamman, S. Popat, S. Shen, An annotated dataset of literary entities, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 2138–2144.
[51] A. Ritter, S. Clark, Mausam, O. Etzioni, Named entity recognition in tweets: An experimental study, in: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 2011, pp. 1524–1534.
[52] L. Derczynski, K. Bontcheva, I. Roberts, Broad Twitter corpus: A diverse named entity recognition resource, in: Proceedings of the 26th International Conference on Computational Linguistics, 2016, pp. 1169–1179.
[53] K. Fukuda, T. Tsunoda, A. Tamura, T. Takagi, Toward information extraction: Identifying protein names from biological papes, in: Proceedings of the Pacific Symposium on Biocomputing, 1998, pp. 707–718.
[54] K. Takeuchi, N. Collier, Bio-medical entity extraction using support vector machines, Artif. Intell. Med. 33 (2) (2005) 125–137.
[55] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, I. Androutsopoulos, SemEval-2015 task 12: Aspect based sentiment analysis, in: Proceedings of the 9th International Workshop on Sementic Evaluation, 2015, pp. 486–495.
[56] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, R. Weischedel, The automatic content extraction (ACE) program tasks, data, and evaluation, in: Proceedings of the 2004 Conference on Language Resources and Evaluation, 2004, pp. 1–4.
[57] R. Weischedel, A. Brunstein, BBN Pronoun Coreference and Entity Type Corpus, Vol. 112, Linguistic Data Consortium, 2005.
[58] G. Crichton, S. Pyysalo, B. Chiu, A. Korhonen, A neural network multi-task learning approach to biomedical named entity recognition, BMC Bioinformatics 18 (1) (2017) 368–371.
[59] X. Wang, X. Song, B. Li, Y. Guan, J. Han, Comprehensive named entity recognition on CORD-19 with distant or weak supervision, 2020, arXiv preprint: arxiv.org/abs/2003.12218.
[60] P. Mazur, R. Dale, WikiWars: A new corpus for research on temporal expressions, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010, pp. 913–922.
[61] N. UzZaman, H. Llorens, L. Derczynski, M. Verhagen, J. Allen, J. Pustejovsky, SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations, in: Proceedings of the 7th International Workshop on Semantic Evaluation, 2013, pp. 1–9.
[62] X. Zhong, A. Sun, E. Cambria, Time expression analysis and recognition using syntactic token types and general heuristic rules, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017, pp. 420–429.
[63] X. Zhong, E. Cambria, Time expression recognition using a constituent-based tagging scheme, in: Proceedings of the 2018 World Wide Web Conference, 2018, pp. 983–992.
[64] B. Strauss, B.E. Toma, A. Ritter, M.-C. de Marneffe, W. Xu, Results of the WNUT16 named entity recognition shared task, in: Proceedings of the 2nd Workshop on Noisy User-Generated Text, 2016, pp. 138–144.
[65] X. Zhong, E. Cambria, A. Hussain, Extracting time expressions and named entities with constituent-based tagging schemes, Cogn. Comput. 12 (4) (2020) 844–862.
[66] T. Paccosi, A.P. Aprosio, KIND: an Italian multi-domain dataset for named entity recognition, 2021, arXiv preprint arXiv:2112.15099.
[67] B. Johansen, Named-entity recognition for norwegian, in: Proceedings of the 22nd Nordic Conference on Computational Linguistics, NoDaLiDa, 2019.
[68] S.D. Dumitrescu, A.-M. Avram, Introducing RONEC–the Romanian Named Entity Corpus, 2019, arXiv preprint arXiv:1909.01247.
[69] T.B. Arnold, J.W. Emerson, Nonparametric goodness-of-fit tests for discrete null distributions, R J. 3 (2) (2011).
[70] D.S. Dimitrova, V.K. Kaishev, S. Tan, Computing the Kolmogorov-Smirnov distribution when the underlying CDF is purely discrete, mixed, or continuous, J. Stat. Softw. 95 (10) (2020) 1–42, http://dx.doi.org/10.18637/jss.v095.i10, URL: https://www.jstatsoft.org/index.php/jss/article/view/v095i10.
[71] J. Lafferty, A. McCallum, F. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: Proceedings of International Conference on Machine Learning, 2001, pp. 281–289.
[72] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780.
[73] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.
[74] D. Nadeau, S. Sekine, A survey of named entity recognition and classification, Lingvist. Investig. 30 (1) (2007) 3–26.
[75] M. Krallinger, F. Leitner, O. Rabal, M. Vazquez, J. Oyarzabal, A. Valencia, Overview of the chemical compound and drug name recognition (CHEMDNER) task, in: BioCreative Challenge Evaluation Workshop, 2015, pp. 2–33.
[76] V. Yadav, S. Bethard, A survey on recent advances in named entity recognition from deep learning models, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 2145–2158.
[77] X. Zhong, Time Expression and Named Entity Analysis and Recognition (Ph.D. thesis), Nanyang Technological University, Singapore, 2020.
[78] X. Zhong, E. Cambria, A. Hussain, Does semantics aid syntax? An empirical study on named entity recognition and classification, Neural Comput. Appl. 34 (11) (2022) 8373–8384.