Argentinian Journal of Applied Linguistics
Vol. 7, No. 1, May 2019, 8-26

Profiling specialized web corpus qualities: A progress report on "Domainhood"

Marina Santini* - RISE Research Institutes of Sweden
Wiktor Strandqvist - RISE Research Institutes of Sweden & Linköping University
Arne Jönsson - RISE Research Institutes of Sweden & Linköping University

(Received 19/12/18; final version received 25/03/19)

This article is an extended version of the following paper:
Strandqvist W., Santini M., Lind L. and Jönsson A. (2018). Towards a Quality Assessment of Web Corpora for Language Technology Applications. In: Read T., Montaner S. and Sedano B. (2018). *Technological Innovation for Specialized Linguistic Domains Languages for Digital Lives and Cultures Proceedings of TISLID'18*. Editions universitaires europeenne.

ABSTRACT

In this article we describe ways to profile the domain specificity, a.k.a. domainhood, of specialized web corpora in English and in Swedish. Several studies have been carried out to measure the "qualities" of general-purpose web corpora. On the contrary, less attention has been paid to the evaluation of specialized or domain-specific web corpora. To fill this gap, in this article we present case studies where we explore the effectiveness of several statistical measures – i.e. rank correlation coefficients (Kendall and Spearman), Kullback–Leibler divergence, log-likelihood and burstiness - to assess domainhood. Our findings indicate that it is possible to profile the domainhood quality of a corpus. However, further research is needed to generalize on the results.
Keywords: corpus evaluation; term extraction; log- likelihood; rank correlation; Kullback-Leibler divergence.

RESUMEN

En este artículo describimos formas de trazar la especificidad del dominio ("domainhood") de los corpus de webs especializados en inglés y en sueco. Muchos estudios se han llevado a cabo para medir las "cualidades" de los corpus de webs de carácter general. Sin embargo, se ha prestado menos atención a la evaluación de corpus de web especializados o de dominios específicos. Para llenar este vacío, en este artículo presentamos estudios de caso donde exploramos la efectividad de diferentes medidas estadísticas, a saber, coeficientes de correlación de rango (Kendall and Spearman), divergencia Kullback–Leibler, probabilidad de registro y *burstiness* – para evaluar la especificidad del dominio. Nuestros resultados indican que es posible perfilar la calidad de dominio de un corpus. Sin embargo, es necesaria una mayor investigación para generalizar en los resultados.
Palabras clave: evaluación de corpus; extracción de términos; probabilidad de registro correlación de rango; divergencia Kullback-Leibler.

* Corresponding author; e-mail: marinasantini.ms@gmail.com, marina.santini@ri.se

WEB CORPORA ARE text collections made of documents that have been retrieved and downloaded from the web. While texts in traditional corpora are hand-picked from several media and agreed upon by a number of experts, web corpora are built with documents available on the web at the time of corpus bootstapping. Traditional corpora are carefully curated and annotated to preserve the original traits of the selected texts, while web corpora can be noisy in several respects, e.g. they might contain damaged characters, problematic symbols, inconsistent punctuation or ungrammatical texts. In short, traditional corpora and web corpora represent different approaches to corpus construction and use.

Traditional corpora are a trove of hand-crafted qualities. However, the added value of web corpora is in their malleability. Similar to traditional corpora, web corpora can be general-purpose or specialized (Barbaresi, 2015) and may serve different purposes, such as linguistic studies (e.g., Schäfer & Bildhauer, 2013; Biemann et al., 2007; Lüdeling et al. 2007) and professional uses (Goldhahn et al., 2012; Baroni et al., 2006). The unique and unprecedented potential of web corpora is that they can promptly and inexpensively account for virtually any domain, topic, genre, register, sublanguage, style and emotional connotation, since the web itself is a gold mine of linguistic and textual varieties. In particular, domain-specific web corpora are widely used in several linguistic disciplines (e.g. in translation studies and lexicography) and they are an important building block of language technology applications (e.g. machine translation, terminology extraction and lexicon induction). Both in linguistics and in language technology, the reliability of the results may depend on the domain representativeness of the web corpus itself.

While bootstrapping a web corpus is common practice (many tools exist, either based on crawling or on search engine queries), the validation of web corpora is still a grey area. With the investigations described in this article, we would like to contribute to the discussion by adding a new perspective to web corpus evaluation. Normally, corpora can be assessed according to several "qualities", for instance corpus balance (in terms of domain, genre, style, register etc.), corpus representativeness (with respect to a purpose), and the like. In this complex scenario, we single out one quality, namely domain specificity, a.k.a. domainhood. Domainhood (Santini et al., 2018) refers to the domain representativeness of a corpus. Here "domain" is defined as the "subject field" or "area" in which a web document is used. For instance, a high frequency of medical terms is a sign that a corpus is a specialized medical corpus. We are aware that domain-specific web corpora might have a different domain granularity, and this is an additional factor to be taken into account.

In this article, we present case studies where we explore the effectiveness of several statistical measures – i.e. rank correlation coefficients (Kendall and Spearman), Kullback–Leibler (KL) divergence, log-likelihood and burstiness - to assess domainhood. The long-term goal is to find suitable metrics that would help assess whether one corpus is more domain-specific than another corpus. This information would speed up any post-editing of specialized web corpora by reducing manual intervention. In this article we empirically

investigate these issues and present two experiments, the first one based on English corpora, and the second one hinged upon Swedish corpora.

The article is organized as follows: in Section 2 ("Related Work"), we discuss previous research; Section 3 ("Experiment 1: Building and Profiling Domain-Specific Web Corpora in English") presents the eCare Term Extractor and the profiling of two medical web corpora in English with similar domain granularity and a similar corpus size; in Section 4 ("Experiment 2: Building and Profiling Domain-Specific Web Corpora in Swedish"), we apply burstiness to pin down the domainhood of two medical web corpora in Swedish that have different domain granularity and a different size; finally in Section 5 ("Conclusion and Future Work"), we draw conclusions and outline future work.

Related Work

When we talk about web corpora, it seems more appropriate to talk about "qualities" rather than a single "quality". Several approaches have been proposed to capture the "qualities" of web corpora (e.g. see Oakes, 2008; Schäfer et al., 2013). However, no standard metrics have been agreed upon for the automatic quantitative assessment of the different "qualities" of web corpora. "Qualities" can be defined as dimensions of variation. Domain, genre, style, register, medium, etc. are well-known dimensions of language variation. In this study, we focus on the dimension of "domain", i.e. the "subject field" in which a web document is used. Our aim is somewhat similar to the one expressed in Wong et al. (2011), where the authors propose a technique, called SPARTAN, for constructing specialized corpora from the web. Our approach is different though, because in order to assess the domainhood quality, we rely on measures that are well-established and easy to replicate. Since in this article we describe comparative experiments based on rank correlation (Kendall and Spearman), KL divergence, log-likelihood and burstiness, in this section we provide a short overview of studies where these measures were used.

The importance of a quantitative evaluation of corpora has been stressed for a long time. In his seminal article, Kilgarriff (2001) motivates his review of approaches to corpus comparison by asking two crucial questions: "how similar are two corpora?" and "in what ways do two corpora differ?". He presents comparative experiments based on several corpora and on several statistical measures. Rayson and Garside (2000) show that log-likelihood can be safely used as a "quick way to find the differences between corpora" and that it is more robust than other measures because it is insensitive to corpus size. Gries (2013) suggests using a Kendall Tau correlation coefficient to determine whether the observed patterns of two corpora show significant correlations. Ciaramita and Baroni (2006) propose using KL divergence to assess the "randomness" or "unbiasedness" of general-purpose corpora. They compare domain-specific sub parts of the British National Corpus (BNC) against the whole corpus and show that KL divergence can reliably indicate the difference between general purpose corpora (random and unbiased) and domain-specific

corpora (biased). Burstiness has been used in information retrieval and in terminology extraction (Church and Gale, 1995; Katz, 1996), and more recently for corpus evaluation (Sharoff, 2017). Burstiness is a measure that can be utilized for inducing specialized lexicon that is not evenly distributed in a corpus but appears "in bursts". Burstiness indicates "how peaked a word's usage is over a particular corpus of documents" (Pierrehumbert, 2012). More specifically, "bursty words are topical words that tend to appear frequently in documents when some topic is discussed, but do not appear frequently across all documents in a collection" (Irvine and Callison-Burch, 2017). While bursty words are feared and filtered out when assessing general-purpose corpora (Sharoff, 2017), we think that they could give a good indication of domain specificity, and for this reason we include burstiness in our experiments.

Experiment 1: Building and Profiling Domain-Specific Web Corpora in English

In the first experiment, we propose a two-step approach. In the first step, we build a term extraction that can automatically identify term candidates in project-specific personas and use cases/scenarios. Personas and use cases/scenarios are narratives that describe a "system's behavior under various conditions as the system responds to requests from stakeholders" (Cockburn, 2000). This type of texts are nowadays normally included in many language technology projects. Personas and use cases/scenarios are relatively short texts - only a few dozen pages − normally based on numerous interviews and observations of real situations and written by domain experts who know how to correctly use terms in their own domain. For this reason, we argue that they are a convenient textual resource to automatically extract term seeds to bootstrap domain-specific web corpora, thus overriding the tedious and somehow arbitrary process normally required to collect term seeds. In our study, we focus on the medical terms that occur in personas and use cases/scenarios written in English for *E-care@home*, a multi-disciplinary project that investigates how to ensure medical care at home for the elderly. We complete this step with the evaluation of the term extractor against a gold standard made of the SNOMED CT terms manually selected by a domain expert from the *E-care@home* personas and use cases/scenarios. SNOMED CT is the largest existing resource of medical terminology. The challenge of this step is to create a "good enough" term extractor based on a relatively small textual resource, a task that is still under-investigated since most of existing term extractors are based on large corpora (e.g. see Nazarenko and Zargayouna, 2009).

In the second step, we use the term seeds extracted in the previous step to *bootcat* a medical web corpus and evaluate its quality. The term "bootcat" means bootstrapping specialized corpora from the web using BootCaT, a tool that was introduced by Baroni and Bernardini (2004). Leveraging on the web to create specialized corpora is a well-established idea (e.g. Baroni & Bernardini, 2004; Kilgarriff et al. 2010), less so their evaluation. For this reason, in Experiment 1 we analyse and test three corpus profiling measures, namely rank

correlation (Kendall and Spearman), KL divergence and log-likelihood. The challenge of this step is to find an empirical answer to the following question: "*can we assess the domainhood quality of a corpus automatically bootstrapped from the web*?".

E-care Term Extractor

Arguably, the use of personas and use cases/scenarios, when available, is a good starting point to automatize the manual process of term seeds selection. The *E-care* term extractor developed for this purpose includes three main components. The first component (*terminology extractor*) uses a shallow syntactic analysis of the text to extract candidate terms. The second component (*terminology validator*) compares each of the candidate terms and their variations against SNOMED CT (International Edition) to produce candidate terms. The third component (*seed validator*) evaluates the performance of the term extractor.

The *terminology extractor* relies on the Stanford Tagger (Toutanova, Klein, Manning, & Singer, 2003) to assign a part-of-speech (POS) tag to each word in the texts. The tagged text is then searched sequentially with each of the selected syntactic patterns shown in Table 1 (cf. Pazienza, Pennacchiotti, & Zanzotto, 2005).

Patterns
(noun)+
(adjective)(noun)+
(noun)(prep)(noun)+

Table 1. Syntactic patterns used to identify terminology

The *terminology validator* takes the candidate terms produced in the previous steps, and matches them against SNOMED CT. If an exact match is not found, each word is stemmed. The stemmed words are permuted, and each permutation is then matched against SNOMED CT once again, this time using wildcards between the words, to allow for spelling variations. Matches are then ranked by DF/IDF scores (cutoff = 200). In this context, DF stands for *term document frequency* and refers to the frequency of a term in a document divided by the document length (i.e. the total number of words in the document). DF is basically like a TF (*term frequency*[1]) but normalized to the document length in order to avoid any bias towards long documents[2]. IDF stands for *inverse document frequency* and it is based on counting the number of documents in the corpus which contain the term in question (Robertson, 2004). Similar to TF-IDF, DF/IDF is a way to reflect how important a term is in a given document.

The *seed generator* generates three terms (i.e. triples) from the cutoff list when they occurred in the same document.

E-care Term Extractor: Results and Discussion

The *E-care* term extractor performance is summarized in Table 2. The *terminology extractor* has an extraction recall of 81.25% on the development set, which is the subset of documents used to optimize the extraction algorithm. When evaluated against the gold standard made of the SNOMED CT terms manually selected by a domain expert from the E-care@home personas and use cases/scenarios, the *terminology validator* achieves the following performance: Precision = 34.2%, Recall = 71%, F1 = 46.2%. These metrics were calculated according to standard formulas, namely Precision = true positives / (true positives + false positives), Recall = true positives / (true positives + false negatives) [3] and F1 = 2 * ((precision * recall) / (precision + recall))[4].

	Metrics	%
Term candidate extraction	Extraction recall	81
Term validation	Precision	34.2
	Recall	71
	F1	46.2

Table 2. Current performance of E-care term extractor

Interestingly, the moderate performance of the current version of the *E-care* term extractor did not affect detrimentally the domainhood quality of the resulting web corpus, as shown in the next subsections.

Corpus Evaluation Metrics

For corpus evaluation, we use metrics based on word frequency lists, namely rank correlation coefficients (Kendall and Spearman), KL divergence and log-likelihood.

1) Correlation coefficients: *Kendall* correlation coefficient (Tau) and *Spearman* correlation test (Rho) are non-parametric tests. They both measure how similar the order of two ranks is. We used the R function "cor.test()" with method="kendall, spearman" to calculate the tests.

2) Kullback–Leibler (KL) divergence (a.k.a. relative entropy): KL divergence is a measure of the "distance" between two distributions. The KL divergence quantifies how far-off an estimation of a certain distribution is from the true distribution. The KL divergence is non-negative and equal to zero if the two distributions are identical. In our context, the closer the value is to 0, the more similar two corpora are. We used the R package "entropy", function "KL.empirical()" to compute KL divergence.

3) Log-Likelihood (LL-G^2): log-likelihood (Dunning, 1993) has been used for corpus profiling (Rayson and Garside, 2000). The words that have the largest log-likelihood scores

show the most significant word-frequency difference in two corpora. Log-likelihood is not affected by corpus size variation.

For the evaluation, we use three web corpora, namely:

- **ukWaCsample** (872 565 words): a random subset of ukWaC (Ferraresi et al., 2008), a general- purpose web corpus.
- **Gold** (544 677 words): a domain-specific web corpus bootstrapped with ***terms manually selected by a domain expert*** from the *E-care@home* personas and use cases/scenarios.
- **Auto** (492 479 words): a domain-specific web corpus collected with ***automatically extracted term seeds*** from the *E-care@home* personas and use cases/scenarios.

Results and Discussion

In this section, we present and discuss the results of our experiments.

Measuring Rank Correlation. We computed the normalized frequencies of the three corpora (words per million) and ranked them (with ties). The plots of the first 1000 top frequencies of the three corpora are shown in Fig. 1. From the plots, we can see that UkwaCsample has very little in common with both Gold and Auto (boxes 1 and 2), while Gold and Auto (box 3) are similar.

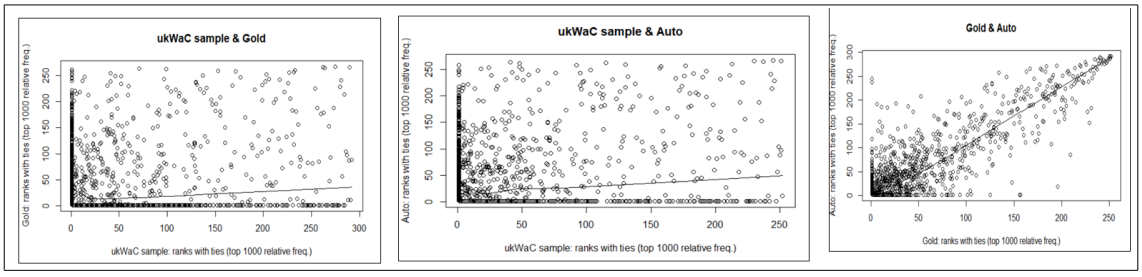

Fig. 1 Plotting 1000 top ranks: (from left to right): ukWaCsample and Gold (box 1), ukWaCsample and Auto (box2), and Gold and Auto (box 3).

When testing the rank correlation (Kendall and Spearman), we observe a statistically significant positive rank correlation between Gold and Auto (see Fig. 2, box 3; Fig. 3, box 3), which means that words in Gold and in Auto tend to have similar ranks. Conversely, the correlation between ukWaCsample and Gold and ukWaCsample and Auto is negative and weak (see Fig. 2, box 1 and box 2; Fig. 3, box 1 and box 2), which essentially means that their ranks follow different distributions.

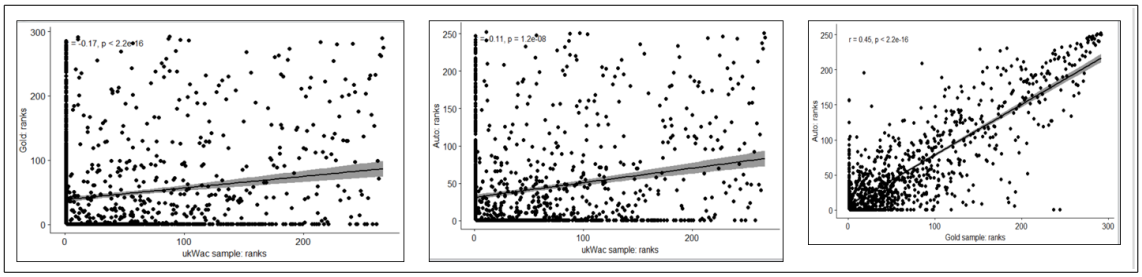

Fig. 2 Kendall Tau: (from left to right):): ukWaCsample and Gold (box 1), ukWaCsample and Auto (box2), and Gold and Auto (box 3).

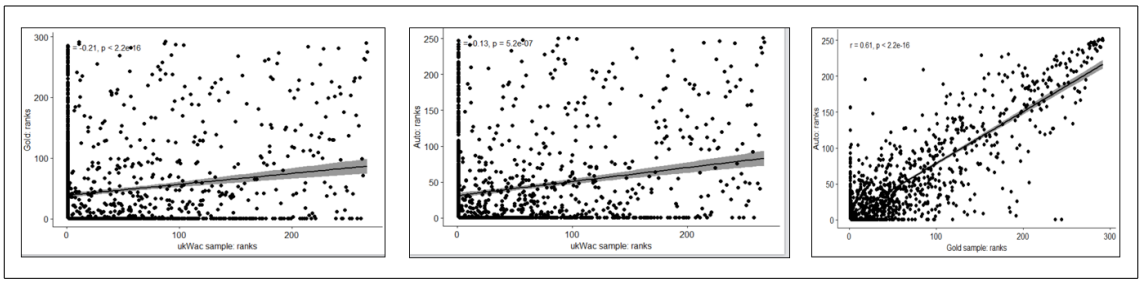

Fig. 3 Spearman Rho: (from left to right):): ukWaCsample and Gold (box 1), ukWaCsample and Auto (box2), and Gold and Auto (box 3).

KL divergence. Before calculating KL divergence, a smoothing[5] value of 0.01 was been added to the normalized frequencies. Results are shown in Table 3. The scores returned by KL distance for ukWacSample vs Gold (row 1) and ukWacSample vs Auto (row 2) − 7.544118 and 6.519677, respectively − are (unsurprisingly) large and indicate a wide divergence between the general-purpose ukWacSample and the domain-specific Gold and Auto. On the contrary, the KL score of 1.843863 indicates that Gold vs Auto (row 3) are similar to each other.

Corpora	**KL scores**
ukWacSample vs Gold	7.544118
ukWacSample vs Auto	6.519677
Gold vs Auto	1.843863

Table 3. KL scores

Log-Likelihood (LL-G^2). We computed log-likelihood scores on smoothed word frequencies. The total log-likelihood scores for the three web corpora (top 1000 words) are shown in Table 3. The larger the log-likelihood score of a word, the more different its distribution in two corpora. The large log-likelihood scores for ukWaCsample vs Gold (453 441.6) and for ukWaCsample vs Auto (393 705.9) indicate that these corpora are remarkably dissimilar if

compared to the much smaller log-likelihood score returned for Gold vs Auto (114 694.2), which suggests that Gold and Auto are similar to each other (see Table 4).

Corpora	Total log-likelihood scores
ukWacSample vs Gold	453 441.6
ukWacSample vs Auto	393 705.9
Gold vs Auto	114 694.2

Table 4. Log-likelihood scores of the three corpora

It is also possible to assess the statistical significance of the individual log-likelihood scores. Normally, a log-likelihood score of 3.8415 or higher is significant at the level of $p<0.05$ and a log-likelihood score of 10.8276 is significant at the level of $p<0.001$ (Desagulier, 2017). Fig. 4 shows the breakdown of the top-ranked log-likelihood scores of three corpora. We take 3.8415 ($p < 0.05$) as a threshold and observe that ukWaCsample vs Gold (box 1) differs very much in the use of words such as "patient" or "patients" and "blood", and in ukWaCsample vs Auto (box 2) these words have a similar behavior. Conversely, these words are not in the top list of Gold vs Auto (box 3). Additionally, the log-likelihood scores in box 3 are much smaller in magnitude, which indicates that the difference between words is less pronounced.

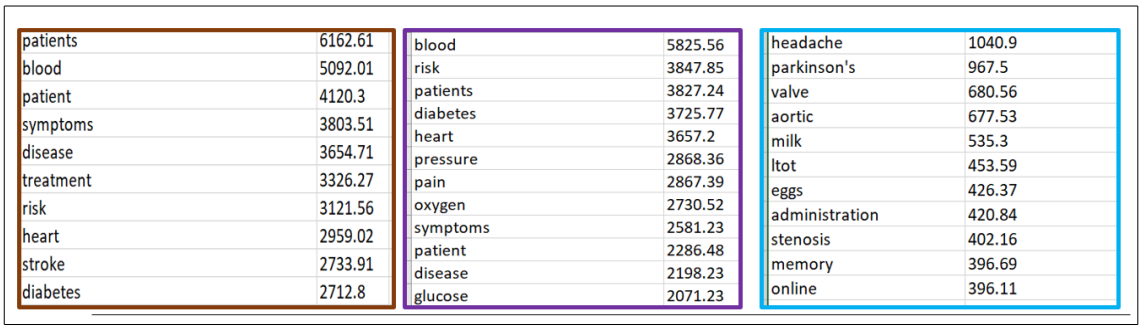

Fig. 4. Top-ranked log-likelihood scores (from left to right): ukWaCsample and Gold (box 1), ukWaCsample and Auto (box 2), and Gold and Auto (box 3).

Experiment 1: Conclusion

We have shown that it is possible to create a reliable term extractor (although the intrinsic evaluation of its performance is moderate) that works well in practice for relatively short texts written by domain experts. When used to bootstrap a web corpus, the automatically extracted term seeds create a corpus whose domain specificity is similar to a corpus bootstrapped with manually selected term seeds. This is an added value because corpus construction can be fully automatized and standardized.

We have seen that well-established measures, such as rank correlation, KL divergence and log-likelihood, do give a coarse but grounded idea of domain specificity. Essentially, they allow for an evaluation of the domainhood quality of web corpus and presumably they could also be used to pre-assess the portability of NLP tools from a domain-specific corpus to another. Similar experiments have also been carried out on Swedish corpora with much the same results (Santini et al., 2018), showing that our approach may become a language-independent standardized step in corpus evaluation practice, if these results will be confirmed by future experiments in other languages.

We can now provide empirical answer to the questions asked above. That is: yes, we can assess the domainhood quality of a corpus automatically bootstrapped from the web. This can be done by using metrics that are well-established and easily replicable, such as rank correlation, KL divergence and log-likelihood. Last but not least, these metrics also help get a coarse but robust indication of topical similarities across corpora.

Experiment 2: Building and Profiling Domain-Specific Web Corpora in Swedish

Since "words are not selected at random" (Kilgarriff, 2005), we assume that the content words included in a corpus represent its content and domain. The corpora that we describe below both belong to the medical domain, but they have been built with slightly different target domains and domain granularity. The target domains are reference lists made of words representative of the domain of interest. As pointed out by Lippincott et al. (2011) "[w]hile variation at a coarser domain level such as between newswire and biomedical text is well-studied and known to affect the portability of NLP systems, there is a need to develop an awareness of subdomain variation when considering the practical use of language processing applications". In this experiment, we investigate whether burstiness can help make sense of subdomain variations or different domain granularities.

Same Domain, Different Granularities

For this investigation, we rely on two web corpora of Swedish texts, namely *eCare_ch_sv_01* and *eCare_uc_sv_02*. Both corpora are components of the eCare web corpus. *eCare_ch_sv_01* is about chronic diseases, while *eCare_uc_sv_02* was built with terminology in English automatically extracted in Experiment 1, and then translated in via SNOMED CT.

eCare_ch_sv_01 was built using 155 terms listed in SNOMED CT (Swedish edition) indicating chronic diseases. The 155 term seeds were selected from a much longer list of chronic diseases compiled by a domain expert and they represent a restricted and fine-grained domain (Santini et al., 2017). The size of this corpus is approx. 700 000 words. This corpus was used in the experiments presented in Santini et al. (2018).

The size of *eCare_uc_sv_02* is approx. 7 million words (6 942 193 tokens). *eCare_uc_sv_02* is, thus, about 10 times larger than *eCare_ch_sv_01*.

Both web corpora are supposed to represent the domain of chronic diseases but with different domain granularities and different corpus sizes. We assume that the domain granularity is more fine-grained in *eCare_ch_sv_01* and coarser in *eCare_uc_sv_02* because of the way the corpora have been bootstrapped. In this experiment, "fine-grained domain" means a very specialized domain where the seeds to bootstrap the corpus are specialized medical terms, e.g. "artrit" (en: arthritis). Conversely, "coarse-domain" refers to a corpus that has been bootstrapped both with specialized medical terms and polysemous words that are often related with diseases, e.g. "dos" (en: dosage) or "akut" (en: acute). The domain granularity is implicitly incorporated in the gold standards, as explained below.

Corpus Seeds and Gold Standards

What is the best way to represent a target domain? This question is complex and arguably the ideal solution depends on the purpose of an application. Here we take a basic approach and represent the target domains as reference lists – the gold standards - that contain the term seeds used to bootstrap the corpora. It makes sense to use domain-specific terms both for bootstrapping a web corpus and for evaluating its domainhood because the terms used as seeds (source terms) should be found in non-trivial proportions to be sure that the corpus is representative of the domain of interest. Here we present two different approaches to gold standard construction. The gold standard used to profile and evaluate *eCare_ch_sv_01* is made only of specialized medical terms, while the gold standard automatically extracted from use cases contains also polysemous words, such as "attack" (en: attack), "extrem" (en: extreme), "fet" (en: fat). The gold standards contain tokenized term seeds, without duplicates. This means that terms like "kronisk anemi" (en: chronic anemia) and "kronisk artrit" (en: chronic arthritis), in the gold standard are represented by three entries, namely "kronisk", "anemi" and "artrit". Both these lists and the top-ranked bursty words were stemmed, stopwords and numbers were removed using the R package Quanteda, without applying any customization to the stoplist and to the stemmer.

The two web corpora are then evaluated against two gold standards. More specifically, *gold_eCare_ch_sv_01* represents the target domain of *eCare_ch_sv_01* and contains 164 unigrams, while the target domain of *eCare_uc_sv_02* is represented by *gold_eCare_uc_sv_02* that contains 248 unigrams.

Burstiness

Several burstiness formulas exist. Here we use the formula from Church and Gale (1995), including the modification proposed by Irvine and Callison-Burch (2017) (i.e. the use of relative frequencies rather than absolute frequencies), namely:

$$B_w = \frac{\sum_{d_i \in D} rf_{wd_i}}{df_w}$$

where *rf* refers to the relative frequency of word *w* in a document, and *df* is the number of documents in which the word *w* appears. Relative frequencies are raw frequencies normalized by document length. In other words, burstiness is given by the sum of the all the relative frequencies of word *w* in the documents of the corpus divided by the number of documents containing the word. Burstiness is essentially the mean of a word in a corpus normalized by the number of documents where the word appears, and it ignores the documents where the word does not appear (Church and Gale, 1995; Katz, 1996). Burstiness differs from measures like TF (term frequency), which denotes the number of times that term *t* occurs in document *d*, or TF-IDF, a weight where TF is normalized by IDF (inverse document frequency)[6]. If compared with profiling measures such as log-likelihood, burstiness is a "self-contained" measure, because it does not need a reference corpus to be calculated, and the top-ranked bursty words can be easily matched against a gold standard representing the target domain.

Assessment of Bursty Words against Gold Standards

Burstiness was calculated separately for *eCare_ch_sv_01* and for *eCare_uc_sv_02*. For each corpus, we sorted the burstiness values by decreasing order and we took the top 2105 bursty words for *eCare_ch_sv_01* (Santini et al., 2018) and the top 21028 bursty words for *eCare_uc_sv_02* (since *eCare_uc_sv_02* is about 10 times larger than *eCare_ch_sv_01*) and matched them against the two gold standards that were described above. We used several metrics to assess the results, namely: intersection, percentage, precision@, Jaccard and Dice coefficients. For precision@ we use two cut-off points, i.e. 2105 for *eCare_ch_sv_01* and 21028 for *eCare_uc_sv_02*.

	Inter	%	Precision@	Jaccard	Dice
ch_sv_01	93	58.1%	0.0359	0.0427	0.0819
uc_sv_02	183	73.7%	0.0111	0.0086	0.0172

Table 5. Assessment of bursty words against gold standards.

Results are shown in Table 5, which reports the intersection between the top-ranked scores and the gold standard (col.2), percentage (col. 3), precision@ (col. 4), Jaccard coefficient (col.5), and Dice coefficient (col. 6). The size of the intersection and the percentage give an intuitive understanding of the overlap between the top-ranked bursty words and the target domains stored in the gold standards. The intersections amount to 58.1% for *eCare_ch_sv_01* and 73.6% for *eCare_uc_sv_02*. We think these figures are promising because when we measured the bursty words extracted from the Swedish National Corpus (called Stockholm-Umeå Corpus or SUC), the intersection with *gold_eCare_ch_sv_01* amounted to one occurrence (Santini et al., 2018), as shown in Table 6.

	Intersection	Jaccard	Dice	Precision@2105
SUC	1	0.000440	0.00088	0.00001

Table 6. Intersection between SUC bursty words and *gold_eCare_ch_sv_01*.

It is also worth noting that burstiness seems to be robust to corpus size variation since we observe that the number of domain-specific words identified increases with the size of the corpus rather than dropping. Apparently, the values of precision@ and those of Dice and Jaccard coefficients do not make justice to the magnitude of the overlap since their calculation takes into account the number of unmatched items, which in our case are many because the gold standards are much shorter than the lists of top-ranked bursty words.

Discussion

Results show that burstiness and the extent to which words with a higher burstiness overlap with gold standards (i.e. reference lists comprising domain-specific vocabulary) can be used to profile and quantify the domain specificity of a (web) corpus. As stated earlier, the burstiness of a word indicates the extent to which its frequency is unevenly distributed across documents within a specialized web corpus. This characterization fits very well the web corpora used in these experiments where domain-specific medical terms appear only in some documents and are not evenly distributed in all the documents of the corpus. We find these results auspicious because burstiness has the potential to "discover" and bring to the surface words that are important and domain- specific, but that could be missed out by other metrics, like log-likelihood, because they are distributed unevenly across a corpus (see also results in Santini et al., 2018). In a situation like this one, also a measure like perplexity, an evaluation metric that is often used to evaluate language models and that is also employed to assess domain adaptation in NLP tasks, could give misleading results, because it can be biased by the number of "unpredictable" bursty words.

In Experiment 2, many bursty words match the gold standards. This is encouraging because burstiness seems to capture the way in which content is distributed in this kind of web corpora. We observe that an intersection of 93 words out of the 160 unigrams listed in

gold_eCare_ch_sv_01 (58.1%) indicates that about 8% of the 2015 top-ranked bursty words belong to the fine-grained domain of 155 SNOMED CT chronic diseases. An intersection of 183 words out to the 248 unigrams listed in *gold_eCare_uc_sv_02* (73.7%) indicates that about 1.2% of the 21028 top-ranked bursty words belong to the coarse-grained domain extracted from eCare use cases (see Table 7).

	eCare_ch_sv_01	***eCare_uc_sv_02***
Coprus size (words)	700 000	7 000 000
Corpus Seeds	155 chronic diseases. Ex: [...] lungemfysem mycetom ozena polyserosit postkardiotomisyndrom Swimmingpooldermatit trumhinneatelektas adhesiv mediaotit aktinomykotiskt mycetom [...]	160 triplets extracted from uses cases. Ex: [...] tremor "parkinsons sjukdom" styrka skakningar "parkinsons sjukdom" styrka urinanalys "fynd som rör viktförlust" lever koma "akut exacerbation av kroniskt obstruktiv luftvägssjukdom" hjärtfrekvens "angina pectoris" angina balanserad uppföljningsstatus insulin "skyddat boende" långtidssyrgasbehandling medicin sarkoidos blåsa "diabetes mellitus typ 2" riskbedömning pillerask "mätning av fysiologiska parametrar" sjukgymnast [...]
Gold Standard (GS)	160 unigrams	248 unigrams
Top-ranked bursty words (BW)	2105 unigrams	21028 unigrams
Intersection: GS & BW	93 words out of 160 (58.1%)	183 words out of 248 (73.7%)
% of BW in GS	approx. 8% [2105:100=160:x]	approx. 1.2% [21048:100=248:x]
% of BW in Intersection	approx. 4% [2105:100=93:x]	approx. 0.9% [21248:100=183:x]

Table 7. Summary table

At this stage of research, we do not make any assumption about the minimum size of intersection that would account for a certain domain granularity, since we need further investigations to find a more principled approach to assess the relation between the size of the corpus, the length of the gold standards, and the cut-off points.

Open Issues

Research on the quantification of domain granularity of corpora bootstrapped from the web is still at the outset and several issues need to be further discussed and investigated.

Domain granularity: we put forwards two working definitions, namely "fine-grained domain" means bootstrapped with specialized medical terms, and "coarse-grained domain" means bootstrapped with both specialized medical terms and more general words.

Evaluation: the quantification using the intersection and percentage is more intuitive than precision@, Jaccard and Dice coefficients. However, further experimentation is needed to establish a balanced and principled relation between the size of the corpus, the length of the gold standards, and the cut-off points.

Cut-off points: the decision about the cut-off points was based on a rule of thumb, but in the future we would rather find more theoretically-grounded threshold settings, for example, the statistical significance of the burstiness scores.

Gold standards: the design of the gold standards is exploratory rather than principled. Discussion with domain experts is ongoing.

Last but not least, here we focus on lexical items because words are easy to pre-process. However, domain specificity certainly includes other aspects, such as special syntactic constructs, stance or sublanguage variations.

Experiment 2: Conclusion

In this experiment, we explored whether burstiness is a suitable measure to profile and quantify domain specificity both for small and large specialized web corpora with different domain granularities. Results show that burstiness can provide an indication of domainhood. We find these results promising because burstiness has the potential to discover terms that are domain-specific but evenly distributed, in a corpus and could easily be ignored by other statistical measures.

However, some open issues need to be further investigated, such as the need for more appropriate evaluation metrics, the quest of less empirical cut-off points, and a more principled design of the gold standards.

Conclusion and Future Work

In this article we have presented two experiments where we have explored measures to assess domainhood in web corpora. Domainhood indicates the degree of domain specificity of a specialized corpus.

In Experiment 1 (English corpora), we have profiled two specialized medical web corpora against a general-purpose web corpus using well-established measures, such as rank correlation, KL divergence and log-likelihood. These measures provide an indicative idea of domain specificity and allow us to assess whether a corpus bootstrapped from the web is satisfactorily domain-specific or whether it needs some amends before being used for linguistic studies or LT applications.

In Experiment 2 (Swedish corpora), we have used burstiness to identify domain-specific terms and lexicon that could give hints about the domain granularity of a corpus. Results show that burstiness can give an indication of the domainhood of a web corpus about diseases, since it helps ferret out terms that are domain-specific, but that could be ignored because of their uneven distribution.

The statistical measures that have been tried out in the experiments seem to be language-independent, since they give similar results for English and Swedish.

We are currently planning several follow-up studies that include comparative experiments between burstiness, perplexity, TF, TF-IDF and topic models on several (web) corpora characterized by different word frequency distributions (e.g. poisson mixtures).

TF can be simply the raw count of a term in a document or it can be the results of different types of normalization. See Wikipedia article about TF-IDF <https://en.wikipedia.org/wiki/Tf-idf>. Retrieved 25 March 2019.

For a different definition of DT, see <https://nlp.stanford.edu/IR-book/html/htmledition/inverse-document-frequency-1.html>. Retrieved 25 March 2019.

See Wikipedia article about Precision and Recall <https://en.wikipedia.org/wiki/Precision_and_recall>. Retrieved 25 March 2019.

See Wikipedia article about F1 <https://en.wikipedia.org/wiki/F1_score>. Retrieved 25 March 2019.

"Probability smoothing is a language modeling technique that assigns some non-zero probability to events that were unseen in the training data. This has the effect that the probability mass is divided over more events, hence the probability distribution becomes more smooth.". (Hiemstra, 2009)

See Wikipedia article about TF-IDF <https://en.wikipedia.org/wiki/Tf-idf>. Retrieved 25 March 2019.

References

Barbaresi, A. (2015). Ad hoc and general-purpose corpus construction from web sources. Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Linguistics. École Normale Supérieure de Lyon (Université de Lyon), France.

Baroni, M. & Bernardini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. LREC 2004 - Fourth International Conference On Language Resources And Evaluation.

Baroni, M., Kilgarriff, A., Pomikálek, J., & Rychlý, P. (2006). WebBootCaT: instant domain-specific corpora to support human translators. In Proceedings of EAMT (pp. 247-252).

Biemann, C., Bildhauer, F., Evert, S., Goldhahn, D., Quasthoff, U., Schäfer, R., ... & Zesch, T. (2013). Scalable Construction of High-Quality Web Corpora. JLCL, 28(2), 23-59.

Church, K. W., & Gale, W. A. (1995). Poisson mixtures. Natural Language Engineering, 1(2), 163-190.

Ciaramita, M., & Baroni, M. (2006). A Figure of Merit for the Evaluation of Web-Corpus Randomness. EACL 2006 - 11th Conference of the European Chapter of the Association for Computational Linguistics.

Cockburn, A. (2000). *Writing effective use cases, The crystal collection for software professionals*. Addison-Wesley Professional Reading. (24th printing, 2012).

Desagulier, G. (2017). *Corpus Linguistics and Statistics with R*. Springer.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, vol. 19, no. 1.

Ferraresi, A., Zanchetta, E., Bernardini, S. & Baroni, M. (2008). Introducing and evaluating ukWaC, a very large Web-derived corpus of English. Proceedings of the 4th Web as Corpus Workshop (WAC-4) "Can we beat Google?".

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In LREC (Vol. 29, pp. 31-43).

Gries, S. Th. (2013). Elementary statistical testing with R. In Manfred Krug & Julia Schlüter (eds.), *Research methods in language variation and change*, Cambridge University Press.

Hiemstra D. (2009) Probability Smoothing. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA.

Irvine, A., Callison-Burch, C.: A comprehensive analysis of bilingual lexicon induction. Computational Linguistics 43(2), 273{310 (2017)

Katz, S.M. (1996). Distribution of content words and phrases in text and language modelling.*Natural Language Engineering, 2*(1).

Kilgarriff, A. (2001). Comparing corpora. *International journal of corpus linguistics*, Vol.6(1).

Kilgarriff, A., Reddy S., Pomikálek J. and PVS A. (2010). A corpus factory for many languages.Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).

Lüdeling, A., Evert, S., & Baroni, M. (2007). Using web data for linguistic purposes. *Language and Computers, 59*, 7.

Nazarenko, A., & Zargayouna, H. (2009). Evaluating term extraction. International Conference Recent Advances in Natural Language Processing (RANLP'09).

Oakes, M. P. (2008). Statistical measures for corpus profiling. In Proceedings of the Open University Workshop on Corpus Profiling.

Pazienza, M., Pennacchiotti, M., & Zanzotto, F. (2005). Terminology extraction: an analysis of linguistic and statistical approaches. In *Terminology Extraction: An Analysis of Linguistic and Statistical Approaches*. Springer.

Pierrehumbert, J.B.: Burstiness of verbs and derived nouns. In: Shall We Play the Festschrift Game?, pp. 99-115. Springer (2012)

Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. *Proceedings of the workshop on Comparing Corpora.*

Santini M., Jönsson A., Nyström M. and Alirezai M. (2017). A Web Corpus for eCare: Collection, Lay Annotation and Learning - First Results. Workshop LTA'17 (Language Technology Applications 2017) co-located with FedCSIS 2017, Prague. In M. Ganzha and L. Maciaszek and M. Paprzycki (eds), Position Papers of the 2017 Federated Conference on Computer Science and Information Systems, Proceedings, Vol. 12. pp. 71-78.

Santini M., Strandqvist W., Nyström M., Alirezai M. & Jönsson A. (2018). "Can we Quantify Domainhood? Exploring Measures to Assess Domain specificity in Web Corpora". TIR 2018 - 15th International Workshop on Technologies for Information Retrieval.

Schäfer, R., & Bildhauer, F. (2013). Web corpus construction. *Synthesis Lectures on Human Language Technologies, 6*(4), 1-145.

Schäfer, R., Barbaresi, A., & Bildhauer, F. (2013). The good, the bad, and the hazy: Design decisions in web corpus construction. Proceedings of the 8th Web as Corpus Workshop.

Sharoff, S. (2017). Know thy corpus! Exploring frequency distributions in large corpora. In Diab, M., Villavicencio, A. (eds.) Essays in Honor of Adam Kilgarriff. Text, Speech and Language Technology Series, Springer (2017).

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation, 60*(5), 503-520.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Vol.1.

Wong W., Liu W., and Bennamoun M. (2011). Constructing specialized corpora through analysing domain representativeness of websites. *Language resources and evaluation*, Vol.45(2).

Acknowledgements

This research was supported by E-CARE@HOME, a "SIDUS – Strong Distributed Research Environment" project, funded by the Swedish Knowledge Foundation. Project website: <http://ecareathome.se/>