Profiling Domain Specificity of Specialized Web Corpora using Burstiness Explorations and Open Issues

Marina Santini, Wiktor Strandqvist, Arne Jönsson

RISE Research Institutes of Sweden
marina.santini@ri.se, wiktor.strandqvist@gmail.com, arne.jonsson@ri.se

Abstract

In this paper we describe an approach to profile the domain specificity of specialized web corpora in Swedish. The proposed approach is based on burstiness. Burstiness is a statistical measure that identifies words with uneven distribution across the documents of a corpus. We apply burstiness to two medical web corpora that have different size and different domain granularity. Results are promising and show that burstiness is an appropriate measure to profile the domain specificity when matched against reference lists (gold standards) that represent the target domains. However, further research is needed to find adequate evaluation metrics, less empirical cut-off points and more principled gold standard design.

1. Introduction

Web corpora are valuable textual resources widely exploited in Language Technology. Leveraging on the web for corpus creation is a well-established idea because bootstrapping corpora from the web is fast and inexpensive. 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. Arguably, traditional corpora and web corpora are complementary and allow for a wide spectrum of possible linguistic, empirical and computational studies and experiments. 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 mine of linguistic and textual varieties.

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 paper, 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 parameters, for instance corpus balance, corpus representativeness, corpus quality, corpus size, and similar. In this complex scenario, we single out one aspect, namely domain specificity, and test whether a statistical measure like burstiness can help profile and quantify it given a reference domain. The long-term goal is to find a suitable metric 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.

Here ”domain” is defined as the ”subject field” or ”area” in which a web document is used. Domain specificity, a.k.a domainhood (Santini et al., 2018), refers to the domain representativeness of a corpus. For instance, a high frequency of medical terms is a sign that a corpus is a specialized medical corpus. However, a domain might have different granularities. 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”. Previous experiments showed that burstiness is a promising measure for the profiling and quantification of domain specificity (Santini et al., 2018). Burstiness is attractive for three main reasons. First, it helps identify words that are frequent in certain documents, but that are unevenly distributed in the corpus as a whole. This characterization is suitable for many specialized web corpora, where domain-specific terms are discussed in some of the documents, but not in all of them. Second, it is a measure based on word frequencies, so it requires very little pre-processing and can be applied to any language. Third, it is easy to understand and implement, since: ”Burstiness is like the mean but it ignores documents with no instances” (Church and Gale, 1995).

2. Previous Work

The importance of a quantitative evaluation of corpora has been stressed for a long time (Kilgarriff, 2001). Although many researchers have worked on the design and assessment of web corpora, no standard metrics have been agreed upon to date.

Currently, research is available on the evaluation of general-purpose web corpora. For instance, Schäfer et al. (2013) focus on the quality of texts, Ciaramita and Baroni (2006) on the representativeness of a web corpus when compared to a traditional corpus, Eckart et al. (2012) highlight the importance of standardized preprocessing steps, and Kilgarriff et al. (2014) show how to evaluate a web corpus for a specific task, namely a collocation dictionary.

Corpora can be assessed according to several criteria.

Domain, genre, style, register, medium, etc. are well-known aspects that affect corpus representativeness. Here we focus on the quality of "domain" and explore ways to profile and quantify domain-specific web corpora. Our aim is somewhat similar to SPARTAN, a technique for constructing specialized corpora from the web by systematically analysing website contents (Wong et al., 2011). However, our purpose is not to analyze the domain-specificity of individual websites as a whole, rather we focus on web pages about chronic diseases retrieved from several web sites by search engines. In recent experiments (Santini et al., 2018), we presented a case study where we explored the effectiveness of different measures - namely the Mann-Withney-Wilcoxon Test, Kendall correlation coefficient, Kullback–Leibler divergence, log-likelihood and burstiness - to assess domainhood. Our findings indicated that burstiness was the most suitable measure to single out domain-specific words. In the next sections, we apply burstiness to two medical web corpora of different size and different domain granularity.

3. Specialized Web Corpora and Domain Granularity

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 (see Section 3.1). The target domains are represented by reference lists (see Section 3.2).

3.1 Same Domain, Different Granularities

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 automatically extracted from the E-care@home's project use cases, i.e. narratives that describe chronic diseases that affect the elderly.

eCare_ch_sv_01 was built using 155 terms listed in SNOMED CT, Swedish edition indicating chronic diseases as seeds. The 155 terms 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).

eCare_uc_sv_02 was created more recently using seed terms automatically extracted from the use cases of the E-care@home project. These use cases describe the chronic ailments that affect the elderly and the recommended treatments. The size of this corpus is approx. 7 million words (6 942 193 tokens). *eCare_uc_sv_02* is, thus, about 10 times larger than *eCare_ch_sv_01* and we use it here for the first time.

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 study, "fine-grained domain" means a very specialized domain where the seeds to bootstrap the corpus are specialized medical terms, e.g. "artrit" (en: arthritis), while "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 (see Section 3.2). Both web corpora were bootstrapped and downloaded with BootCat (Baroni and Bernardini, 2004), which is currently based on Bing or Google. Using regular search engines (like Google, Yahoo or Bing) and seeds to build a corpus is handy, but it also has some caveats that depend on the design or distortion of the underlying search engine (Wong et al., 2011). These caveats affect the content of web corpora since it might happen that irrelevant documents are included in the collection, especially when searching for very specialized terms. Since manual and qualitative inspections are often prohibitive, the automatic assessment of the domain specificity of a corpus bootstrapped from the web is potentially very useful.

3.2 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 (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 domain-representative. 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 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.

4. Burstiness

Burstiness indicates "how peaked a word's usage is over a particular corpus of documents" (Pierrehumbert, 2012) and helps identify words that are important in certain documents, but that are "unevenly distributed in the corpus as

a whole" (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 in some kind of web corpora, like the eCare corpus.
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:

$$B_w = \frac{\sum_{d_i \in D} rf_{w_{d_i}}}{df_w} \quad (1)$$

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 ignoring the documents where the word does not appear (Church and Gale, 1995; Katz, 1996).

Burstiness differs from measures like TF (Term Frequency) – which is simply the frequency of occurrence of a word normalized by document length – and TF*IDF where the TF is normalized by IDF (Inverse Document Frequency), which takes the log of the total number of documents in a corpus (irrespective of the presence or absence of the word *w*) divided by the number of documents containing the word *w*. If compared with more traditional 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.

5. Experiments

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 in Section 3.2. 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*.

Table 1: Assessment of bursty words against gold standards

	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

Results are shown in Table 1, 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 show a promising 58.1% for *eCare_ch_sv_01* and 73.6% for *eCare_uc_sv_02*. It is also encouraging to note 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 the two 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.

5.1 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 to what extent 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. We find these results promising because burstiness has the potential to "discover" and bring to the surface words that are important and domain specific, but that are distributed unevenly across a corpus. 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. In this situation, a measure like perplexity, an evaluation metric used to evaluate language models and often also to assess domain adaptation in NLP tasks, could give misleading results, because of the number of "unpredictable" bursty words.
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. 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.

5.2 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: in this study, 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, in these experiments 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.

6. Conclusion and Future Work

In this paper, 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 gives a good indication of the domainhood. We find these results promising because burstiness has the potential to discover terms that are domain specific, but that are not 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.
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). In the future, we plan to use burstiness not only to assess domainhood, but also for document indexing, terminology induction and for removing irrelevant documents from a web corpus.

Acknowledgements

We thank the reviewers for their useful comments. 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/ Lists of seeds, gold standards and other material is available at: http://santini.se/eCareCorpus/home.htm

References

M. Baroni and S. Bernardini. 2004. Bootcat: Bootstrapping corpora and terms from the web. In *LREC*.

Kenneth W Church and William A Gale. 1995. Poisson mixtures. *Natural Language Engineering*, 1(2):163–190.

M. Ciaramita and M. Baroni. 2006. A figure of merit for the evaluation of web-corpus randomness. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*.

Thomas Eckart, Uwe Quasthoff, and Dirk Goldhahn. 2012. The influence of corpus quality on statistical measurements on language resources. In *LREC*, pages 2318–2321. Citeseer.

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

S. M Katz. 1996. Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, 2(1):15–59.

Adam Kilgarriff, Pavel Rychlỳ, Milos Jakubicek, Vojtech Kovár, Vit Baisa, and Lucia Kocincová. 2014. Extrinsic corpus evaluation with a collocation dictionary task. In *LREC*, pages 545–552.

A. Kilgarriff. 2001. Comparing corpora. *International journal of corpus linguistics*, 6(1):97–133.

A. Kilgarriff. 2005. Language is never, ever, ever, random. *Corpus linguistics and linguistic theory*, 1(2):263–276.

T. Lippincott, D. Ó Séaghdha, and A. Korhonen. 2011. Exploring subdomain variation in biomedical language. *BMC bioinformatics*, 12(1):212.

J. Pierrehumbert. 2012. Burstiness of verbs and derived nouns. In *Shall We Play the Festschrift Game?*, pages 99–115. Springer.

M. Santini, A. Jönsson, M. Nyström, and M. Alireza. 2017. A web corpus for ecare: Collection, lay annotation and learning. first results. In *Proceedings of the 2nd International Workshop on Language Technologies and Applications (LTA17)*. FedCSIS.

M. Santini, W. Strandqvist, M. Nyström, M. Alirezai, and A. Jönsson. 2018. Can we quantify domainhood? exploring measures to assess domain-specificity in web corpora. In *International Conference on Database and Expert Systems Applications*, pages 207–217. Springer.

R. Schäfer, A. Barbaresi, and F. Bildhauer. 2013. The good, the bad, and the hazy: Design decisions in web corpus construction. In *8th Web as Corpus Workshop*, pages pp–7.

S. Sharoff. 2017. Know thy corpus! Exploring frequency distributions in large corpora. In Mona Diab and Aline Villavicencio, editors, *Essays in Honor of Adam Kilgarriff*. Text, Speech and Language Technology Series, Springer.

W. Wong, W. Liu, and M. Bennamoun. 2011. Constructing specialised corpora through analysing domain representativeness of websites. *Language resources and evaluation*, 45(2):209–241.