Towards a Quality Assessment of Web Corpora for Language Technology Applications

Wiktor Strandqvist
RISE Research Institutes of Sweden & Linköping University

Marina Santini
RISE Research Institutes of Sweden

Leili Lind,
RISE Research Institutes of Sweden & Department of Biomedical Engineering & Medical Informatics, Linköping University

Arne Jönsson
RISE Research Institutes of Sweden & Linköping University

Abstract

In the experiments presented in this paper we focus on the creation and evaluation of domain-specific web corpora. To this purpose, we propose a two-step approach, namely the (1) the automatic extraction and evaluation of term seeds from personas and use cases/scenarios; (2) the creation and evaluation of domain-specific web corpora bootstrapped with term seeds automatically extracted in step 1. Results are encouraging and show that: (1) it is possible to create a fairly accurate term extractor for relatively short narratives; (2) it is straightforward to evaluate a quality such as domain-specificity of web corpora using well-established metrics.

Keywords: corpus evaluation, term extraction, log-likelihood, rank correlation, Kullback-Leibler distance.

1. INTRODUCTION

Creating reliable domain-specific web corpora to be used in language technology (LT) applications can be daunting, knowing that the reliability and the performance of the final applications depend on the quality and appropriateness of the underlying corpora. Web corpora are normally used in many LT applications such as ontology creation from texts, paraphrase detection, lay-specialized terminology and so on. In our study, we investigate

a practical approach to create and validate the quality of domain-specific corpora bootstrapped from the web.

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. These texts are narratives that describe a "system's behavior under various conditions as the system responds to requests from stakeholders" (Cockburn, 2000) and 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 SNOMED CT terms. 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* (Baroni and Bernardini, 2004) a medical web corpus and evaluate its quality. Leveraging on the web to create corpora is a well-established idea

(e.g. Baroni & Bernardini, 2004; Kilgarriff et al. 2010). However, while bootstrapping a web corpus is considered to be common practice, corpus evaluation is still a grey area. Currently, there is little research available on this topic and approaches are not standardized, so it is not possible to compare results. In our study, we analyse and test three corpus profiling measures, namely rank correlation (Kendall and Spearman), Kullback–Leibler Divergence, and log-likelihood. *The challenge of this step is to find empirical answers to the following questions:*

1. *What is meant by "quality" of a web corpus?*
2. *How can we assess the quality of a corpus automatically bootstrapped from the web?*
3. *What if a bootstrapped web corpus contains documents that are NOT relevant to the target domain?*
4. *Can we measure the domain-specificity of a corpus*?

2. 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", that we define as the "subject field" or "area" 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 domain-specificity quality, we rely on measures that

are well-established and easy to replicate. Since in this paper we describe comparative experiments based on rank correlation (Kendall and Spearman), Kullback-Leibler distance and log-likelihood, in this section we provide a short overview of studies where these measures were used.

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 Kullback-Leibler distance to assess the"randomess" or "unbiasedness" of general-purpose corpora. They compare domain-specific sub parts of the BNC against the whole BNC corpus and show that KL divergence can reliably indicate the difference between general purpose corpora (random and unbiassed) and domain-specific corpora (biassed).

3. *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 to SNOMED CT to produce candidate terms. The third component is a *seed validator*.

The *terminology extractor* uses 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 syntactic patterns (Pazienza, Pennacchiotti, & Zanzotto, 2005) presented in Table 1.

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

Table 1. Syntactic patterns used for term recognition

The *terminology validator* takes the candidate terms produced in the previous step 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 word, to allow for spelling variations. Matches are then ranked by DF/IDF scores (cutoff = 200).

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

3.1 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. When evaluated, the *terminology validator* achieves the following performance: Precision = 34.2%, Recall = 71%, F1 = 46.2%.

	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 quality of the resulting web corpus. This means that our approach is effective and help create a domain-specific corpus without any manual intervention.

4. CORPUS EVALUATION METRICS

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

3) 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).

4) Log-Likelihood (LL-G^2): LL-G^2 (Dunning, 1993) has been used for corpus profiling (Rayson and Garside, 2000). The words that have the largest LL-G^2 scores show the most significant word-frequency difference in two corpora. LL-G^2 is not affected by corpus size variation.

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

1. **ukWaCsample** (872 565 words): a random subset of ukWaC, a general-purpose web corpus (Ferraresi et al., 2008).
2. **Gold** (544 677 words): a domain-specific web corpus collected with hand-picked term seeds from the *E-Care* personas and use cases/scenarios.
3. **Auto** (492 479 words): a domain-specific web corpus collected with automatically extracted term seeds from the *E-Care* personas and use cases/scenarios (see Section 3).

4.1. 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.

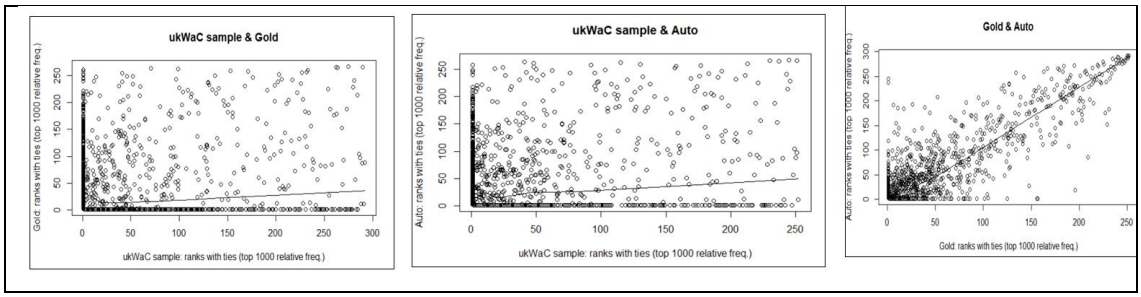

Fig. 1 Plotting 1000 top ranks: (from left to right): ukWaCsample and Gold (box 1), ukWaCsample and Auto (box2), and Gold and Auto (box 3).

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.

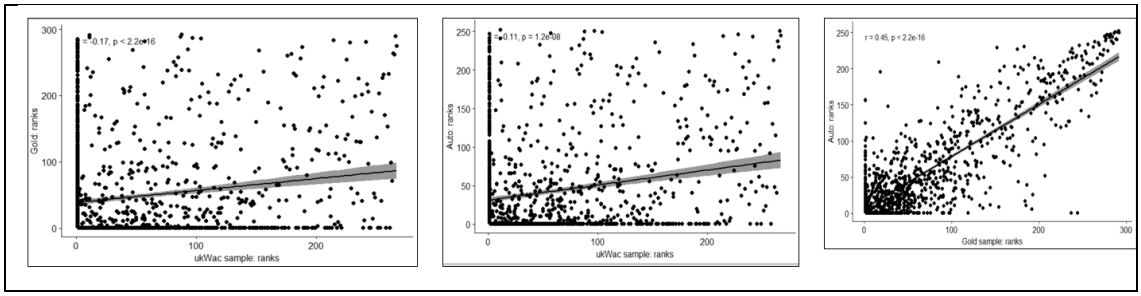

Fig. 2 Kendall Tau: (from left to right):): ukWaCsample and Gold (box 1), ukWaCsample and Auto (box2), and Gold and Auto (box 3).

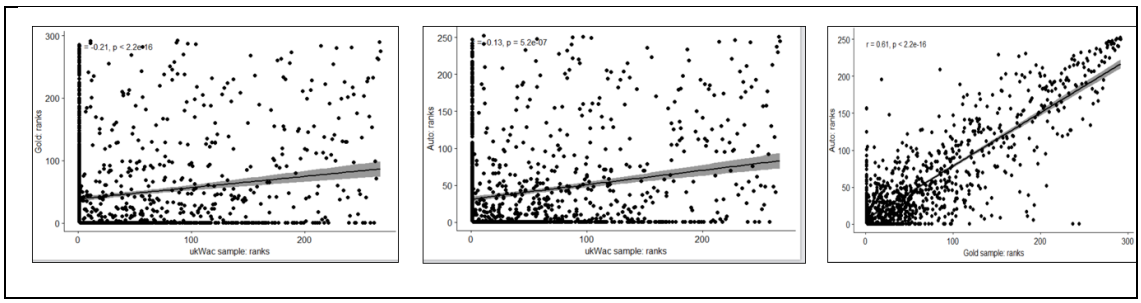

Fig. 3 Spearman Rho: (from left to right):): ukWaCsample and Gold (box 1), ukWaCsample and Auto (box2), and Gold and Auto (box 3).

Kullback-Leibler Divergence. Before calculating KL divergence, a smoothing value of 0.01 was been added to the normalized frequencies. Results are shown in Table 2. 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 2. KL scores

Log-Likelihood (LL-G2). We computed LL-G^2 scores on smoothed word frequencies. The total LL-G2 scores for the three web corpora (top 1000 words) are shown in Table 3. The larger the LL-G^2 score of a word, the more different its distribution in two corpora. The large LL-G^2 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 LL-G^2 score returned for Gold vs Auto (114 694.2), which suggests that Gold and Auto are more similar to each other.

Corpora	**Total LL-G^2 scores**
ukWacSample vs Gold	453 441.6
ukWacSample vs Auto	393 705.9
Gold vs Auto	114 694.2

Table 3. LL-G^2 scores of the three corpora

It is also possible to assess the statistical significance of the individual LL-G^2 scores. Normally, a LL-G^2 score of 3.8415 or higher is significant at the level of $p<0.05$ and a LL-G2 score of 10.8276 is significant at the level of $p<0.001$ (Desagulier, 2017). Fig. 4 shows the breakdown of the top-ranked LL-G2 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 LL-G2 scores in box 3 are much smaller in magnitude, which indicates that the difference between words is less pronounced.

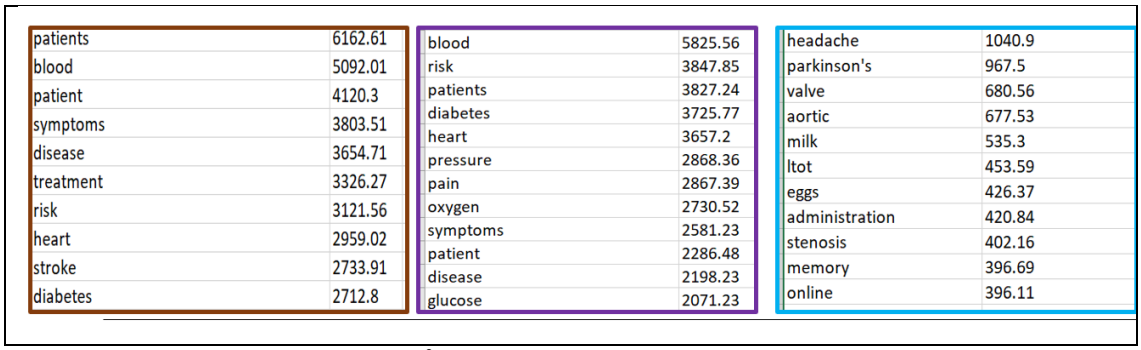

Fig. 4. Top-ranked LL-G^2 scores (from left to right): ukWaCsample and Gold (box 1), ukWaCsample and Auto (box 2), and Gold and Auto (box 3).

5. CONCLUSION

We have shown that it is possible to create a fairly accurate term extractor 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 quality is similar to a corpus bootstrapped with hand-picked term seeds. This is an added value because corpus construction can be accelerated and standardized.

We have seen that well-established measures − such as rank correlation, KL divergence and log-likelihood (LL-G2 scores) − DO give a coarse but grounded idea of domain-specificity. Essentially, they allow for an evaluation of the quality of a domain-specific web corpus and can also be used to pre-assess the portability of NLP tools from one domain-specific corpus to a different corpus belonging to another domain. 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 (intrinsic evaluation metrics).

We can now provide empirical answers to the questions asked in the Introduction. Namely, (1) in these experiments, "quality" means high density of medical terms related to certain illnesses described in the personas and use cases/scenarios; (2) we can assess the quality of a corpus automatically bootstrapped from the web by using metrics that are well-established and easily replicable; (3) we can get a coarse but robust indication of the similarities across corpora; (4) we can measure the domain-specificity of a corpus and assess whether it is satisfactorily domain-specific or whether the corpus needs some amends before being used for LT applications.

REFERENCES

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

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?".

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.

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).

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.

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

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., 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.

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 [KK-STIFTELSEN, Diarienr: 20140217]. Project website: <http://ecareathome.se/>.