Unconventional Usage of Entropy in the Field of Web Usage Data Preprocessing and Machine Translation Evaluation

Michal Munk and Ľubomír Benko

Abstract This paper focuses on an unconventional usage of entropy. On one side it deals with preprocessing phase, especially the session identification using the Reference Length method. Entropy, in this case, offers an alternative to determining the ratio of auxiliary pages that is important for this method. With the approach introduced in this paper, the need of a sitemap becomes void. On the other hand, the paper looks at entropy in the case of reliability analysis of Machine Translation metrics. In this case, entropy offers also an alternative mean to validate the metrics.

Keywords Entropy · Data preprocessing · Reference length · Machine translation

1 Introduction

Entropy offers a lot of possibilities to many fields of study. This paper focuses on Information Entropy introduced by Shannon and its unorthodox usage in two fields—Web Usage Mining and Machine Translation evaluation. Entropy can be used as a measure of disorder, where lower entropy means order and higher entropy on the contrary disorder. Following the definitions by Shannon [1], entropy can be used as a measure of uncertainty in a data set.

The rest of the paper is structured as follows: in Sect. 2 is summarized the related work of other authors about entropy and maximum entropy. In the field of

M. Munk (✉) · Ľ. Benko
Faculty of Natural Sciences, Department of Informatics,
Constantine the Philosopher University, Nitra, Slovakia
e-mail: mmunk@ukf.sk

Ľ. Benko
e-mail: lubomir.benko@gmail.com

Ľ. Benko
Institute of System Engineering and Informatics,
University of Pardubice, Pardubice, Czech Republic

K. Ntalianis and A. Croitoru (eds.), *Applied Physics, System Science and Computers*, Lecture Notes in Electrical Engineering 428,
DOI 10.1007/978-3-319-53934-8_34

Web Usage Mining was the aim of the paper to create an alternative way to calculate the ratio of auxiliary pages for the session identification method Reference Length. This is dealt with in Sect. 3. To validate Machine Translation metrics were usually used Cronbach's alpha or Standardized alpha. Section 4 deals with another aim of the paper, to analyze entropy as an alternative mean to evaluate the metrics. Subsequently, the discussion and future work are offered in the last section.

2 Related Work

The first concept of entropy originates from thermodynamics [2], where it was used to provide a statement of the second law of thermodynamics on the irreversibility of the evolution, i.e. an isolated system cannot pass from a state of higher entropy to a state of lower entropy [3]. Shannon (1948) was the first to re-define entropy and mutual information, for this purpose he used a thought experiment to propose a measure of uncertainty in a discrete distribution based on the Boltzmann entropy of classical statistical mechanics [3]. Entropy can be described as a measure of the expected content of the information or uncertainty probability distribution. It is also described as the degree of disorder or randomness in a system. Based on Shannon's definition [1, 4], given a class random variable C with a discrete probability distribution $\{p_i = Pr[C = c_i]\}_{i=1}^{k}$, $\sum_{i=1}^{k} p_i = 1$ where c_i is the ith class. Then the entropy $H(C)$ is defined as $H(C) = -\sum_{i=1}^{k} p_i \log p_i$, while the function decreases from infinity to zero and p_i takes values from interval 0–1 [1, 4].

E.T. Jaynes formulated [5] the principle of Maximum Entropy and transformed that way entropy to a modeling tool. Maximum Entropy is used to estimate unknown parameters of a multinomial discrete choice problem, whereas the Generalized Maximum Entropy includes noise terms in the multinomial information constraints [3].

3 Entropy in the Field of Web Usage Data Preprocessing

Data preprocessing is a crucial part of Web Usage Mining and based on another experiment [6] especially session identification can prove important. Authors in [7] combine the maximum entropy model for the recommendation system. Their results showed that the recommendation system can achieve better accuracy than standard Markov model for page recommendation. It also provided a better interpretation of web users' navigational behavior. Authors in [8] focused on comparing the performance of maximum entropy with Naïve Bayes and Support Vector Machine, where the entropy outperformed both of them.

In the experiment [6] was the assumption about the ratio of auxiliary pages estimated for the session identification method Reference Length based on the sitemap. Based on the ratio can be determined the cutoff time $C = \frac{-\ln(1-p)}{\lambda}$ [6]. The

sitemap can offer an accurate estimate of the ratio but can be problematic if the examined web portal was changed in the meantime. In that case is the log file the only possible option to get the necessary data. One option would be to extract the sitemap from the log file using a complicated algorithm.

The experiment was conducted on a log file of a course of virtual learning environment (VLE) portal. The log file was prepared using standard data mining techniques same as in [9]. The final log file was imported to a database where were conducted several experiments involving the entropy, on the basis of which could be distinguished navigation pages from the content pages from one another. The aim was to create an algorithm that would be able to calculate entropy for a specific page on the basis of a random variable Length that represents the length of the time spent on each web page of the portal. With the use of the algorithm was created the variable Relative Time, which represented the relative time spent on the page by the user. From this variable was derived MaxEnt for each page and created a new data matrix (Table 1) that contained MaxEnt of each page. Subsequently was calculated the average length of all accesses on the web portal and served as the cutoff time that divides the pages to auxiliary and content pages. Pages with smaller MaxEnt than the MaxEnt of the average length of time spent across the whole portal will be classified as auxiliary pages. On the opposite pages with higher MaxEnt will be content pages.

In the log file of the course of VLE portal were identified 58 pages. Using the algorithm were 10 pages classified as auxiliary and the rest (48 pages) was classified as content pages (Table 2). Therefore the ratio of auxiliary pages of the web portal is 17.24%. Another option for specifying the time threshold of time spent on the web portal were quartiles. The time was calculated by $Q_{STT} = Q_{III} + 1.5Q$, where Q_{III} represents the upper quartile and Q represents the quartile range. Using the quartiles were 9 pages classified as auxiliary and 49 pages were identified as content pages. Compared to the calculation of the ratio of auxiliary pages based on the sitemap (15.34%) that was based on previous research [6], is the ratio calculated based on MaxEnt average time spent on the whole portal higher by almost 2%. But the ratio calculated based on MaxEnt quartiles of the time spent on the whole portal was almost similar (15.51%) to the sitemap calculation. In comparison to the subjective estimate of the ratio (25%), that was determined by the administrator of the web portal, the estimates calculated from the sitemap or MaxEnt offer more accurate results. The influence of more accurately calculated ratio of auxiliary pages in the session identification phase was described in more detail by authors in [6].

Table 1 Data matrix extended by maximum entropy

URL ID	Length	Relative Time	MaxEnt
58450	24	0.0039636	7.9789487
661	138	0.0001486	12.7156915
69022	48	0.0008319	10.5307314
69022	6	0.0001039	13.2311711
⋮	⋮	⋮	⋮

Table 2 Ratio of auxiliary pages based on different calculations

	Auxiliary pages	Content pages	Ratio (%)
Subjective estimate	–	–	25
Sitemap calculation	29	189	15.34
MaxEnt of average	10	48	17.24
MaxEnt of quartiles	9	49	15.51

Table 3 Statistics of automatic metrics of error rate

	Metrics-total correlation	Alpha if deleted	Metrics-total accuracy entropy
PER	0.878	0.934	0.934
WER	0.845	0.964	0.852
CDER	0.958	0.869	0.895

4 Entropy in the Field of Machine Translation Evaluation

Entropy can be used not only in the phase of preprocessing but also for analysis of reliability. Authors experiment with entropy also in the field of Machine Translation. Authors in [10] offer a complex survey of data selection methods in Machine Translation. They also describe articles that focused on cross-entropy which has become the most commonly used approach in data selection. Authors in [11] introduce a novel framework based on maximum entropy for word alignment. Based on the experiment the authors improved the alignment quality and translation quality as measured by standard reliability metrics.

PER, *WER*, and *CDER* metrics are called metrics of error rate, i.e. the higher values of these metrics, the lower the translation quality. On the other hand, metrics *Precision*, *Recall*, *F-measure*, and *BLEU*s are called metrics of accuracy, i.e. the higher values of these metrics, the better the translation quality. The automatic metrics defining Machine Translation error rate and representing the automatic evaluation of Machine Translation are considered highly reliable based on the direct estimation of reliability. The aim of this experiment was to assess the reliability of the automatic evaluation of machine translation for inflectional languages using traditional methods and entropy; in this case on-line statistical machine translation system was used.

All items (Table 3) correlate (*Avg inter-metrics correlation*: 0.885) with the total score of evaluation and after their eliminating the coefficient of reliability has not increased except the *WER* metric (*Cronbach's alpha*: 0.947; *Standardized alpha*: 0.950). After elimination of the metric *WER*, the coefficient of reliability—*Cronbach's alpha* increased from 0.947 to 0.964, but that it is negligible.

For the *entropy* calculation (Table 3), in the case of analysis of automatic metrics characterizing the error rate of Machine Translation evaluation, individual metrics in comparison over accuracy metrics were used. *Entropy* was calculated for each sentence analyzed using the specific metrics and for the comparison was used the

average entropy of all sentences. From the Shannon definition of entropy [1], if the *entropy* is closer to 1, then the system is more irregular. The results of the *entropy* for each of the error rate metric relate with the coefficient of reliability—*Cronbach's alpha*.

Even though the reliability analysis of automatic metrics characterizing the error rate of Machine Translation evaluation showed that the *WER* metric is the most deviated from the other automatic metrics of automatic Machine Translation evaluation (PER or CDER) in translation quality assessment. It is understandable seeing that, the *WER* metric is very strict to syntax errors (word order).

The estimations of the entropy of automatic metrics of accuracy were similarly calculated as in the case of metrics of error rate. Also, in this case, was used the average entropy of all sentences for each metric and the results relate with the coefficient of reliability—*Cronbach's alpha* with negligible variations. In the case of *entropy*, it also showed that the metric *BLEU-4* deviates the most from the other metrics.

We explain this by the fact that *BLEU-4* metric measures a score of a sequence of four words (including articles and prepositions) and sometimes it is very complicated to achieve an output with such a sequence by systems of machine translation.

5 Discussion and Future Work

As has been shown in this paper it is possible with the use of Maximum Entropy to divide the pages of the web portal to auxiliary and content pages. This can be used in the method Reference Length of session identification of data preprocessing of log files. The ratio of auxiliary pages for the Reference Length method can be then calculated with similar accuracy as if calculated from the sitemap. Since it is possible to work also with historical data, there could not be available an appropriate sitemap of the web portal of that time. Therefore the possibility to estimate the ratio of the auxiliary pages without the sitemap of the web portal is beneficial. An important role plays also a thoroughly made log file cleaning from unnecessary data because poorly cleaned log may generate an inaccurate ratio of auxiliary pages. The experiment was realized only on log file of a course of VLE portal, the future work could be focused also on web portals with anonymous accesses because of a bigger variance of the structure of such portals. Also these kind of portals (for example e-shops) contain bigger log files that could prove more difficult to analyze for the algorithm. Future work could be also focused on determining the size of the page content. It would be assumed that the high entropy would suggest the high proportion of the content of a particular page. That would inform the web portal administrator of content-rich pages. On basis of such information, the page could be divided into more pages with less content and thus improve browsing of the web portal for users. The principle could be based on research [12] which calculated page rank for each page of the web portal.

The next aim of the paper was focused on verifying the reliability of the automatic metrics for Machine Translation evaluation. Three different measures of reliability were used—*Cronbach's alpha*, *Standardized alpha,* and *entropy*—to estimate reliability. *Cronbach's alpha* and *Standardized alpha* were very similar, i.e. individual automatic metrics for Machine Translation evaluation have the same variability. The use of *entropy* provided alternative means to validate the reliability of metrics and the results relate to results of *Cronbach's alpha* and *Standardized alpha*.

Acknowledgements This work was supported by the Slovak Research and Development Agency under the contract No. APVV-14-0336 and Scientific Grant Agency of the Ministry of Education of the Slovak Republic (ME SR) and of Slovak Academy of Sciences (SAS) under the contracts No. VEGA-1/0559/14 and by the project No. UGA VII/3/2015 Modelling the behavior of web users depending on time.

References

1. Shannon, C.E.: A mathematical theory of communication. ACM SIGMOBILE Mob. Comput. Commun. Rev. **5**, 3 (2001)
2. Clausius, R.: On the Motive Power of Heat, and on the Laws which Can Be Deduced from It for the Theory of Heat. Dover (1960)
3. Holzinger, A., Hörtenhuber, M., Mayer, C., Bachler, M., Wassertheurer, S., Pinho, A.J., Koslicki, D.: On entropy-based data mining. In: Holzinger, A., Jurisica, I. (eds.) Interactive Knowledge Discovery and Data Mining in Biomedical Informatics: State-of-the-Art and Future Challenges, pp. 209–226. Springer, Berlin (2014)
4. Lima, C.F.L., de Assis, F.M., de Souza, C.P.: A Comparative Study of Use of Shannon, Rényi and Tsallis Entropy for Attribute Selecting in Network Intrusion Detection (2012)
5. Jaynes, E.T.: Information theory and statistical mechanics. Phys. Rev. **106**, 620 (1957)
6. Munk, M., Benko, Ľ., Gangur, M., Turčáni, M.: Influence of ratio of auxiliary pages on the pre-processing phase of Web Usage Mining. E+M Ekon. a Manag. **3,** 144–159 (2015)
7. Jin, X., Zhou, Y., Mobasher, B.: A maximum entropy web recommendation system. In: Proceeding of the eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining—KDD'05, p. 612. ACM Press, New York (2005)
8. Wang, H., Wang, L., Yi, L.: Maximum entropy framework used in text classification. In: 2010 IEEE International Conference on Intelligent Computing and Intelligent Systems, pp. 828–833. IEEE (2010)
9. Benko, Ľ., Reichel, J., Munk, M.: Analysis of student behavior in virtual learning environment depending on student assessments. In: ICETA 2015: 13th International Conference on Emerging eLearning Technologies and Applications, Stary Smokovec, November 26–27, 2015. pp. 33–38. IEEE, Stary Smokovec, Danvers (2015)
10. Eetemadi, S., Lewis, W., Toutanova, K., Radha, H.: Survey of data-selection methods in statistical machine translation. Mach. Transl. **29**, 189–223 (2015)
11. Tomeh, N., Allauzen, A., Yvon, F.: Maximum-entropy word alignment and posterior-based phrase extraction for machine translation. Mach. Trans. **28**, 19–56 (2014)
12. Kapusta, J., Munk, M., Drlík, M.: Analysis of differences between expected and observed probability of accesses to web pages. In: Hwang, D., Jung, J., and Nguyen, N.-T. (eds.) Computational Collective Intelligence. Technologies and Applications SE-68, pp. 673–683. Springer International Publishing (2014)