A Tool for Automatic Extraction of Information from Company Web Sites in the Field of International Servitization

Gábor Berend
University of Szeged
berendg@inf.u-szeged.hu

Christian Stadlmann
University of Applied Sciences Upper Austria
christian.stadlmann@fh-steyr.at

Stefan Mang
University of Passau
stefan.mang@uni-passau.de

Margarethe Überwimmer
University of Applied Sciences Upper Austria
margarethe.ueberwimmer@fh-steyr.at

Abstract

This paper introduces the first prototype of a tool developed as part of an interdisciplinary applied research effort. Our long-term goal is to develop such a service which is capable of providing valuable information aiding decision making regarding international servitization. Our platform is currently able to identify the industry for some company based on the contents of its web site. Key terms, including the extraction of multiword expressions – that we deem as relevant industry-specific terms – and named entities are also part of our demo application. The platform also incorporates a preliminary information retrieval component, which returns the most similar company websites included in our database of companies. Our demo is accessible from https://rgai.inf.u-szeged.hu/prosperAM/demo.

1 Introduction

There has been a recent interest in the computational processing and treatment of financial and business documents (Hahn, Hoste, and Tsai 2018; Chen et al. 2019; Hahn, Hoste, and Zhang 2019; Mahmoud El-Haj et al. 2019). Extracting automatically information from the myriad of semi,- and unstructured data sources – including corporate web sites – can provide a valuable source of information to decision makes in industry.

In this paper, we introduce the first working component of an online tool which will be integrated into a platform for providing information to companies planning to export their services to some foreign market. Our long term goal is to provide companies such a platform which can aid them in assessing their perspectives for exporting services to foreign markets. We think that a first step towards this goal is to become able to present companies their competitors. To this end, we developed an online tool which can categorize the business profile of a company based on its website, then also provides a ranked list of the most similar companies included in our database. Key industry-related terms – also extracted from the corporate websites – are currently also determined by our tool.

Due to the interdisciplinary nature of our project, we next provide a brief overview of the theory of international servitization, then detail our technical contributions.

2 International Servitization

In the last decades services provided by manufacturing companies have moved into the focus of both practitioners and researchers (Oliva and Kallenberg 2003; Baines, Lightfoot, and Smart 2011). Supplementing products with value enhancing services is seen as a clever manoeuvre to reach a competitive advantage (Gebauer, Gustafsson, and Witell 2011), profitability and economic stability (Bandinelli and Gamberi 2011). According to Service Dominant Logic additional services are central for direct value enhancement of the physical product, which are themselves viewed as a mere distribution mechanism for service delivery (Vargo and Lusch 2004). However, the success of servitizing manufacturers remains debated as research shows mixed results (Gebauer 2005; Wang, Lai, and Shou 2018). The existence of service business results even in higher bankruptcy risks for manufacturing firms (Benedettini, Neely, and Swink 2015). These failure risks are particularly increased if services are provided in different countries (Zarpelon Neto 2015). In international business further challenges emanate from outside the company as changes in technology, regulations, competition or customer demands. Moreover, failure can derive from wrong managerial decision-making as about local investments, disadvantageous contracts with customers or network partners or falling behind resident competitors (Benedettini, Neely, and Swink 2015).

Hence, when implementing services the performance of manufacturing companies may vary in different regions (Szász et al. 2017). Indeed, manufacturing firms with the same extent of servitization may achieve diverse performance levels in different geographic regions or if they operate in distinct industrial market sectors (Wang, Lai, and Shou 2018). Therefore, it may be concluded that the regional market conditions are decisive for success or failure in sertivitization (Baines et al. 2017). Above, the global competition has changed dramatically in the last decade. In 2007, globally more than 30 percent of manufacturing firms have already been servitized (Neely 2008). In 2018 though, a radi-

cal increase of servitization has been identified in many economic areas around the globe as in China where the number of servitized manufacturing firms increased from 17 percent to 38, which is similar to the German level of servitization (Mastrogiacomo, Barravecchia, and Franceschini 2019).

However, the understanding of this external context is still undeveloped in research and there is a strong need of a descriptive comprehension of the broader environment and social aspects of servitization as well as for prescriptive studies about the external conditions influencing the right time to implement or adapt a service strategy of manufacturing companies (Baines et al. 2017). From managerial point of view, a continuous monitoring of markets and their characteristics are fundamental for mitigating failure risks and recognizing potential market opportunities.

Finding an approach to successfully engage within a market is a difficult investigative and planning challenge due to its nature of an ill-structured problem Grünig (2017). Thus, two heuristic principles should be applied for solving this analytic problem, i.e. the rule of factorization breaking down the issue into sub-problems and the procedure of modelling for developing proven solution methods (Grünig 2018). Following Grünig (2017) it is out of purpose to examine a large number of country and industry market combinations. Therefore, the focus of this paper is specifically on the US market as it is one of the most promising regions for European manufacturing companies (International Trade Center 2019).

Moreover, various market characteristics are important for the success of servitization as the political, economic, competitive, social, technological, environmental, industrial, and regulatory components (Baines et al. 2017). Hence, the developed prototype specifically focuses on industry-specific and competitive elements which are starting points for a better understanding of the foreign market for servitization. This is done by automatic processing of company web sites which may be competitors, service or network partners or potential leads for servitization activities of manufacturing firms.

3 The technical contributions

We next describe the current functionalities of our platform and provide their technical implementation details.

3.1 Classifying websites according to SIC codes

Standard Industrial Classification (SIC) provides a fine-grained hierarchical categorization of industrial activities. A small sample of the entire categorization hierarchy can be seen in Table 1.

We used the `fasttext` (Joulin et al. 2016) Python package for predicting the most appropriate SIC code for a given company based on its corporate web site. In order to train our model, we relied on the 2017 version of a US business database[1] which contains information on several millions of companies, including their addresses, contact person names and most importantly, the URL of their website along with their SIC.

[1] https://www.uscompanieslist.com/

Table 1: A sample excerpt of the SIC categories

SIC code	Industry
1000	Metal Mining
1040	Gold and Silver Ores
1090	Miscellaneous Metal Ores
⋮	⋮
1731	Electrical Work
2000	Food and Kindred Products
2011	Meat Packing Plants
⋮	⋮

Not all companies have a website included in the database, furthermore a fair share of the URLs were no longer affiliated with the company any more, e.g. they were selling the domain instead of providing content related to its onetime corporate owner. In order to reduce the noise from the training procedure, we applied a handful of heuristics which made semi-automatic removal of non-relevant (URL, SIC) pairs possible. In the end we were left with approximately 400K corporate URLs that we could train our SIC-classifier on.

The websites that we used for training were processed only for their opening page, and the same applies for our model when it makes inference to unseen URLs.

3.2 Extracting industry terms

Multi-word expressions (MWEs) are such white space delimited sequences of tokens that bear some idiosyncratic behavior, such as *oil refinery* and *steel manufacturing*. Of course, not all MWEs are relevant from an industry point of view, e.g. *cold war* and *privacy policy*. Nonetheless we make the simplifying assumption that MWEs can be treated as such expressions which can be usefully applied to further characterize and refine company profiles beyond their SIC.

To this end, we trained an MWE detector based on the WIKI50 training dataset (Vincze, Nagy T., and Berend 2011). For determining the multiword expressions (aka. industry terms), we use the same fast conditional random field (CRF) based sequence classification architecture that we introduced earlier for POS tagging and named entity recognition in (Berend 2017). Our implementation is based on the highly-efficient CRFsuite package (Okazaki 2007).

3.3 Basic IR component

Our platform currently offers a basic information retrieval component as well. As mentioned already in Section 3.1, we have compiled a collection of corporate websites consisting of nearly 400K URLs. Upon determining the most likely SIC for a website entered, we also perform a quick ranking of those corporate websites that belong to the same SIC and return those companies which has the highest overlap to the queried company in terms of their industry terms.

A screenshot from our application can be seen in Figure 1. The application itself is publicly available from https://rgai.inf.u-szeged.hu/prosperAM/demo.

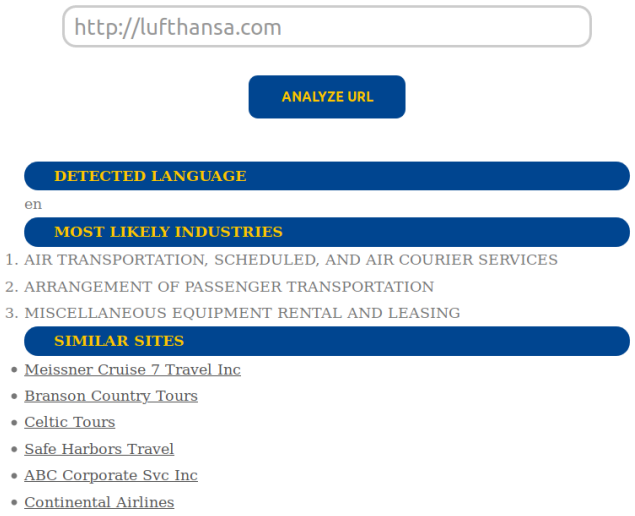

Figure 1: A screenshot from the online service

4 Conclusions and future work

We have introduced the first component of our platform for providing help for companies upon making business decisions for their international servizitation. Our tool is currently able to identify the industry of a company based on its corporate website and returns a ranked list of similar corporate websites. We are planning to improve our results by handling multilingual documents as well, and integrating a large scale web crawling mechanism instead of building the information retrieval component on a fixed pool of companies.

Acknowledgements

This research has been partly conducted in the project "Progressing Service Performance and Export Results of Advanced Manufacturers Networks", no CE1569 ProsperAM-net. The project has been supported by the European Fund for Regional Development in the framework of Interreg CENTRAL EUROPE 2019-2022. Moreover, we like to acknowledge the early ideation and financial contributions of Philipp Schachinger of Salesbeat GmbH.

References

Baines, T.; Ziaee Bigdeli, A.; Bustinza, O. F.; Shi, V. G.; Baldwin, J.; and Ridgway, K. 2017. Servitization: Revisiting the state-of-the-art and research priorities. *International Journal of Operations & Production Management* 37(2):256–278.

Baines, T.; Lightfoot, H.; and Smart, P. 2011. Servitization within manufacturing: Exploring the provision of advanced services and their impact on vertical integration. *Journal of Manufacturing Technology Management* 22(7):947–954.

Bandinelli, R., and Gamberi, V. 2011. Servitization in oil and gas sector: outcomes of a case study research. *Journal of Manufacturing Technology Management* 23(1):87–102.

Benedettini, O.; Neely, A.; and Swink, M. 2015. Why do servitized firms fail? a risk-based explanation. *International Journal of Operations & Production Management* 35(6):946–979.

Berend, G. 2017. Sparse coding of neural word embeddings for multilingual sequence labeling. *Transactions of the Association for Computational Linguistics* 5:247–261.

Chen, C.-C.; Huang, H.-H.; Takamura, H.; and Chen, H.-H. 2019. *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*.

Gebauer, H.; Gustafsson, A.; and Witell, L. 2011. Competitive advantage through service differentiation by manufacturing companies. *Journal of Business Research* 64(12):1270–1280.

Gebauer, H. 2005. Overcoming the service paradox in manufacturing companies. *European Management Journal* 23(1):14–27.

Grünig, R. 2017. *Developing International Strategies*. Berlin, Heidelberg: Springer, 2nd ed. 2017 edition.

Grünig, R. 2018. *The Strategy Planning Process: Analyses, Options, Projects*. 2nd ed. 2018 edition.

Hahn, U.; Hoste, V.; and Tsai, M.-F. 2018. *Proceedings of the First Workshop on Economics and Natural Language Processing*. Melbourne, Australia: Association for Computational Linguistics.

Hahn, U.; Hoste, V.; and Zhang, Z. 2019. *Proceedings of the Second Workshop on Economics and Natural Language Processing*. Hong Kong: Association for Computational Linguistics.

International Trade Center. 2019. International trade in services statistics by service. Exports 2000-2018. http://www.intracen.org/itc/market-info-tools/statistics-export-service-country/. Accessed: 2019-11-15.

Joulin, A.; Grave, E.; Bojanowski, P.; and Mikolov, T. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Mahmoud El-Haj; Rayson, P.; Young, S.; Bouamor, H.; and Ferradans, S. 2019. *Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019)*. Turku, Finland: Linköping University Electronic Press.

Mastrogiacomo, L.; Barravecchia, F.; and Franceschini, F. 2019. A worldwide survey on manufacturing servitization. *The International Journal of Advanced Manufacturing Technology* 103(9-12):3927–3942.

Neely, A. 2008. Exploring the financial consequences of the servitization of manufacturing. *Operations Management Research* 1(2):103–118.

Okazaki, N. 2007. Crfsuite: a fast implementation of conditional random fields (crfs). *URL http://www. chokkan. org/software/crfsuite*.

Oliva, R., and Kallenberg, R. 2003. Managing the transition from products to services. *International Journal of Service Industry Management* 14(2):160–172.

Szász, L.; Demeter, K.; Boer, H.; and Cheng, Y. 2017. Servitization of manufacturing: the effect of economic con-

text. *Journal of Manufacturing Technology Management* 28(8):1011–1034.

Vargo, S. L., and Lusch, R. F. 2004. Evolving to a new dominant logic for marketing. *Journal of Marketing* 68(1):1–17.

Vincze, V.; Nagy T., I.; and Berend, G. 2011. Multiword expressions and named entities in the wiki50 corpus. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, 289–295. Hissar, Bulgaria: Association for Computational Linguistics.

Wang, W.; Lai, K.-H.; and Shou, Y. 2018. The impact of servitization on firm performance: A meta-analysis. *International Journal of Operations & Production Management* 38(7):1562–1588.

Zarpelon Neto, G. 2015. What problems manufacturing companies can face when providing services around the world? *Journal of Business & Industrial Marketing* 30(5):461–472.