How to Boost Customer Relationship Management via Web Mining Benefiting from the Glass Customer's Openness

Frederik Simon Bäumer and Bianca Buff
Semantic Information Processing Group, Paderborn University, Warburger Str. 100, 33098 Paderborn, Germany

Keywords: Customer Relationship Management, Web Mining, Local Grammars.

Abstract: Customer Relationship Management refers to the consistent orientation of a company towards its customers. Since this requires customer-specific data sets, techniques such as web mining are used to acquire information about customers and their behavior. In this case study, we show how web mining can be used to automatically collect information from clients' websites for Customer Relationship Management systems in Business-to-Business environments. Here, we use tailored local grammars to extract relevant information in order to build up a data set that meets the required high quality standards. The evaluation shows that local grammars produce substantial high-quality results, but turn out to be too rigid in some cases. In summary, our case study demonstrates that web mining in combination with local grammars is suitable for Business-to-Business CRM as long as the information demand can be defined precisely and the requested information is available online.

1 INTRODUCTION

The goal of Customer Relationship Management (CRM) is to establish, maintain and make use of successful customer relationships (Link, 2001; Chen and Popovich, 2003; Zeng et al., 2003). In order to achieve this, CRM combines and generalizes customer information to provide helpful answers to a great number of business-relevant questions (Buttle and Maklan, 2015). Although CRM has long been perceived as business strategy in the Business-to-Consumer (B2C) sector, it can be applied in both Business-to-Business (B2B) and B2C environments (Kurt et al., 2018; Buttle and Maklan, 2015). Its implementation in companies can range from a single marketing tool to an enterprise-wide strategy.

Regardless of the conditions of use, the acquisition and maintenance of the underlying customer database is very important for a successful CRM. According to Kurt et al. (2018), the care of necessary customer data is challenging because customer data is not often stored in CRM systems but in separate files. This is particularly the case in B2B environments, where key account managers or consultants manage a smaller number of clients and where personal loyalty has a high priority (Kurt et al., 2018). This results in incomplete, outdated and conflicting data sets, which are no operation basis for an efficiently working CRM. Modern CRM systems are not only seeded with information provided by customer service and sales staff, but also gather data from e-commerce systems, access logs or behavior analysis tools. Information is automatically collected, processed and enriched, both in B2C and B2B environments. A common example are personal usage profiles on web shops. In this case study, we focus on clients' company websites as data source for CRM systems in the B2B context. In the following, we explain how Web (Content) Mining (WM) can be used in B2B CRM and how local grammars (LGs) can be applied to create high quality data sets (Gross, 1993). Sec. 2 presents related work. Based on this, the show case (Sec. 3) and a system design for a WM system will be presented (Sec. 4). The system will be evaluated (Sec. 4.3.3) and its results are then discussed (Sec. 5). Finally, a research outlook is given in Sec. 6.

2 RELATED WORK

In the following, we describe existing work with focus on CRM in B2B companies as well as on WM and LGs, which can be used in CRM to create and enrich underlying databases (Gross, 1997). Besides very different definitions of CRM, numerous CRM goals exist (Kurt et al., 2018). The primary goal of CRM can be described as an increase in corporate value and corporate success, whereby the three secondary goals

Bäumer, F. and Buff, B.
How to Boost Customer Relationship Management via Web Mining Benefiting from the Glass Customer's Openness.
DOI: 10.5220/0007828301290136
In *Proceedings of the 8th International Conference on Data Science, Technology and Applications (DATA 2019)*, pages 129-136
ISBN: 978-989-758-377-3

(1) improvement of customer loyalty, (2) improvement of customer profitability and (3) improvement of customer acquisition are named (Kurt et al., 2018). In this case study, we focus on improving customer profitability through the acquisition of CRM-relevant information from the World Wide Web (WWW). This information extraction (IE) and monitoring task can be done manually or with computational assistance. Since it is important to identify profitable customers and perceive business-relevant changes as quickly as possible and to react adequately to them, extensive automation is desirable. At this point, WM reveals its value for CRM.

The idea of WM is to use the wide range of information offered by the WWW for a multitude of questions. However, WM is defined ambiguously in the literature (Zhao and Bhowmick, 2003; Kosala and Blockeel, 2000; Cooley et al., 1997). Within the scope of this work, WM can be defined as IE using the WWW as data source with the goal to collect predefined information about B2B clients. However, WM can be used and implemented in many different ways. Some existing approaches use pattern-based WM to build databases for question answering or to build web IE systems (Zhang and Lee, 2002; Chang et al., 2001). A follow-up of this approach is the use of (linguistic) grammars to extract information, relations, etc. from natural language (NL) texts. As finite state machines, such grammars are used in many fields of computational linguistics. “From the linguistics point of view, finite state machines are adequate for describing relevant local phenomena in language research and for modeling some parts of NL, such as its phonology, morphology, or syntax” (Pajić et al., 2013). Therefore, LGs describe sequences of words forming semantic units and syntactic structures. In addition, they provide information about the morpho-syntactic properties of the elements described therein, which can be syntactically or semantically shaped (Geierhos, 2010). Therefore, they are well suited for the analysis of the structure of proper names, can deal with domain-specific knowledge and sub-languages and capture the interdependencies of the lexemes among themselves.

In general, LGs are visualized in the form of graphs. The combination of parameterized graphs with a dictionary can be extremely effective for the syntactic analysis of simple sentences (Lee and Geierhos, 2011). Graphs are very user-friendly and intuitive representations for local grammars, which are far superior to equivalent formalisms such as regular expressions. With the help of various drawing programs, local grammars can easily be created, extended and edited.

While Iftene and Balahur-Dobrescu (2008) apply grammar induced patterns on Wikipedia to extract relations between named entities, Pajić et al. (2013) use finite state machines for large Web Monitoring. Therefore, the potential main benefit of automatic data acquisition is the following: It saves time and reduces CRM administration costs. Furthermore, it is not practicable for e.g. the sales staff to acquire and maintain information for each single customer manually – even in B2B context with a smaller number of customers – since they are not CRM experts or knowledge engineers. This refers especially to information that does not occur in everyday business processes. An example is meta data such as the number of employees of a client’s company to assess the sales potential. However, existing work is mostly dedicated to WM as an effective technique to extract business value from companies’ own websites, web shops and social network channels – mostly with focus on B2C companies (Mouthami et al., 2013; Kietzmann et al., 2011; Culnan et al., 2010). But for an overall understanding of the business market, some companies also need to monitor activities and analyze information on their competitors social media channels and business websites in order to increase competitive advantage (Sheng et al., 2005; Croll and Power, 2009). For example, He et al. (2013) use WM to analyze user-generated texts on social media sites of the three largest U.S. pizza chains in order to learn more about the communication behavior of competing companies. The better a company is informed about its competitive activities, the better it can manage its own activities. Therefore, there exist business software solutions that already offer WM and also implement interfaces to CRM systems in the B2B area. This usually includes the extraction of non-sector-specific information such as address data, the names of the managing director or other executives. Depending on the information need of companies, however these generalized approaches can be too short-sighted. Basically, the tools, which are necessary to conduct WM in the context of B2B CRM, are not new: Machine Learning (ML) and WM techniques have already been applied on IE tasks on user-generated content (Tsytsarau and Palpanas, 2012; Freitag, 2000, 1998) and used in CRM (Cioca et al., 2013). LGs as IE technique have also proven themselves (Gross, 1993; Geierhos, 2010). Even though the individual techniques taken by themselves are well established in various contexts and use cases, we are not aware of a work that collectively applies them as a combined B2B WM system to a real business case.

3 SHOW CASE

In the following, we will take a closer look at the use of WM in B2B CRM systems to discuss the applicability in this context. As show case, we present the experiences we have gained in cooperation with a medium-sized special tool production and distribution company for craft businesses.

The company is active in the B2B sector and supplies craft businesses with focus on e-commerce but still relying on traditional sales. The sales staff receive client-related information from a CRM system, which contains both online sales and information about direct sales, client communication, meta information such as number of employees, business areas in which the client's craft businesses operates and further notes. As this is a highly competitive market, it is necessary to react quickly to relevant changes among clients. If there are indications that, for example, a client is entering a new business area (which is covered by the company) or that the client's staff is growing (additional tools and equipment might be needed), the sales department is automatically informed.

However, since the number of clients is constantly growing, manual maintenance of the CRM database is no longer possible. While the sales staff is used to update information about the clients manually, this will be increasingly automated in the future. In the scope of this paper, we focus on the following meta information, which is to be stored in the CRM system:

1. Foundation of the client's business (e.g. 1995)
2. Number of employees (e.g. 20)
3. Business areas (e.g. "painters and varnishers")

This is a selection of information that the company uses to analyze the potential of clients and to make the sales department as effective as possible. We have chosen these three types of information because they can be expressed very differently in NL and therefore have to be collected in different manner. While, for example, the company foundation is given almost entirely as (numerical) date (e.g. "*1998*", "1 February *1998*"), the number of employees varies in format ("A large team of *10* employees", "we are more than *fourteen* hard-working men").

This circumstance must be taken into account during WM. The overall assumption is that customer's potential can be read off at least in part from this information. For instance, a growing number of employees awaits selling more equipment and tools. In addition, an expansion or change in operations, for example from painting to painting and refurbishment, may also indicate that new equipment can be sold. So far, this information has been entered into the CRM system by the sales staff. For this purpose, some rough estimates have been made: The number of employees, for example, can be estimated from the number of company vehicles. The year of foundation often results from publicly accessible documents and the classification into business areas can be derived from the sold products. However, this is a time-consuming task.

For this reason, alternatives were searched for and it turned out that more than 60% of the clients have a company website (already available in CRM) and that often the searched information is available on these websites. In order to automatically acquire the information needed to maintain the CRM system, it is necessary to develop a WM software that automatically searches and extracts the information. We give insight into our work on a WM system in the following.

4 SYSTEM DESIGN

The system to be developed should be able to extract a predefined information from a given website of a client. All relevant information bits should be collected and transmitted to sales employees, who can finally check it and transfer it into the existing CRM system or reject it. Since the data in the CRM system must be of high quality, manual verification of the information is an absolute requirement by any company. However, in order to keep the maintenance time as short as possible, a very high extraction quality (precision) should be guaranteed.

4.1 Data Set

As a starting point for our work we have received 2,000 website URLs of existing clients. Of these addresses, only 1,680 (84%) were still available online. This does not shed a good light on the current CRM database. In order to learn more about the way in which the required information is presented online, the websites were analyzed manually and corresponding texts were stored.

In addition, we searched the WWW for further websites of craft businesses and gathered the relevant text. This way, a text corpus of 6,571 sentences (118,805 tokens, 19,056 types) was created. Examples for such sentences are: "*When the company was founded on 1st December 1992, only three men were employed.*" or "*Today, an average of 100 employees work on major projects throughout Germany*" (Translated from German).

Table 1: Number of annotations in the training corpus per class.

Class	Annotations	Sample sentences
No. Employees	1,778	Unser Team besteht zur Zeit aus **23** Mitarbeitern *Our staff currently consists of 23 employees*
Founding year	2,516	Seit **1996** sind wir Ihr Malerbetrieb in Berlin *Since 1996 we are your painting company in Berlin*
Business area	4,736	Zum Beruf des **Malers** gehört auch Wissen über **Brandschutz** *The profession of a painter also includes knowledge about fire protection*

4.2 Data Acquisition via Web Crawling

In general, a crawler is a software application that systematically processes a defined, but not necessarily limited, dataset *ad infinitum* with regard to defined goals and based on configured instructions for action. The term (World Wide) Web Crawler refers to a tool whose targeted data set results from the content of the WWW (Castillo, 2004). The goals behind web crawling are diverse: While global search engines aim at the complete indexing of existing web resources, so-called focused crawlers limit the results with regard to given criteria (Menczer et al., 2004). This can be a Top Level Domain (e.g. ".de"), a subject area (e.g. painting tools) or the relevance to an information need. To extract information from websites and to monitor websites for changes, a focused web crawler is required that iteratively accesses given URLs from the CRM database, downloads the source codes (HTML) of all relevant pages and stores the collected data with a timestamp for further processing (cf. Fig. 1). However, the relevance criterion is difficult to define, basically all sub-pages of a website are relevant, which can potentially contain relevant information. During the initial analysis of clients' web pages during text corpus creation, we were able to create a list of URL patterns, which served as a blacklist to speed up the crawling process. For example, it is not necessary to access support pages ("*/faq/*") or privacy pages ("*/datenschutz/*") because they mostly do not contain relevant information. Furthermore, we limited the crawling depth to a maximum of four sub-pages and the maximum number of pages per client to 150. Even with these limitations, the average acquisition time is 3.5 minutes per website, which is mainly due to weak web servers[1].

We have tried classification techniques to detect and ignore irrelevant sub-pages, but the relevant information is often represented on sub-pages with completely different content, so we had too many false positives and there was a considerable loss of information. Therefore, all pages that are not blacklisted are currently passed to the IE component.

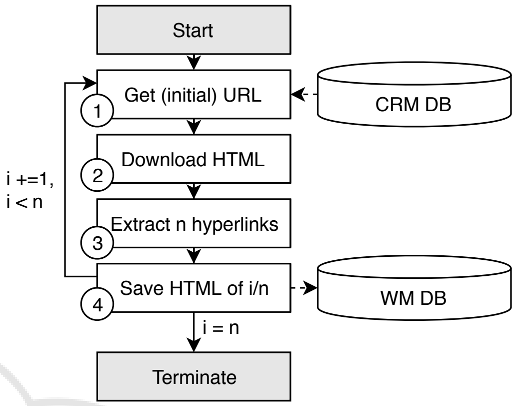

Figure 1: Basic functionality of the web crawler.

4.3 Information Extraction

The automatic transformation of HTML-source-code into a semantically structured document (template) is challenging and requires text preprocessing such as HTML-stripping and sentence splitting in order to achieve good results (cf. Sec. 4.3.1). In the following, IE is conducted by means of LGs. While ML methods have proven to be robust in the context of IE in the past (Chang et al., 2006; Dollmann and Geierhos, 2016), LGs are convincing due to their high precision (Bsiri et al., 2008), since the linguistic context of searched information can be defined very precisely (Gross, 1993; Geierhos, 2010). As Bsiri et al. (2008) show, LGs can also achieve a high recall on user-generated content from the WWW, but require extensive preprocessing. By choosing LGs, we hope to meet the high demands on data quality made by our practice partner.

4.3.1 Data Preprocessing & Data Annotation

We applied data preprocessing to each accessed page of a client's website which is structured into the following steps: (1) *HTML-removal* in order to get the plain text including removing boilerplate content, such as navigation, headers and footers[2], (2) *Paragraph cleaning* via regular expressions to remove

[1] It is necessary to reduce the crawling speed in order not to attract negative attention, cause damage or be banned.

[2] We used: https://pypi.org/project/jusText/.

paragraphs that are too short (less than four words), (3) *Sentence splitting*[3] as IE has to be done on a sentence basis in order to make the annotation of the relevant information for the practice partner as fast and simple as possible. The initial annotation of the sentences with regard to the requested information (cf. Sec. 3) was carried out by the research team. We have token-based annotations in three classes (“*Employees*”, “*Foundation*” and “*Keyword*”). It is very important for the resulting WM system that it is continuously fed with new annotated data to improve. For this reason, we have developed a convenient graphical annotation interface which enables company employees to annotate single tokens from sentences regarding the matching class by simply drag'n'drop the tokens. Table 1 shows the number of annotations for each class.

The annotation of tokens for the company foundation and the annotation of keywords for later allocation to business areas is not challenging due to the low number of structural variants. Years are usually given in two-digit (“*99*”), four-digit (“*1999*”) notation or as a date (“*09.02.1988*”). Keywords are diverse in their lexical form, but consist mainly of one or two words. The number of employees is more difficult to annotate: On the one hand, the numbers are often written out (“*ten* skilled workers”), which increases the number of variants. On the other hand, teams are often subdivided with regard to their function (“The team consists of *3* painters and *2* electricians”) or given as a rough estimation “*more than 20*”. In these cases, all relevant tokens are annotated.

4.3.2 Local Grammars

We employed LGs[4] to transform the preprocessed text into a semantically structured form. This means that the original text is automatically enriched with annotations that allow the information to be extracted. This task is conducted by a multitude of LGs that allow the modeling of a high degree of syntactic variation. This flexibility could not be reached by standard string search, where a list of collected phrases would be matched with the document, since minor morpho-syntactic variation prevents the match. Here, bootstrapping with LGs (Gross, 1999) is a much more promising method, as illustrated by the following example (translated from German).

- “The story of our company, which has ***350*** **qualified employees**, began in 1927.”
- “I have a team of ***twenty*** **men**.”
- “The company employs **more than *hundred* people** worldwide in 18 national companies.”

Though these three phrases are different on the morpho-syntactic level, they contain the same type of information. By using a LG, it is possible to efficiently represent all three phrases (cf. Fig 2). In the following, we describe the process of LG modeling by using the example “number of employees”.

Altogether, the sentences related to employees in the corpus are composed of 25,406 tokens (4,306 types). In order to be able to derive patterns, we have examined the context in which the relevant information is embedded in more detail. We have found a total of 34 general identifiers such as “*Mitarbeiter*” (employees), “*Beschftigte*” (coworkers), “*Fachkrfte*” (skilled workers), “*Mnner*” (men), “*Leute*” (people) and “*Personen*” (persons) as well as 270 specific job identifiers such as “*Maler*” (painter), “*Lackierer*” (varnisher), “*Zimmermann*” (carpenter), “*Dachdecker*” (roofer) and “*Heizungs- und Sanitrtechniker*” (heating and sanitary technician). The context is further defined by location or company specifications such as “*Firma*” (firm), “*Unternehmen*” (company) and “*Branche*” (sector). Examples for frequent syntactical constructions are:

- “{*who*} employ(s) {*adv.*} {***number***} {*identifier*}”
- “{*identifier*}: {***number***}”
- “{***number***} {*identifier*} are employed by {*who*}”

Particularly frequent adverbs are “*zurzeit*” (presently) and “*inzwischen*” (meanwhile). Another specialty are words, which relativize the concrete number of employees. Examples are “*circa*”, “*rund*” (around) oder “*durchschnittlich*” (average) and “*insgesamt*” (in total). Because LGs support lists, lexica and sub-graphs, the lexical variants can easily be considered. However, only what is known can be modeled. So there will always be constructions that have not yet been taken into account. With our current LG (cf. Fig. 2), we are able to cover 88% of the annotated sentences in the corpus (cf. Tab. 2).

Table 2: Covered sentences in the training corpus [%].

Class	**Coverage**
No. Employees	0.88
Founding year	0.91
Business area	0.85

Some cases cannot be covered because no additional context is given. This is, for example, the case if the number of employees is not embedded in a sentence but given as a single number with no context.

[3] We used: https://spacy.io/.

[4] We used the multilingual corpus processing suite Unitex/GramLab, https://unitexgramlab.org.

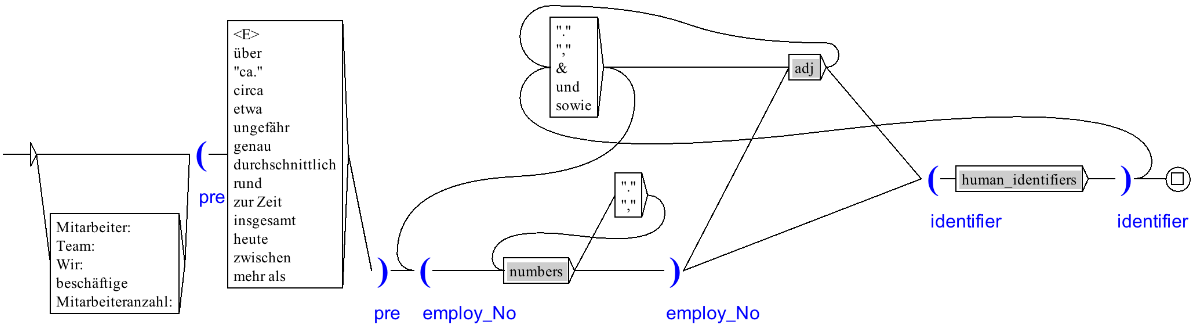

Figure 2: LG for extracting numbers of employees (simplified).

4.3.3 Evaluation

To evaluate the quality of our WM system regarding the LGs, we designed a manually annotated test corpus composed of 1,000 sentences. This test corpus allows us to provide values for recall and precision of the automatically retrieved results.

Table 3: Evaluation of the LGs performances [%].

Class	Precision	Recall	F-score
No. Employees	0.89	0.72	0.80
Founding year	0.91	0.75	0.82
Business area	0.82	0.47	0.60

Table 3 shows promising results of precision (87% on average) and recall (65% on average). The expected low recall value can be explained by uncovered syntactic constructions and missing entries in the word lists, which have to be constantly maintained and extended. However, the precision suffers above all from lexical ambiguities that can arise when the LGs are weakened in the sense of a higher recall as can be seen in the following example: “[...] Lukas 32 angestellter Maler [...]” (“[...] Lukas 32 employed painter [...]”). Here, 32 employees are detected, since the adjective “angestellter” (employed) is equated with the noun “Angestellter” (employee) in the LG. The argument that the initial sentence is grammatically questionable does not apply, because this is a basic assumption for user-generated content that LGs have to face. In our case, since the recognized information is checked again, such an error can be intercepted. However, it is necessary to consider such cases in future work.

Moreover, knowing that IE is nowadays essentially influenced by ML methods, we have also applied an established method to our topic as a comparative value. For short text like the sentences used in this paper, the MIT Information Extraction (MITIE) library[5] was already used in the past (Geyer et al., 2016; King, 2009). MITIE is an open source NLP library focused on Named Entity Extraction and is known for state-of-the-art statistical ML. Applied to our test corpus, MITIE[6] could convince with a good average recall value (0.82), but had a comparatively low precision (0.67). This result may be due to the small amount of data in the corpus and is not suitable for a final judgment. In fact, MITIE shows very good results in the extraction of keywords. Its exclusive usage can be promising and sensible in an area where precision is less important due to the large number of information bits per website and the creation of LGs is too time-consuming due to the number of variants.

4.4 Monitoring & Data Integration

The monitoring currently includes an automatic revisit of the websites once a week. The system stores extracted information in a database to detect changes when pages are revisited. If no change is detected during revisit, no data is transferred to the CRM system. However, if a change is detected (this also includes if the website is not accessible), a change notification is sent for review. A notification contains information about the source (URL), the paragraph, the sentence, and the extracted information. In this way, the employee can disambiguate and review information (e.g. conflicting foundation data), merge information bits (e.g. several entries on the number of employees), and detect overseen information that occurs in the same paragraph. As exchange format, we chose JSON, which seemed to be suitable because of the lower overhead compared to XML. A client application was developed on the part of the practice partner, which receives the JSON and presents it to an employee in a user-friendly way. The employee makes the final decision to transfer the extracted information to the CRM system. If the extraction result is erroneous, the employee can reject it. In this case, an error is reported to the IE system and stored in the

[5] We used: https://github.com/mit-nlp/MITIE/.

[6] Dictionary: 200,000 words, No. Features: 271.

database. On the one hand, this is valuable information for the revision of the LGs and on the other hand, it is very important training data for further work with ML techniques. This way, we have fulfilled both the initial requirement for strict data reviewing and created further training data by the help of an expert-in-the-loop approach.

5 DISCUSSION

The approach described in this case study is work in progress but very promising: information stored or to be stored in CRM can be automatically updated or added. The use of LGs proves to be effective because the achieved quality of information can be regarded as very high. However, challenges remain that require further work: The IE system reliably extracts the type of information we are looking for. However, this type of information occurs in different contexts on the clients' websites. For example, there are websites that present the clients' company's history and contain several (sometimes contradictory) data on the foundation and expansion of the company. Furthermore, sentences such as "*1997: The company celebrates its 120th anniversary with seven employees, including office workers and trainees*" include an indication of when the company was founded (1877), but the automated identification is challenging. The situation is similar for statements such as "*For more than 60 years*", which do not allow any temporal mapping. Complex syntactic and semantic constructions such as "*In the meantime our team includes 5 journeymen, two trainees as well as my wife Nicole and my daughter Stefanie in the office*" are also very difficult to grasp, both for LGs and ML techniques. Here it could turn out to be an advantage that experts review the sentences and can interpret them before the data is transferred into the CRM. In the future, such records are to be extracted and transferred as flagged candidates.

From the point of view of B2B CRM, we can report that 83% of the processed websites could be assigned to a business area on the basis of the extracted keywords. In addition, we were able to find and extract information on the foundation of the respective company in 31% of all processed websites. In comparison, the information on the size of the team could be extracted much less often automatically, only in 12% of the cases such information was extracted in live operation.

6 CONCLUSION

As could be shown in this case study, WM combined with suitable IE can lead to noteworthy results in B2B CRM contexts. In order for such an implementation to be successful, some challenges and assumptions have to be taken into account. The challenge for LGs is of course the poor quality of the NL texts and the lack of context. While the poor text quality can be partly compensated by a weakening of the LGs (which probably damages the precision), a missing context cannot be compensated because a LG cannot apply on the one hand and on the other hand cannot disambiguate important information. Assumptions that have to be made are that the relevant information can be defined as a kind of pattern and is available online on the company's website. For example, addresses and tax numbers can easily be acquired from public websites, especially thanks to their structured character, whereas individual product names or general product characteristics are more difficult to extract. In future work, we will create more LGs to acquire further information for the B2B CRM. Moreover, we will improve the reliability of the current LGs by improving the coverage of the used linguistic resources (e.g. domain-specific dictionaries).

REFERENCES

Bsiri, S., Geierhos, M., and Ringlstetter, C. (2008). Structuring Job Search via Local Grammars. *Advances in Natural Language Processing and Applications. Research in Computing Science*, 33:201–212.

Buttle, F. and Maklan, S. (2015). *Customer Relationship Management: Concepts and Technologies*. Routledge, 3 edition.

Castillo, C. (2004). *Effective Web Crawling*. PhD thesis, University of Chile.

Chang, C.-H., Kayed, M., Girgis, M. R., and Shaalan, K. F. (2006). A Survey of Web Information Extraction Systems. *IEEE Transactions on Knowledge and Data Engineering*, 18(10):1411–1428.

Chang, C.-H., Lui, S.-C., and Wu, Y.-C. (2001). Applying Pattern Mining to Web Information Extraction. In Cheung, D., Williams, G. J., and Li, Q., editors, *Advances in Knowledge Discovery and Data Mining*, pages 4–15, Berlin, Heidelberg, Germany. Springer.

Chen, I. J. and Popovich, K. (2003). Understanding customer relationship management (CRM): People, process and technology. *Business Process Management J.*, 9(5):672–688.

Cioca, M., Ghete, A. I., Cioca, L. I., and Gîfu, D. (2013). Machine Learning and Creative Methods Used to Classify Customers in a CRM Systems. In *Innovative Manufacturing Engineering*, volume 371 of *Applied Mechanics and Materials*, pages 769–773.

Cooley, R., Mobasher, B., and Srivastava, J. (1997). Web Mining: Information and Pattern Discovery on the World Wide Web. In *Proc. of the 9th IEEE Int. Conf. on Tools with Artificial Intelligence*, pages 558–567.

Croll, A. and Power, S. (2009). *Complete Web Monitoring: Watching your visitors, performance, communities, and competitors*. O'Reilly Media.

Culnan, M., McHugh, P., and I. Zubillaga, J. (2010). How Large U.S. Companies Can Use Twitter and Other Social Media to Gain Business Value. *MIS Quarterly Executive*, 9:243–259.

Dollmann, M. and Geierhos, M. (2016). On- and Off-Topic Classification and Semantic Annotation of User-Generated Software Requirements. In *Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing*, pages 1807–1816, Austin, TX, USA.

Freitag, D. (1998). Information Extraction from HTML: Application of a General Machine Learning Approach. In *Proc. of the 15th National/10th Conf. on AAAI/IAAI*, AAAI '98/IAAI '98, pages 517–523, Menlo Park, CA, USA. AAAI.

Freitag, D. (2000). Machine Learning for Information Extraction in Informal Domains. *Machine Learning*, 39(2):169–202.

Geierhos, M. (2010). *BiographIE - Klassifikation und Extraktion karrierespezifischer Informationen*. Linguistic Resources for Natural Language Processing. Lincom.

Geyer, K., Greenfield, K., Mensch, A., and Simek, O. (2016). Named Entity Recognition in 140 Characters or Less. In *6th Workshop on Making Sense of Microposts (#Microposts2016)*, pages 78–79.

Gross, M. (1993). Local grammars and their representation by finite automata. Data, Description, Discourse, Papers on the English Language in honour of John McH Sinclair, pages 26–38. Harper-Collins.

Gross, M. (1997). *The Construction of Local Grammars*. Finite-state language processing. The MIT Press.

Gross, M. (1999). A Bootstrap Method for Constructing Local Grammars. In *Proc. of the Symposium on Contemporary Mathematics*, pages 229–250. University of Belgrad.

He, W., Zha, S., and Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *Int. J. of Info. Management*, 33(3):464 – 472.

Iftene, A. and Balahur-Dobrescu, A. (2008). Named Entity Relation Mining Using Wikipedia. In *In Proc. of the 6th Int. Language Resources and Evaluation Conf.*

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., and Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, 54(3):241 – 251. SI: Social Media.

King, D. E. (2009). Dlib-ml: A Machine Learning Toolkit. *J. Mach. Learn. Res.*, 10:1755–1758.

Kosala, R. and Blockeel, H. (2000). Web Mining Research: A Survey. *SIGKDD Explor. Newsl.*, 2(1):1–15.

Kurt, A., Glauser, O. J., and Weiss, E. (2018). *Optimierung des Customer Relationship Managements in B2B-Märkten*, pages 207–224. Springer Fachmedien Wiesbaden, Wiesbaden, Germany.

Lee, Y. S. and Geierhos, M. (2011). Buy, Sell, or Hold? Information Extraction from Stock Analyst Reports. In Beigl, M., Christiansen, H., Roth-Berghofer, T. R., Kofod-Petersen, A., Coventry, K. R., and Schmidtke, H. R., editors, *Modeling and Using Context*, pages 173–184, Berlin, Heidelberg. Springer Berlin Heidelberg.

Link, J. (2001). *Grundlagen und Perspektiven des Customer Relationship Management*, pages 1–34. Springer, Berlin, Heidelberg, Germany.

Menczer, F., Pant, G., and Srinivasan, P. (2004). Topical Web Crawlers: Evaluating Adaptive Algorithms. *ACM Trans. Internet Technol.*, 4(4):378–419.

Mouthami, K., Devi, K. N., and Bhaskaran, V. M. (2013). Sentiment analysis and classification based on textual reviews. In *Proc. of the 2013 Int. Conf. on Info. Communication and Embedded Systems*, pages 271–276.

Pajić, V., Vitas, D., Lažetić, G. P., and Pajić, M. (2013). WebMonitoring Software System: Finite State Machines for Monitoring the Web. *Computer Science and Info. Systems*, 10(1).

Sheng, Y. P., Mykytyn, Jr., P. P., and Litecky, C. R. (2005). Competitor Analysis and Its Defenses in the e-Marketplace. *Commun. ACM*, 48(8):107–112.

Tsytsarau, M. and Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24(3):478–514.

Zeng, Y. E., Wen, H. J., and Yen, D. C. (2003). Customer relationship management (CRM) in business-to-business (B2B) e-commerce. *Info. Management & Computer Security*, 11(1):39–44.

Zhang, D. and Lee, W. S. (2002). Web Based Pattern Mining and Matching Approach to Question Answering. In *Proc. of Text REtrieval Conf. 2002*.

Zhao, Q. and Bhowmick, S. S. (2003). Sequential Pattern Mining : A Survey. *ITechnical Report CAIS Nayang Technological University Singapore*, 1:26.