Aalto University
School of Business

INFORMATION EXTRACTION FROM PROCUREMENT CONTRACTS

Master's Thesis
Esa Toikka
Aalto University School of Business
Information and Service Management
Spring 2021

Aalto University
School of Business

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Abstract of master's thesis

Author Esa Toikka		
Title of thesis Information extraction from procurement contracts		
Degree Master of Science in Economics and Business Administration		
Degree programme Information and Service Management		
Thesis advisors Pekka Malo, Sammeli Sammalkorpi		
Year of approval 2021	**Number of pages** 67	**Language** English

Abstract

The increasing amount of text data produced by companies demands automating the extraction of key information from contracts. It removes the manual step of inserting contract metadata into contract management software. This improves the efficiency and meaningfulness of work for procurement professionals as they will have growing amount of time in more creative and value-creating tasks.

Previous information extraction research has considered many fields but left a gap in complete procurement contract documents. This study aimed to contribute to filling that gap by experimenting methods of extracting contract metadata from multi-paged non-standard documents. The best methods for extracting contracting parties, effective dates, payment terms, liabilities, and contracting periods were investigated. The automation can be done utilizing methods of natural language processing. A tool for automatic extraction was created applying named-entity recognition algorithms and hand-crafted rules. Quantitative results were obtained from the tool. These results were discussed in semi-structured interviews with professionals from the case company Sievo Oy.

A key finding of this study was that the combination of methods used in the extraction performs well with contracting parties, while with other items more work needs to be done. Despite the varying results from the extraction, the interviews exposed an alternative way of easing the work of finding key information from contracts. It was suggested that instead of completely automating the task, the software could act as an assistant for its users. How this possibility is utilized and what is the reception of the users remains to be seen.

Keywords information extraction, natural language processing, contract management, procurement analytics

Aalto-yliopisto
Kauppakorkeakoulu

Aalto-yliopisto, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Maisterintutkinnon tutkielman tiivistelmä

Tekijä Esa Toikka

Työn nimi Tiedon eristäminen hankintasopimuksista

Tutkinto Kauppatieteiden maisteri

Koulutusohjelma Tieto- ja palvelujohtaminen

Työn ohjaajat Pekka Malo, Sammeli Sammalkorpi

Hyväksymisvuosi 2021 **Sivumäärä** 67 **Kieli** englanti

Tiivistelmä

Yritykset tuottavat kasvavassa määrin tekstidataa. Tähän kehitykseen vastaamaan tarvitaan automaattista avaintietojen eristämistä sopimustiedostoista. Automatisointi poistaa aikaa vievän työvaiheen, jossa sopimuksen metatietoja syötetään käsin sopimustenhallintajärjestelmiin. Tällä tavoin voidaan parantaa hankintatoimen tehokkuutta ja työntekijöiden kokemaa työn merkityksellisyyttä, kun työaikaa voidaan käyttää luoviin ja arvoa tuottaviin tehtäviin.

Aiemmassa tiedon eristämistä koskevassa tutkimuksessa on käsitelty monenlaisia tekstityyppejä, mutta kokonaiset hankintasopimukset ovat jääneet vaille huomiota. Tämän tutkimuksen tarkoituksena oli olla täyttämässä tätä aukkoa etsimällä menetelmiä, joilla voidaan eristää metatietoja monisivuisista epästandardeista sopimuksista. Parhaiksi todettuja menetelmiä tiedon eristämiseen sovellettiin sopijaosapuoliin, voimaantulopäivään, maksuehtoihin, vastuisiin ja sopimuksen kestoon. Automaatiossa hyödynnettiin luonnollisen kielen käsittelyn menetelmiä. Kehitetyssä ohjelmassa käytettiin nimentunnistuksen algoritmeja ja käsin määriteltyjä sääntöjä. Ohjelma tuotti määrällisiä tuloksia, joista keskusteltiin case-yritys Sievo Oy:n asiantuntijoiden kanssa puolistrukturoiduissa haastatteluissa.

Tutkimuksen keskeisten löydösten pohjalta sopivaksi valikoitunut menetelmien yhdistelmä toimi hyvin sopijaosapuolten kohdalla. Muiden eristettävien nimikkeiden kohdalla työtä on vielä edessä. Vaihtelevista tuloksista huolimatta haastatteluissa keskeiseksi teemaksi nousi vaihtoehtoinen toimintatapa, joka helpottaisi avaintietojen eristämistä sopimuksista. Haastateltavat nostivat esiin mahdollisuuden käyttää ohjelmistoa avustajana täyden automatisoinnin sijaan. Nähtäväksi jää, miten tätä mahdollisuutta hyödynnetään, ja millaisen vastaanoton se saa käyttäjiltä.

Avainsanat tiedon eristäminen, luonnollisen kielen käsittely, sopimustenhallinta, hankintatoimen analytiikka

Acknowledgements

Throughout the thesis writing process I have gotten a good deal of support from various sources. First of all, thank you Pekka Malo for acting as my thesis supervisor. Your feedback was valuable and highly appreciated.

This thesis was made in collaboration with Sievo Oy. I am thankful to Sievo for the opportunity and to Sammeli Sammalkorpi for the high-quality advisory during the process. It was great working on a topic that both is interesting and has a real-life business application. Having my background in language studies it was fun working with natural language processing. Many thanks to Hanna-Leena Lindblom and Samuli Vuorinen for getting me going with the topic and to Matan Shenhav for reviewing my methods and code.

I would like to acknowledge all my awesome colleagues at Sievo. I am especially thankful for Ville Väisänen, Aki Haanpää, Tomi Paatsalo, Eemil Oikari, and Niko Kinnunen for the companionship during the time we spent together. Additionally, I want to thank all my classmates and the whole Aalto network for providing the framework for studies and student life. Thank you, KY, for the support and community feeling.

Finally, I am extremely grateful to all the people around me, family and friends. Deepest thanks to my loving wife for the support through the thesis process. And to anyone I did not mention yet but is reading this: thank you and happy reading!

Espoo 22.2.2021

Esa Toikka

Contents

Tables

Figures

1 Introduction

It is a well-known phenomenon that the amount of textual data in the world is growing fast. Blogs and social media posts are topics on their own while enterprises produce lots of emails and legal documents. Such documents as contracts have been traditionally printed on paper and sent back and forth between the parties for reviewing and finally signing. These growing piles of paper have then been archived in physical files in the shelves of offices and storages. One of the biggest creators of contracts within a company is the procurement function as they maintain relationships with the suppliers. Currently, the contracts are often sent by email or within a platform which allows collaboration. Storing of new documents is also done digitally. Moving to digital contract storages and contract management software involves a huge amount of work if the old physically stored contracts are to be digitized. It is not only scanning the papers but also adding metadata to the system so that the main contents of the contract can be found with a quick glance. This can become burdensome for the acting people and act as a barrier of fully benefitting from the contract management software as some of the required metadata can be incorrectly added or left completely empty. Therefore, automation of the contract digitization is required. This study concentrates on finding a way of automating the key information extraction from procurement contracts in the context of procurement analytics software provider Sievo Oy. The following subsections will shed light on the motivation behind the study, the research questions, and the structuring of the rest of the thesis.

1.1 Motivation

As mentioned above, the number of contractual documents maintained by enterprises can be enormous. Large companies can have thousands of active contracts and tens of thousands of expired ones. Furthermore, a single document can contain dozens of pages. The documents can be in non-standard format which makes it difficult for those who need the key information of a contract to find it with a quick look. Hence, companies often utilize contract management software in storing and analyzing contracts. Both active and expired contracts can provide value in terms of analytics on development of contracting and purchasing. However, the metadata collected from the contracts is frequently manually sought for and can therefore contain human errors. Automation of the information extraction would mitigate these errors as well as increase the speed of the task. Furthermore, if it were done automatically the procurement professionals would have more time for more creative tasks and feel their work as more meaningful (Costas and Kärreman, 2016).

There is a lot of evidence that the technology of natural language processing is beneficial in the field of extracting information from textual data. Information extraction has been researched especially in the medical and biological fields in studies such as detecting medications of patients from clinical notes (Jagannathan *et al.*, 2009). Those fields have been extensively studied, as recognized by Wang *et al.* (2018). They found 263 studies in clinical information extraction applications published between 2009 and 2016. However, there are some studies that are more closely related to contracts. Zhang and El-Gohary (2016) suggested that a rule-based information extraction approach could be used in checking whether construction agreements are compliant with regulatory documents. Lee *et al.* (2019) have also studied rule-based information extraction in the construction domain. They stated that using rule-based approach can lead to high accuracy, but is very laborious, as the rules need to be updated continuously. Hence, their model can be used to supplement human work and not fully automate it. Nevertheless, there are only few studies when it comes to procurement contract information extraction in general, which all however suffer from the limitation of taking only certain sections of contracts into consideration while ignoring complete contract documents (Chalkidis *et al.*, 2017; Chalkidis and Androutsopoulos, 2017; Chalkidis *et al.*, 2019). Still, that work provides an established baseline for the level of results that can be expected to be achieved with the combination of advanced machine learning models and hand-crafted rules.

Inspired by Lavelli *et al.* (2008), named entity recognition algorithms are utilized in this study as applicable. Those are pretrained models for finding certain data elements from text. The types of data can include for instance dates, names of organizations or people, and geographical locations. Combining the algorithms with hand-crafted rules seems to be a promising approach to the automation of contract information extraction (Chalkidis *et al.*, 2017). However, it still remains to be answered whether these methods are applicable in practical solutions as such or could simpler methods reach acceptable results to be ready for a product. Thus, it is valuable to investigate the different methods and their performance. The research questions related to these topics are introduced in the following subsection.

1.2 Research questions

Motivated by the characteristics mentioned in the previous subsection there are two main objectives for this study. The first objective is to explore the possibility of automating the contract information extraction task by developing a tool which takes contract documents as inputs and after processing provides the relevant metadata items as outputs. The second objective is to find out whether such tool is indeed needed and has potential to be developed further towards a fully functioning product. Driven by these goals, the following research question was formatted:

- What is the best way of automatically extracting contract information?

The following sub-questions were formatted to the main research question two first of which are related to the methods of information extraction while the third one to the potential of the solution:

i. Are named entity recognition algorithms a feasible solution?
ii. Should hand-crafted rules be utilized?
iii. Is there future potential for the best solution found?

Utilizing these questions, this study aims to contribute to filling the gap of the procurement contract domain in information extraction research by exploring the possibilities and potential of automation of the task. This research also considers full contract documents instead of pre-extracted relevant sections. Furthermore, the aim is not completely theoretical as this study is used in exploring the potential of a practical business case.

1.3 Structure of the thesis

The thesis consists of seven sections. This section has introduced the topic of the research. The rest of the thesis is structured as follows. In section 2 the relevant previous research is reviewed. The subsections discuss procurement, contracts, software as a service business model, the technology of natural language processing, as well as work automation and meaningful work. Section 3 presents the case of this study. It includes an introduction to the case company and the problem as well as to the data used in the study. In section 4 the research methodology is presented. Both quantitative and qualitative aspects are discussed. Section 5 reviews the results from the quantitative part of the study. In section 6 the quantitative results are discussed and connected to the qualitative results from the interviews. Section 7 completes the thesis with conclusions and future research possibilities.

2 Literature review

In this section the topics relevant to this study are introduced based on previously published academic literature. The section is divided into five subsections. Firstly, procurement as a function within companies is presented. The strategic importance of procurement is highlighted, and the basic purchasing process is shed light on. Secondly, the domain of business contracts is introduced. In this subsection contracts are discussed in general, then the process of getting into contract adapted from Monczka *et al.* (2009) is presented, and finally the area of contract management is considered. Thirdly, the concept of software as a service is discussed due to its relevance for the field of contract management tools. Fourthly, the idea of natural language processing is considered from multiple perspectives. The basic process of natural language processing is introduced and the applications relevant to this study namely information retrieval and information extraction are discussed. Lastly, automation of work and how it affects perceived meaningfulness at work are discussed. Wrzesniewski *et al.* (1997) stated that finding meaningfulness from work is one of the key motivators for people to work in the first place. Automation can have both positive and negative effects on that (Rintala and Suolanen, 2005).

2.1 Procurement

Procurement is one of the basic building blocks of a manufacturing company. Good products cannot be manufactured without good vendors and the selection of vendors is in the core of the procurement function (Weber *et al.*, 1991). Procurement has typically been perceived as one of the three key phases of the supply chain, production and distribution being the other two. Nowadays the discussion revolves more around supply chain management than each of its parts (Thomas and Griffin, 1996). The raison d'être of procurement as a function, as stated by Lewis (1946), is "to make the needed raw materials, fabricated parts, and supplies available to the factory, and at reasonable cost". This includes all kinds of sub-steps such as determining purchased amounts, examining the received goods, and settling suppliers' claims, while the key steps are negotiating with suppliers and controlling inventories. Bals *et al.* (2019) found that the ten most important future competencies for procurement professionals include such as automation, big data analytics, computer literacy, and process optimization.

Companies have traditionally optimized their internal procurement processes, while literature suggests more close intercompany relationships to be built. Long term buyer-vendor relationships lead to savings for all parties (Thomas and Griffin, 1996). The procurement processes of companies were recognized to be critical for business already during the World War II as there were constant disruptions in the supply chain (Lewis, 1946). The role of procurement is highlighted by the fact that the function is in charge of most of the company's costs. Furthermore, procurement function manages a network of external suppliers and partners that can be of a strategic importance (Bals *et al.*, 2019). Also, the strategic importance of procurement was brought into spotlight by Kraljic (1983) in his article that is foundational for procurement where he introducesd a framework for assessing the strategic importance of different kinds of purchases. Another underlining factor in the strategic role of procurement is having close connections to production and finance departments. However, it is always the importance that management gives to procurement that determines how strategic it is regarded (Lewis, 1946).

According to Monczka *et al.* (2009, 42-69) the basic purchasing process can be constituted from six phases as illustrated in Figure 1. It starts by identifying a need for a purchase either by forecasting or finding a completely new need. The need may arise as recurring from inventory levels or as a project, for instance, a single subassembly for a new prototype or a social media marketing campaign. After the need arises the next step is to clarify the need. In this phase procurement professionals are working in very close collaboration with other functions in the organization. While procurement people are qualified in the process itself the needed product or service might be new to then. Thus, for instance research and development professionals need to clarify the specifications of the product they are willing to have.

The third phase consists of identifying and selecting a supplier. There are two alternatives in this stage: firstly, the product or service is bought from a supplier with which the buying organization already has a contract in place or secondly, a supplier that the company is not in a contract with is selected. The first one is a safe alternative when applicable. The second option can be completed in a good or a bad manner. The good one is to include procurement organization in the identification and selection of a supplier, while in the case of a bad manner procurement function is bypassed and the function needing the product or service buys it directly from a supplier. The latter way is called maverick purchasing, which includes significant risks as the spending is not under procurement function’s radar and the suppliers are not verified to be compliant and sustainable. On the presumption that it is done correctly, the supplier selection process is not a quick task to complete. It can include either competitive bidding or extensive negotiations after which evaluation of alternative suppliers can be done and the selection itself finished.

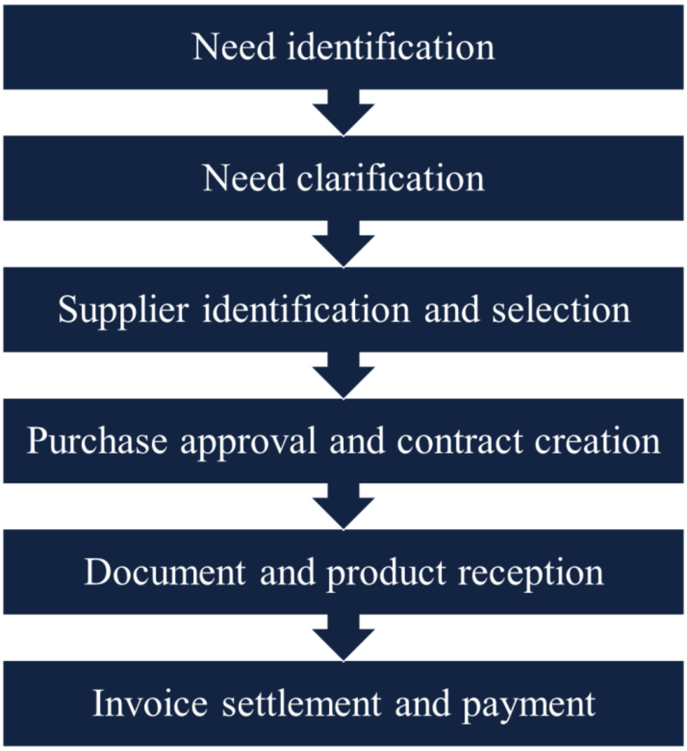

Need identification
Need clarification
Supplier identification and selection
Purchase approval and contract creation
Document and product reception
Invoice settlement and payment

Figure 1. Purchasing process.

The fourth stage of the procurement process includes approval of the purchase and creation of a contract with the selected supplier. The contract can be either a long-term contract which allows repetitive purchases from the supplier or a one-time purchase order. Purchase orders are agreements and thus also legally binding. Contracts being in the spotlight of this study are introduced more extensively in their own section. In both cases, the needed product or service is described as well as the terms and conditions of the agreement.

After the agreement falls effective, the purchasing company is expected to receive the agreed product or service. In the receiving process the recipient examines that the result is as expected, and all required documents have been delivered. In case of physical products, these documents can include for instance packing lists from the supplier and bills of lading from the logistics provider. Should the need arise, the buying company gives feedback to the supplier with a receiving discrepancy report. If he received goods are complete, the receiving department makes sure they are transported to the correct order placer.

The last step of the purchasing process, as stated by Monczka *et al.* (2009, 42-69), is to settle the invoice and complete the payment to the supplier. The invoice details are delivered to the accounts payable department in which the payment is created. At the same time the procurement function maintains a record of supplier transactions and their success in order to analyze the relationship with every separate supplier. The performance and relationship should be measured and managed continuously in order to make the most out of the partnerships. In this task the contract management and other procurement analytics tools are used, more of which in the later sections. Next, the contracts mentioned in the fourth stage of purchase process are introduced more thoroughly. The subsection will also provide context to the data used in this study.

2.2 Contracts

Cooperation between companies, whether it is buyer-seller relationship or strategic partnership, is most often coordinated with contracts (Ring and Van de Ven, 1992). Hart and Holmström (1986, 1) argued that all the trade is mediated by a contract of some sort. However, the form of contracts may vary. Formal contracts create binding relationships between contracting parties and are thus beneficial in most transactions between two companies (Kwok *et al.*, 2008). A formal contract should be so specified that an objective third party can verify it afterwards. That enforcing by an external party is what makes the contract tightly binding (Baker *et al.*, 2002). If the world was perfect contracts wouldn't be needed, but as there are imperfections in the world contracts are made for both parties to know their rights and responsibilities (Monczka *et al.*, 2009, 500-501).

Contracts often consider the aspects of time, quantity, quality, and price and the uncertainty around the topics (Giannoccaro and Pontrandolfo, 2004). They take both legal and economic aspects into account (Milosevic *et al.*, 1995). Some of the key points of information written in the contract are contracting parties, effective date, payment terms, liabilities, and contracting period. On top of that there are usually many clauses defining general terms, scopes, deliveries, and different specifications. Additionally, it is important to define the rights and responsibilities of parties as well as the confidentiality of specified information (Monczka *et al.*, 2009, 501-504).

One way of looking into theory of contracts is from the point of view of uncertainty or hazards. Contracts are used to deal with possible hazards in exchange. Commonly seen hazards are specificity of assets, difficulty of measurement, and uncertainty of technological development (Poppo and Zenger, 2002). Contracts cannot be completed to include all possible events in the world. However, parties are expected to be reasonable in their behavior even if they technically would not need to. Being reasonable builds trust and reputation (Hart and Holmström, 1986, 102).

Contracts are often made for a limited period of time and then renegotiated which increases the amount of documents companies possess (Hart and Holmström, 1986, 91). Committing to long-term contracts leads to lower transaction costs than making repeated one-off contracts, while short-term contracts can be experienced as safer choices due to unknown future (Ring and Van de Ven, 1992). Contracts can lead to long-term relationships with trust and collaboration, even if from different point of view they can be considered as signs of distrust (Poppo and Zenger, 2002). Long-term contracts can be used to lock in partners to relationships, which provides security of supply (Hart and Holmström, 1986, 53). This works for both parties and creates predictability for the future, whether it considers cash flows or material flows.

Contracts play an essential role in determining whether a commercial relationship has been successful or a failure. They are also a cheap way to avoid long and costly legal fights, as long as enough attention is paid in the process of contract formulation (Monczka *et al.*, 2009, 500). A common rule of contracts is that they need to be kept. The party who breaks a contract needs to compensate the damage caused (Hart and Holmström, 1986, 59). Contracting parties try to achieve optimized benefits and costs by negotiating a contract (Milosevic *et al.*, 1995). However, as the customization of contracts can take time and be expensive, some managers rely more on informal relational governance, while Poppo and Zenger (2002) suggested that they should be used as complements rather than substitutes. They state that complex contracts should be formed only if the violation of the contract comes with a high cost. As relational governance is not so formal, it relies on trustworthiness and close collaboration. Still, it does not come without a cost, as it ties people and their time in maintaining the collaboration and trust. Nevertheless, companies should have a process of negotiating a contract when needed.

2.2.1 Contract creation process

Monczka *et al.* (2009, 464-470) introduced the procedure of getting into a contract as a five-step process as shown in Figure 2. Firstly, the need for a purchase is identified or forecasted. This can happen for instance by identifying a need for a new material or if the inventory level of a material hits a predetermined reorder point. Secondly, it needs to be decided whether negotiations are needed or not. Plentiful negotiations might be unnecessary if the procured materials are commodities or there are no hazards such as those introduced earlier by Poppo and Zenger (2002) related to the purchase. If the negotiations are deemed unessential the third and the fourth step can be bypassed. Thirdly, Monczka *et al.* (2009, 464-470) proposed that negotiations require thorough planning. This includes identifying potential parties and gathering all the available information. Information gathering can include sending out requests for proposal (RFP) and requests for quotation (RFQ) for the potential suppliers to fill out. Those are effective in gathering necessary information in predetermined form.

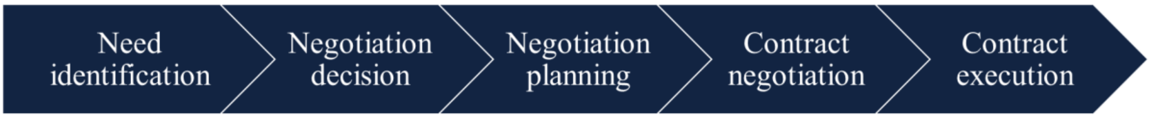


Figure 2. Contract negotiation process.

Fourthly, the negotiation phase takes place. Parties included in the negotiation are optimizing their outcome by demanding things important to them and compensating it by giving concessions in less important topics (Milosevic *et al.*, 1995). When negotiating on long-term partnerships it is increasingly important to try to find a solution that is beneficial for all parties so that the relationship can grow stronger. In the negotiation phase the wording of the contract is also discussed and drafts are sent back and forth other for reviewing. Lastly, based on the division of Monczka *et al.* (2009, 464-470) the process is finalized by executing the contract. When the agreement has been reached and signed all parties are bound to the contract having their own responsibilities. All parties should keep track on the performance of the other parties and give feedback if needed. This is where the daily work of contract management begins.

2.2.2 Contract management

As mentioned above, contracts are in the very core of businesses. They are needed in dealing and building relationships with suppliers and partners (Kwok *et al.*, 2008). Thus, they need to be stored safely and managed actively. Procurement professionals are regularly working with contracts which develops their skills in creating and managing them (Monczka *et al.*, 2009, 500). Many companies store and manage their contracts in a certain electronic document management system to increase business process efficiency (Sprague Jr., 1995).

The process of managing contracts starts already in the negotiation phase. Contract drafts can be exchanged between the contracting parties repeatedly. However, once the satisfying wording has been achieved and the contract signed the major work for procurement professionals in contract management begins. It is their responsibility to daily make sure that the contracts are being fulfilled, and in case of breaches resolve the conflicts (Monczka *et al.*, 2009, 500-501).

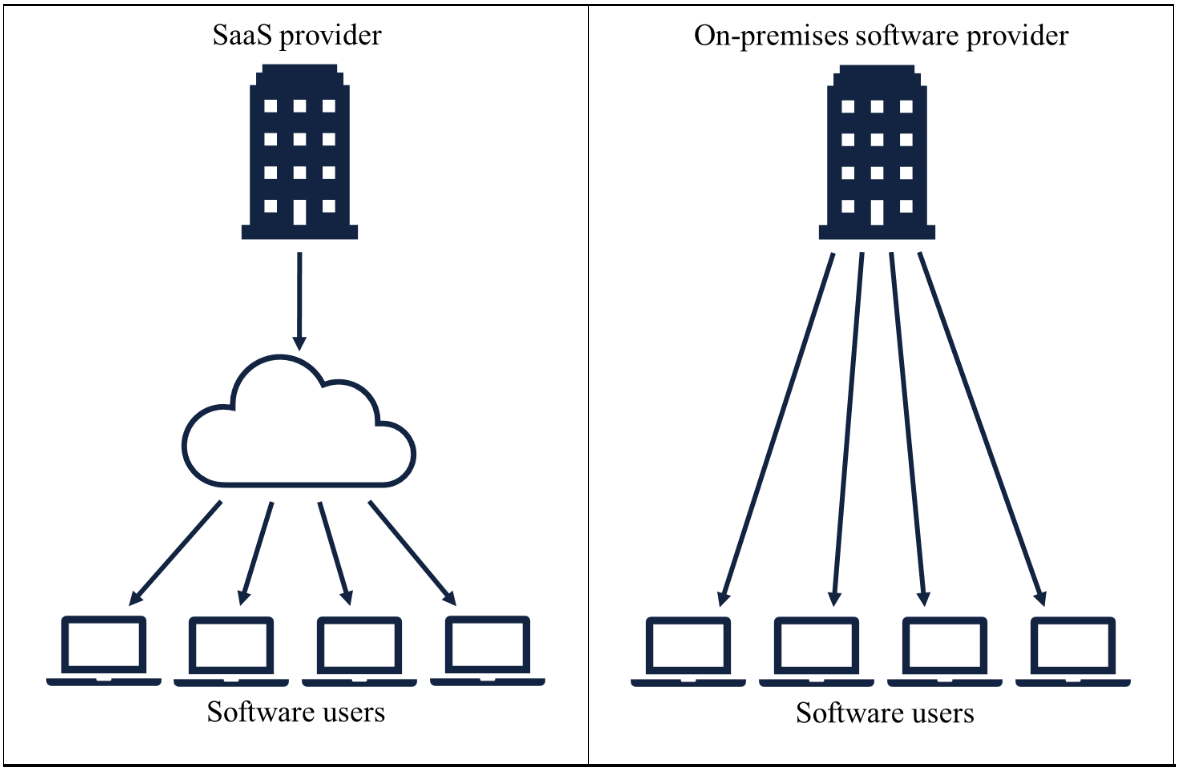
When concentrating on contracts, document management tools are called contract management systems. Contract management systems are part of the group of Supplier Relationship Management tools together with such tools as spend analysis, e-sourcing, supplier performance, and cost reporting. These tools can be used in making decisions on future actions (Monczka *et al.*, 2009, 687-688). Storing contractual documents being the basis of the systems, depending on the software many managing actions can be taken within the tool. The expiration and payment of contracted activities can be followed, and contracts can be renewed or terminated (Chalkidis and Androutsopoulos, 2017). Contract management software can also be used to actively follow the whole lifecycle of a contract, including communicating with contracting parties and digitally signing newly negotiated contracts. So, it can be used both internally and in collaboration with partners. The software can include a workflow for contracts in which each step has an assigned user who must act in order to move the contract forward. The contract management tool can also be configured to send automated reminder emails if the contract should be renegotiated or a person has not completed the tasks assigned to them (Kwok *et al.*, 2008). Contract management software delivers real-time data and analytics for the daily users and can help in keeping track of volumes, due dates, changes in terms, and overall compliance on contracts. With these tools the full potential of contracts can be realized (Monczka *et al.*, 2009, 692-693).

Kwok *et al.* (2008) argued that it is small and medium-sized businesses that should utilize external software for contract management while larger corporations can build one on their own. However, if the contractual information is combined with other procurement data it is the large companies that have most value to gain. Still, extracting the information from contracts that is needed in decision-making is often big manual effort when big enterprises have thousands of active contracts (Chalkidis and Androutsopoulos, 2017). Thus, automating the task of contractual information extraction is extremely beneficial. Software as a service providers bring feasible solutions for companies of all sizes, as their software can be either given out of the box or customized based on customer's needs and the prices they are willing to pay (Kwok *et al.*, 2008). As the case company of this study operates in the business of software as a service, next subsection sheds light on this particular topic.

2.3 Software as a service

While traditionally software is bought to be operated on company’s own premises, software as a service (SaaS) is delivered to be used without changes in ownership and possession. The client buys a right to use the software from the provider (Turner *et al.*, 2003). Software as a service is an integration of software to an online infrastructure which enables access to the software independent of place and time (Sääksjärvi *et al.*, 2005). Software that is delivered as a service instead of an on-premises solution assists procurement organizations in moving costs from capital expenditure to operational expenditure. The software is not hosted on the client’s own servers but off-premises and is used via web. It removes the need for internal maintenance of software and frees more resources to be used in core operations (Godse and Mulik, 2009).

Sääksjärvi *et al.* (2005) recognized that from the client’s point of view the biggest risks of SaaS are possibility of an information security breach, more standardized and less tailored solution, and uptime issues concerning online platforms. The biggest benefits, on the other hand, are short implementation period, mitigated need for software version management, and availability of software anytime and anywhere. The client organizations have also the possibility to access the so-called best-of-breed solutions, which could be too expensive to afford as on-premises software. The possibility to update and manage the software remotely benefit both provider and client. SaaS providers can also roll out updates and fixes to all users at once as they do not need to travel physically to the users. This possibility is illustrated in Figure 3. Godse and Mulik (2009) argued that the intuitiveness of the software's user interface as well as the availability of help and support are within the most important factors to the users when using a software as a service. The costs of procuring a software as a service consists of two factors: annual subscription fee and one-off implementation fee. On top of that additional consultation and configuration fees can occur.



a) Software as a service based system

b) On-premises based system

Figure 3. Software as a service and on-premises software.

The size of a client company does not affect the adoption and usage rate of a SaaS solution. Therefore, large companies are more attracting to software providers as there lies a higher revenue potential. It is beneficial for SaaS providers to standardize their solution as much as possible to be able to serve higher number of clients more easily (Benlian *et al.*, 2009). Moving from software product to SaaS business model can leverage the expansion to new geographical areas if the provider can solve the risks associated with the change (Lassila, 2006). The biggest risks for SaaS providers include high investment in the beginning of business and issues with technical solutions when scaling up the business. The barrier of entry especially when moving from on-premises software product to SaaS business can be high. The benefits for SaaS provider include the predictability of cash flows due to the recurring payments as well as economies of scale due to standardization (Sääksjärvi *et al.*, 2005). Godse and Mulik (2009) stated that SaaS providers should be able to integrate to other systems, scale their availability even when the number of users is peaking, and be reliable at all times. One of the key points from client's point of view is that the provider is extremely reliable in the information security.

2.4 Natural language processing

Natural language processing (NLP) is a process of utilizing machine reading in getting value out of text. It includes a plethora of sub-fields for different tasks. Some of the most well-known applications are machine translation (Cho *et al.*, 2014), chatbots (Li *et al.*, 2016), and voice assistants (Hoy, 2018). Other applications, and more relevant to this study, are information retrieval and information extraction (Riloff and Lehnert, 1994). Characteristically, NLP solutions need plenty of prior information to make sense out of complex interdependencies (Ratinov and Roth, 2009). The four basic steps of NLP systems are part-of-speech (POS) tagging, chunking, named entity recognition (NER), and semantic role labeling (SRL) (Collobert *et al.*, 2011). In this section, each of the steps are introduced shortly. NER is given most attention due to its highest relevance for this study. After that, the NLP applications of information retrieval and information extraction are presented due to their close linkage to this study.

2.4.1 Part-of-speech tagging

Part-of-speech (POS) is the syntactic role of a word in text, for instance, singular noun or numeral. POS tagging assigns the corresponding tags to the words (Collobert *et al.*, 2011). An example of a tagged sentence is shown in Figure 4 including such tags as personal pronoun (PRP), past tense of a verb (VBD), word to (TO), base form of a verb (VB), determiner (DT), and singular form of a noun (NN). The task of POS tagging itself is not a new phenomenon. The basis was laid already by Chomsky (1965). He combined linguistics with statistical models by studying whether finite-state Markov process could be used to structure the grammar of English sentences. The results were encouraging even though only simple sentences were used.

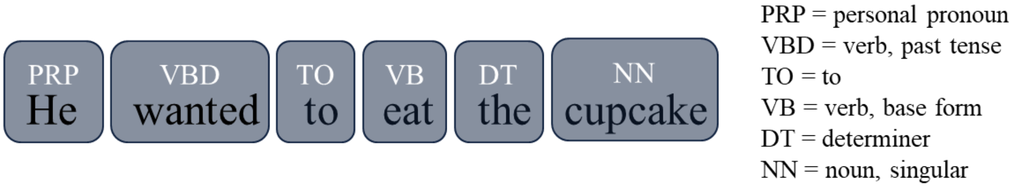


Figure 4. Part-of-speech tagging.

In the early work, Church (1989) proposed a stochastic model for tag assigning. He clarified that context is needed in order to determine the correct POS tag for a word. However, his model was unidirectional meaning that the role of a word is determined only by the words before or after the target word, and not both ways. Toutanova et al. (2003) used a bidirectional dependency network to consider a wider context for each word. They achieved superior results from using these combined features in tag assigning.

2.4.2 Chunking

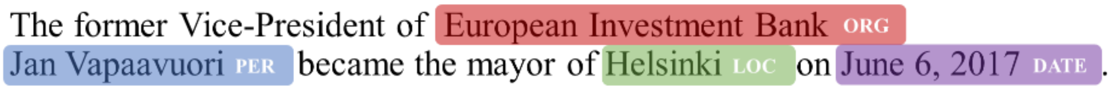
In the task of chunking parts of a sentence are assigned with labels based on their syntactic role. What differentiates chunking from POS tagging is that phrases resulted from chunking can consist of multiple words, while POS tags are assigned to each separate word. (Sang and Buchholz, 2000) However, it is also possible to label each individual word in chunking based on their location in the phrase, whether it is in the beginning (B), inside (I), in the end (E) of a phrase, or even outside the scope of the chunks (O) (Collobert *et al.*, 2011).

An example of a chunk is a verb phrase (VP). In a sentence *She has been jogging a lot* the chunk *has been jogging* would be assigned the VP label, wherein separate words would be labeled as B-VP, I-VP, and E-VP, respectively.

2.4.3 Named entity recognition

Named entity recognition (NER) is also part of the natural language processing. In the task, chunks of text in a larger context are labeled into predetermined categories based on their semantic information (Collobert *et al.*, 2011). The task is two-fold: first, the entities are identified from text, and then they are classified into the most suitable categories (Ekbal and Saha, 2016). The term of Named Entities was coined by Grishman and Sundheim (1996) for the use of sixth Message Understanding Conference (MUC-6). It first covered the categories of people, organizations, and geographic places. According to Nadeau and Sekine (2007), later the scope has been extended to include such entities as dates, currencies, and numbers. Named entities can also vary based on the texts in question. While names of legal norms and regulations are important in legal research (Leitner *et al.*, 2019), medical researchers might be interested in genes and viruses (Zhou *et al.*, 2004). An example of a sentence with tagged named entities is illustrated in Figure 5. There is an organization

(ORG), a personal name (PER), a geographic location (LOC), and a date (DATE) found in the sentence.



The former Vice-President of European Investment Bank Jan Vapaavuori became the mayor of Helsinki on June 6, 2017.

Figure 5. Named entity recognition.

A central task for a NER system is to find entities that are not yet familiar to it (Nadeau and Sekine, 2007). In the early days, to create a new named entity recognition tool, a lot of manual annotation work was needed (Grishman and Sundheim, 1996). Early systems were based on hand-crafted rules while more novel ones rely on statistical and machine learning methods (Nadeau and Sekine, 2007). NER tasks can also be handled with a combination of rules and machine learning (Ekbal and Saha, 2016). Zhang and Johnson (2003) suggest that high quality rules might still be necessary in capturing difficult linguistic features. Nowadays there are both commercial and open-source tools for the task, so that the whole tool is not needed to be built from scratch. Usually, the tools include large base dictionaries of entities (Collobert *et al.*, 2011). The importance of prior and external information in NER tasks was also noted by Ratinov and Roth (2009).

The systems for NER usually consist of sequential prediction problems (Ratinov and Roth, 2009). In these problems, for instance, previous predictions, word types, and words around the predicted chunk can be used in estimating the conditional probabilities (Zhang and Johnson, 2003). The probabilities are then used to evaluate the most suitable category for the text chunk. As mentioned above, external knowledge can enhance the performance of an NER system. Two main ways of providing this information are unlabeled text and gazetteers (Ratinov and Roth, 2009). Miller *et al.* (2004) introduced a model with which the NER system could learn the category of a text chunk based on the similarity of its context to other chunks. For instance, words *January* and *December* are assigned to the same cluster based on their similar context of occurrence. This so-called word clustering could be taught with unlabeled text and then be populated to new texts.

On the other hand, gazetteers are external dictionaries from which the NER system can perform lookups (Ratinov and Roth, 2009). Gazetteers as themselves are not extensive enough to act as a single source of information for NER systems, and they are also difficult to maintain. For these reasons, Kazama and Torisawa (2007) developed an automated gazetteer extractor for the whole English Wikipedia. That worked relatively well due to its extensiveness and constant content updating, but Ratinov and Roth (2009) argued that the best results have been obtained by combining gazetteers with machine learning algorithms. By contrast, Lample *et al*. (2016) stated that rule-creation and gazetteer are domain and language specific as well as too complex to maintain. Thus, they created neural machine learning models without external information that can be applied to various languages and domains.

Krishnan and Manning (2006) noted a problem with many of the previously built NER systems. The models had utilized only relatively small context of words in assigning tags to text chunks, which could lead to the same entity mentioned elsewhere in the text being categorized differently. Using only local information the tags may include inconsistencies. Thus, the non-local information in the text must be considered for better accuracy. This problem can be addressed with penalizing the algorithm for inconsistent categorizations (Finkel *et al.*, 2005). However, the consistency problem is not trivial due to ambiguity of words. For instance, *Washington* can carry different meanings in varying contexts. It can be a geographical location (*Washington D.C.*), last name (*George Washington*), and a sports team (*Washington Capitals*) all of which can be referred to with the same single word.

Even though the context of the text chunk is important, some simple language independent linguistic features have been widely studied and can easily enhance the performance of NER systems (Zhang and Johnson, 2003). The linguistic information can include such features as affixes, POS tags, the case of the letters, and predetermined trigger words. The features can be applied to both the chunk in question and the ones around it. The underlying technologies in NER systems are manifold. Varying algorithms have been utilized in them, hidden Markov models (Zhou *et al.*, 2004), perceptrons (Buitinck and Marx, 2012), and conditional random fields (Lample *et al.*, 2016) being among the most popular ones.

2.4.4 Semantic role labeling

Verbs are in the center of semantic role labeling (SRL). In SRL tasks, text chunks that act as arguments for verbs are labeled (Collobert *et al.*, 2011). The labels are based on the semantic roles of chunks such as being an actor, patient, or attribute for the verb (Palmer *et al.*, 2005). The available roles in a sentence are affected by the type of the verb as not all types of verbs can have every argument (Korhonen and Briscoe, 2004). Figure 6 contains an example sentence with labelled semantic roles.

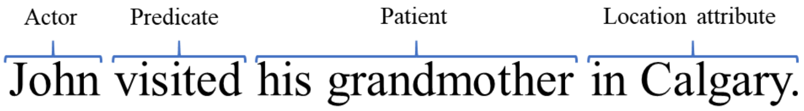


Figure 6. Semantic role labeling.

As shown in the Figure 6, the predicate verb can be surrounded by such labels as actor and patient, while the sentences often contain other attributes as well, for instance location. Gildea and Jurafsky (2002) used statistical classifiers to automatically complete the SRL task. They noted that one of the biggest challenges is that a similar looking verb can have different roles depending on the context. However, if the task can be done successfully it can prove useful in further applications such as text understanding and information extraction (Palmer *et al.*, 2005). Some of the best results in SRL have been achieved by utilizing the results from NER, POS tagging, and chunking in the task (Pradhan *et al.*, 2003). One application that can benefit from mentioned techniques is information retrieval. Next subsection contains further information on that application.

2.4.5 Information retrieval

The amount of textual information available has been increasing rapidly due to internet and electronic documents (Califf and Mooney, 1999) which has made it impossible for humans to go through all the text when in search of relevant information. Information retrieval systems answer to this problem by returning relevant documents from collection based on user’s needs and questions (Strzalkowski, 1995), which is why it has also been called document retrieval or text retrieval (Voorhees, 1999).

Basic information retrieval systems are based on indexing and matching. Indexing includes going through the text and selecting representing keywords for the document (Strzalkowski, 1995). The indexing can be done by tokenizing the text into words based on white space, removing common words such as pronouns and prepositions (stop words), and by stemming the words, i.e., removing suffixes. In the matching step, the keywords gotten from indexing are compared to the search terms of the system’s user. An example of a basic information retrieval system is any of the internet search engines, such as Google or Bing

(Voorhees, 1999). Also, the synonyms of the keywords can be used in fetching the set of documents the user is interested in (Moens, 2006, 12-13). The closely related application of information extraction is introduced in the next subsection.

2.4.6 Information extraction

Information extraction is the sub-field of natural language processing utilized in this study. It is a process for analyzing text data and finding both relationships and semantic entities from it (Grishman, 2015). It is not to be mixed with previously introduced information retrieval, which is used for classifying and indexing individual documents from large collections (Freitag, 2000), while in information extraction certain predetermined information is sought from inside the documents (Lee *et al.*, 2019). However, Moens (2006, 13) suggested that the results from information extraction can be used in enhancing the accuracy of the information retrieval tools. Information extraction itself is never the end goal but a way of finding information relevant to decision making (Moens, 2006, 226).

Information extraction exists to solve a problem with four characteristics: information is requested, solution to the request is in unstructured format, due to the laboriousness of the task humans are not able to process the amount of data, and due to the unstructured nature computers cannot query the information directly (Moens, 2006, 1-2). The process of information extraction can be used to both summarizing text by turning predetermined useful information into structured format (Cardie, 1997) and filling slots in a form (Lavelli *et al.*, 2008; Freitag, 2000). This process is visualized in Figure 7.

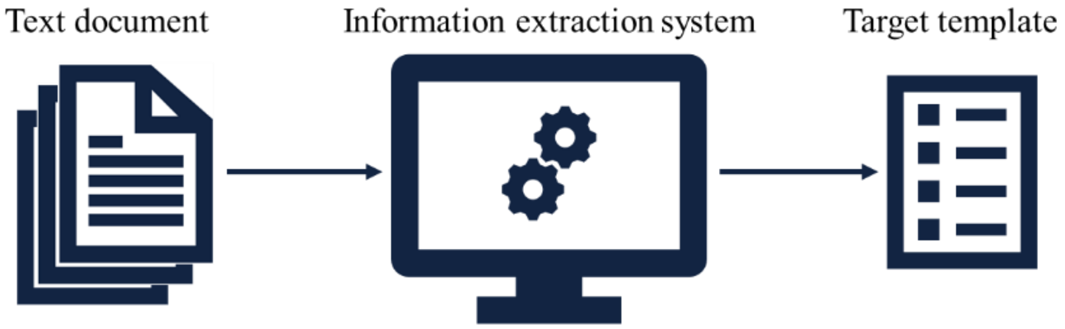


Figure 7. Information extraction process.

Information extraction is applied to problems where the outcome of the extraction is somewhat defined, some specific information is tried to be found within the text, but the exact wording and location are not known. In the process the unstructured information is converted into structured by extracting relevant information and formatting it into predetermined structure (Moens, 2006, 7-9). Cardie (1997) divided the information extraction process into five stages: tokenization and tagging, sentence analysis, extraction, merging, and template generation. In the first phase the textual data is divided into small pieces called tokens, which then are given tags based on their parts of speech. In the second phase if necessary, the tokens are combined into phrases to find their meaning in the context of a sentence. Next, the extraction itself is conducted for the predetermined information, for instance, date and participants of an event. After that, it is sometimes necessary to do merging of tokens to combine synonyms. Lastly, the extracted information is filled into a pre-structured template.

In the context of business-related texts, such as contracts, the structured and unstructured data exist in parallel. These kinds of combined unstructured and structured documents are suboptimal for automation of information extraction as even though the structuredness of part of the data eases the search for important data, the contract as a whole might not be in standardized form. Layouts and orders of tables may vary between contractual documents. When it comes to the legal domain, the field of information extraction has been extremely understudied. There are lots of text documents with all the required information available, but the wordings and structures of the documents can be difficult and unclear for both people and machines to understand. Contracts lie in the crossroads of business and legal domains. Companies are willing to constantly follow all the available information about their business, including information about their contracts which thus makes the contractual domain an interesting field to be studied. (Moens, 2006, 213-215)

There are only few contract related information extraction studies. Chalkidis *et al.* (2017) experimented various methods for extracting contractual information. They divided each contract into relevant extraction zones from which they searched for the desired information. This division into sections can be achieved with somewhat standardized contract formats, but for non-standard formats it is not that simple. Their methods included both hand-crafted rules and machine learning algorithms such as logistic regression and support vector machines. They also investigated the possibility to use NER algorithms in the task. The best results were achieved by combining different methods. Later on, Chalkidis and Androutsopoulos (2017) used deep learning methods with the same data set and achieved promising results. However, such aspects as the speed of performance and possibility to use the methods in practice were ignored. Going still forward, Chalkidis *et al.* (2019) found that using additional features such as POS tags did not improve the results of their deep learning contract element extraction. They also stated that using generic corpora as a background information did not increase the performance, but it would be worth experimenting with pre-training the models with relevant contract data.

The information extraction tasks can be extremely time-consuming to do manually, due to which machine learning can be a useful tool (Califf and Mooney, 1999). Still, Jackson *et al*. (2003) argued that the system must achieve an appropriate performance to be accepted by the users. If the information extraction system is inaccurate, it requires manual correcting afterwards, which can be as laborious as doing the whole task by hand. The automation of work and how that affects the experienced meaningfulness are discussed in the next subchapter.

2.5 Work automation and meaningful work

World Economic Forum (2020) concluded in their *The Future of Jobs Report 2020* that the rate of automation in work will increase from current 33 % in 2020 to 47 % in 2025. Currently, the field of procurement is experiencing the effects of digital transformation and work automation. Bals *et al.* (2019) suggested that one of the important competencies of procurement professionals in the future is the ability to automate their work. However, not all tasks can be automated easily. Most optimal tasks for automation are those that are routine-like and include only little cognitive challenges, as illustrated in Figure 8. Furthermore, the steps within the task should be able to be defined relatively easily (Asatiani and Penttinen, 2016). Automating the more time consuming and less interesting work, the meaningfulness of the job increases. People can concentrate on the more purposeful tasks or work more in interpersonal relationships (Smids *et al.*, 2019).

People can find three kinds of motivational reasons in their work. Some focus on financial rewards, while others are more interested in advancement and career development. The third motivation is finding meaningfulness from work (Wrzesniewski *et al.*, 1997). Martela and Riekki (2018) argued that the psychological factors of meaningfulness in work are autonomy, competence, relatedness, and beneficence. Autonomy is the feeling of being able to affect something in the work instead of being pressured from the outside. Competence is about capabilities, meaning that one is aware of their professionalism in some tasks. Relatedness concerns connectedness to others and belonging to a community. Beneficence means contributing to the society in a way that benefits others and thus builds one's own well-being.

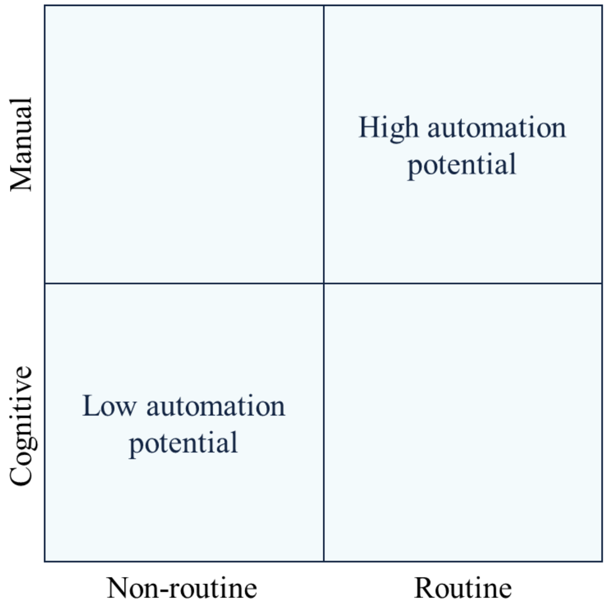


Figure 8. Assessing potential automation of a task (adapted from Asatiani and Penttinen, 2016).

Technology has always affected jobs, but never decreased the number of jobs in total. As the content of work changes, people need to adapt their skills accordingly. Automating the tasks that people are unwilling to do or not so good at frees their time to more interesting and suitable tasks (Cascio and Montealegre, 2016). Moreover, automation of boring work can lead to developing new skills (Smids *et al.*, 2019). While routine tasks are automated, employees have the chance to be part of designing their roles. This so-called job crafting motivates people in their daily work (Rintala and Suolanen, 2005).

There are also tasks that if automated can lead to less meaningful work if those tasks are the most interesting and complex (Smids *et al.*, 2019). If automation leads to fewer social interactions with colleagues, it can become a negative factor. This is one of the reasons people have divided opinions about automation and robotics. People having the attitude that robots and computers are there to help, not to take their job, have the most positive experiences in automation of some of their tasks (Rintala and Suolanen, 2005). The more significant the tasks are perceived the better people perform in them. Furthermore, having better knowledge on how their work benefits others increases the motivation to perform well

(Grant, 2008). Rintala and Suolanen (2005) had a similar view to other literature, that automation often leads to learning new skills. This also leads to the division of thoughts around automation as people can, but also must, learn to keep up with the development. Some find it stressful others enjoy it.

3 Case introduction

In this section the general characteristics of the case study are discussed. First, the case company Sievo Oy (Sievo) and their procurement analytics software are introduced. Especially the contract management solution is regarded due to it being in the spotlight of this study. Then, the problem of the case is described. After that, the data used in the study is presented.

3.1 Sievo

Sievo is a leading procurement analytics software provider established in 2003. They offer software in the areas of spend analysis, savings lifecycle, materials forecasting, procurement benchmarking, and contract management. Combining this offering they provide best-of-breed procurement analytics solutions to world-class procurement organizations in companies such as Deutsche Telekom, Carlsberg, and Kellogg's. Even though the tools are associated to procurement, also finance organizations benefit from using these tools. The collaborative use of procurement analytics solutions between procurement and finance is also motivated by procurement being in charge of most of the spend and finance being interested in all the money flowing in and out, as also noted by Bals *et al.* (2019).

The cornerstone of all procurement analytics is spend analysis. It is the routine of examining procurement spending in order to achieve better results in measures such as decreased costs, increased efficiency, and improved relationships with suppliers. With spend analysis tools procurement organizations can spot new savings opportunities and manage their risks. Once the savings opportunities have been identified, the savings lifecycle tools come into use. Sievo offers tools for managing the portfolio of savings initiatives and measuring the savings that have realized. Also closely related to spend analysis is materials forecasting. Using forecasting tools, procurement and finance professionals can focus on the future opportunities instead of only reporting the past performance. Furthermore, the benchmarking solutions provided by Sievo build on their big customer base in spend analysis. Their peer benchmarking tool gives users access in anonymized spend data from other companies, while the market benchmarking tool compares the paid prices to the market indices.

The data for all the above-mentioned tools comes from varying sources. Most usual sources are companies' enterprise resource planning software, financial software, or other internal systems. Besides those, the data can be further enriched with third-party data such as indices and supplier provided data but also other internal data such as contracts. The flow of data in Sievo's procurement analytics software is visualized in Figure 9. Contract management is the one of the solutions offered by Sievo. It is also the most relevant to this study. As mentioned already earlier, the basis of contract management tools is storing contractual documents on a safe and centralized platform. Unlike non-specialized storages such as physical shelves or digital folders, contract management software allows easy search and analysis of contractual data. Furthermore, linking this data to information about spending enables seeing big picture and opportunities of for instance better contract term negotiations. Adding new contract documents in the Sievo's contract management tool takes only minutes and can lead to large benefits for its users. Contract metadata can be stored in the system in order to see the key information with a quick glance. This is also the context of the problem in this study, more of which in the next subsection.

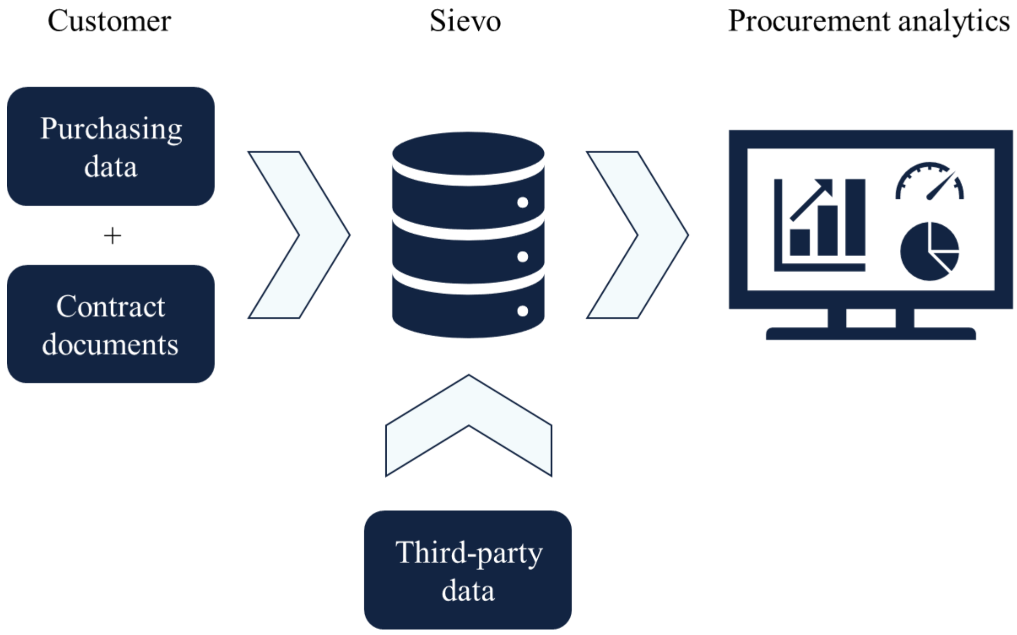


Figure 9. Data flow in Sievo.

3.2 Problem

As mentioned above, one of the benefits of contract management software for procurement professionals is getting a quick overview of a contract without skimming through multi-page text documents. However, the users must be aware of the contents of the documents they upload into the system as they need to manually input metadata for the contracts. Within the metadata some of the most important data points are contracting parties, effective date of the contract, payment terms, contracting period, and liabilities agreed in the contract. Even if the contract management tool enables finding the key information quickly once the contract has been uploaded, the upload process can be somewhat laborious if the user is not too familiar with the structure of the contract. Inputting the metadata to open text fields in the system reduces the time used in more productive work but proves beneficial in the long run. However, the saved time could be used in more important tasks if the manual work could be reduced by automating the filling of the metadata fields. As large companies can have hundreds of active contracts and thousands of expired ones the global time saving would be enormous. The benefit would become visible especially when implementing the contract management software as plenty of active and expired contract documents are uploaded at once. Furthermore, the automation of manual steps would increase the convenience of the software use and thus improve the meaningfulness of procurement work.

The benefits of automating the manual inputting of contract metadata have been described above. Therefore, the goal is to improve the usability of the contract management software and increase the efficiency and work meaningfulness. In order to achieve this, a solution which automatically goes through the contract documents, spots the correct pieces of information, and inputs them into the open text fields needs to be developed. This study concentrates on the stage of finding the key information. The earlier mentioned data points of contracting parties, effective date of the contract, payment terms, contracting period, and liabilities are included in the scope. As the software will need further development to be actual production ready solution, the integration to Sievo’s contract management software will be discussed within the company later. However, the potential of becoming part of the software is discussed in the later sections. The following subsection concentrates on describing the data used in this study to solve the above introduced problem.

3.3 Data

The data used in this study is dichotomous. Firstly, real-world contract documents are used in both developing the software and quantitatively assessing its success. Secondly, interviews with three Sievo employees are used to get qualitative feedback on the solution and its potential. These data types are introduced in the following subsections.

3.3.1 Quantitative data

The contracts used consist of freely available sample documents from LexPredict (2017) and documents provided by Sievo. Contracts from different sources were selected to include variation in the form of documents. All of the documents were readily in machine-readable format, meaning that they were not scanned images of text documents, but actual text recorded in copiable and readable form. In a more sophisticated and production-ready solution, some kind of optical character recognition (OCR) tool would be needed to complement the information extraction. In short, OCR tools recognize text from within images and turns it into machine readable text. These images can be for instance the above-mentioned scanned contracts. A sophisticated OCR technology would be compulsory to allow the inputs as scanned documents, which is the case for most contracts. A broader spread of digital signatures would remove the need for OCR.

In total 52 contract documents were used in this study. The documents from LexPredict (2017) were readily in .txt file format, while those supplied by Sievo were in .pdf format and thus needed to be transformed into .txt for the software to be able to process all files in a similar manner. The documents consisted of agreements between two parties. They were from the fields of business such as software licensing and construction projects. All of the contracts utilized were written in English. Further developed and more universal tools could be able to handle documents in other languages too, but for the sake of simplicity only English contracts were regarded.

3.3.2 Qualitative data

After having results from the qualitative analysis, interviews within Sievo were conducted to collect qualitative data of the success and potential of the tool created in the study. The interviews were semi-structured to allow flexibility in both asking questions and answering to them. Semi-structured interviews have a predefined outline, but they tolerate variance in the order of asking questions and give the opportunity to ask follow-up questions about previous answers. The interview type was chosen due to differing profiles of the interviewees and to allow questions that have not been planned before-hand if they emerge during a conversation. This also makes the data they provide non-standard as the interviews can have possible sidelines. (Barratt *et al.*, 2011)

Three interviews were conducted in total. All of them were held in Finnish. The interviewees work in key roles at Sievo when it comes to their contract management software. Their job titles were Product manager, Software engineer, and Key account manager. All the interviews were held online on Microsoft Teams video conference platform due to the COVID-19 pandemic situation. Each interview was booked a 45-minute slot. The interviews were recorded and transcribed to be analyzed by the author. The more detailed information about the interview methods as wells as methods concerning quantitative data is presented in the next section.

4 Research methodology

As the data was dichotomous, so was the methodology in this research. The research can be divided into quantitative and qualitative parts. So according to Bryman (2006), this study is a multi-strategy research. This kind of combination of qualitative and quantitative research was conducted to enhance the validity of results by benefitting from the strengths of both methods (Greene *et al.*, 1989). In this section both of these methods of study will be introduced in detail.

4.1 Quantitative methods

The objective of the research was to study the possibility of being able to automatically extract information from contracts. The tool created for the purpose in this research was developed with Python programming language. However, before moving to the programming phase, the contracts needed to be pre-screened. Both the screening of contracts and the structure of the program are introduced in the following subsections.

4.1.1 Gold standards

In this first stage of the empirical part of this study, the objective data points to be extracted from the contractual documents was discussed with experts in the case company. As mentioned before, the data points selected were contracting parties, effective date of the contract, payment terms, contracting period, and liabilities. These were selected due to their regular appearance in contracts as well as their diversity in both data type and context of existence. The first three data points were discussed to be in the main focus of this research, while the latter two would work as interesting but not as important points of exploratory analysis.

Out of the collection of contracts, 26 documents were randomly picked into the training set. These contract documents were read with human eyes and the correct results for the selected pieces of information to be extracted were sought. Independent of the field of study, these data points manually flagged as correct are commonly referred to as gold standards (see e.g. Smith, 2002; Snow *et al.*, 2008; Chalkidis *et al.*, 2017). This manual work makes the phase time consuming and laborious due to which the amount of data used in the research was left considerably small. After the gold standards and information about their context of existence were collected, the outline of the program could be structured.

4.1.2 Programming foundations

In this subsection the structure of the developed program is presented. As mentioned in the opening of this section, Python was selected as the programming language for this study. The selection was made due to the powerful open-source natural language processing libraries available for Python as well as due to its familiarity to the author. The program design was created on Jupyter Notebook, which is a browser-based platform for programming.

Extracted item	**Keywords**
Contracting parties (sample)	Oyj, AB, BVBA, LTD, Inc, Incorporated, Sarl
Effective date	effective, start date, made as of, entered into
Payment terms	payment term, terms of payment, term of payment, after recei, from recei, completion, advance payment, at closing
Contracting period	contract period, contracting period, valid period
Liabilities	liabilit, liable

Table 1. Lists of keywords.

In the beginning of designing the program, inspired by Chalkidis *et al.* (2017) heuristics of the contexts of existence for each extracted data point were created. Combining their research methods and the training data set collected in the previous phase, lists of keywords were compiled. The keyword list for contracting parties was provided by the case company, as they had previously collected an extensive list of legal entity types. A sample of this list is presented in Table 1, while the comprehensive list can be found in Appendix 1. For the other data points the lists were created entirely based on the sample data. The complete keyword lists for effective date, payment terms, contracting period, and liabilities are shown in Table 1. These lists were gathered from the 26 contracts by going through the surroundings of each gold standard and finding the common denominators as well as variations of wordings. Also, discussions with professionals from the case company were utilized. After the creation of the heuristics, a pipeline for one document going through the extraction was built. The contents of the pipeline are visualized in Figure 10 and introduced more thoroughly in the next subsection.

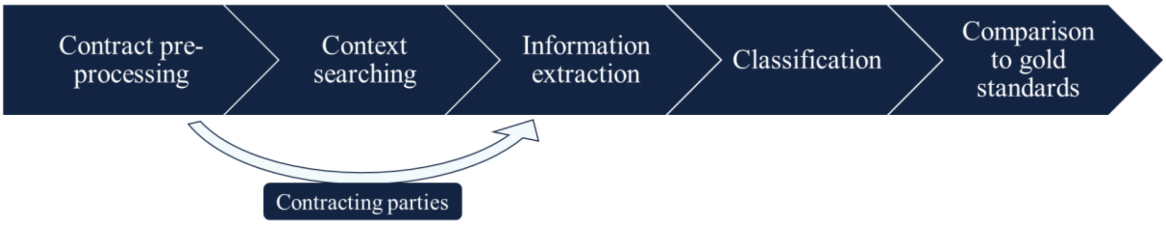


Figure 10. Pipeline for a contract.

4.1.3 Pipeline for a contract

The pipeline for extracting a desired item from contracts was built to process one file at a time. Then the pipeline was repeated for all the files. As shown in Figure 10, the first step in the pipeline was the preprocessing of the file. This was done in order to achieve standardized and clean documents. The files were in UTF-8 format and were normalized using NFKD algorithm. To get the file into a concise and standard form, two consecutive line breaks were replaced with a text " DOUBLE_NEWLINE ", including the single whitespaces around the words. Many of the files had empty or half empty pages in which case this cleanup made the file length much shorter. Single line breaks were further replaced with a single white space. This was a sufficient cleanup at this point of the program. At a later point some further standardization was made for smaller pieces of texts, as will be discussed later. Up until this point the pipeline was similar for all the extracted items but going forward some differentiation was created based on the method of searching for the items in the text. The differences between the methods are summarized in Table 2. However, the general idea behind the pipeline stays the same and will be introduced as such. The exceptions will be remarked and the process for each extracted item presented later.

For all other data points but contracting parties the next step was to find the relevant contexts of appearance within the contract document. This was done based on the lists of keywords presented earlier. It was defined for the software to seek for the keywords and if found, collect the contexts consisting of the keyword and 70 characters both before and after it into a list. The number of characters was determined based on the training data and manual searching of the gold standards. Further cleanup was then conducted for these contexts. All the punctuation and extra white spaces were removed so that there was only one whitespace between each word, and the text was transformed into uppercase form. This standardization was done to achieve the comparability to the manually sought gold standards.

After the contexts were formatted to a standard form, the extraction of the desired data point within the context was done. For payment terms, contracting period, and liability it was done with hand-crafted rules, while for effective date named-entity recognition algorithms were used. Once the extractions had been finished, the found results were classified to either correct or incorrect and compared to the gold standards. The comparison was done based on exact matches, meaning that if the item classified as correct matched the golden standard only partially it was not regarded as matching. A deviation from this rule were the payment terms. The reasoning for that is explained in the subsection dedicated to payment terms. The specifications of extraction and classification methods are introduced in the following subsections with other special features used for each data point.

Extracted item	**Contract pre-processing**	**Context searching**	**NER**	**Rule-based extraction**	**Order classification**	**Frequency classification**	**Gold standards**
Contracting parties	x		x		x	x	x
Effective date	x	x	x		x		x
Payment terms	x	x		x	x		x
Contracting period	x	x		x	x		
Liabilities	x	x		x	x		

Table 2. Comparison of extraction methods for different items.

4.1.4 Contracting parties

As mentioned earlier, the extraction of contracting parties differs from other items in a way that the step of finding relevant contexts was skipped over. This was done due to the fact that there are powerful NER algorithms that can find the organization names within a file with relative ease. In this study NER algorithms from two libraries, *spaCy* (spaCy, 2015) and *flair* (flair, 2018), were utilized. On the one hand, *spaCy* was chosen since it is widely used and pursues to be a fully industrial level natural language processing library. It has also proven to be a well-performing all-round NLP library (Al Omran and Treude, 2017). On the other hand, *flair* was chosen due to it having achieved superior results in NER research (Akbik *et al.*, 2018). The process of extracting contracting parties is presented below in Figure 11.

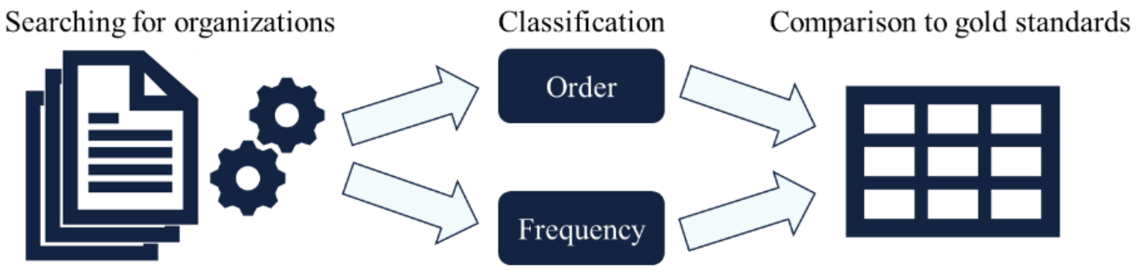


Figure 11. Process for extracting contracting parties.

So, as the context finding phase was not done for contracting parties, the whole document worked as the relevant context. Thus, the approach was to find all the organization names within the contract and then classify them to either correct or incorrect. Both of the NER algorithms had an inbuilt and pretrained feature of searching for organizations. For *spaCy* its *en_core_web_sm* model was used, which is a convolutional neural network algorithm trained on OntoNotes data (Weischedel *et al.*, 2013). When it comes to *flair,* a *ner-ontonotes* model utilizing a recurrent neural network algorithm was selected since it has been trained with the same OntoNotes data. Once the algorithms had searched through the file and stored organization names, these named entities were cleaned the same way the contexts were for other data points. So, the names were transformed into uppercase and punctuation and extra whitespaces were removed. Then these found organizations were filtered with the keyword list consisting of legal entity types to make sure the potential candidates for the correct results do not include such generic words as “SUPPLIER”, “CONTRACTOR”, and “BUYER” which the algorithms could categorize as organizations based on their context of appearance within the contract.

After the cleaned list of potential contracting parties had been achieved, the classification to positive and negative results was made. In the case of contracting parties this was another point of difference comparing to the other extracted items. The classification was done with two methods. Firstly, as the contracts were all made between two parties, they were classified based on their order of appearance in the contract. As Chalkidis *et al*. (2017) observed, contracts often have some sort of introduction section within which the contracting parties are presented. This finding supports the idea of classifying based on order. So, in this method, the two first non-identical organization names found were classified as correct ones, while all the rest were classified as incorrect ones. Secondly, based on the pre-screening of the contracts, a classification was made based on the frequency of the organization name appearing in the contract. This was simply done by calculating the sum of appearances for each individual organization and classifying the two most referred ones as correct results, while the rest were classified as incorrect. In the last step, the gold standards were compared to the classification results. It was recorded whether the gold standard appeared in the positives, negatives, or neither of them. Since the whole document was searched in only contracting parties, the frequency-based classification was not relevant for the other extracted items. However, there were similarities in the method for extracting effective dates, which will be presented in the next subsection.

4.1.5 Effective date

The process for extracting the effective dates was largely similar to the one for contracting parties. So, the first step as in every other extracted item was to preprocess the file. However, there were some differences when it comes to context finding and classification phases. As mentioned earlier the collected keyword list was used to capture the possible contexts of appearance within the contracts, which was not done to contracting parties. In the case of effective date, the keywords were short and simple referring to the date.

The NER algorithms used were the same in effective date as in contracting parties, but in the case of effective date, they did not search through the full contract document but rather through all of the found relevant contexts. The pretrained models of *spaCy* and *flair* are able to search also dates in many formats, which made them extremely useful also in this data point. Not going through the whole documents means that it would not have been applicable to do the classification based on the frequency of occurrences, as only small pieces of the text were captured. Furthermore, while executing the manual prework with the contracts it was noticed that the effective date is most often mentioned only once. Therefore, it was decided for the sake of simplicity that the classification is done based on the order of appearance within the contract. The same classification method was used also for payment terms as presented in the next subsection.

4.1.6 Payment terms

The extraction of payment terms was largely based on keyword matching. While searching for the relevant contexts the keywords contained both words found near the payment terms and parts of the payment terms themselves. Again, these keywords were collected from the sample of contracts pre-screened before looking at the test data set. Compared to the effective date, the wordings were much more diverse, and thus the list of keywords was also longer.

When looking at the extraction itself, NER algorithms could not be utilized as such. Even if both *spaCy* and *flair* can find cardinal numbers – which are often included within payment terms – the payment terms can be expressed also in various verbal ways, for instance “ADVANCE PAYMENT”. It must be noted that NER algorithms could be trained to find also other entities than what they have in their in-built models. However, this was out of the scope of this study. Due to this, the payment term extraction was done by further utilizing keyword matching and regular expressions. The most commonly occurring way of expressing payment terms is a combination of a cardinal number and either word “DAYS” or “MONTH(S)”. All the diverging expressions found were programmed as their own cases. The sample data also included such phrases as “UPON DELIVERY” and “UPON COMPLETION”.

In the case of payment terms, the classification was done solely based on the order of appearance. So, the first expression that matched the hand-crafted rules was classified as positive while the rest matching ones as negatives. Also, the phrases were compared to the gold standards based on partial match rule. However, the partial match was here strictly defined so that the gold standard must be completely included within the phrase. So, there could be extra characters in the expression, but the gold standard must exist there in its entirety and in a perfectly similar form as was manually extracted from the test data to ensure that the correct context was regarded. As will be presented below, contracting period and liabilities were in many ways different from the first three data points. However, due to their similarity to each other they will be considered together in the next subsection.

4.1.7 Liabilities and contracting period

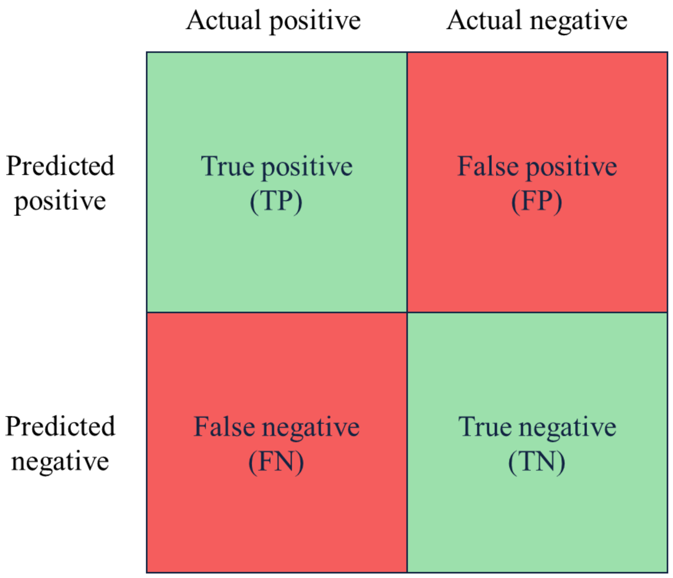
The two objects of exploratory studying – contracting period and liabilities – were processed similarly to each other. They both had their own short list of keywords with which the contexts were searched for. These lists were not as comprehensive as for the three data points addressed earlier. This was due to the fact that it was challenging to find the correct results even manually, so generic keywords were utilized. As for the payment terms, both contracting period and liabilities were extracted from contexts based on further keywords and pattern matching. On one hand, for contracting period, a combination of a number and some temporal expression was searched, as inspired by Chalkidis *et al*. (2017). On the other hand, for liabilities, defining as detailed heuristic was not feasible. Thus, it was decided to use a simple keyword “LIAB” to catch such variations as “liability”, “liabilities”, and “liable”.

As with all but the contracting parties, the classification for the two extracted items in question was done exclusively based on the order of occurrence due to the restrictions of searching from contexts and the object commonly appearing in only one place. The last step with all the other extracted items was to compare to the gold standards. However, as mentioned above it proved to be challenging to manually find the gold standards from the contracts. Thus, the comparison to the correct results was not done, but the success in the fields of contracting period and liabilities was measured with the number of contexts found and by qualitatively reviewing the found contexts. The next subsection, however, discusses the quantitative measures used in this study to evaluate the success of the contract information extraction tool.

4.1.8 Quantitative evaluation criteria

In this subsection the evaluation criteria of the results achieved from the quantitative part of the study are introduced. The measures discussed here are later considered with the quantitative results. However, before presenting the measures some key terms need to be defined. These are true positive (TP), false positive (FP), true negative (TN), and false negative (FN). True positive means that an example is correctly classified as positive. False positive is a negative example which is incorrectly classified as positive. True negative is an example that is correctly classified as negative. False negative is a positive example that is erroneously classified as negative. (Davis and Goadrich, 2006)

The terms defined above were used in deducting confusion matrices. Confusion matrix contains all the above information in a compact form. Figure 12 visualizes a confusion matrix for a binary classification problem, which was the case in this study. Its rows are used in storing the numbers of predicted positives and negatives, while its columns store the numbers of actual positives and negatives. In the cross-sections lie the amounts of true positives, false positives, false negatives, and true negatives. The green color in the figure illustrates the correct classifications (TP in top left and TN in bottom right corner), while the red color stands for incorrect classifications (FP in top right and FN in bottom left corner). (Powers, 2008)



	Actual positive	Actual negative
Predicted positive	True positive (TP)	False positive (FP)
Predicted negative	False negative (FN)	True negative (TN)

Figure 12. Confusion matrix.

The key metrics used in measuring the results in this study can be derived from the numbers stored in confusion matrices. These metrics are precision, recall, F1-score, accuracy, and false positive rate. Precision measures how many of the examples classified as positive are actual positives (Davis and Goadrich, 2006), and is calculated as

$$precision = \frac{TP}{TP + FP}. \tag{1}$$

Recall is a metric for the proportion of actual positive examples that are correctly classified as positive (Powers, 2008). It is calculated from the numbers as

$$recall = \frac{TP}{TP + FN}. \tag{2}$$

F1-score is the harmonic mean of recall and precision, which is often used as the main point of interest in comparing different classification methods (Forman, 2003). F1-score does not take true negatives into account, which is not an issue in the case of information extraction where the number of negatives is countless, but finding the true positives is the real target. Its equation can be written as

$$F1 - score = \frac{2 \cdot precision \cdot recall}{(precision + recall)}. \tag{3}$$

As the main goal of the tool built in the study is to reliably find the true positives, utilizing false positive rate in analyzing the results is justified. It measures the proportion of found negative examples that have been classified incorrectly as positive (Powers, 2008) and can be written in the form of

$$false\ positive\ rate = \frac{FP}{FP + TN}. \tag{4}$$

The value for all above-presented measures ranges between 0 and 1. In precision, recall, and F1-score higher value is primarily better, while in false positive rate smaller value is interpreted as better. Still, making conclusions based on one value only may result to erroneous results as the measures are dependent on for instance the balance of the dataset and each other. In this subsection the quantitative metrics used in evaluation of the tool developed in this study have been presented. These will be discussed further in the later chapters when presenting the results and findings. The next subsection will focus on the qualitative methods of this study.

4.2 Qualitative methods

This subchapter discusses the qualitative methods used in this study. As mentioned earlier, three interviews with employees from Sievo were conducted in order to measure the success of the study and to review the potential of the tool as a basis for a production ready solution. The details of the interviews are listed in Table 3. A semi-structured interview method was chosen to be appropriate in the case of this study. The outline for the interviews is presented in Appendix 2 and the methods are introduced more closely in this subsection.

In selecting the interviewees, an expert sampling method was used. It can be classified under purposive sampling methods. Expert sampling gives researcher the freedom to choose the interviewees based on their qualities and expertise. This is a suggested way of sampling when the research is conducted in a field that has been studied fairly little and the future research possibilities are being assessed. (Etikan *et al.*, 2016)

Interviewee's job title	**Date of interview**	**Length of interview**
Key account manager	December 9, 2020	36 min 16 sec
Software engineer	December 10, 2020	40 min 4 sec
Product manager	December 10, 2020	38 min 31 sec

Table 3. Interview details.

The results from the quantitative part of the study were used as a basis for the interviews. Greene *et al.* (1989) called this the development purpose. They stated that the development purpose of using multi-strategy approach uses "the results from one method to help develop or inform the other method". This is exactly what was done in this study, as the results from the quantitative method were collected first and the interviews were held afterwards to discuss and assess these results.

The interview method, as mentioned earlier, was selected to be semi-structured. This method allows the interviewer to catch on the topics emerging in the answers of the interviewees (Barratt *et al.*, 2011). Asking questions about topics coming to prominence within the interview was vital in a study like this, where there is no clear quantitative answer to a question what good enough result is. Also, the expertise of the interviewees was diverse even though they were all very much familiar with the topic of contract management, so adapting to the emergent answers was crucial in order to get as comprehensive understanding of the thoughts of the interviewees as possible.

All interviews were held in Finnish due to that being a common language besides English for all interviewees and the interviewer. Thus, an extra care was needed to keep the original contents and emphases as close to original as possible when translating the transcripts to English. The interviews were conducted via Microsoft Teams software and recorded. The recordings were then transcribed into text format. Eventually, analysis of the expert opinions was made both by finding common themes arising from the interviews but also by noting the distinctions in responses.

5 Quantitative results

The results from the quantitative part of the study are presented in this section. The extracted items are considered one after another. However, the liabilities and contracting period are again discussed in a combined subsection due to their similarity in terms of both extraction and results. The results presented in decimal number format have been rounded to four decimal places.

5.1 Contracting parties

The first item extracted was the contracting parties to which this subchapter focuses on. As mentioned earlier, two NER algorithms were utilized in the extraction. Furthermore, two methods of classification were employed in choosing the correct contracting parties from the list of candidates. The results from all four combinations of these methods are concluded in Table 4. The best result in all measures is bolded. The confusion matrices for the four methods of contracting parties are shown in Figure 13.

Measure	**spaCy_order**	**spaCy_frequency**	**flair_order**	**flair_frequency**
Precision	0,3333	0,3333	0,7059	**0,7451**
Recall	0,7500	0,7500	0,9000	**0,9500**
F1-score	0,4615	0,4615	0,7912	**0,8352**
False positive rate	0,3846	0,3846	0,1667	**0,1444**

Table 4. Results for contracting parties.

An example of lists of candidates extracted by *spaCy* and *flair* is presented in Table 5. In this case the correct contracting parties were "VIGGLE INC" and "SFX ENTERTAINMENT INC". Both algorithms found the correct ones to their lists. However, only the combination of *flair* and classification based on frequency of occurrences was able to handle this contract completely correctly. Both methods utilizing *spaCy* and *flair* combined with order-based classification classified "VIGGLE INC" correctly as positive, but instead of "SFX ENTERTAINMENT INC" chose "F K A FUNCTION X INC" as positive due to its earlier occurrence in the contract.

spaCy	**flair**
VIGGLE INC	VIGGLE INC
F K A FUNCTION X INC	F K A FUNCTION X INC
SFX ENTERTAINMENT INC	SFX ENTERTAINMENT INC
LLC	BLUE SPIKE LLC
VIGGLE INC	SFX ENTERTAINMENT INC
	VIGGLE INC

Table 5. Example lists of contracting party candidates.

The results for both classification methods combined with *spaCy's* NER algorithm were completely similar. They were able to correctly find and classify 15 (TP) out of 52 contracting parties. Additionally, they found five more correct contracting parties but classified them incorrectly as negatives (FN). Combining the true positives and false negatives, *spaCy* was able to find 38,46 % of the correct contracting parties. Furthermore, the algorithm listed 78 false candidates for contracting parties, out of which 30 were incorrectly classified as positives (FP) and 48 correctly as negatives (TN). Calculated from these numbers displayed in the confusion matrices in Figure 13, the precision was 0,3333. It means that one third of the examples classified as positive were actual positives. Recall for models utilizing *spaCy* was 0,75 indicating that three quarters of actual positives found were classified as positives. The F1-score calculated from the two previous numbers was 0,4615. Computing the false positive rate for both models resulted in 0,3846, meaning that a bit less than two out of five actual negatives were classified as positive.

Focusing on the combination of *flair's* NER algorithm and classification based on order of occurrence, 40 out of 52 correct contracting parties were found. That is 76,92 % of all the correct contracting parties. 36 out of the 40 were classified as positives (TP) and four as negatives (FN). On top of that the model collected 90 actual negatives out of which 15 were classified as positives (FP) and 75 as negatives (TN). The mentioned numbers are presented in Figure 13. Thus, the precision of the model was 0,7059 while the recall was 0,9. Calculating the F1-score based on these numbers resulted in the value of 0,7912. One out of six actual negatives was classified as positive, eventuating in the false positive rate of 0,1667.

	Actual positive	Actual negative
Predicted positive	15 (TP)	30 (FP)
Predicted negative	5 (FN)	48 (TN)

a. spaCy_order

	Actual positive	Actual negative
Predicted positive	15 (TP)	30 (FP)
Predicted negative	5 (FN)	48 (TN)

b. spaCy_frequency

	Actual positive	Actual negative
Predicted positive	36 (TP)	15 (FP)
Predicted negative	4 (FN)	75 (TN)

c. flair_order

	Actual positive	Actual negative
Predicted positive	38 (TP)	13 (FP)
Predicted negative	2 (FN)	77 (TN)

d. flair_frequency

Figure 13. Confusion matrices for contracting parties.

Having used the same NER algorithm, the combination of *flair* and frequency-based classification method found the exact same actual positive and actual negative examples. The differences in the results were thus in the classification. As shown in Figure 13, 38 out of the 40 found actual positives were classified as positives (TP), while the remaining two were classified as negatives (FN). Furthermore, 13 out of the 90 actual negatives were classified as positives (FP) and 77 as negatives (TN). The precision for this model was 0,7451 and the recall was 0,95, which further led to the F1-score of 0,8352. The false positive rate was 0,1444.

5.2 Effective date

This subsection presents the results for the second extracted item, namely the effective date. As mentioned earlier, hand-crafted heuristics were created for finding the contexts of effective dates within the contract documents. Subsequently, NER algorithms from both *spaCy* and *flair* were utilized in the extraction. Finally, the classification of found examples was made only based on order of occurrence. This resulted in two methods as combinations of algorithm used in the extraction and order-based classification. The extraction results for both of these methods are summarized in Table 6. The confusion matrices are presented in Figure 14. One example of found context with the keyword "effective" is shown below. From that context, both NER algorithms were able to extract the date "October 12, 2000".

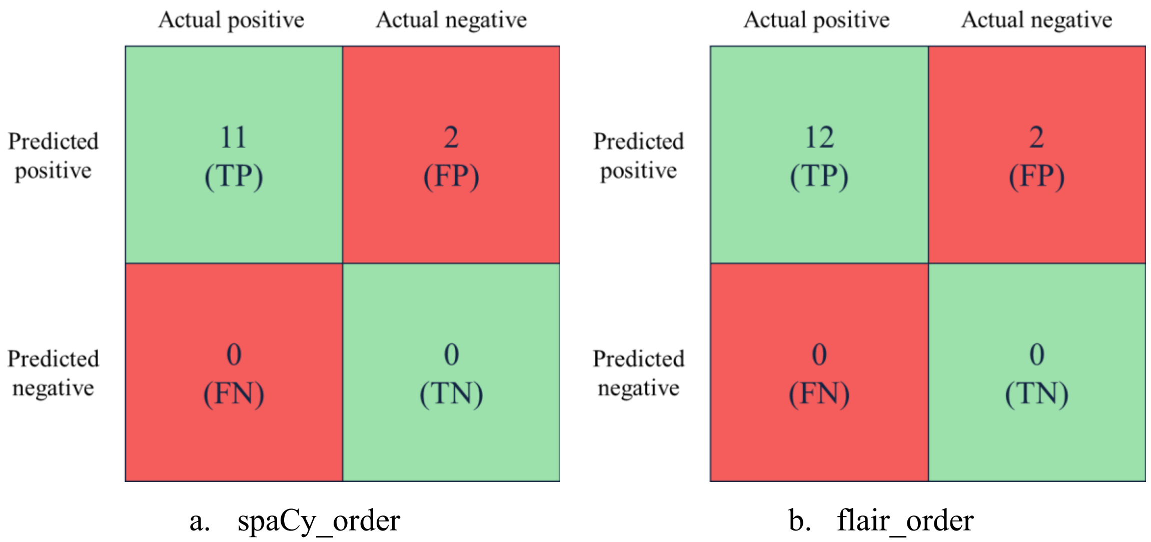
> *and conditions of this Agreement. WHEREAS, the Agreement shall become effective as of October 12, 2000, the Closing Date as defined in the Purchase…*

Measure	**spaCy_order**	**flair_order**
Precision	0,8462	**0,8571**
Recall	**1,0000**	**1,0000**
F1 score	0,9167	**0,9231**
False positive rate	**1,0000**	**1,0000**

Table 6. Results for effective date.

Combination of heuristics and *spaCy's* NER algorithm was able to find 11 correct effective dates from the 26 contracts, which is 42,31 % of all the correct ones. All of these 11 were classified as positive (TP). On top of that the model found two actual negatives which both were classified as positives (FP). This means that there were neither false negatives (FN) nor true negatives (TN). These numbers collected in the confusion matrix in Figure 14 led to precision being 0,8462 and recall being 1. Calculating from these, the F1-score of the model was 0,9167. As there were no true negative classifications the false positive rate was also 1.

Considering the *flair's* NER algorithm, it found 12 out of the 26 correct effective dates. That is 46,15 % of the corrects. The classification results were rather similar to *spaCy's*, as shown in Figure 14. All 12 found correct effective dates were classified to positives (TP), and furthermore, two actual negatives were classified found and classified as positives (FP). Again, no false negatives (FN) or true negatives (TN) were extracted from within the contexts. The precision calculated from these numbers was 0,8571 and the recall was 1. This led to the F1-score being 0,9231. The false positive rate was again 1 due to missing true negatives.



	Actual positive	Actual negative
Predicted positive	11 (TP)	2 (FP)
Predicted negative	0 (FN)	0 (TN)

a. spaCy_order

	Actual positive	Actual negative
Predicted positive	12 (TP)	2 (FP)
Predicted negative	0 (FN)	0 (TN)

b. flair_order

Figure 14. Confusion matrices for effective date.

5.3 Payment terms

The third item extracted from the contracts was payment terms, which is in the focus of this subsection. As with the effective date, also here the finding of contexts based on keywords was utilized. However, the solution for payment terms was different from the previous ones in the method of extraction. No ready-made algorithms were used, but the extraction was completely based on hand-crafted rules. The classification was made based on order of occurrence. Thus, there was only one model, results of which are concluded in Table 7. The confusion matrix for the model is presented in Figure 15. An example of a context found with the keyword “after recei” was

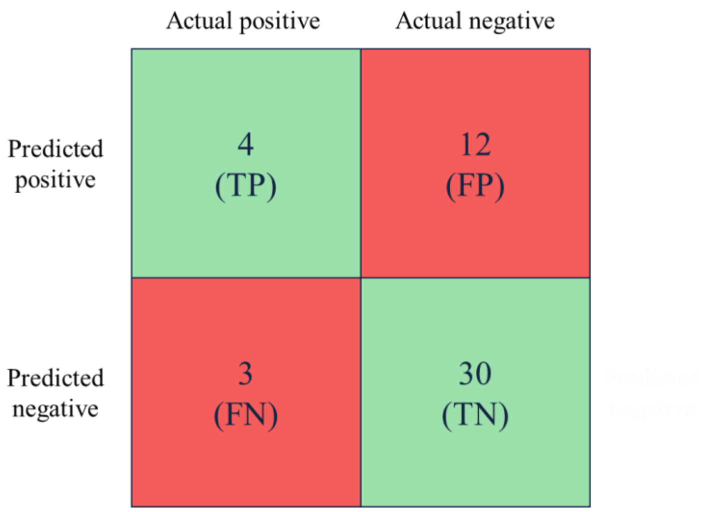
> *The Owner shall pay to the General Contractor, within ten (10) days after receipt of the Application for Payment, the amount determined pursuant to…*

From that context the hand-crafted regular expression rules were able to extract the candidate “ten (10) days” by defining that there should be a combination of any word, any number, and either word “days” or “month”. It was also taken into account that there can be some words between each of the three key factors, which can be seen from another example extracted from the same contract: “fifteen (15) working days”. After removing the punctuation and transforming the text into uppercase the candidate was “TEN 10 DAYS”.

Measure	Rules
Precision	0,2500
Recall	0,5714
F1 score	0,3478
False positive rate	0,2857

Table 7. Results for payment terms.

The model for extracting payment terms was able to find 7 out of 26 correct payment terms, which sums up to 26,92 %. Out of these 7 correct ones, 4 were classified as positives (TP) and 3 as negatives (FN). Additionally, 42 actual negative examples were found. 12 out of those were classified incorrectly as positives (FP) while the remaining 30 were classified correctly as negatives (TN). Computed from these numbers, the precision of the payment terms extraction model was 0,25 and recall was 0,5714. The F1-score derived from those was 0,3478. The false positive rate was 0,2857.



	Actual positive	Actual negative
Predicted positive	4 (TP)	12 (FP)
Predicted negative	3 (FN)	30 (TN)

Figure 15. Confusion matrix for payment terms.

5.4 Liabilities and contracting period

In this subsection, the results for the final items of extraction are considered. Liabilities and contracting period had similar keyword-utilizing context finding method as payment terms. Additionally, keyword matching was used in extracting the examples from the contexts. For liabilities, candidates were found from every contract. The number of examples found varied by contract between one and 53. Examples extracted from contracts include such as

> *MENTAT shall have no liability to indemnify…*

> *CISCO SHALL HAVE NO LIABILITY HEREUNDER FOR…*

> *Theriac of Littlestown, LLC, a limited liability company ("Owner")…*

When it comes to contracting period, there was only one context found with the predetermined keywords from all the contracts. Furthermore, this context was irrelevant due to it was only defining that

> *All monthly, quarterly, and annual periods refer to calendar months, quarters, and years (not contract period related).*

Thus, it can be concluded for contracting parties that there were no relevant results found. This chapter has presented the results from the quantitative phase of the study. Next chapter will focus on the analysis of these results as well as on connecting them to the outcomes of the conducted interviews.

6 Discussion

Now that the quantitative results have been presented, findings from them are discussed and connected to the qualitative outcomes from the interviews. Firstly, this section focuses on the analysis of the results and the comments from the interviewees regarding them. Secondly, the analysis of the future potential directions concerning the solution built in this study are considered.

6.1 Analysis of results

As mentioned, this subsection sheds light on the findings from the quantitative analysis and links it to the discussions with the interviewees. Additionally, other discoveries from the interviews are considered. The items extracted from the contracts are discussed in their own subsections as was done when presenting the results.

6.1.1 Contracting parties

The extraction of contracting parties resulted in the best outcome of all items. Using the combination of *flair's* NER algorithm and the classification based on frequency of occurrences outperformed the other method combinations in all measures. Also using the order-based classification for *flair* resulted in better outcome than either of the *spaCy's* combinations. With *spaCy's* NER algorithm the results were completely similar to each other independent of the method of classification. The most successful method was able to correctly classify 38 positive examples out of the 40 it found and from the total of 52 actual contracting parties existing in the contracts. It did not achieve a result close to perfection with the F1-score of 0,8352. Chalkidis *et al.* (2017) achieved 0,9 with a combination of sliding window support vector machine, hand-crafted rules, and post-processing rules. Also, Chalkidis and Androutsopoulos (2017) reached an F1-score as high as 0,94 with a more sophisticated combination of bidirectional long short-term memory layer, another long short-term memory layer, and a logistic regression layer in their recurrent neural network model. The models used in this study were out-of-the-box solutions with no manual configurations. Another difference to the mentioned literature is that both mentioned studies utilized restriction of extraction context for contracting parties too. The confidence level of the positive classifications should be higher than the one achieved in this study in order to automatically populate the extracted data points into the metadata fields, as one of the interviewees said:

> *"The results are tricky, as the contracting parties are the information that is also easy for the user to fill in. Then again, searching for payment terms or the others takes more time from a human. At least in this study the results were not so good that they would directly provide value for the user. – – If [the information] is found confidently enough it is good, but if the confidence level is bad this would make more harm for everyone. That is often the problem in things like this. If the confidence level were good it would create value for the customer, as the users would not need to search for the information in the contracts themselves." Key account manager*

The other interviewees agreed that as such the solution for contracting parties is not production ready. However, it was discussed in two of the interviews that the level of confidence needed for the tool depends on the intended use of the tool, whether it exists to fully automate the population of metadata fields in the software or work as an assistant for the users. This would work also for other fields than just contracting parties.

> *"For sure the more the tool was developed the better the confidence would get. Then again, it depends a bit on how this would be utilized. If it were to be used so that it actually automatically recognizes the correct values the confidence level would need to be somewhat high in which case this would be beneficial for only contracting parties. However, this could be also used as an assistant. Instead of trusting that you always get the correct answer it could propose you something which in some cases would be beneficial." Product manager*

> *"It should be found out whether this would work so that there would be three or four alternatives which software would suggest, from which you could choose the correct one or fill it in manually if it did not suggest the correct one. It would ease the work a bit, even though it would be nicest if [the software] could fill [the fields] automatically. But then again, that is much more challenging." Software engineer*

With contracting parties, the extraction speed was an issue. Running the 26 test data contract files through the pipeline took approximately two hours with *spaCy* and three hours with *flair*. There were three explaining main factors for the slowness. Firstly, the contexts for the contracting party extraction were not limited but included the whole documents. For the other metadata items this was not the case finding contexts based on keywords was included in their pipelines. Secondly, it was not possibly to run the NER algorithms for entities of only particular type. The need in this case would have been to restrict the entity tagging to organization names. However, neither of the algorithms allow such configuration. It was only possible to filter the results based on the entity type of organization names after searching for all pre-trained entities within the contract. Thirdly, the software was run locally on a laptop with limited computational power. Using cloud-based computations would overcome this issue.

The speed issue, which luckily occurred only with contracting parties, could be mitigated relatively easily by limiting the context of extraction to the first pages of the contracts. As noted also by Chalkidis *et al.* (2017) that is most often the location for the contracting parties. However, that would cause losing the benefit of classifying based on the frequency of occurrences as it is not certain that the names of contracting parties occur more than once within the first pages. The solution needs to balance between speed and confidence. All of the interviewees agreed that in order to create value for the user the application needs to be faster in inputting the information than the users themselves. Nevertheless, speed is not the only objective the automation is there to fulfill. It can create comfort of use and give a good impression which is beneficial from the sales point of view.

> *"If searching for contracting parties is slow [the tool] does not help, but if it were fast it would be beneficial. If you think about it, filling in for example payment terms is fast as it is just a number, while typing in a party can be slower. Even if [the automation] did not bring actual benefits at least it would give the impression that this system is smart." Key account manager*

6.1.2 Effective date

The results for extracting effective date were contradictory. Methods utilizing *spaCy* and *flair* found 11 and 12 correct effective dates, respectively. So, less than half of the 26 of the actual correct ones were found. However, neither of the methods found more than two incorrect effective dates which makes the precision relatively good. As the models did not classify anything as negatives the recall, F1-score, and false positive rate are not relevant in the results. Hence, the method using *flair* can be stated to have performed better based on its precision of 0,8571 and the 46,15 % of the correct effective dates found. Chalkidis *et al.* (2019) have developed the best model found in the literature for extracting effective date, which included a convolutional neural network and reached an F1-score of 0,959, precision of 0,969, and recall of 0,951. However, those results are not completely comparable to the ones achieved in this study, as their data set consisted of contract headers that included the effective date instead of complete contracts.

Even though a lot of correct effective dates were not found at all, in absolute terms both models performed well. They did not find many negative examples from the contracts and classify them as positives. Having high number of false positives would be the worst-case scenario from the user's perspective as it would lead to incorrect metadata in the contract management software. Hence, having a good level of confidence is vital for the solution to be beneficial.

> *"But if you think about it from a practical point of view: approximately half of the dates were found. The worst thing would be that [the tool] would find an incorrect date and classifies it as positive. So, some kind of confidence level could be included. If the software is sure that the result is correct it would classify it as positive and if it is not sure it would classify nothing as positive, but the user would be notified to search for the correct one manually. If you think about it that way, it does not look that bad."*
> *Key account manager*

6.1.3 Payment terms

This subsection focuses on the results for extracting payment terms. NER algorithms could not be utilized in this item as such, so the extraction was done based on hand-crafted rules. Due to the small size of the sample data based on which the rules were created they were not comprehensive enough to find most of the correct payment terms. Only seven of the 26 payment terms were found and out of these seven only four were classified as positives. The remaining three were classified as negatives due to the same contract having another example that satisfied the rules and appeared before the correct payment term. In total 12 negative examples were classified as positive. The achieved F1-score was as low as 0,3478. Even though the extraction of payment terms cannot be found in earlier literature, for other contract elements the F1-scores have mostly been over 0,7 (Chalkidis and Androutsopoulos, 2017; Chalkidis *et al.*, 2017; Chalkidis *et al.*, 2019).

The precision of 0,25 was too low for the tool to be used in fully automating the extraction and filling the fields in the template. However, on top of including a bigger sample data to build more comprehensive rules, two options of developing the tool further were discussed in the interviews. Firstly, as discussed already earlier the tool could work as a basis for a template filling assistant. The software could give a list of possible payment terms out of which the user would select the correct one. Secondly, the results from the software could be validated against a list of possible values. Companies often use a limited number of payment terms in their contracts. The second approach could be used in two different ways. Either by comparing the found values to the ones listed or by in the first place searching for only the terms accepted by the company. The values could then be either automatically filled or suggested to the user.

> *"It would be interesting to investigate an approach where the results would be validated against a list. – – It is different to search for a value from scratch than if there are let's say three or four values and then [the software] would search which one would get a hit. It could then be processed further so that the more contracts have been filled even without knowing anything about the terminology, if it finds out that this value has often been selected with this terminology the information would be used in developing the algorithm." Product manager*

The automatic extraction of payment terms was considered important in the value creation for the users. As with all the items, for the users of the metadata it does not make any difference whether the data is inputted automatically or manually, but for the ones inputting the data automation would make sense, as long as the results are accurate.

> *"From our perspective it doesn't make any difference whether the information is inputted automatically or by the user. The data can be similarly processed further anyway. From the customer's point of view, it is truly valuable if the information is [automatically] found. It would allow that the person who uploads the files into the system doesn't need to know how to find for example payment terms. If you have searched for them many times you improve but it's not trivial the first time you do it. If there was an algorithm to find the payment terms it would definitely be very*

> *advantageous. – – For the user that uploads one contract it is important that the one specific contract goes correctly.” Key account manager*

6.1.4 Liabilities and contracting period

The last two extracted contract items were liabilities and contracting period. As in previous sections, they are discussed in the same subsections due to their many similarities. It was mentioned already earlier that with these two contract items it was difficult to define the gold standards. The issue was also discussed with the interviewees. Therefore, no such scoring was done as for other items.

> *“[Contracting period] is not necessarily unambiguously mentioned but it can be conditional. I mean, it probably is possible to find these too, but it would require a huge machine learning exercise concentrating only on this topic and a lot more contract documents.” Software engineer*

Even though there was only one context found with the pre-determined keywords, according to the interviewees, the contracting periods could be more clearly defined for the software to find as there are some common ways of expressing them in the contracts. However, the wordings can vary a lot as the period can be expressed for instance as a date of expiration, number of valid years or months, or the contract can be valid until either party terminates it. There were even contracts in which the contracting period was not mentioned at all due to the contract covering a one-off project to be implemented. Chalkidis *et al.* (2017) achieved much better results for extracting contracting periods with their combination of logistic regression and hand-crafted rules which implies that in the future their methods could be replicated more closely to reach a positive result in also this field.

With liabilities there were similar issues as well. The definition of what is to be extracted should be clearer in order to mediate the information to the software. Still, there were many potential contexts found where the liabilities were discussed within the contracts. It holds out hope for developing a more accurate tool in the future even for liabilities extraction. Nevertheless, before that a well-functioning solution for the easier items should be developed.

> *"It should be brainstormed which items are easier to find based on these experiences and not jump directly to the liabilities and other more difficult ones. This should be started from those low hanging fruit." Software engineer*

This subsection has covered the analysis of the results achieved with the information extraction tool. The following subsection focuses on the potential future development of the solution.

6.2 Future potential

In this subsection the future potential of the contract information extraction tool developed during this study is discussed. The interviews work as the primary source for this analysis. All interviewees agreed that there is a need for such solution and that there clearly is potential in the tool built in this study if developed further. The need arises from two things. Firstly, the automation can increase the speed of adding new documents to the contract management software and thus free the user's time to other tasks. Secondly, it can decrease the number of human errors in the contract metadata.

> *"The automatic recognition speeds up the processes of digitizing old contracts or third-party contracts a lot. A large number of contracts can be digitized a lot faster when these certain data points are recognized automatically. This could also be utilized when the contracts are already in the system with the inputted metadata, so that this would be used in validating the data. It speeds up the process and decreases the number of errors. That's where it comes very useful. –– The more metadata fields there are the more laborious it becomes for the used to fill them. When a new contract is created but the fields need to be manually filled it can lead to not filling all fields or making human errors." Product manager*

The overall approach of this study to the problem was considered reasonable. Positive results were achieved with out-of-the-box functionalities of NER algorithms as well as rules created from a small amount of sample contracts. Nevertheless, numerous ideas for the future development were found in the interviews. Some of them were discussed in the analysis of the results. For instance, the issue with the extraction speed of contracting parties should be tackled before the tool can be taken into use. Also, the utilization of this tool as an assistant instead of fully automated template filling solution was mentioned as well as the possibility to validate the results of the tool against a pre-defined list of accepted values. Furthermore, using a threshold in the confidence level of automatically extracted items was a topic arising in the interviews.

One potential way of enhancing the results of all the extracted items which was discussed with the interviewees would be benefitting from the fact that companies often use a template for most of their contracts. It would ease the finding of items if the users flagged the locations of the items from a commonly used contract template. Utilizing company-wide templates could work relatively well without NER component in the solution. Only the format of the result would somehow be needed to be ensured. However, this would not solve the problem of contracts made with other templates. Furthermore, the template should then be defined per customer, which could lead to non-standard configurations.

> *"If a large part of the contracts of a company are created on a same template it would reduce the work from the algorithm when it could be told that this is our template and that is the location where the contracting parties are usually found in. Or that this is the place where the contracting period is usually filled in to. –– So, it is much easier to search for all the things if you know the approximate place where they should be found from. But surely the topic of this study was a bit different." Product manager*

Even if the results were divided into good success of contracting parties and less successful other contract elements, the objectives of the study were met in exploring the solution and reviewing the potential of this kind of a tool. These findings preliminarily suggest that there is no shortcut to a working solution. More sophisticated methods would need to be investigated, which is consistent with the findings of Chalkidis *et al.* (2017). The following quote from an interview concludes the results well.

> *"In my opinion it can be clearly seen that there is [potential]. It can be seen from the results of contracting parties that if you can come up with the right questions you can get started with relatively ready-made solutions." Software engineer*

7 Conclusions

This study has presented a tool for automated information extraction tool for procurement contract documents and discussed its performance. The study was conducted in collaboration with the case company Sievo Oy in the context of their procurement analytics software. This section concentrates on concluding the research. The section is divided into two subsections discussing firstly the most important findings and secondly the limitations of the study as well as the future research possibilities.

7.1 Main findings

The key findings of the study are concluded in this subsection. The research question regarded a way of automatically extracting information from procurement contracts. The individual extracted items were considered separately. Based on the experiments of the study, the best solution found for extracting contracting parties was utilizing named entity recognition algorithm from *flair* library (flair, 2018). However, as discussed, the algorithm was slow in processing multi-paged documents, which implies that the context for searching the item should be restricted to the first pages of the contracts. Contracting parties were the most mature item in terms of the rate of finding correct results from contracts as a large part of correct parties were extracted with ready-made algorithms. For other items less than half of the correct ones were found with even the best methods and thus the results were not very encouraging.

In extracting the effective dates from contracts, a combination of finding a context based on hand-crafted rules combined with *flair* was the best performing method. For payment terms the extraction method used was completely based on hand-crafted regular expression rules. The investigation of extracting liabilities and contracting period was tentative and exploratory. Both of them had a lot of variations in the wordings which made the extraction unsuccessful due to the rules based being on small sample data. However, with more sophisticated models Chalkidis *et al.* (2017) have shown that it is possible to extract even the items that had worst performance in this study. When it comes to the sub-questions concerning the methods of extraction, it is thus concluded that using both named-entity recognition algorithms and hand-crafted rules are potential solutions to include in the extraction process depending on the extracted item.

Notwithstanding the discordant results, the people interviewed for the study saw future potential of the tool. A central finding in this research emerged in these interviews. It suggested that from the perspective of the business case, the complete automation of extraction and template filling is not the only solution. Using the tool as an assistant to the contract management software users by suggesting prospective correct items would both ease and quicken their work as well as create an impression of the software being smart. This would be an alternative solution to the one experimented in this study. The need for a solution in assisting contract management software users is clearly supported by the findings from the interviews. The approach used in this study provides useful information about the possibilities of using NER algorithms and hand-crafted rules in contract information extraction. The insights gained will help the case company in determining the development path of automated contract metadata finding and extraction. The next challenge is to advance towards a production-ready solution. This study also contributed to the increasing amount of research around automated information extraction by filling the gap of considering complete procurement contracts. The possibilities of future research are discussed in the final subsection of the thesis together with the limitations of this study.

7.2 Limitations and future research

Due to the nature and methodology of the research there are limitations that need to be addressed. Firstly, because of the data being partly confidential the numbers are not reproducible as such. Also related to the reproducibility and confidentiality, the tool developed in this study will not be publicly available. Secondly, as discussed already before, both the sample data and the test data batches used in the creation of hand-crafted rules were considerably small. This was due to the laboriousness of manually annotating the gold standards in the datasets. Including a larger dataset would lead to more variation in the wordings and thus allow creation of more comprehensive rules. Thirdly, the study was limited to contract documents which were already in machine-readable format and written in English. Hence, scanned contracts which are still the most common form of contracts were not considered. Utilizing them would require a reliable optical character recognition tool to be used first. The emergence of digital signatures brings both relief and inconveniences for the tool development. Digitally signed tools are readily in a machine-readable format, but they can be secured with an encryption which will then need to be bypassed in order read the document with the software. Also, having contracts in various languages would require separate pre-training of models for each language. There are generic models available for some languages, but for more exotic ones the model training would require a tremendous effort.

This study leads to further research possibilities. From the point of view of the case company, it would be valuable to investigate whether the named entity recognition algorithms could be trained to find also other extracted items. Training the algorithms in question to find for instance payment terms has not been previously studied. Furthermore, a prospective topic is to research incremental learning algorithms within this context. Developing a software based on a limited dataset which would suggest the user a handful of potential correct answers could be combined to incremental learning algorithms based on user selections. The performance of these algorithms in the context of contracts is still nonexistent. A third topic of possible future research would consider the user needs and reception towards the idea of a template filling assistant. All of the presented ideas would fill research gaps in the contract information context as well as have practical value for the case company.

References

Akbik, A., Blythe, D. and Vollgraf, R. (2018) "Contextual string embeddings for sequence labeling", *Proceedings of the 27th International Conference on Computational Linguistics,* pp. 1638-1649.

Al Omran, F. and Treude, C. (2017) "Choosing an NLP library for analyzing software documentation: a systematic literature review and a series of experiments", *IEEE/ACM 14th International Conference on Mining Software Repositories*, pp. 187-197.

Asatiani, A. and Penttinen, E. (2016) "Turning robotic process automation into commercial success–Case OpusCapita", *Journal of Information Technology Teaching Cases*, 6(2), pp. 67-74.

Baker, G., Gibbons, R. and Murphy, K.J. (2002) "Relational Contracts and the Theory of the Firm", *The Quarterly Journal of Economics*, 117(1), pp. 39-84.

Bals, L., Schulze, H., Kelly, S. and Stek, K. (2019) "Purchasing and supply management (PSM) competencies: Current and future requirements", *Journal of purchasing and supply management*, 25(5), pp. 1-15.

Barratt, M., Choi, T.Y. and Li, M. (2011) "Qualitative case studies in operations management: Trends, research outcomes, and future research implications", *Journal of Operations Management*, 29(4), pp. 329-342.

Benlian, A., Hess, T. and Buxmann, P. (2009) "Drivers of SaaS-adoption–an empirical study of different application types", *Business & Information Systems Engineering*, 1(5), pp. 357.

Bryman, A. (2006) "Integrating quantitative and qualitative research: how is it done?", *Qualitative research*, 6(1), pp. 97-113.

Buitinck, L. and Marx, M. (2012) "Two-stage named-entity recognition using averaged perceptrons", *International Conference on Application of Natural Language to Information Systems, pp. 171-176.*

Califf, M.E. and Mooney, R.J. (1999) "Relational learning of pattern-match rules for information extraction", *Proceedings of the National Conference on Artificial Intelligence*, pp. 328-334.

Cardie, C. (1997) "Empirical methods in information extraction", *AI Magazine*, 18(4), pp. 65-79.

Cascio, W.F. and Montealegre, R. (2016) "How technology is changing work and organizations", *Annual Review of Organizational Psychology and Organizational Behavior*, 3, pp. 349-375.

Chalkidis, I. and Androutsopoulos, I. (2017) "A deep learning approach to contract element extraction", *Frontiers in Artificial Intelligence and Applications*, 302, pp. 155-164.

Chalkidis, I., Androutsopoulos, I. and Michos, A. (2017) "Extracting contract elements", *Proceedings of the International Conference on Artificial Intelligence and Law*, pp. 19-28.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P. and Androutsopoulos, I. (2019) "Neural Contract Element Extraction Revisited", *Proceedings of the Document Intelligence Workshop of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, pp. 1-4.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y. (2014) "Learning phrase representations using RNN encoder-decoder for statistical machine translation", *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 1724-1734.

Chomsky, N. (1965) "Three models for the description of language", *IRE Transactions on Information Theory*, pp. 113-124.

Church, K.W. (1989) "Stochastic parts program and noun phrase parser for unrestricted text", *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2, pp. 695-698.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. and Kuksa, P. (2011) "Natural language processing (almost) from scratch", *Journal of Machine Learning Research*, 12, pp. 2493-2537.

Costas, J. and Kärreman, D. (2016) "The bored self in knowledge work", *Human Relations*, 69(1), pp. 61-83.

Davis, J. and Goadrich, M. (2006) "The relationship between Precision-Recall and ROC curves", *Proceedings of the 23rd international conference on Machine learning*, pp. 233-240.

Ekbal, A. and Saha, S. (2016) "Simultaneous feature and parameter selection using multiobjective optimization: application to named entity recognition", *International Journal of Machine Learning and Cybernetics*, 7(4), pp. 597-611.

Etikan, I., Musa, S.A. and Alkassim, R.S. (2016) "Comparison of convenience sampling and purposive sampling", *American journal of theoretical and applied statistics*, 5(1), pp. 1-4.

Finkel, J.R., Grenager, T. and Manning, C. (2005) "Incorporating non-local information into information extraction systems by Gibbs sampling", *ACL-05 - 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 363-370.

flair. (2018). "flairNLP/flair," Retrieved from https://github.com/flairNLP/flair on Feb 25, 2020.

Forman, G. (2003) "An extensive empirical study of feature selection metrics for text classification", *Journal of machine learning research*, 3(Mar), pp. 1289-1305.

Freitag, D. (2000) "Machine learning for information extraction in informal domains", *Machine Learning*, 39(2), pp. 169-202.

Giannoccaro, I. and Pontrandolfo, P. (2004) "Supply chain coordination by revenue sharing contracts", *International Journal of Production Economics*, 89(2), pp. 131-139.

Gildea, D. and Jurafsky, D. (2002) "Automatic labeling of semantic roles", *Computational linguistics*, 28(3), pp. 245-288.

Godse, M. and Mulik, S. (2009) "An approach for selecting Software-as-a-Service (SaaS) product", *CLOUD 2009 - 2009 IEEE International Conference on Cloud Computing*, pp. 155-158.

Grant, A.M. (2008) "The significance of task significance: Job performance effects, relational mechanisms, and boundary conditions.", *Journal of applied psychology*, 93(1), pp. 108.

Greene, J.C., Caracelli, V.J. and Graham, W.F. (1989) "Toward a conceptual framework for mixed-method evaluation designs", *Educational evaluation and policy analysis*, 11(3), pp. 255-274.

Grishman, R. (2015) "Information Extraction", *IEEE Intelligent Systems*, 30(5), pp. 8-15.

Grishman, R. and Sundheim, B.M. (1996) "Message understanding conference-6: A brief history", *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 466-471.

Hart, O.D. and Holmström, B. (1986) "The theory of contracts", *Working papers*, 418, Massachusetts Institute of Technology, Cambridge, MA.

Hoy, M.B. (2018) "Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants", *Medical Reference Services Quarterly*, 37(1), pp. 81-88.

Jackson, P., Al-Kofahi, K., Tyrrell, A. and Vachher, A. (2003) "Information extraction from case law and retrieval of prior cases", *Artificial Intelligence*, 150(1-2), pp. 239-290.

Jagannathan, V., Mullett, C.J., Arbogast, J.G., Halbritter, K.A., Yellapragada, D., Regulapati, S. and Bandaru, P. (2009) "Assessment of commercial NLP engines for medication information extraction from dictated clinical notes", *International journal of medical informatics*, 78(4), pp. 284-291.

Kazama, J. and Torisawa, K. (2007) "Exploiting Wikipedia as external knowledge for named entity recognition", *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pp. 698-707.

Korhonen, A. and Briscoe, T. (2004) "Extended lexical-semantic classification of English verbs", *Proceedings of the Computational Lexical Semantics Workshop at HLT-NAACL 2004*, pp. 38-45.

Kraljic, P. (1983) "Purchasing must become supply management", *Harvard business review*, 61(5), pp. 109-117.

Krishnan, V. and Manning, C.D. (2006) "An effective two-stage model for exploiting non-local dependencies in named entity recognition", *COLING/ACL 2006 - 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 1, pp. 1121-1128.

Kwok, T., Nguyen, T. and Lam, L. (2008) "A software as a service with multi-tenancy support for an electronic contract management application", *Proceedings - 2008 IEEE International Conference on Services Computing, SCC 2008*, 2, pp. 179-186.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. and Dyer, C. (2016) "Neural architectures for named entity recognition", *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, pp. 260-270.

Lassila, A. (2006) "Taking a service-oriented perspective on software business: How to move from product business to online service business", *IADIS International Journal on WWW/Internet*, 4(1), pp. 70-82.

Lavelli, A., Califf, M.E., Ciravegna, F., Freitag, D., Giuliano, C., Kushmerick, N., Romano, L. and Ireson, N. (2008) "Evaluation of machine learning-based information extraction algorithms: Criticisms and recommendations", *Language Resources and Evaluation*, 42(4), pp. 361-393.

Lee, J., Yi, J.-. and Son, J. (2019) "Development of Automatic-Extraction Model of Poisonous Clauses in International Construction Contracts Using Rule-Based NLP", *Journal of Computing in Civil Engineering*, 33(3).

Leitner, E., Rehm, G. and Moreno-Schneider, J. (2019) "Fine-Grained Named Entity Recognition in Legal Documents", *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11702 LNCS, pp. 272-287.

Lewis, H.T. (1946) "This Business of Procurement", *Harvard business review*, 24(3), pp. 377-393.

LexPredict. (2017). "LexPredict/lexpredict-contraxsuite-samples," Retrieved from https://github.com/LexPredict/lexpredict-contraxsuite-samples/tree/master/agreements on Jan 23, 2020.

Li, J., Monroe, W., Ritter, A., Galley, M., Gao, J. and Jurafsky, D. (2016) "Deep reinforcement learning for dialogue generation", *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pp. 1192-1202.

Martela, F. and Riekki, T.J. (2018) "Autonomy, competence, relatedness, and beneficence: A multicultural comparison of the four pathways to meaningful work", *Frontiers in psychology*, 9, pp. 1157.

Miller, S., Guinness, J. and Zamanian, A. (2004) "Name tagging with word clusters and discriminative training", *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pp. 337-342.

Milosevic, Z., Berry, A., Bond, A. and Raymond, K. (1995) "Supporting business contracts in open distributed systems", *Proceedings of the International Workshop on Services in Distributed and Networked Environments*, pp. 60-67.

Moens, M. (2006) "Information extraction: Algorithms and prospects in a retrieval context", Springer, Dordrecht.

Monczka, R.M., Handfield, R.B., Giunipero, L.C. and Patterson, J.L. (2009) "Purchasing and supply chain management", Cengage Learning, Mason, OH.

Nadeau, D. and Sekine, S. (2007) "A survey of named entity recognition and classification", *Lingvisticae Investigationes*, 30(1), pp. 3-26.

Palmer, M., Gildea, D. and Kingsbury, P. (2005) "The proposition bank: An annotated corpus of semantic roles", *Computational linguistics*, 31(1), pp. 71-106.

Poppo, L. and Zenger, T. (2002) "Do formal contracts and relational governance function as substitutes or complements?", *Strategic Management Journal*, 23(8), pp. 707-725.

Powers, D.M.W. (2008) "Evaluation evaluation", *Frontiers in Artificial Intelligence and Applications*, 843-844.

Pradhan, S., Hacioglu, K., Ward, W., Martin, J.H. and Jurafsky, D. (2003) "Semantic role parsing: Adding semantic structure to unstructured text", *Third IEEE International Conference on Data Mining*, pp. 629-632.

Ratinov, L. and Roth, D. (2009) "Design challenges and misconceptions in named entity recognition", *CoNLL 2009 - Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pp. 147-155.

Riloff, E. and Lehnert, W. (1994) "Information Extraction as a Basis for High-Precision Text Classification", *ACM Transactions on Information Systems (TOIS)*, 12(3), pp. 296-333.

Ring, P.S. and Van de Ven, A. H. (1992) "Structuring cooperative relationships between organizations", *Strategic Management Journal*, 13(7), pp. 483-498.

Rintala, N. and Suolanen, S. (2005) "The implications of digitalization for job descriptions, competencies and the quality of working life", *Nordicom Review*, 26(2), pp. 53-67.

Sääksjärvi, M., Lassila, A. and Nordström, H. (2005) "Evaluating the software as a service business model: From CPU time-sharing to online innovation sharing", *IADIS international conference e-society*, pp. 177-186.

Sang, E.F. and Buchholz, S. (2000) "Introduction to the CoNLL-2000 shared task: Chunking", *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*, pp. 127-132.

Smids, J., Nyholm, S. and Berkers, H. (2019) "Robots in the Workplace: a Threat to—or Opportunity for—Meaningful Work?", *Philosophy & Technology*, pp. 1-20.

Smith, S.M. (2002) "Fast robust automated brain extraction", *Human brain mapping*, 17(3), pp. 143-155.

Snow, R., O'connor, B., Jurafsky, D. and Ng, A.Y. (2008) "Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks", *Proceedings of the 2008 conference on empirical methods in natural language processing,* pp. 254-263.

spaCy. (2015). "spacy.io," Retrieved from https://spacy.io/ on Feb 25, 2020.

Sprague Jr., R.H. (1995) "Electronic document management: Challenges and opportunities for information systems managers", *MIS Quarterly: Management Information Systems*, 19(1), pp. 29-49.

Strzalkowski, T. (1995) "Natural language information retrieval", *Information Processing and Management*, 31(3), pp. 397-417.

Thomas, D.J. and Griffin, P.M. (1996) "Coordinated supply chain management", *European Journal of Operational Research*, 94(1), pp. 1-15.

Toutanova, K., Klein, D., Manning, C.D. and Singer, Y. (2003) "Feature-rich part-of-speech tagging with a cyclic dependency network", *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 173-180.

Turner, M., Budgen, D. and Brereton, P. (2003) "Turning software into a service", *Computer*, 36(10), pp. 38-44.

Voorhees, E.M. (1999) "Natural language processing and information retrieval", *Information Extraction: Towards Scalable, Adaptable Systems*, 1714, pp. 32-48.

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S. and Liu, H. (2018) "Clinical information extraction applications: A literature review", *Journal of Biomedical Informatics*, 77, pp. 34-49.

Weber, C.A., Current, J.R. and Benton, W.C. (1991) "Vendor selection criteria and methods", *European Journal of Operational Research*, 50(1), pp. 2-18.

Weischedel, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., Xue, N., Taylor, A., Kaufman, J., Franchini, M., El-Bachouti, M., Belvin, R. and Houston, A. (2013). "OntoNotes Release 5.0," Retrieved from https://catalog.ldc.upenn.edu/LDC2013T19 on Jun 9, 2020.

World Economic Forum. (2020). "The Future of Jobs Report 2020," Retrieved from https://www.weforum.org/reports/the-future-of-jobs-report-2020/in-full/infographics-e4e69e4de7 on Dec 17, 2020.

Wrzesniewski, A., McCauley, C., Rozin, P. and Schwartz, B. (1997) "Jobs, careers, and callings: People's relations to their work", *Journal of research in personality*, 31(1), pp. 21-33.

Zhang, J. and El-Gohary, N.M. (2016) "Semantic NLP-Based Information Extraction from Construction Regulatory Documents for Automated Compliance Checking", *Journal of Computing in Civil Engineering*, 30(2).

Zhang, T. and Johnson, D. (2003) "A robust risk minimization based named entity recognition system", *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pp. 204-207.

Zhou, G.D., Zhang, J., Su, J., Shen, D. and Tan, C.L. (2004) "Recognizing names in biomedical texts: A machine learning approach", *Bioinformatics*, 20(7), pp. 1178-1190.

Appendix 1: Keywords for contracting parties

&Co
(publ)
AB
Abp
AC
ACE
AD
ADSITz
AE
AG
AJ
akc.spol
Aktiengesellschaft
AL
AMBA
Anl
ANS
AO
Apb
ApS
AS
ASA
ASBL
AULC
AVEE
AVV
Ay
BA
BC
Berhad
Bhd
BK
BL
BM
BO
Bpk
bt
BV
BVBA
C.porA
CA
CC
Ccc
Cedel
Cia
Cía

ec
EE
EEG
EEIG
eG
ehf
EI
EIRL
eK
Ek.för
ELP
EO
EOOD
EP
EPE
EPP
ESV
ET
EU
EURL
eV
F:ma
Fa
Familiengesellschaft
FCP
FIE
FKF
FMBA
FOP
GbR
GCV
Gen
GesbR
GesmbH
gGmbH
GIE
GIU
gk
GmbH
gmk
GP
GS
gsk
Gte
GUP
HAAO

kv
Ky
l.p
Lda
LDC
Limited
LLC
LLLP
LLP
LP
LŞ
LTC
Ltd
Ltda
LTDA-EPP
Ltee
Ltée
MB
mbH
MChJ
ME
MEPE
Mts
MTÜ
nk
NL
NSULC
NT
NUF
NV
Nyrt
OAJ
obrt
OD
OE
OEG
OG
ohf
OHG
OK
ONG
oo
OOD
OOO
OPG
ops

RHF
RIC
RL
rp
rs
Rt
rV
ry
S.A.
s.a.r.l
S.Coop
S.Cra
S.de
S.enC
S.enC.
s.l
SA
SAA
SAB
SAC
SAD
SAdeCV
SAE
SAFI
Sagl
SAICA
SAL
SAOC
SAOG
sapa
SAPI
SAPIdeCV
SARF
SARL
SàRL
SAS
SASU
SAU
sc
SCA
SCC
SCCL
SCE
SCeI
SCI
SCOP

Societe
SOCIETEENCOMMANDITE
SOPARFI
SP
sp.j
sp.k
sp.p
Sp.zo.o.
SpA
SPE
spj
spk
spol
spp
SPRL
SPRLU
SrK
SRL
sro
ss
stG
Şti
Şube
SVM
szöv
T:mi
Tbk
td
TDV
tk
TLS
TNHHMTV
TOB
TOV
TÜ
UA
UAB
UD
UG
UK
Ultd
Unlimited
Unltd
UÜ
VAT
VEB

CIC
CIN
CIO
CNCTY
Co
CoLtd
Comm.V
Comm.VA
company
CONGTYCP
Coop
co-op
Corp
Corporation
CRL
CS
CtyCP
CtyTNHH
CTYTNHHMTV
CUSIP
CV
CVA
CVBA
CVoA
CXA
Cyf
DA
DAT
dba
dd
DECV
DNNN
dno
DNTN
doo
DOOEL
DTNN
e.Kfm
e.Kfr
EAD
EBVBA

HB
hf
HTX
HUF
IBC
IK
IKE
IKS
ILP
Inc
Incorporated
IP
IS
ISIN
IVS
jdoo
jp
jtd
jy
KAS
Kb
KD
KDA
kdd
KEG
KF
kft
Kft.
KG
KGaA
kht
kk
Kkt
Kol
KolG
Koll.Şti
Kom
KomAG
Koop
ks
KT

osk
OÜ
OVEE
Oy
Oyj
PartG
PartGmbBH
PC
PCLtd
PE
PEEC
PFA
PL
Plc
PLLC
PLT
PMA
PMDN
porA
PP
PPU
PPUH
PrC
PrK
Prp
PS
PSU
PT
Pte
PtK
Pty
Pty.
PUH
Pvt
PvtLt
PvtLtd
QK
QMJ
RAO
RAS
rf

SCP
SCpA
SCRI
scrl
SCS
SD
Sde
SdeRL
SdeRLdeCV
Sdn
SdnBhd.
SE
SECS
SEDOL
SEM
SENC
SenNC
SEP
sf
SGPS
SGR
Sh.A
Sh.p.k
SIA
SIC
SICAF
SICAV
SICC
Sicovam
SK
SKA
SL
SLL
SLNE
SLP
SM
SMBA
snc
SOC
Soc.Col
SOC.COOP.R.L.

venture
VOF
VoG
vos
VšĮ
VZW
WA
wIG
WKN
WPK
XT
YA
yCía
yk
YoAJ
ZAT
ZO.O
zoo
Zrt
zs
ABEE
AE
EE
IKE
OBEE
OE
OEE
AAT
AO
BAT
ET
КА
КТ
HP
OAO
OB
OOO
PO
TAA
TOO
ХК

Appendix 2: Structure of the interviews

General

- Go through the interview guidelines
    - Interview will be recorded, but only for the purpose of creating a transcript
    - Only the researcher will have access to both the recording and the transcript
    - Name will not be mentioned in the thesis, but the position at Sievo will
    - Do you have any questions at this point?
- Start recording
- Go through the following components of the study:
    - Objectives
    - Qualitative methods and purpose of this interview in assessing the quantitative results as well as the potential of future development
    - Quantitative methods
    - Results from the quantitative part of the study

Questions

- Is this kind of a solution needed? Who would it bring value to?
- What is your opinion on the methods used? Do you have any development ideas?
- What is your opinion on the overall results achieved?
- How about one item at a time?
- How do you experience the future potential for this solution?
- Is there anything else you would like to add regarding this study?