UiT The Arctic University of Norway

Faculty of Science and Technology
Department of Computer Science

Keeping Up with the Market: Extracting competencies from Norwegian job listings

Anton Garri Fagerbakk

Masters thesis in Computer Science May 2021

This thesis document was typeset using the *UiT Thesis LaTeX Template*.
© 2021 – http://github.com/egraff/uit-thesis

To Ebba and my caretakers at the Toukoul Orphanage

“I particularly hope that you will conclude the merit of the ideas I present outweigh my defects as a writer.”
–Philip A. Fisher

Abstract

The Norwegian labour market is under continuous change because of fast-paced innovation in technology. It is therefore vital for educational institutions curricula to reflect the changing requirements to keep the population hireable and provide employers with a highly adaptable workforce. There are no complete systems that let us analyse and extract this information about the labour market efficiently. Therefore, there is a need for tools to keep up with the labour market changes and to enable efficient analysis on large Norwegian job listing data sets.

In this project, we developed an algorithm that extracts the skills, competencies and knowledge from Norwegian job listing data. Our evaluation results show that we manage to extract skills from the jobs listings, but not to the extent of our defined requirements. This is caused by language ambiguity and semantic differences between our data sets, which significantly impacted our results.

We conclude that our algorithm has not fully solved the complex problem at hand but that our project has contributed with open-source code and processed open access data sets. Furthermore, through the development of the algorithm and analysis of the data sets, we have laid the foundation for future work and proposed how to develop solutions for understanding the fast-paced and continuous change in the Norwegian labour markets.

Acknowledgements

First, I would like to thank my supervisor, Edvard Pedersen, and my co-supervisors, Lars Ailo Bongo and Harald Groven (Kompetanse Norge) for their continuous guidance throughout this project.

I also want to thank Mona Mathisen and Vidar Berg at Kompetanse Norge for approaching me to work with Kompetanse Norge on this project.

I would also like to thank all of my classmates for the company and good times, especially during the long days and nights at the Department of Computer Science.

Finally, I would like to thank my girlfriend Margherita Falavigna, who has been my support system, even while delivering and defending her PhD thesis.

Contents

List of Figures

List of Tables

1

Introduction

The labour market is under continuous change because of fast-paced innovation in technology. Therefore, it will be increasingly challenging to educate and train the population for the quickly changing labour market. Pajarinen et al. [1] estimated that one-third of jobs are at risk in Norway due to computerisation. Frey and Osbourne [2] also shows that as much as 47% of U.S employees are in the high-risk category for jobs lost to computerisation or technological innovations. The unemployed population will require a new way to meet the new labour market conditions. Therefore, there is a need for new tools to analyse job listings to understand how the required skills, competences and education change. It is also vital that the sought after competencies from the labour market are reflected in the education given by learning institutions. Changing educational curricula is a lengthy process; hence, it is essential to have a continuous overview and analysis of the required skills and competencies in the labour market to expedite educational or other legislative processes. To understand these changes, we need a thorough analysis of the market. In Norway, this analysis is the responsibility of Kompetanse Norge, a government directorate under the Ministry of Education responsible for ensuring that the public has the information required to obtain an education, and that the labour market has access to a competent and highly adaptable workforce.

In the Nordic countries, public agencies produce or have archived a significant amount of high-quality open access data. In Norway, government agencies work with each other and have centralised solutions and mandates for personal or public data. According to the Labour Market Act of 2004, The Labour and

Welfare Administration collects and stores all public job listings. This database has free text that describes the job listing from the perspective of the employer. The job listing text can vary from technical, education and certification specific subjects and skills wanted by the employers to general knowledge of written and spoken language and wanted character traits. The free text is accompanied by 27 fields of structured metadata that describes the job listing. The listing data is licensed with a Creative Commons licence (CC-BY 4.0) and is accessible to the public. NAV has since 2002 published all job listing data resulting in a data set with over three million job listings per 2021. We believe that these job listings can provide novel insights into the labour market and the competencies required to participate. However, these data sets are large and continuously growing, so efficient tools must be created and maintained to extract the required information and changes in the data.

The thesis aims to deliver an algorithm to extract skills and help enable analytical work on the job listings. We will also make these developments available as an open-source contribution that Kompetanse Norge and others can continue to build. This system will work as a proof of concept for the Norwegian language within the job listing domain using state of the art Norwegian Natural Language Processing (NLP), cloud services and available data sets.

In this thesis, the word skill, or skills, will encompass skills, competencies, abilities, knowledge and behaviours. So whenever the word skill, or skills, is mentioned, it refers to either or all of these words.

1.1 Challenges

Commonly used libraries and state of the art cloud providers have developed NLP solutions. They offer easy accessible APIs with out of the box NLP solutions for different languages to large scale data processing pipelines to create ML models for NLP. However, these solutions do not completely help us solve the problems as listed below. Here are some of the main challenges that we face for this project:

1. State of the art and out of the box solutions have not yet fully solved how to work with free and unstructured text within specific domains, like extracting skills from job listings. We lack out of the box solutions for different languages which a user or developer can easily interface with to solve general or domain problems like high accuracy Name Entity Recognition, domain-level stop words or word category extraction, which is relevant for our project.

2. The accuracy offered by open-source NLP libraries will often reflect the popularity of a language or an extraordinary effort done by private persons or government agencies for which the language is being used. Therefore, more minor languages usually require more effort to achieve accuracy, similar to what the English language is provided with in open source libraries or out of the box services.

3. Currently, humans manually annotate, label or extract data from Norwegian data sets. However, these data sets are for specific projects and domains, which do not intersect with our project, some of which are not public. Manual labour can be accurate but inefficient, especially when the workload and digital job listings are continuously proliferating.

4. Another challenge is that a lot of time and effort is spent on data wrangling and preprocessing so that analytical work can be done, especially in domain-specific texts.

5. For the Norwegian job listing text, there is no apparent structure we can lean on. The listing texts vary a lot as there is no standard way to write a job listing, furthermore, the respective occupations have very different information to convey to an expected candidate.

1.2 Limitations of Existing Solutions

As mentioned in the section 1.1, cloud providers like IBM, Google Cloud, AWS, and Microsoft Azure offer AI, ML and NLP services. Using their readily available APIs, we can get sentiment analysis, key phrases and language detection services for the Norwegian language and many more for other languages. The cloud solutions also offer the possibility to set up an AI or ML infrastructure for more complex solutions for training massive data sets to create models and classifiers. For example, Azure offers an NLP pipeline with a drag and drop interface to preprocess and clean large data sets with excellent built-in functions for statistic analysis. However, most built-in NLP functions from the cloud providers mainly support a handful of the most popular languages. For example, Azure's ML NLP infrastructure only supports the five most major western languages. Therefore, we are not able to exploit pre-built cloud solutions for the Norwegian language and domain level problems.

Azure provide a medical text analysis prototype service exclusively for the English language [3]. For Azure developing the service is powered by the fact that medical language and jargon are usually homogeneous and the terms, diseases, and diagnoses have a defined taxonomy. This gives the text more structure

which in turn makes it easier to analyse and develop solutions. However, the medical text service is only in the preview stage, and Azure states that it should not be used in production. The service works best with smaller amounts of washed data since there are limits to the data amount one can apply to the service.

There are not many out of the box solutions for NLP and analyses for different languages, especially domain-specific solutions. LinkedIn has worked with and developed competence matching algorithms for employees and employers. LinkedIn has extensively researched and developed solutions to extract and match skills and competencies in the structured and unstructured text by using LinkedIn profiles and job advertisements on their platform and other job-related data sources. They have published and shared some of their research and findings in academic journals [4][5][6][7][8], but not the source code or data sets that they used. It is not available for other developers.

1.3 Solution & Contributions

The solution and contributions for this thesis are:

- The request analysis outline interesting problems, questions and use cases regarding the data set. The use cases provided by Kompetanse Norge help us understand the data set and the needs of the people who work with the data. The requirements derived from the use cases let us understand the domain problems and the required steps to solve them. This solves the problem of analysing the data for anyone interested in working with the data set.

- Our data wrangling efforts by cleaning and publishing the data sets. This solves the problem of spending time to clean future data sets and expedites data analysis. The data sets ; ground truth data, job listing data and ESCO data are made available on Kaggle as open-access.

- Description of the annotated ground truth data and an example of how to procure more of these data sample for future work. Initial evaluation of the algorithm and system using the ground truth job listings. [1]

- Developed a skill extraction algorithm for job Norwegian job listings using state of the art NLP. Contributing open source code based on modular design for ease of continued contributions or development within the

1. Provided by Kompetanse Norge

domain.

There are challenges and use cases that are not solved in this project, but based on the requirement analysis, we believe that the contributions we have made are essential to solve the problems of this thesis and future use cases.

2

Requirements Analysis

In this chapter, we look at some standards and frameworks which help define the data sets we use in this project and then analyse the use cases provided by Kompetanse Norge to derive implementation requirements to solve the use cases.

2.1 Data Sources & Standards

Both the US and EU have developed frameworks and standards to help develop systems to organise occupations, competencies, and abilities regarding the labour market globally. The organisation of the standards has been both for research purposes and to ease the implementation of future information systems that require a labour market framework.

2.1.1 O*Net

"The O*NET Program is the nation's primary source of occupational information. Valid data are essential to understanding the rapidly changing nature of work and how it impacts the workforce and the US economy. O*NET is a project by the US Department of Labour/Employment and Training Administration. O*NET is a database of occupational content, from occupational characteristics to worker requirements."

O*NET has a content model that describes occupations in terms of skills, knowledge, abilities required, how to perform the work and the tasks' descriptions. The O*NET content is widely used in academic research and private development for labour market analysis. The O*NET database contains information and taxonomies for occupations, educations and skills, which can be used to develop tools to map structured or unstructured texts within the domain.

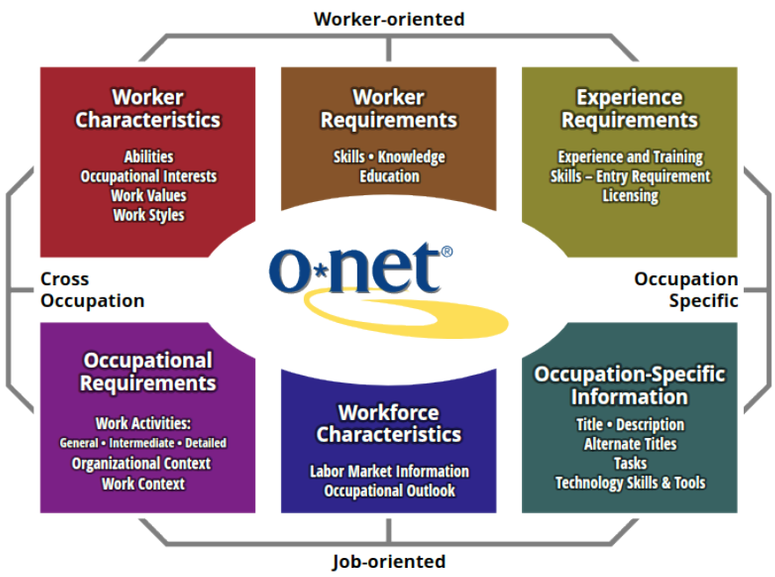

Figure 2.1: O*NET Content Model [9]

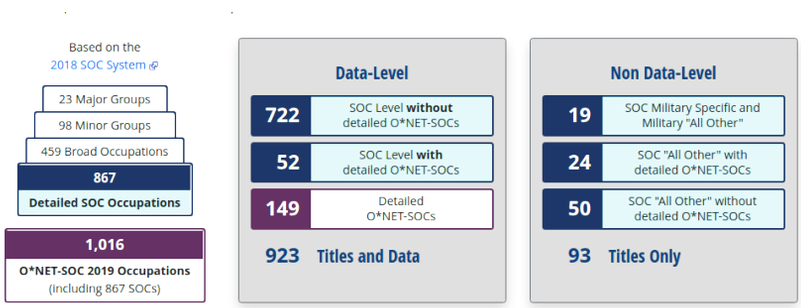

Figure 2.2: O*NET Taxonomy [10]

2.1.2 ISCO

International Standard Classification of Occupations (ISCO) is one of the international classification standards for occupations. It provides an international standard and framework for countries that have not yet defined or classified occupations. ISCO helps with statistical applications and the development of

national systems.

ISCO has several definitions for occupations, skills and skill levels. The latest ISCO-08 standard offers classification of these definitions and the hierarchies related to the skills, occupations and skill levels. The ISCO standard holds 400 occupational categories.

As seen in figure 2.3 ESCO has a link to ISCO's defined hierarchies which helps ESCO with occupational definitions.

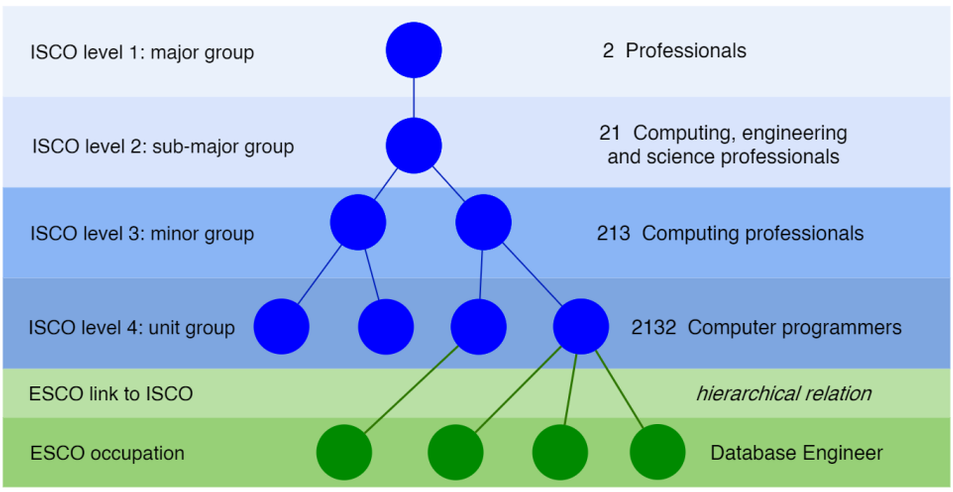

Figure 2.3: ISCO Taxonomy [11]

2.1.3 ESCO

European Skills, Competences, Qualifications and Occupations (ESCO) is EU's multilingual version of ISCO and O*NET. It is a classification of European skills, competencies, qualifications and occupations. It is developed as part of the EU project "Europe 2020 strategy and the New Skills Agenda for Europe" but it is an ongoing effort to map the skills and abilities in the changing labour and educational market. The difference between ESCO and O*NET/ISCO is that ESCO is multilingual and currently supports 27 languages, primarily European, including Norwegian, to different capacities.

ESCO currently has 13485 skills in their skill pillar. The skills pillar is structured in a hierarchy with four sub-classifications. Each classification is characterised by a different concept. The concepts are:

- Knowledge

- Skills
- Attitudes and values
- Language skills and knowledge

In addition to the hierarchy, the subsets of skills can be accessed through:

- A transversal skill hierarchy
- A collection of languages
- A collection of digital skills

Most importantly, ESCO has 13485 skills, almost 3000 occupations and several other features as seen in figure 2.4 in Norwegian available to the public for download on their web page. With these terms of occupations and skills brought to us by ESCO and ISCO, we as developers can train classifiers, machine learning models for occupation, skill or other types of prediction, as well as skill matching algorithms based on the terms found in job listing text and terms defined in the ESCO and ISCO standard.

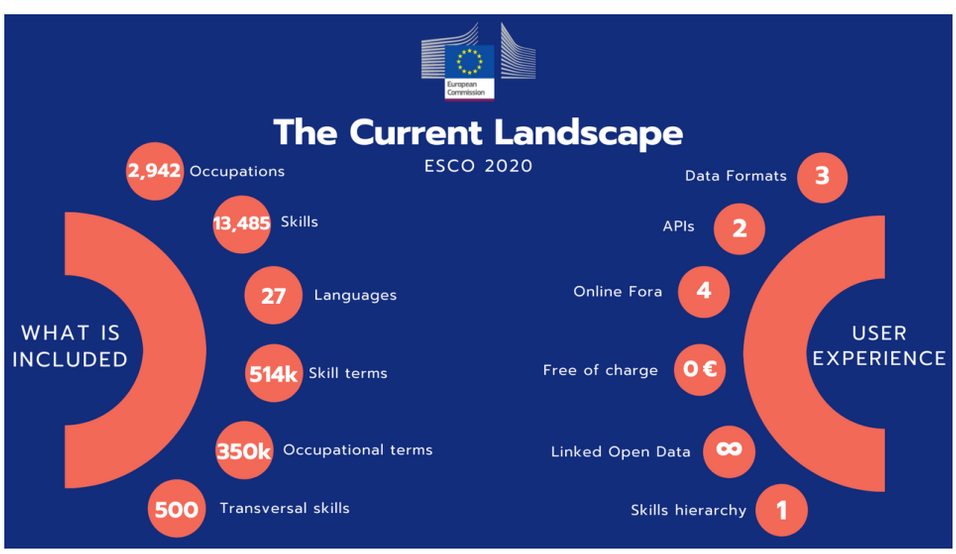

Figure 2.4: ESCO Landscape [12]

2.1.4 Kompetanse Norge

Kompetanse Norge (KN) is a directory under the Ministry of Education. Their mission is to continuously educate Norway's population through correct and

necessary information on acquiring education and or occupations. KN develops different public tools like www.utdanning.no, which is the official Norwegian national education and career portal. Utdanning.no offers information and tools regarding how to achieve an education from 8000 different educational possibilities or acquire a particular job or profession. Statistisk Sentralbyrå, or Statistics Norway, has developed the STYRK occupational taxonomy, a derivative of ESCO. The creation of STYRK occupational taxonomy comes partly from the fact that the ESCO or ISCO standard can be underwhelming for nations, and therefore it makes sense to build their framework on top of the established standards. The Norwegian STYRK98 occupational standard has 6000 different occupations which employers must categorise their employees into when they are registered as employees versus the 400 available ISCO occupations.

2.2 Use Cases

Let us analyse the use cases that Kompetanse Norge has defined for this data set. The use cases outline the information they want to extract, but it is up to us to analyse how to solve and define the requirements for each use case. The process for analysis is:

- Get use cases by Kompetanse Norge.
- Analyse use case and create requirements.
- Analyse the requirements to find the central and important elements to solve.

2.2.1 Use Cases

1. Which occupations have more vacancies past ten years?
2. Which occupation has declined in ten years?
3. Which education descriptions are most frequently mentioned in job vacancies for a given title or occupation code?
4. Which skills are most frequently mentioned in job vacancies for a given title or occupation code?
5. Which skills are most frequently listed as required (“må kunne... “/” må beherske”) for a given title or occupation?

6. Which skills are disproportional more frequently mentioned in job vacancies in a given municipality than the whole country?

7. Which sub-fields are most common within a given occupation code? E.g. “IT-developer” may have “front-end developer”, “back-end developer”, “web designer”, “dev-ops.”

We break down the use cases to further analyse the requirements and what is required to fulfil them.

2.3 Use Case Analysis

We create requirements from the use cases depending on the use case analysis, and we state what a satisfactory result to the requirement would be. High accuracy, in this case, means a hit rate of higher or equal to 90%.

2.3.1 Use Case 1

Which occupations have more vacancies past ten years?

- We need to find the occupation in the job listing data to solve the requirement. This use case is satisfactorily solved when we have a solution with high accuracy to determine the correct occupation either from the free text or the occupational metadata. The correct occupation would be defined as the correct occupation within the category as defined by STYRK or a similar standard.

2.3.2 Use Case 2

Which occupation has declined in ten years?

- This use case requires us to use the timestamp in the structured data for each job listing. This use case is satisfactorily solved when we can present a time analysis from the use case question.

2.3.3 Use Case 3

Which education descriptions are most frequently mentioned in job vacancies for a given title or occupation code?

- The occupations education requires a classifier or a data source to extract the educational level. Like the occupation use case, to solve this a resulting solution should give us the educational level either from the free text or occupational code with good accuracy. The correct education would be defined as the education that is correct within the category defined by NUS[13] or a similar standard.

2.3.4 Use Case 4

Which skills are most frequently mentioned in job vacancies for a given title or occupation code?

- This use case requires us to have a way to extract skills from the job listing. This use case is considered solved when we can extract skills with high accuracy from a free text job listing.

2.3.5 Use Case 5

Which skills are most frequently listed as required (“må kunne... “/” må beherske”) for a given title or occupation?

- The use case requires us to define how "required" is worded as this can be hard to define and contextualise in the text. This use case is solved when we can find the required skill from the free text with high precision.

2.3.6 Use Case 6

Which skills are disproportional more frequently mentioned in job vacancies in a given municipality than the whole country?

- For this use case, we need several features like skill extraction combined with the job listing’s structured data. This use case is solved when the use case for skill extraction is solved.

2.3.7 Use Case 7

Which sub-fields are most common within a given occupation code? E.g. “IT-developer” may have “front-end developer”, “back-end developer”, “web designer”, “dev-ops.”

- This use case is complex and requires us to have a definition or taxonomy for occupations and their respective skills. This use case is solved when the skill use case is solved and result in high precision results for a given occupation code.

2.3.8 Use Case Complexity

From the use cases and their requirements, we can categorise them as either complex or non-complex, regarding the solutions to produce a satisfactory result.

- Non-complex use cases
  - Which occupations have more vacancies in the past ten years?
  - Which occupations has declined in the last ten years?
- Complex use cases
  - Which education descriptions are most frequently mentioned in job vacancies for <word in title field/occupation-code>?
  - Which skills are most frequently listed as required (“må kunne... “/” må beherske”) for a given title or occupation?
  - Which skills are disproportionally more frequently mentioned in job vacancies in a given municipality, compared to the whole country?
  - Which sub-fields are most common within a given occupation-code? E.g. “IT-developer” may have “front-end developer”, “back-end developer”, “web designer”, “dev-ops.”

In this situation, non-complex use cases are defined as use cases or questions that can be solved using readily available solutions.

Complex use cases are the use cases or questions requiring more complex solutions that are not readily available. More complex means that we must use complex methods and algorithms to solve more nuanced problems like extracting skills or occupation from a job listing text. These complex use cases can require solutions that correlate to the challenges mentioned in the introduction, which require domain data or ML models that have not yet been developed.

2.4 Thesis Focus

As seen in the 2.3.1 and 2.3.8 several use cases require sophisticated solutions to yield a wanted result. In the use cases, the result requires us to have an algorithm that can extract specific data like occupation or skills. In this thesis, we focused on solving the use cases where we needed to extract a skill. The problem of extracting skills from the job listing data is present in several use cases, and there is also a commonality of being able to extract some information about a category. Therefore extracting skills seems like an essential first step that can lend itself to future development.

/3

Data Set

3.1 NAV

The Norwegian Labour and Welfare Administration (NAV) organise social and financial aid for the Norwegian people on behalf of the state. The NAV system is vast and entails a third of the Norwegian governments annual budget. Therefore, we will only refer to their systems which aligns with our work. The unemployed can apply for welfare checks, which the state pays; NAV and the state have an incentive to solve joblessness, and they do so by helping jobless people apply for jobs through their system. This means that they host job listings from public and private companies and organisations and host the applicants' CV online.

In this project, NAV is also the source for our job listing data. They archive and publicly publish job listing data annually. Since 2002 NAV has published over three million job listings, and every public job has to be listed in the NAV's portal. The public data is used by researchers, policymakers, and other organisations or individuals interested in analysing the Norwegian labour market. NAV also does analysis and research on the data that they collect in their services.

3.2 Job Listing Data

To fully understand the design choices, use cases and underlying problems, we need to understand the data. The data-set consists of 26 fields of structured data and one field of unstructured free-text data.

3.3 Listing Fields

As shown in table 3.1, the meta-data is very detailed for each job listing. The most important fields for our development are "stilling_id" and "beskrivelse". The "beskrivelse" field is the description that holds the free text, and the "stilling_id" is the unique identifier that ties all the meta-data to a particular listing.

3.3.1 Styrko8 and Occupations

Yrkeskode_styrko8 and specifically STYRK08 is the taxonomy that NAV uses to categorise occupations [13] [14]. The STYRK codes make it easier to map the meta-data to the specific occupations and occupational categories. Each STYRK code also has a name and is a subset of a STYRK main category. For example, a listing as an engineering job in the data-set; Title: "Senior Ingeniør Konstruksjonsteknikk" translates to senior construction engineer. The STYRK08 code is 2142, STYRK08 name: "Sivilingeniører (bygg og anlegg)", which translates to civil construction engineers, and the main STYRK category is: "Ingeniør- og ikt-fag" which translates to engineering and IT. We also have "yrkesbetegnelse", which translates to the occupational designation. "Yrkesbetegnelse" will sometimes coincide with the STYRK08 name but can also differ as yrkesbetegnelse is specific regarding the job listing role while STYRK08s name can be more categorical.

Name	Example of Listing	Description
stilling_id	1894669	Unique listing ID (INT)
tittel	Senior Ingeniør Konstruksjonsteknikk	Listing title (STRING)
antall_stillinger	1	Number of jobs (INT)
yrkeskode_styrko8	2142	Occupation code (INT)
yrkeskode_styrko8_navn	Sivilingeniører (bygg og anlegg)	Occupation name (STRING)
yrkeskode_hovedkategor		Occupation category (STRING)
regdato	2009-02-05	Date registered (DATE)
siste_publisert_dato	2009-12-31 00:00:00	Last updated (DATE)
kommunenr	0301	Municipality code (INT)
kommunenavn	Oslo	Municipality name (STRING)
fylkenr	03	County code (INT)
fylke	Oslo	County name (STRING)
landkode	NO	Country code (STRING)
land	Norge	Country (STRING)
yrkesbetegnelse	Sivilingeniør (bygg og anlegg)	Occupation Category (STRING)
orgnr	991706992	Organisation code (INT)
org_navn	SEATOWER AS	Organisation name (STRING)
aktiv_flagg	1.0	Active flag (FLOAT)
org_nace	74909	Organisation code (INT)
hovedenhet	891697872	Main organisation (INT)
hovedenhet_navn	SEATOWER AS	Main organisation name (STRING)
hovedenhet_sektor	Privat og offentlig næringsvirksomhet	Main organisations sector (STRING)
tilleggskriterium	Dagtid, Fast stilling, Heltid	Extra criteria (STRING)
beskrivelse	Har du tung erfaring innen design av konstruksjoner (...)	Description (STRING)
sprak	no	Language (STRING)
kilde	Reg av arb.giver på nav.no	Source (STRING)
nav_enhet_kode	0334	NAV unit code (INT)

Table 3.1: All listing fields and an accompanying example

3.3.2 Geographical Data

To analyse the data and answer question on a municipality, county or even country basis, we can use six fields encompassing geographical location data. These fields are "komunenr", "kommunenavn", "fylkenr", "fylke", "landkode" and "land".

3.3.3 Date and timestamps

Fields like "regdato", which translates to registration date, helps us understand when the data was listed and opens for time series analysis or other time-related analysis. The job listing can be listed several times in theory, but we have not observed duplicate listings in our data set. A company or organisation could also re-use an old listing text if they look for another employee in an identical occupation or role. Therefore we have to assume that all listings are unique as long as they have different listing ids.

3.3.4 Miscellaneous Fields

We have many essential fields regarding job listing registration for official records in the data set, but not particularly interesting regarding the project requirements. These are the fields describing the organisation listing the job, the language of the listing, the listing source, and which NAV unit it belongs to.

3.3.5 Description

Up until now, all of the fields we have mentioned have been standardised and structured. The field "beskrivelse" is a free text description of the listing. The "beskrivelse" field is unstructured and non-standard. Individual employers write the description text in a manner that they think conveys the most critical aspect of the job. Therefore, each text can be structured differently, but there are usually similar phrases and terminology in most job listings. As with a general job listing, this description field should hold essential information about the job. It is here that we expect to find the necessary skills or abilities from potential candidates. This field is however optional, and some listings do not utilise this field. The description can be empty or have a URL link to another website where the employer hosts the essential information about the listing. The informational level in the listing varies for different occupations. Since some occupations have implied education like a physician, there is no reason for the employer to express the importance of having said education in the job listing. For other occupations that do not require formal education, there might be more emphasis on wanted or expected skills or education. In occupations like front-end developer, where there are many different frameworks, programming languages and other specific fields, there is often the expectation or want for the candidate to have specific technical skills or certifications. Since the data set has several hundred occupations, we encounter different job listing texts, which make the description field highly challenging for information retrieval.

3.4 ESCO Ontology

The ESCO data set contains 13845 different skills and knowledge classifications. With the additional language skills, ICT skill collection, and skill groups, we have 14151 skills, competencies, languages and knowledge. This collection is available in 27 different languages, including the Norwegian language, in text form through ESCO files.

3.4.1 ESCO Skills

The skill fields described in the 3.2 give us an overview of the different fields. We have focused mainly on the "conceptType" and "preferredLabel" data fields as they hold the essential information for the implementation.

Concept Type

The concept type details what concept the particular skill adheres to. The skills are classified into two different labels, "Knowledge" and "Skill/Competences".

Preferred Label

The "preferredLabel" data field contains a short description or label in one to seven words about the said skill. Skills labelled with the concept type "knowledge" usually have shorter descriptions making them more label-like in classification.

Name	Description
conceptType	Overall concept type (STRING)
conceptUri	URL to ESCO definition of concept (STRING)
skillType	Skill type (STRING)
reuseLevel	Skill re-usability level (STRING)
preferredLabel	Preferred skill label (STRING)
altLabels	Alternative skill labelling (STRING)
hiddenLabels	Hidden labels (STRING)
status	Status of data field (STRING)
modifiedDate	Date field was modified last (DATE)
scopeNote	Note on scope of skill (STRING)
definition	Definition of skill (STRING)
inScheme	Concept scheme URL (STRING)
description	Verbose description of skill (STRING)

Table 3.2: Description of ESCO skill fields

3.5 Ground Truth Data

3.5.1 Data Set Evaluation

[1] The ground truth data is sample of 100 job listings taken from the NAV public job listing data set. We start to extract the data with an SQL query for extracting a random sample of job listings from the database nav_ledigestillinger in Norwegian, having a length longer than a threshold value.

```
SELECT
        CONCAT(tittel, " - ", beskrivelse) AS text,
        stilling_id,
        tittel AS title,
        CHAR_LENGTH(beskrivelse) AS description_len
FROM nav_ledigestillinger
# only show texts longer than threshold value,
# here: median length of description field
WHERE CHAR_LENGTH(beskrivelse) > 1645 # i.e. median length
AND sprak = "no" # only Norwegian lang rows
ORDER BY RAND() # randomise result set
LIMIT 100 # Show only top N results
;
```

1. Ground truth data and its explanation is written and provided by Kompetanse Norge

	Total Skills	Word Count	Total N-Grams
Skills	2057	6630	34647
Job Listing Text	284	5926	10761

Table 3.3: Annotated data for a job listing

3.5.2 Manual Tagging Software

There are several commercial software for labelling training data for ML. For our project, the staff at Kompetanse Norge used the open-source web application Label studio. Label studio (version 1.0) currently supports more than 40 different annotation templates, covering both text, table, image, audio, and video data. Label studio stores completed training data as JSON files.

3.5.3 Labelling Data

26.6 percent of job listings in the database contain less than 1000 characters in the description field (beskrivelse). Most of these contain only a link to another job listing web site. Job listings shorter than the median length of job listings were excluded from the sample.

Example job listing:

"WE ARE RECRUITING ADVISOR/SENIOR ADVISOR - COMPLIANCE Deadline for application is 11th October 2015. Please see www.nbim.no for more details and how to apply." (stilling_id 9240510)

Such listing, which refer to an external URL for application, doesn't contain much useful information for NLP analysis. If we limit the training data set to listings longer than a certain threshold.

The median text length of the log listing data set was 1645 characters. A random unweighted sample of job listings from the database was drawn within the population of listings longer than the median length. A manually created training data set from a sample of 100 job listings. These were tagged with the following three categories;

- Language: Words in the job listing referring to which language skills required for doing the job.

- Education Level: Descriptions of which level of education was regarded as necessary for the job.

- Skill: "Developed capacities" the employee needs in order to be able to do the job.

Figure 3.1: Labelling a job listing position for work as a physician

Staff at Kompetanse Norge created a training data set with manually tagged data. An example of labelling **??**.

4

Design and Implementation

Based on the use cases provided by Kompetanse Norge and the requirements we have derived from said use cases, we have developed and designed an architecture to handle text data and skill extraction. The first phase required us to implement a way to process text into a standardised format to make skill extraction efficient. In the second phase, we needed to create an algorithm to extract these skills from text using the data sources we processed in the first phase.

In the experiment and evaluation, we used the manually annotated data set provided by Kompetanse Norge, which provided the listing id, listing text, the annotated skills called "Listing Skills" and the educational level of a listing.

4.1 Phase One - Text Data Cleaning

The cleaning phase consists of removing unwanted characters and standardising the text format to give higher matching precision during the skill extraction phase. The data cleaning step is independent of the extraction phase.

To understand how the cleaning works, let us look at the different states of a job listing text before, during and after text cleaning.

"Sykepleier — - 69,75% sykepleier Demensenheten er en spesialenhet innenfor pleie- og omsorg i kommunen. Enheten har tilsammen 51 sykehjemsplasser som er lokalisert i Brumunddal og Moelv med tilsammen 41 plasser, samt med 10 plasser som er lokalisert på Sundheimen på Helgøya. Avdelingene er funksjonsinndelt i grupper med egne postkjøkken og oppholdsrom. I tillegg har enheten dagsenter i Moelv og Brumunddal. Enheten har 1 årsverk som fagkonsulent. Vi søker 69,75% stilling som sykepleier/vernepleier er ledig for tiltredelse snarest. Arbeidsoppgaver Stillingen er turnusstilling med arbeid hver 3. helg og er for tiden knyttet til avd. i Moelv. Kvalifikasjoner Det kreves autorisasjon som sykepleier eller vernepleier. Arbeidets art krever erfaring fra å jobbe med personer med alvorlig demens og stor grad av utfordrende og uforutsigbar atferd. Søker må ha erfaring med aktiv bruk av miljøtiltak og skjerming. Søker må ha en trygg og faglig forankret tilnærming til pasientene. Søker må vise til en tydelig visjon om eget bidrag i avdelingen. Søker må ha erfaring fra teamarbeid og ha en stor grad av løsningsfokus. Søker til forsterket skjermet enhet må påregne å jobbe ved de andre avdelingene i Demensenheten i perioder, samt å kunne veilede personalet i omsorgsdistriktene. Arbeidets art krever førerkort klasse B, disponere egen bil i tjenesten, samt god fysisk helse. Personlig egnethet vil bli tillagt stor vekt. Annet Årslønn i 100% stilling: Sykepleier, st.kode 7174: Fra kr 354 000,- til kr 405 100,-. Vernepleier, st.kode 6455: Fra kr 354 000,- til kr. 405 100,- Vi oppfordrer til å levere søknad elektronisk via kommunens adresse: www.ringsaker.kommune.no under ledige stillinger. Ta kontakt og vi er gjerne behjelpelige dersom du trenger hjelp til å søke elektronisk. Attester og vitnemål medbringes ved eventuelt intervju. I henhold til offentleglova § 25, kan opplysninger om søker unntas i offentlig søkerliste dersom søkeren selv ber om det. Søkere som ønsker at navn skal unntas i offentlig søkerliste, må oppgi begrunnelse. Dersom begrunnelsen for fritak ikke er tilstrekkelig, vil søker bli kontaktet. Stillingtype: Fast stilling Kontakter: , Avdelingsleder, , Leder demensenheten, Søknad merkes: REF. For fullstendig utlysningstekst og elektronisk søknadsskjema: "

Figure 4.1: Original state of job listing 6835001

Job Text Cleaning

The job listings are scraped directly from the website they are hosted on before they become publicly available. NAV posts and hosts the job listing data. There are HTML tags and other "noise" which we need to remove. The Norwegian language has a lot of hyphenated and compound words. The hyphenated words and the hyphen character are essential regarding the context of the text, which exposes some interesting problems regarding matching strings. For example, the word "3d-modeller", which translates to "3d models", in Norwegian, has a numerical character and a hyphen. Therefore, using regular expressions methods to clean all numerical and non-word characters from text strings will remove the meaning and context of the word. This means that numerals and hyphens are kept in the job listing text. Usually, libraries or cloud solutions will remove all non-word characters from the text when running their cleaning

methods unless specified not to do so. We choose to write our cleaning method for this project as is non-problematic and can be essential to capture data set specific contextual differences like hyphenated words in Norwegian.

This figure shows the difference of the state after the cleaning step of the job listing as above.

sykepleier 69 75 sykepleier demensenheten er en spesialenhet innenfor pleie- og omsorg i kommunen enheten har tilsammen 51 sykehjemsplasser som er lokalisert i brumunddal og moelv med tilsammen 41 plasser samt med 10 plasser som er lokalisert på sundheimen på helgøya avdelingene er funksjonsinndelt i grupper med egne postkjøkken og oppholdsrom i tillegg har enheten dagsenter i moelv og brumunddal enheten har 1 årsverk som fagkonsulent vi søker 69 75 stilling som sykepleier vernepleier er ledig for tiltredelse snarest arbeidsoppgaver stillingen er turnusstilling med arbeid hver 3 helg og er for tiden knyttet til avd i moelv kvalifikasjoner det kreves autorisasjon som sykepleier eller vernepleier arbeidets art krever erfaring fra å jobbe med personer med alvorlig demens og stor grad av utfordrende og uforutsigbar atferd søker må ha erfaring med aktiv bruk av miljøtiltak og skjerming søker må ha en trygg og faglig forankret tilnærming til pasientene søker må vise til en tydelig visjon om eget bidrag i avdelingen søker må ha erfaring fra teamarbeid og ha en stor grad av løsningsfokus søker til forsterket skjermet enhet må påregne å jobbe ved de andre avdelingene i demensenheten i perioder samt å kunne veilede personalet i omsorgsdistriktene arbeidets art krever førerkort klasse b disponere egen bil i tjenesten samt god fysisk helse personlig egnethet vil bli tillagt stor vekt annet årslønn i 100 stilling sykepleier st kode 7174 fra kr 354 000 til kr 405 100 vernepleier st kode 6455 fra kr 354 000 til kr 405 100 vi oppfordrer til å levere søknad elektronisk via kommunens adresse www ringsaker kommune no under ledige stillinger ta kontakt og vi er gjerne behjelpelige dersom du trenger hjelp til å søke elektronisk attester og vitnemål medbringes ved eventuelt intervju i henhold til offentleglova sect 25 kan opplysninger om søker unntas i offentlig søkerliste dersom søkeren selv ber om det søkere som ønsker at navn skal unntas i offentlig søkerliste må oppgi begrunnelse dersom begrunnelsen for fritak ikke er tilstrekkelig vil søker bli kontaktet stillingtype fast stilling kontakter avdelingsleder leder demensenheten søknad merkes ref for fullstendig utlysningstekst og elektronisk søknadsskjema

Figure 4.2: State of job listing 6835001 after cleaning

4.1.1 Lemmatisation

Lemmatisation lets us get the dictionary form of a word and standardises the text format before string matching, which gives us a better chance to match words and strings. To lemmatise the text, we use the SpaCy NLP library with a lemmatisation feature for different languages.

This figure shows the state after the lemmatisation step.

['sykepleier', '69', '75', 'sykepleier', 'demensenhete', 'er', 'en', 'spesialenh', 'innenfor', 'pleie', 'og', 'omsorg', 'kommune', 'enhet', 'ha', 'tilsammen', '51', 'sykehjemsplass', 'som', 'er', 'lokalisere', 'brumunddal', 'og', 'moelv', 'med', 'tilsammen', '41', 'plass', 'samt', 'med', '10', 'plass', 'som', 'er', 'lokalisere', 'på', 'sundheime', 'på', 'helgøya', 'avdeling', 'er', 'funksjonsinndelt', 'gruppe', 'med', 'egne', 'postkjøkke', 'og', 'oppholdsrom', 'tillegg', 'har', 'enhet', 'dagsenter', 'moelv', 'og', 'brumunddal', 'enhet', 'ha', 'årsverk', 'som', 'fagkonsulent', 'vi', 'søke', '69', '75', 'stilling', 'som', 'sykepleier', 'vernepleie', 'er', 'ledig', 'for', 'tiltredelse', 'snar', 'arbeidsoppgave', 'stilling', 'er', 'turnusstilling', 'med', 'arbeid', 'hver', 'helg', 'og', 'er', 'for', 'tid', 'knytte', 'til', 'avd', 'moelv', 'kvalifikasjon', '-PRON', 'kreve', 'autorisasjon', 'som', 'sykepleier', 'eller', 'vernepleie', 'arbeidets', 'art', 'kreve', 'erfaring', 'fra', 'jobbe', 'med', 'person', 'med', 'alvorlig', 'demens', 'og', 'stor', 'grad', 'av', 'utfordrende', 'og', 'uforutsigbar', 'atferd', 'søke', 'må', 'ha', 'erfaring', 'med', 'aktiv', 'bruk', 'av', 'miljøtiltak', 'og', 'skjerming', 'søke', 'må', 'ha', 'en', 'trygg', 'og', 'faglig', 'forankre', 'tilnærming', 'til', 'pasient', 'søke', 'må', 'vise', 'til', 'en', 'tydelig', 'visjon', 'om', 'eget', 'bidrag', 'avdeling', 'søke', 'må', 'ha', 'erfaring', 'fra', 'teamarbeid', 'og', 'ha', 'en', 'stor', 'grad', 'av', 'løsningsfokus', 'søke', 'til', 'forsterke', 'skjerme', 'enhet', 'må', 'påregne', 'jobbe', 'ved', 'de', 'annen', 'avdeling', 'demensenhete', 'periode', 'samt', 'kunne', 'veilede', 'personale', 'omsorgsdistrikte', 'arbeidets', 'art', 'kreve', 'førerkort', 'klasse', 'disponere', 'egen', 'bil', 'tjeneste', 'samt', 'god', 'fysisk', 'helse', 'personlig', 'egnethet', 'vil', 'bli', 'tillegge', 'stor', 'vekt', 'annen', 'årslønn', '100', 'stilling', 'sykepleier', 'st', 'kode', '7174', 'fra', 'kr', '354', '000', 'til', 'kr', '405', '100', 'vernepleie', 'st', 'kode', '6455', 'fra', 'kr', '354', '000', 'til', 'kr', '405', '100', 'vi', 'oppfordre', 'til', 'levere', 'søknad', 'elektronisk', 'via', 'kommunens', 'adresse', 'www', 'ringsaker', 'kommune', 'no', 'under', 'ledig', 'stilling', 'ta', 'kontakt', 'og', 'vi', 'er', 'gjerne', 'behjelpelig', 'dersom', 'du', 'trenge', 'hjelp', 'til', 'søke', 'elektronisk', 'attest', 'og', 'vitnemål', 'medbringe', 'ved', 'eventuell', 'intervju', 'henhold', 'til', 'offentleglova', 'sect', '25', 'kan', 'opplysning', 'om', 'søke', 'unnta', 'offentlig', 'søkerliste', 'dersom', 'søker', 'selv', 'be', 'om', '-PRON', 'søker', 'som', 'ønske', 'at', 'navn', 'skal', 'unnta', 'offentlig', 'søkerliste', 'må', 'oppgi', 'begrunnelse', 'dersom', 'begrunnelse', 'for', 'fritak', 'ikke', 'er', 'tilstrekkelig', 'vil', 'søker', 'bli', 'kontakte', 'stillingtype', 'fast', 'stilling', 'kontakte', 'avdelingsleder', 'lede', 'demensenhete', 'søknad', 'merke', 'ref', 'for', 'fullstendig', 'utlysningstekst', 'og', 'elektronisk', 'søknadsskjema']

Figure 4.3: State of job listing 6835001 after lemmatisation

ESCO Text Cleaning

The content described in 3.2 lets us extract the skills and store them in a DataFrame(DF) using Pandas Python library. By iterating through the DF, where each line represents a skill, we can pass the skill to a method that uses the SpaCy NLP library. We send in the language as a parameter to specify which NLP model to use and the skill, or text, which we want to lemmatise. The SpaCy NLP library has part-of-speech tagging (POS) for different languages, and we can use it to extract the lemma form for every word in a text that we pass to a filtering function. After we are done filtering the word's lemma form, we write the lemmatised word to the DF, written to a file containing the original skill text and the lemma form of that skill.

4.2 Phase Two - Skill Extraction

In the skill extraction phase, we iterate through a list of ground truth job listings. For each iteration, we do these steps:

1. SpaCy NLP & Lemmatisation
2. N-gram Creation
3. Format Ground Truth Skills
4. Skill Matching - ESCO
5. Store Results

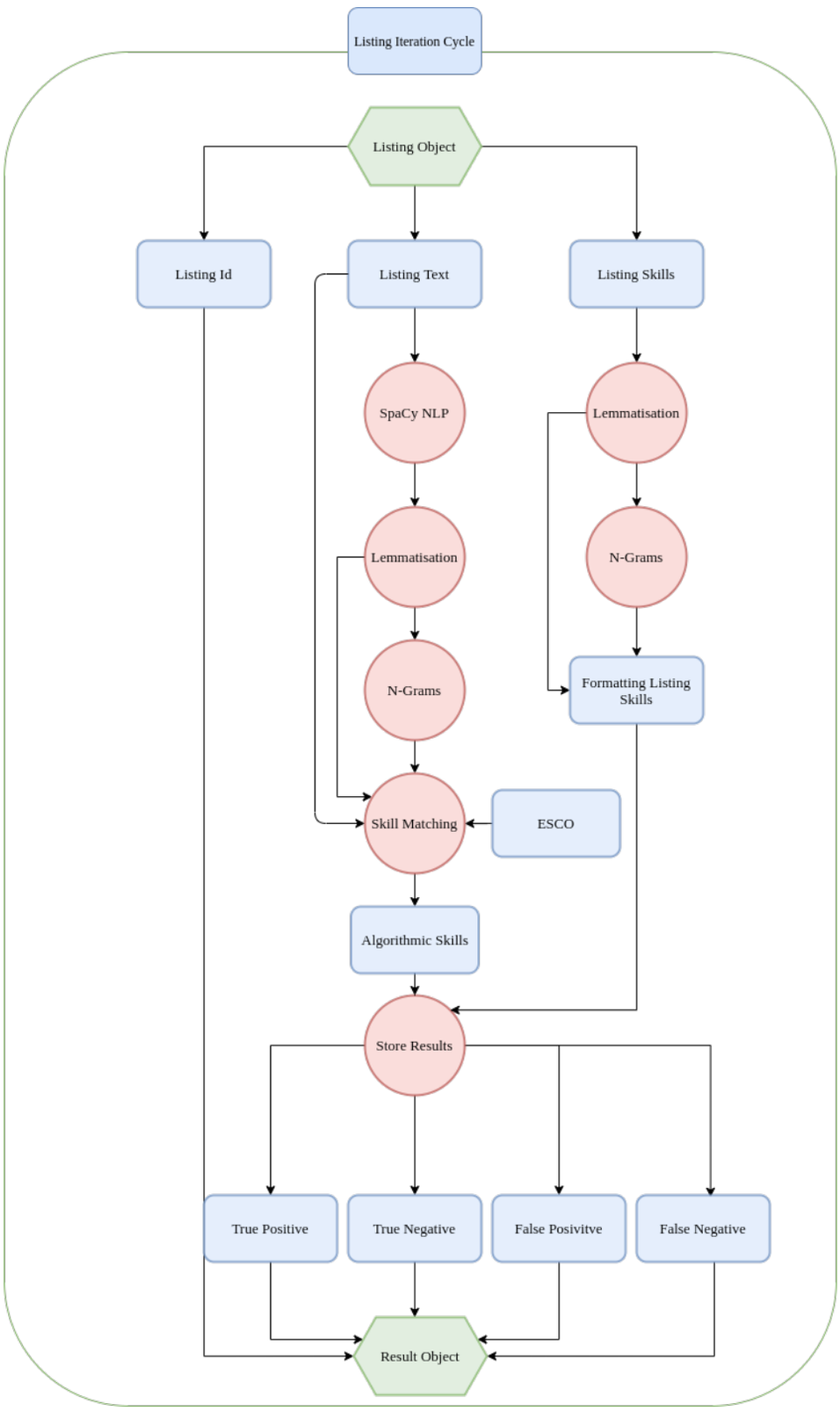

Figure 4.4: Listing Iteration Cycle

4.2.1 SpaCy NLP & Lemmatisation

We gather the listing description, listing id, and ground truth skills from the job listing object. Next, we take the job description and pass it to the SpaCy NLP function, which turns the description text into a SpaCy document which lets us utilise SpaCy functionalities like POS and NER. The document is then passed to the lemmatisation method described earlier, and we get the lemmatised version of the job listing text back. The lemmatised text is passed to the n-gram function.

4.2.2 N-grams

Our skills, abilities and values are not necessarily one word or two worded sentences but can go all the way up to n words, where n is the number of words in the text passed to the function. This means that making a strict string comparison of one word will be difficult against a skill that is eight words long. Therefore, we need to create n-grams and match the combinations of words from the job listing against our skills.

The job listing text in the original and lemmatised form is made into n-grams from size 2 to 8.

After getting both the original and lemmatised job listing text, we construct n-grams with length n, with a minimum of 2 words, for bots sets of texts.

4.2.3 Ground Truth Skills

We do the same procedure for the ground truth listing skills as we did for the job listing description by producing lemmatised and n-grams versions of the skills. This leaves us a lemmatised list of the ground truth skills, an n-gram list and the original skill text. The three different sources will be used when generating the results to improve accuracy.

4.2.4 Skill Matching

We want to extract skills from the job listing text in the skill matching step using our matching algorithm. We use the matching method and pass three arguments: the text we want to match, in our case, a word from the job listing text or an N-gram, the ESCO language we want to match it against and the threshold of similarity between words. After all words in the text have been iterated over, we remove duplicate skills and return the acquired skills in a

list.

The process of the algorithm is as follows; first, we fetch the ESCO skills in the language we passed to the method; we then iterate through the text that has been passed to the matching method to match it against the ESCO skills. We have two conditions to filter and match against the ESCO data:

- We check if the job listing word is in the skill we match.
- We make a strict comparison of the two strings.

For both of these options, we also use the fuzzy string matching technique that looks at the similarity between the words, also called word distance or Levenshtein distance[15]. The fuzzy matching method lets us pass our similarity threshold as a parameter. We found that 90% similarity is efficient to find similar words; if we lower the threshold further, we get many false positives, but this also differentiates regarding sentence or word length. Therefore we will discard a text match if it does not have a 90% word similarity.

N-gram Matching

As mentioned in section 4.2.2, we have created n-grams from 2 to n n-grams from the job listing text to match the job listing skill text. We reuse the method for ESCO matching, but instead of matching words, we pass the n-grams to the method.

Store Results

When the matching is done, and we have gotten a list of skills, we store the results in an object. We store five things in the listing result object.

- True Positive(TP): The skills found in both the algorithmic and the annotated result.
- True Negative(TN): The skills that are not found in the annotated nor algorithmic results. This does not apply to this project and will always be zero.
- False Positive(FP): The algorithmic skills found in the experiment but not in the annotated data.
- False Negatives(FN): The annotated data that are not found in the text.

The result object is then stored in a list for metric calculations of our experiments.

4.3 Modularity

The work for this project is prototype-based which means that we want to see if our implementation is a viable way to solve our use cases. While solving these problems, the focus has been on keeping the implementations and architecture as modular as possible. For us, prototyping means discarding non-working modules while keeping the ones that do, which is why modularity is necessary for future work. The architecture or the program steps can be used as general algorithms to solve problems within the skill domain. However, future researchers or developers can easily exchange text classifiers, ML models, or data sources for extracting information from other domains. It is also possible to reuse specific parts of the cycle. We want to contribute to further development in this area for Kompetanse Norge or anyone interested in this problem space.

4.4 Unused Methods

Some methods yielded unsatisfactory results in the experiment and development phase, which is why they are unused in the final implementation. However, conceptually they are viable, and we will discuss this in the results and discussion.

4.4.1 Translation

After cleaning the job listing text, we use the Azure Translation API service to get the job listing text back in different languages, e.g. English or Spanish. Since ESCO data contains many languages, we can use the translated job listing text to enrich our results using different languages to catch sublime differences between languages. This will be further explained in the Language Enrichment section.

This is how the text looks like after it has been translated from Norwegian to English.

nurse 69 75 nurse dementia unit is a special unit within nursing and care in the municipality the unit has a total of 51 nursing home places located in brumunddal and moelv with a total of 41 places as well as with 10 places located at sundheimen on the weekend island the wards are functionally divided into groups with their own mail kitchens and living areas in addition, the unit has a day centre in moelv and the brumunddal unit has 1 man-year as a professional consultant we are looking for 69 75 positions as a nurse social educator is vacant for accession as soon as possible the position is internship with work every 3 weekends and is currently associated with dept in moelv qualifications it is required authorisation as a nurse or social educator the nature of the work requires experience from working with people with severe dementia and a high degree of challenging and unpredictable behaviour the applicant must have experience with active use of environmental measures and shielding the applicant must have a safe and professionally anchored approach to the patients seeking must show a clear vision of their own contribution in the department seeking must have experience from teamwork and have a large degree of solution focus applying to reinforced sheltered unit must expect to work at the other departments of the dementia unit during periods as well as being able to supervise the staff in the care districts the nature of the work requires driver's license class b dispose of own car in the service as well as good physical health personal suitability will be given great weight second annual salary in 100 position nurse st code 7174 from NOK 354 000 to NOK 405 100 social educator st code 6455 from NOK 354 000 to NOK 405 100 we encourage you to submit an application electronically via the municipality's address www ringsaker municipality no during vacancies contact us and we are happy to help if you need help applying for electronic certificates and diplomas are included in the event of an interview in accordance with the Freedom of Information Act sect 25, information about the applicant may be exempted from the public applicant list if the applicant asks for the applicants who wish to have their name exempted from the public applicant list must provide justification if The reason for exemption is not sufficient, the applicant will be contacted position type permanent position contacting the head of the dementia unit application marked ref for complete call text and electronic application form

Figure 4.5: Listing 6835001 translated to English

Language Enrichment

The language enrichment lets us pick up additional skills from the job listing that was not found in the original language and enrichs the results. After the language has been checked against the ESCO data and matched, we save the skills from the different languages, but we also translate the newfound skills back to the original language to add to our matching results. The language enrichment lets us exploit ESCO's many languages and the sublime difference between said languages.

4.4.2 Skill NER

Using the ESCO skills, we are limited to the number of skills ESCO and ISCO have classified. Using POS and NER, we can look at the word type and its

entity to determine if it is a skill without a definitive data source. By using SpaCy or other NLP models, we get POS and NER capabilities. SpaCy lets us see what entity type a word has after generating a SpaCy Doc by passing the job listing text to the SpaCy NLP function. By iterating through every word, we can first filter out all words that are not relevant based on grammatical or entity category. A word in the text with the entity type of a location and person or a word smaller than two characters is not relevant. Grammatically we only allow nouns, verbs or proper nouns because we prefer the word categories that define skills or abilities. After we have filtered the words we lemmatise them.

The words, or skills, are then in their final form and does not need to be further matched.

5

Methodology

The main goal of the evaluation is to asses the implementation and the design choices. To evaluate the solution, we focus on two aspects; the quantitative evaluation will tell us about the extraction precision and accuracy regarding the ground truth data. The qualitative evaluation gives us a deeper understanding of why we get the results we do and provides some context of the different parts of the implementation and data sets used. Language implementations are hard to evaluate solely quantitatively as languages are ambiguous, so a qualitative evaluation is critical to see if the implementation quantitative results make sense. We can evaluate the similarity by studying the results from the algorithmic skill extraction versus the ground truth data.

5.1 Evaluation

The evaluation is done by combining the skill result from the matching method into one list. We fetch the ground truth skills for the duplicate listing and use set theory operations to get the quantitative results. We generate metrics on how the matching did regarding how many true positives, true negatives, false positives, and false negatives we have gotten when matching the skills.

5.2 Experimental Setup

To evaluate the solutions, we use the ground truth skills that Kompetanse Norge manually extracted from a sample of 100 job listings. By manually extracting the ground truth, we will have an exact number of skills to evaluate to get the quantitative result. The ground truth data also have the text from the said listing with the correlating listing id. Therefore, the ground truth data will be the sole data source for the experiments.

6

Results & Discussion

After using our matching algorithm, we have gotten a list of result objects. To understand the performance and result of the algorithm, we have generated some quantitative metrics from the results list. We also analyse and discuss the annotated skills and the algorithmic skills in a specific job listing to see how the results differentiate and why.

6.1 Quantitative Results

The results are represented with four different outcomes:

- True Positive(TP): The skills found in both the algorithmic and the annotated skills.
- True Negative(TN): The skills that are not found in the annotated nor algorithmic skills. This scenario does not apply to this experiment and will always be zero.
- False Positive(FP): The skills found by the algorithm but that are not present in the annotated skills.
- False Negatives(FN): The annotated skills not found in the text.

Label	Number of skills
True Positive	3
True Negative	0
False Positive	281
False Negative	2054

Table 6.1: Total amount of skills from experiment results

Label	Percentage
Accuracy	0,128
Precision	1,056
Recall	0,146
Specificity	0
F1 Score	0,256

Table 6.2: Result Metrics

As seen in the result figure 6.1, we have not achieved high accuracy or precision regarding the ground truth data.

From the 6.2 we can see that our accuracy and precision is low. Our F1 score is also quite close to zero, which indicate that the experiment is not robust.

6.2 Discussion

6.2.1 Result

Our quantitative results are not as expected, but evaluating languages quantitatively can lack nuance. Therefore, it is essential to analyse the results from a qualitative perspective to understand why the experiment yielded these results.

To understand our results, we will look at a listing and compare if the algorithmic results and the annotated ground truth data make sense when compared qualitatively rather than quantitatively.

6.2.2 Skill Similarity

As we know, a word or a sentence can have multiple meanings and can be interpreted in different ways. In our case, a sentence, or skill, can have the same meaning but can be described by different words. This makes literal or strict

string matching problematic since our algorithmic skills can have the same semantic meaning as the annotated skills, but be written differently.

Let us look at a nursing job listing with id 5050499. Figure 6.3 show us the result for this particular listing.

	Number of Skills
True Positive	2
True Negative	0
False Positive	16
False Negative	119

Table 6.3: Metrics for listing 5050499

There are 121 annotated skills and 18 algorithmic skills for this particular listing. They are categorised into categories that represent the annotated skills. The result is the 121 skills divided into nine categories, and there are a lot of identical or similar skills as we can see from table 6.4. We also want to look at the similarities between the algorithmic skills and the annotated skills to see if there are semantic matches between them that can explain the results.

We have defined semantic matches as words or sentences with the same meaning but that are worded differently or have partial string matching. For example, "miljø sykepleie" and sykepleier both have partial-string similarity with "sykepleier" and semantic similarity since they both are about nursing.

We will take a basis in the 18 algorithmic skills found by the algorithm and see if we find semantic matches in the annotated ground truth data.

Category	Annotated Skills
Data	beherske enkle dataprogrammer, datakunskaper, gode datakunnskaper, god ikt-kompetanse, digitale systemer, vedlikehold og utvikling av digitale hr systemer
Language & Communication	god muntlig og skriftlig fremstillingsevne, gode kommunikasjons og samarbeidsegenskaper,gode samarbeids- og kommunikasjonsevner, kommunisere godt med pasienter deres pårørende og dine kollegaer, gode kommunikasjonsevner, evne til å motivere og kommunisere med elevene, evne og vilje til formidling
Certifications	krever førerkort klasse b, førerkort klasse b, tilfredsstillende politiattest, barneomsorgsattest, medlemskap i den norske kirke, tilfredsstillende politi- og tuberkulinattest, godkjent politiattest, politiattest
Leadership	ledererfaring, lederstøtte, gode lederegenskaper, solid lederkompetanse erfaring, bred erfaring fra ledelse, initiativrik, ledelse, personalledelse, prosjektledelse, utdanning innen fagområdene hr personal ledelse arbeidsrett
Teamwork	evne til å arbeide selvstendig og i team, erfaring fra teamarbeid, arbeide selvstendig og i team med ansatte og frivillige, interesse for teamarbeid, trives med å jobbe i et tverrfaglig team, kunne etablere gode samarbeidsrelasjoner, gode samarbeidsevner, evne til samarbeid, evne til å samarbeide med elever foresatte ansatte
Personal Capabilities	ha en stor grad av løsningsfokus, god fysisk helse, dyktig kundebehandler, pålitelige, pliktoppfyllende, flink til å tilegne deg kunnskap, villig til å lære, fylte 20 år, ansvarsbevisst, flink til å motivere, engasjert, strukturert, fleksibel, ha respekt for den enkelte pasient, erviceinnstilt, personlig egnethet, personlig egnethet vektlegges, stort engasejment, selvstendig, løsningsorientert, ansvarsfull, evne til å planlegge egen tid, helhetlig tenkning, god gjennomføringsevne, solid kunnskap om fagområdet, gjennomføringsevne, integritet, evne til helhetlig tenkning, positiv fleksibel og løsningsorientert, ønske om å jobbe tverrfaglig, løse problemstillinger i tråd med pasientenes egne mål, signalkompetanse, godkjenning av signaltegninger, kompetanseoverføring, relevant kompetanse, blid og motivert, analytiske evner, interesser for fagområdet, sosial legning, systematisk og nøyaktig, stor arbeidskapasitet, serviceinnstilt, like å bygge relasjoner, selvstendig og initiativrik, pålitelig og strukturert, høy arbeidskapasitet, sosiale ferdigheter, forstå mennesker, leveringsdyktig, kunnskaper om lov og avtaleverk, egnethet for arbeid med barn og unge med psykiske lidelser, holde seg faglig oppdatert, evne til selvstendig arbeid
Education	miljøterapeut sykepleier, sykepleie, miljøterapeut sykepleier, vernepleie, teologi livssynsfag, sosionom, teknisk og eller merkantil utdanning, spesialpedagog allmenlærer, fastlege
Knowledge & Experience	erfaring fra å jobbe med personer med alvorlig demens, erfaring med aktiv bruk av miljøtiltak og skjerming, klinisk arbeid, forskning, generell kunnskap om jernbaneteknikk, kjennskap til jernbaneverkets regelverk, kunnskap om ulike typer sikringsanlegg, erfaring fra prosjektering, erfaring i bruk av tradisjonelle prosjekteringsverktøy, erfaring fra spesialisthelsetjenesten, kjennskap til erp-systemer, erfaring fra tilsvarende arbeid med ungdom, rekruttering, musikk og engelsk, tuberkulinprøve, diagnostisering, behandling av barn og unge, veiledning av lis, herunder sykefravær- og personaloppfølgning

Table 6.4: Categories of annotated skills for job listing 5050499

Algorithmic	Matching Annotated
organisasjonsstruktur	
sykepleiere	miljøterapeut sykepleie, sykepleie
ledere	ledererfaring, solid lederkompetanse erfaring, gode lederegenskaper
arbeidsrett	utdanning innen fagområdene hr personal ledelse arbeidsrett
sykepleier	miljøterapeut sykepleie, sykepleie
lederegenskap	ledererfaring, solid lederkompetanse erfaring, gode lederegenskaper
miljø	erfaring med aktiv bruk av miljøtiltak og skjerming
norsk	
utdanning	teknisk og eller merkantil utdanning, utdanning innen fagområdene hr personal ledelse arbeidsrett
arbeide selvstendig	evne til selvstendig arbeid
hjelpemiddel	
planlegge	evne til å planlegge egen tid
personalledelse	personalledelse
engelsk	musikk og engelsk
leder	ledererfaring, solid lederkompetanse erfaring, gode lederegenskaper
prosjektledelse	prosjektledelse
lederegenskaper	ledererfaring, solid lederkompetanse erfaring, gode lederegenskaper
teologi	teologi livssynsfag

Table 6.5: Listing skills with semantic similarities: listing 5050499

As we can see in table 6.5 many of the algorithmic and annotated skills are semantically similar or have partial-string similarity. This means that if our algorithm could evaluate semantic similarities between strings or words, we could better the accuracy. The listing had 13 duplicate skills, and the 6.4 table also shows that many strings are almost duplicates. An example is where a sentence has exchanged the sequence of words like "even til å arbeide selvstendig og i team" and "arbeide selvstendig og i team med ansatte og frivillige", which means that the person should be able to work independently and within a team setting. However, the words have been rearranged or removed/added. The annotated data is much more verbose and has in general more skills, similar or not, than the ESCO data for this listing.

6.2.3 ESCO Standard

Since ESCO tries to define and label as many skills as possible, the definitions must be unique, which leads to a very condensed skill definition. As we have mentioned before, most skills are defined with one to eight words. We see this in the ESCO data table 3.2 where each skill has a preferred label, the condensed version of the skill, versus the description, which is a more verbose explanation. This makes sense when we need to create short and digestible definitions, but language is ambiguous, and for machine interpretation, it becomes problematic when we can have so many semantic similarities. If we look at the 6.5 table, we can see that the algorithmic skill "leder" and "lederegenskaper" has three qualitatively associated annotated skills for the one word. These annotated skills have the same semantic meaning that the algorithmic skill has but is described in several ways in the job listing text. This could partly be the reason for the experiment results.

6.2.4 Ground Truth Data

The process of curating the ground truth data by annotating skills differentiate from the ESCO standardisation. The ESCO standard is developed by a committee of people who chose how to label skills from sentences or words and their respective synonyms. In comparison, a singular person chose and annotated the ground truth data in Kompetanse Norge. This means that the annotated data and ESCO data are defined on different criteria. The ground truth data will be susceptible to the biases of the one person annotating the ground truth data defines as a skill versus the committee choosing the ESCO standards. This will likely lower the probability of getting a match between the job listing and the ESCO data, leading to lower accuracy and precision.

As seen in the experiment metrics table 6.2, accuracy and precision are low. It is important to note that there are 6.7 times more annotated skills than algorithmic skills in the job listing table 6.5; this means that there is quite a discrepancy between the ratios of skills found in the job listing, which is reflected in the experiment result metrics. This will naturally skew the accuracy percentage because even if we matched all of the algorithmic and annotated skills, we would still have 103 annotated skills left. The reason might be that the ESCO data set is too condensed and therefore too limited for this evaluation and experiment as it cannot find all of the skills to give us a proper match. The reason can also be that our annotated data is too verbose. From the 6.5 and 6.4 table, we can see that several of the skills from the annotated data have semantic overlap with the algorithmic results. However, we can also see that the algorithm does not find essential and unique skills as listed in the 6.4 table. Therefore, we can assume that both our data sets are limited regarding the

needs of the evaluation.

'erfaring fra å jobbe med personer med alvorlig demens', 'erfaring med aktiv bruk av miljøtiltak og skjerming', 'erfaring fra teamarbeid', 'ha en stor grad av løsningsfokus', 'krever førerkort klasse b', 'god fysisk helse', 'personlig egnethet', 'dyktig kundebehandler', 'pålitelige', 'pliktoppfyllende', 'flink til å tilegne deg kunnskap', 'villig til å lære', 'beherske enkle dataprogrammer', 'fylte 20 år', 'initiativrik', 'ledelse', 'klinisk arbeid', 'forskning', 'helhetlig tenkning', 'god gjennomføringsevne', 'gode kommunikasjonsevner', 'kunne etablere gode samarbeidsrelasjoner', 'solid kunnskap om fagområdet', 'bred erfaring fra ledelse', 'gjennomføringsevne', 'integritet', 'evne til helhetlig tenkning', 'gode kommunikasjons og samarbeidsegenskaper', 'gode samarbeidsevner', 'positiv fleksibel og løsningsorientert', 'ønske om å jobbe tverrfaglig', 'løse problemstillinger i tråd med pasientenes egne mål', 'evne til å arbeide selvstendig og i team', 'datakunskaper', 'personlig egnethet vektlegges', 'signalkompetanse', 'godkjenning av signaltegninger', 'kompetanseoverføring', 'relevant kompetanse', 'generell kunnskap om jernbaneteknikk', 'kjennskap til jernbaneverkets regelverk', 'kunnskap om ulike typer sikringsanlegg', 'erfaring fra prosjektering', 'tilfredsstillende politiattest', 'barneomsorgsattest', 'medlemskap i den norske kirke', 'førerkort klasse b', 'disponerer bil', 'teologi livssynsfag', 'arbeide selvstendig og i team med ansatte og frivillige', 'blid og motivert', 'ledererfaring', 'personalledelse', 'prosjektledelse', 'formulere deg muntlig og skriftlig', 'analytiske evner', 'interesser for fagområdet', 'sosial legning', 'interesse for teamarbeid', 'erfaring i bruk av tradisjonelle prosjekteringsverktøy', 'fastlege', 'gode samarbeids- og kommunikasjonsevner', 'miljøterapeut sykepleier', 'evne til samarbeid', 'erfaring fra spesialisthelsetjenesten', 'vernepleie', 'sosionom', 'sykepleie', 'miljøterapeut sykepleier', 'gode kommunikasjonsevner', 'ansvarsbevisst', 'flink til å motivere', 'engasjert', 'strukturert', 'fleksibel', 'ha respekt for den enkelte pasient', 'serviceinnstilt', 'gode samarbeidsevner', 'trives med å jobbe i et tverrfaglig team', 'personlig egnethet', 'personlig egnethet vektlegges', 'gode datakunnskaper', 'stort engasejment', 'kommunisere godt med pasienter deres pårørende og dine kollegaer', 'selvstendig', 'løsningsorientert', 'fleksibel', 'gode samarbeidsevner', 'ansvarsfull', 'evne til å planlegge egen tid', 'teknisk og eller merkantil utdanning', 'gode datakunnskaper', 'kjennskap til erp-systemer', 'systematisk og nøyaktig', 'gode kommunikasjonsevner', 'stor arbeidskapasitet', 'gode samarbeidsevner', 'serviceinnstilt', 'evne til å motivere og kommunisere med elevene', 'erfaring fra tilsvarende arbeid med ungdom', 'god ikt-kompetanse', 'evne til å motivere og kommunisere med elevene', 'sosiale ferdigheter', 'evne til å samarbeide med elever foresatte ansatte', 'like å bygge relasjoner', 'selvstendig og initiativrik', 'pålitelig og strukturert', 'høy arbeidskapasitet', 'vedlikehold og utvikling av digitale hr systemer', 'lederstøtte', 'herunder sykefravær- og personaloppfølgning', 'rekruttering', 'digitale systemer', 'utdanning innen fagområdene hr personal ledelse arbeidsrett', 'forstå mennesker', 'leveringsdyktig', 'kunnskaper om lov og avtaleverk', 'god muntlig og skriftlig fremstillingsevne', 'tilfredsstillende politi- og tuberkulinattest', 'spesialpedagog allmenlærer', 'musikk og engelsk', 'godkjent politiattest', 'tuberkulinprøve', 'politiattest', 'diagnostisering', 'behandling av barn og unge', 'veiledning av lis', 'holde seg faglig oppdatert', 'egnethet for arbeid med barn og unge med psykiske lidelser', 'gode samarbeidsevner', 'evne til selvstendig arbeid', 'evne og vilje til formidling', 'solid lederkompetanse erfaring', 'gode lederegenskaper'

Figure 6.1: Annotated skills from listing 5050499

6.3 Beyond Data Limitations

As discussed in the qualitative analysis, the ESCO data will limit our skill matching capabilities. How do we solve the limitation of available data? There is a possibility to make selections in free text by using NLP techniques such as named entity recognition and part of speech tagging as mentioned in section 4.4.2. Combining these makes it possible for us to filter the text and words based solely on grammatical structures in language, which are always present. The technique's accuracy is dependent on machine or deep learning models to train on massive data sets for the language it is parsing and preferably using contextual domain data for higher accuracy. As mentioned in section 4.4.2 POS tells us what grammatical category a word is in, and depending on the ML model, it can give us more profound knowledge about the context the word has in relation to the rest of the sentence. By analysing how a skill appears grammatically in a text, we could extract it without the assistance of an explicit data source.

As mentioned in the introduction, a minor language like Norwegian has less support and interest than a more prominent language like English. For the English language, there are better NLP models for NER and POS. By having an NLP model for NER and POS with higher accuracy, we could have been less dependent on the data limitations of the ESCO data set, but developing these language models require expertise, large data sets and possibly a workforce. The National Library of Norway has released the "NoTraM - Norwegian Transformer Model", a Norwegian NLP model that boasts high accuracy. They have not yet published their results academically, but these could be implemented in future work.

6.4 Language Ambiguity

A commonality for our use cases and questions is how we can extract skills, educational level, occupation, or another essential category. From the qualitative analysis, the ESCO standard, ground truth data and the discussion of data limitations, we encounter the original problem of machine interpretation of language, which is language ambiguity. Using the Levenshtein distance algorithm, we can algorithmically find similarities between two words or texts. Levenshtein's algorithm looks at two words and calculates the difference between two words using single-character edits until they are the same. The calculation returns the distance, which are the necessary edits required to make the words similar. The resulting distance can be interpreted as a threshold between two texts or skills in our matching algorithm. This lets us account for spelling mistakes or word inflexion. However, there is no similar algorithm

that can give us semantic similarities between two sentences or texts. Having an algorithm that could compute the semantic difference or similarities between words or two texts would help us differentiate semantic duplicates. In theory, if we had a dictionary of synonyms for all words and a mapping of all popular phrases and sayings, we could compute all the possible permutations for a sentence and synonyms for a word and calculate a similarity score, just like Levenstheins distance, but for semantic similarities. However, a solution that could solve a severe problem like language ambiguity could effectively solve most NLP problems and seriously move the needle on general artificial intelligence, which is not the scope of this project. Therefore we will struggle with language ambiguity, where skills semantically mean the same but are differently defined until we have solved this problem.

6.4.1 Domain Specific - Structure in Free Text

Even though our domain is specified to the Norwegian job listing, they are spread across many different occupations and differently expressed texts. In the introduction, we mentioned the Azure Medical service [3] which analyses medical texts. The medical texts and reports have a precise vocabulary that limits the possible outcomes of words, synonyms, and other language expressions. This limitation can provide some structure or make it easier to map words and context to a specific disease or diagnosis.

For example, if we would have chosen a specific occupation like nursing in our data set, we would likely have a similar language across listings. We could have used NLP techniques like generating a domain-specific stop word list to filter out vocabulary that is not relevant from the nursing texts. If we combined this with the POS and NER capabilities from a high accuracy model like we discussed in section 6.3, we could get a higher accuracy when extracting skills from job listing texts.

6.5 Related Works

6.5.1 Colombo et. al

In Colombo et al. [16] the authors have implemented a classification system using ML with the ESCO/ISCO standards to classify occupations based on required skills found in Italian web job listings. The web job vacancies are structured with tagged HTML fields that let them extract skills explicitly. They also create a taxonomy for skills and map it to the ESCO standard. During their discoveries, they also analyse the importance of hard and soft skills listed in

job listings and their importance regarding automated occupation.

6.5.2 de Ridder

In de Ridders master thesis [17] the author implemented a text mining solution for Dutch job vacancies for ICT personnel. The author used NLP techniques that look at specific grammatical categories for filtering skills in free text using POS and NER. The author also developed a neural network to map the skill and grammatical categories and their combinations with n-grams in any given job listing. The results are shown in the figure **??** [17]. These results are good and show that POS and NER grammatical filtering are viable options, especially with even better NLP models.

Table 4-B. Results of neural network models on the test data: 100 unseen vacancies.

	Unigram	Bigram	Trigram	Tetragram	Pentagram	Hexagram
True Positive	2914	792	353	53	13	0
True negative	3069	1178	627	132	0	0
False positive	1086	380	96	3	0	0
False negative	182	149	132	73	20	7
Precision	73%	68%	79%	94%	100%	
Recall	94%	84%	73%	42%	40%	
F	82%	75%	76%	58%	57%	
Accuracy	83%	79%	81%	71%	40%	

Figure 6.2: de Ridder performance

/7

Conclusion

The thesis describes the use cases, requirements, data sets, implementation and design of our skill extracting algorithm. The result of our experiment of the developed skill extracting algorithm versus the ground truth data shows lower precision and accuracy than expected. We, therefore, conclude that our data limitations prevent us from solving the use case and requirements for extracting skills satisfactorily. The reason is that there is an explicit limitation between the ESCO and the ground truth data. These data sets are created in vastly different ways and under different criteria, which ultimately causes semantic differences that significantly impact the results.

To help with future work, we have suggested several ways to improve the results through different strategies. These strategies can be; utilising better Norwegian ML and NLP models, utilising larger skill data sets, producing more ground truth data and evaluating the algorithm qualitatively and quantitatively for a deeper understanding of use cases or problems for the data set.

As a novel prototype project, it is clear that there are several challenges and requirements as outlined in our introduction and analysis of the use cases, but this thesis makes several key contributions towards future solutions with;

- Processed and published data sets and open-source code.
- Analysis of said data sets, their use cases and requirements.

- The implementation and modular design enables reuse and modifiable code which can easily integrate suggested future work ideas.

- Analysis and discussion of the data sets and the result of the implemented algorithm.

References

[1] Mika Pajarinen, Petri Rouvinen, and Anders Ekeland. Computerization threatens one-third of finnish and norwegian employment. *ETLA Brief No 34,* 04 2015.

[2] Carl Benedikt Frey and Michael A Osborne. The future of employment: How susceptible are jobs to computerisation? *Technological forecasting and social change*, 114:254–280, 2017.

[3] Microsoft Azure. Azure cognitive servies. Available at `https://docs.microsoft.com/en-us/azure/cognitive-services/text-analytics/how-tos/text-analytics-for-health?tabs=ner`.

[4] Rohan Ramanath, Hakan Inan, Gungor Polatkan, Bo Hu, Qi Guo, Cagri Ozcaglar, Xianren Wu, Krishnaram Kenthapadi, and Sahin Cem Geyik. Towards deep and representation learning for talent search at linkedin. *https://arxiv.org/corr*, abs/1809.06473, 2018.

[5] Sahin Cem Geyik, Qi Guo, Bo Hu, Cagri Ozcaglar, Ketan Thakkar, Xianren Wu, and Krishnaram Kenthapadi. Talent search and recommendation systems at linkedin: Practical challenges and lessons learned. *https://arxiv.org/corr*, abs/1809.06481, 2018.

[6] Sahin Cem Geyik, Vijay Dialani, Meng Meng, and Ryan Smith. In-session personalization for talent search. *https://arxiv.org/corr*, abs/1809.06488, 2018.

[7] Viet Ha-Thuc and Shakti Sinha. Learning to rank personalized search results in professional networks. *https://arxiv.org/corr*, abs/1605.04624, 2016.

[8] Viet Ha-Thuc, Ganesh Venkataraman, Mario Rodriguez, Shakti Sinha, Senthil Sundaram, and Lin Guo. Personalized expertise search at linkedin. *https://arxiv.org/corr*, abs/1602.04572, 2016.

[9] O*NET. onet content model. Available at https://www.onetcenter.org/content.html.

[10] O*NET. onet taxonomy. Available at https://www.onetcenter.org/taxonomy.html.

[11] ISCO. Isco taxonomy. Available at https://i2.wp.com/r-posts.com/wp-content/uploads/2020/07/ESCO_ISCO_hierarchy.png?resize=590%2C299.

[12] ESCO. Esco landscape. Available at https://i.postimg.cc/1RHhzdWw/1-Landscape-ESCO.png.

[13] NAV. Nus2000 dokumentasjonsrapport. Available at https://www.ssb.no/utdanning/_attachment/94898?_ts=13cb49d6358.

[14] NAV. Standard for yrkesklassifisering (styrk-08). Available at https://www.ssb.no/a/publikasjoner/pdf/notat_201117/notat_201117.pdf.

[15] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union, 1966.

[16] Emilio Colombo, Fabio Mercorio, and Mario Mezzanzanica. Ai meets labor market: Exploring the link between automation and skills. *Information Economics and Policy*, 47:27–37, 2019.

[17] EA Eline de Ridder. Analysis of skills needed for ict employees in the netherlands.