UNIVERSITY OF TWENTE.

Faculty of Electrical Engineering,
Mathematics & Computer Science

Ontology Matching using Background Knowledge and Semantic Similarity

Sathvik Guru Rao

M.Sc. Thesis

April 2022

Company Supervisors:

Linda Oosterheert
Roos Bakker

TNO
Kampweg 55
3769 DE Soesterberg

University Supervisors:

Dr. Shenghui Wang
Dr. Nicola Strisciuglio
Dr. Mariet Theune

Faculty of Electrical Engineering,
Mathematics and Computer Science
University of Twente

P.O. Box 217
7500 AE Enschede
The Netherlands

Abstract

Ontology matching (OM) aims at combining the ontologies created by different organizations of the same domain to help in information exchange. OM has been a topic of research and development for a long time. The latest techniques of Artificial Intelligence (AI) like Natural Language Processing (NLP) and Machine Learning (ML) are used to find a better match for the entities of the ontology. However, the labor market hasn't explored ontology matching using these techniques yet. This thesis focuses on the matching of occupations between two occupational ontologies called ESCO and O*NET, which represent the European and American occupations, respectively, by leveraging semantic similarity between the occupations calculated using a language model called XLNet.

In this thesis, we worked on techniques to improve the matching process by incorporating domain-specific knowledge. A state-of-the-art language model - XLNet is used to create contextual embeddings of the occupations' information which is then used to find the semantic similarity between the occupations of ESCO and O*NET. This is used as a baseline to understand the impact of using domain-specific knowledge in the matching process. Domain-specific knowledge is used in two ways: *i) by extending the XLNet model's vocabulary with domain knowledge* and *ii) by using a domain-specific ontology as a helper ontology which can bridge the gap between the two ontologies*. The number of occupations specified by each ontology varied greatly which resulted in the issue of many-to-one matching. This motivated the next stage, which was to establish a semantic relationship between the occupations to make the relationship between them more informative. These relations are based on the taxonomic structure of ESCO and the semantic similarity score between occupations.

Due to the lack of ground truth matching, our research results were evaluated by a domain expert. The accuracy was calculated based on the matches with highest semantic similarity between the occupations. The generic XLNet model performed well in finding an accurate O*NET occupation match with an accuracy of 69% for a sample of 200 ESCO occupations. The matching method using XLNet model with extended vocabulary and matching with the help of domain-specific ontology has an accuracy of 59% and 34% respectively. A detailed analysis of the relationships established between the occupations provided an insight into where the O*NET occupation can be positioned inside ESCO occupation categories which can facilitate in ontology merging.

Acknowledgment

Firstly, I would like to thank God for blessing me and giving me the strength to overcome the tough times and finish this thesis. Then, I would like to thank TNO for giving the opportunity to work in the Skills Matching project which gave me a deeper understanding of the labor marketing. I would like to thank Linda and Roos for their support and feedback throughout this project. They have helped academically and personally when I lacked motivation and guided me in the right direction. I would also like to thank my university supervisors, Shenghui, Mariet and Nicola for giving me honest feedbacks which helped me improve. Finally, I would like to thank my family and friends who have been with me and supported me with all my decisions.

Contents

Chapter 1

Introduction

The labor market is constantly evolving and new job vacancies are opening daily. Countries and organizations have been using classification systems to standardize occupations and occupational groups. An occupational classification has three main purposes which are to help with statistical data collecting, labor market analysis, and career planning and job search [1]. Occupational needs as well as occcupation-specific skill requirements are stored in these classification systems. These classifications are represented in a machine-readable manner using an ontology. The ontologies are typically represented as concepts, attributes, and relationships which give information about their meaning and their relations [2]. These occupational ontologies are employed in a variety of applications such as searching, matching, and analytic tools which use Natural Language Processing (NLP) and Machine Learning (ML) methods. Rentzsch et al. [3] have described some of the applications of occupational ontologies as well as continuing efforts such as the SkillLab, an EU Skills Profile Tool for Third Country Nationals.

When job seekers desire to change careers, they need the current trends and job information to make judgments, and a labor market ontology can assist them in this case. Unfortunately, this data does not come from a single source or format. Each country has a unique labor market that defines occupations from a specific point of view. Moreover, job internationalization is on the rise [4] which makes it harder to capture all of the information collected. Aligning the ontologies of various countries can aid in the exchange of information between them.

The International Standard Classification of Occupations (ISCO) and the Standard Occupational Classification (SOC) are the two primary labor market ontologies. While ISCO classifies the majority of broad categories depending on skill level, SOC classifies based on the tasks performed, independent of the degree of education required [5]. Most countries have their occupational ontologies based on either ISCO or SOC. For example, the United States has O*NET, Canada has the National Occupation Classification (NOC), both of which are based on SOC, and the European Union has the ESCO occupation ontology, which is based on ISCO. Even though the ontologies are in the same domain - occupations - they differ in terms of label, description, and skill requirements. The alignment between two of these ontologies, ESCO and O*NET, is the topic of this thesis.

1.1 Motivation

This thesis project falls under the Skills Matching project 2.0[1], which is in cooperation between TNO (Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek, English: Netherlands Organization for Applied Scientific Research), UWV (Uitvoeringsinstituut Werknemersverzekeringen, English: Employee Insurance Agency), CPB (Centraal Planbureau, English: Bureau for Economic Policy Analysis), and CBS (Centraal Bureau voor de Statistiek, English: Statistics Netherlands). The primary focus of the project is to make a move towards matching jobs based on skills instead of qualifications. The Dutch labor market is trying to find a balance between the demand and supply of labor. Employees, often feel they are not making full use of their skill-set in the ongoing jobs. Employers on the other hand, are unable to find the right employees. With the advances in technology these days in terms of automation and digitalization, these mismatches in the job sector are only increasing. It is crucial to find a way in which employees are matched to the right job. This can be achieved by matching employees to jobs on the basis of their skills instead of qualification. The aim of this project is to investigate how different existing occupational ontologies relate to each other and the possibility to align them. The ESCO and O*NET occupational ontologies work best for this thesis because of the following reasons :

- The UWV developed CompetentNL[2] in collaboration with CBS and TNO to define the Dutch labor market. The goal of the public version of CompetentNL, according to UWV, is to be able to link it to ESCO and O*NET.
- The ESCO ontology classifies occupations by sector, while O*NET defines occupations based on the work activities performed by workers. The alignment of these two ontologies can aid in identifying occupations with specific skills that can be applied in other sectors, which is a major focus of the Skills Matching Project.

[1] https://tinyurl.com/2svkjacs

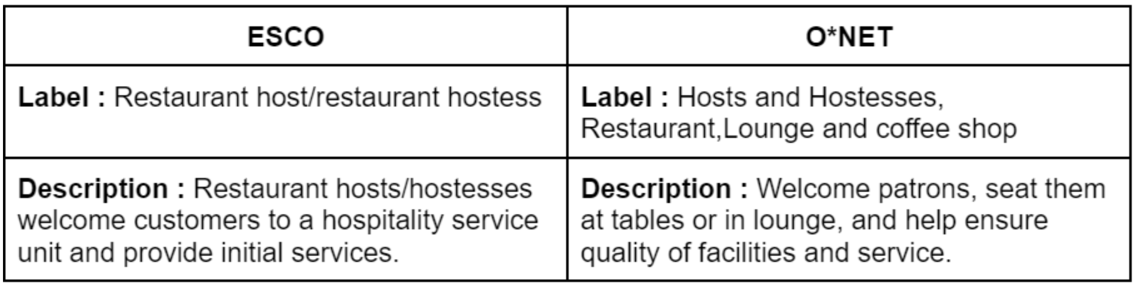
[2] https://www.werk.nl/arbeidsmarktinformatie/skills/competentnl-standaard-voor-skills-in-nederland

- The alignment of the ESCO and O*NET ontologies, which describe occupations in the European Union and the American labor market, can help internationalize and close the gap between countries. This will make it easier for job seekers to search for work on the other side of the border.

Therefore, this thesis focuses on the mapping between the ESCO and O*NET by establishing a relation between occupations.

1.2 Problem Statement

The primary goal of this project is to find which occupations in ESCO corresponds to which occupations in O*NET and establish a relationship between them. This is also known as a *crosswalk* which connects the occupation of one ontology to the occupation of another ontology. Some countries publish official crosswalks to other systems of different countries or between older and newer versions but, the crosswalk between ESCO and O*NET is not developed officially by O*NET or ESCO. The crosswalks that currently exists are discussed in section 3.4. ESCO and O*NET describes the labour market domain by structuring, and describing occupations and the required skills with the help of labels and descriptions in natural language. While comparing these two ontologies, three types of heterogeneity were encountered. First, Syntactic heterogeneity, because ESCO is defined in SKOS[3] format and O*NET is a database [6]. Second, Conceptual Heterogeneity, because ESCO classifies the occupations based on sector and O*NET classifies the occupations based on work activities. Also, ESCO has more granular level of information when compared to O*NET. For example, the occupation `Lawyer` is narrowed down to `Corporate Lawyer` in ESCO but O*NET defines only one `Lawyers` occupation. Third, Terminological heterogeneity, because the titles of the occupations in ESCO and O*NET differ even though they are defining the same occupation. An example is given in the table below. As we know that they are



ESCO	O*NET
Label : Restaurant host/restaurant hostess	**Label :** Hosts and Hostesses, Restaurant,Lounge and coffee shop
Description : Restaurant hosts/hostesses welcome customers to a hospitality service unit and provide initial services.	**Description :** Welcome patrons, seat them at tables or in lounge, and help ensure quality of facilities and service.

Figure 1.1: ESCO and O*NET occupation

[3] https://www.w3.org/TR/skos-primer/

from different countries, it is built from different perspectives and have different structures. An alignment between the concepts establishes a relation between the occupations. Neutel et al. [6] and Kanders et al. [7] used text similarity between occupations' descriptions to find a match for the concepts of ESCO occupation. These crosswalks of ESCO-O*NET has shown that the manual work can be reduced using automatic matching with the help of semantic similarity. Some of the limitations of these crosswalks are:

1. Occupations considered: Kanders et al. [7] has selected occupations which are directly linked to the ISCO unit groups. These are the occupations of level 5, further granular level occupations are not considered . Neutel et al. [6] has used the most specific occupations that do not have any children occupation. This means that not all the occupations defined by ESCO are matched to O*NET in either of these crosswalks.

2. Mapping Relations: The crosswalk developed by [7] finds exact matches of occupations with a confidence level but did not establish any other relation. The ESCO-O*NET alignment developed by [6] also found exact matches and other relations were found manually.

1.3 Research Goals and Questions

Given the limitations of the current crosswalk between ESCO and O*NET ontologies, it would be interesting to integrate Natural Language Processing (NLP) developments such as using a domain-specific language model with ontology matching techniques that can aid in improving the matching between ESCO and O*NET and, ideally, find more correct matches. These innovations in NLP are explored in detail in chapter 3.

Some difficulties were encountered while searching for a match for each ESCO occupation. The number of occupations defined by the ESCO and O*NET ontologies differs significantly. This difference limits the possibility of identifying a one-to-one precise match for each of the 2942 ESCO occupations, and it was also discovered that more number of ESCO occupations were matched with the highest semantic similarity to one O*NET occupation resulting in many-to-one matching. As a result, the objective was set to establish a relationship for each match that could provide more information about the match than just whether it was correct or incorrect. Considering the conceptual heterogeneity, the relationship can provide more information about how the occupations are related instead of finding one-to-one matching. Also, ontology merging and/or evolution can be aided by establishing a more informative semantic relationship.

Based on the limitations described in section 1.2 and the goals defined previously, the research questions for this thesis are as follows :

RQ1: *How can we improve matching between ESCO and O*NET occupations using domain-specific background knowledge?*
This question can be answered by answering the following sub-questions.

(a) How does the generic XLNet language model perform in finding an O*NET occupation match to the ESCO occupation?

(b) How does ontology matching using domain-specific language model perform against ontology matching using generic language model trained on general data?

(c) How does using a domain-specific background knowledge as a helper ontology in ontology matching compare to ontology matching using direct semantic similarity between the occupations ?

RQ2: *How does using various metadata like skills and alternate label related to the occupations improve matching between ESCO and O*NET occupation?*

RQ3: *How can we establish different types relations between ESCO and O*NET occupations using semantic similarity and taxonomic structure of ontology?*

(a) How can the taxonomic structure of the target ontology(ESCO) be used in establishing different types of relations between the occupations?

(b) How does semantic similarity score support in establishing different types of relations between occupations?

The research question RQ1 and the derived sub-questions will answer the effect of using domain-specific background knowledge in two different ways. The results of **RQ1.a** is used as the baseline result with which the results of **RQ1.b** and **RQ1.c** are compared. **RQ2** is another attempt in improving the matching between ESCO and O*NET occupations by considering more information related to the occupations. The **RQ3.a and b** answer the usage of semantic similarity score and taxonomical structure of the ESCO in establishing a relation.

1.4 Structure of Thesis

This section includes a broad overview of the report as well as a reading guide. Chapter 2 helps in getting to know the concepts that are used in this project. Chap-

ter 3 covers the related works from other authors. It begins with an overview of current ontology matching techniques in general, then discusses the existing systems in the methods that are used in this project, and moves on to the ontology matching techniques that exist specifically in the labor market.
Chapter 4 explains the content of ESCO and O*NET ontologies that is used in chapter 5 - methodology. This chapter details all the methods that are used to answers the research questions. The methods include the process of matching occupations from ESCO and O*NET and then the method used to establish a relation.
Chapter 6 presents the experimental setup that is carried out in order to answer the research questions. In addition, the evaluation metrics are also discussed in this chapter.
Chapter 7 gives the results of the experiments and a detail analysis of the results and the relations that are established are discussed with the support of examples.
Finally, in chapter 8, the research is concluded by revisiting the research questions and the results. It also outlines the limitations and possible improvements for future works.

Chapter 2

Background

This chapter presents the background knowledge that is necessary to understand the concepts of ontology matching using semantic similarity and the semantic relations between the concepts of the ontologies. The chapter starts with a brief introduction to ontology and different ontology matching techniques. It is then followed by brief introduction to NLP which is used in ontology matching.

2.1 Ontology and Ontology Matching

An ontology is the representation of a domain of knowledge. Ontology is commonly used in the fields of information science and Artificial Intelligence to refer to a machine-readable representation of a concept in the real-world [8] that is used to model knowledge about individuals, their attributes, and their relationships to other individuals [2].

Ontologies have become widely used on the Internet. Ontologies on the Internet span from huge taxonomies that categorize Web pages to categorizations of items and their characteristics, for example to store and classify medical records. Ontologies are created with a variety of tools and information at various levels of details. Many ontologies have been built in domains that generate a huge amount of data. With the increasing number of ontologies within a domain, there is a problem of interoperability that can be addressed by ontology matching. Ontology-based system designers are frequently required to integrate many ontologies, either to enforce reuse and avoid having multiple ontologies of the same topic or to interconnect numerous relevant ontologies. This raises the issue of heterogeneity. According to Euzenat et al. [9], the types of heterogeneities are:

syntactic heterogeneity: This type of heterogeneity occurs when the ontologies are expressed in different languages or use different types of formal logic. This can be resolved by translating into a common language or by finding equivalence.

Terminological heterogeneity: This type of heterogeneity occurs when the ontologies refer to the same concepts in different names which is due to the usage of different natural languages.
Pragmatic heterogeneity: This type of heterogeneity occurs when the entities are interpreted differently by different people due to the context. This heterogeneity is difficult to resolve by computers.
Conceptual heterogeneity: This type of heterogeneity occurs when the domain of the ontology is built different with respect to coverage or granularity or perspective.

Ontology heterogeneity is the primary step to take in order to achieve interoperability. This can be done with Ontology Matching. The definition of ontology matching changes with different authors. In this thesis, the definition of Euzenat et al. [9] is used, which says that Ontology Matching is the process of finding a relationship between the entities of two ontologies. Ontology matching has 4 main purposes :
1. Achieving interoperability between different ontologies.
2. Sharing knowledge in more granular level.
3. Getting information more flexibly.
4. Obtaining a global ontology having different purposes.
Euzenat et al. [9] divided the ontology matching techniques into two categories:

I. Element-Level Techniques
These techniques consider the ontology entities in isolation and then compare two entities. The different techniques are
1.String-based: This technique considers the name and description of entities as a sequence of letters and compares the strings.
2.Language-based: This is based on NLP techniques to extract the meaning from texts and then compare the concepts.
3.Constraint-based: This technique deals with the internal constraints applied to the entities such as types, the cardinality of the attributes and keys.
4.Informal Resource-based: This technique uses external resources to deduce the relations between the entities of ontologies based on how they are related to the informal resources that are tied to ontologies.
5.Formal Resource-based: This technique uses external ontologies like domain-specific ontology, Linked data and other resources.

II. Structure-Level techniques
These techniques consider the entities and their relations to compare with entities of other ontology.The different techniques of structure-level are :

1.Graph-based: This technique considers the input ontologies as labeled graphs and the similarity is calculated between the pair of nodes from each input ontology and their positions.
2.Taxonomy-based: This technique is similar to the Graph-based technique but considers only the specialization relations. Specialisation connects concepts that are already similar (being understood as a subset or superset of each other), their neighbors may be similarly related as well.
3.Model-based: This technique is also called semantically grounded which considers the semantic interpretation behind the entities. If the two entities are the same, then they will share the same interpretations.
4.Instance-based: This technique considers instances of classes and compares if the classes are similar or not. This is based on set theoretic reasoning or statistical techniques which group the items together.
In the next section, the different relations that can be used to define a match is explained.

2.2 Semantic Relations

The semantic relation types that are used in this thesis are defined by the Simple Knowledge Organization System (SKOS) [10]. SKOS is a data model that allows classification systems to be shared and linked. One of the reasons that the SKOS semantic relations are used here is to retain the relation from the target ontology i.e, the ESCO which is also defined in the SKOS format. This makes it easy to evaluate and reuse when matching with different classifications. These relations are inherent in the meaning of the related concepts of the two ontologies.
The SKOS identifies two types of semantic relations: *hierarchical* and *associative*. A hierarchical relationship between two concepts denotes that one is more general (*broader*) than the other (*narrower*). An associative relationship between two concepts implies that they are naturally ”related”, but that one is not more general than the other [10]. The relations are used to demonstrate that two concepts from distinct schemas have comparable meanings despite the fact that they are modeled using different principles.
There are five mapping properties namely, `skos:closeMatch`, `skos:exactMatch`, `skos:broadMatch`, `skos:narrowMatch` and `skos:relatedMatch`. The properties `skos:broadMatch` and `skos:narrowMatch` are used when specifying hierarchical relations and the property `skos:relatedMatch` is used to state associative mapping relation between two concepts. `skos:broadMatch` and `skos:narrowMatch` are the sub-properties of the semantic relations `skos:broader` and `skos:narrower` respectively.

`skos:closeMatch` : This type of relation is used when two concepts are sufficiently similar and can be used interchangeably in some information retrieval applications. This relation cannot be used as a transitive property.
`skos:exactMatch`: This is a sub category of `skos:closeMatch` which indicates a higher degree of confidence compared to `skos:closeMatch` and can be used interchangeably in a wide range of information retrieval applications.

2.3 Natural Language Processing

As the ontology's concepts are described in natural language text, different ontologies use different terms in representing the information which lead to ambiguity and NLP can be used to address this problem. Word embeddings is the vector representation of a sequence of words in a vector space. These vectors represent the sequence of words by capturing the syntactic and semantic regularities. Static word embedding models like Word2Vec [11], GloVe [12], and fastText [13] generate context-free word embeddings, which represent a word without taking into account its context. Contextual word embeddings, on the other hand, depict words by taking into account the words that surround it. The semantics of words in different contexts are captured by contextual word embeddings.
Language Modeling (LM) is the process of finding the probability of a word occurring given the sequence of words. In this way, the model is able to learn the information from the entire sentence. The learning of the words can be unidirectional or bidirectional ways. ELMo [14], ULMFiT [15] learns the word representation by parsing the sequence of words from both directions with the help of underlying Recurrent Neural Networks (RNN) architecture. These models could not capture long-term dependencies, so the Long Short Term Memory (LSTM) model was used to capture long-term dependencies. Although these model helped in capturing the context of long sequences, they were slow as they were fed word-by-word to the model and very hard to paralellize. Vaswani et al. [16] addressed this problem by developing transformers model. This is explained in detail in the next section.

2.3.1 Transformer Models

In 2017, Vaswani et al. [16] developed the transformers model which uses an encoder-decoder architecture. The Transformer's key characteristic is that it employs attention, a concept that aids in capturing relationships between words in a sentence and has significantly reduced the training time. The transformers architecture is given in figure 2.1, which consists of *encoder* and *decoder*.

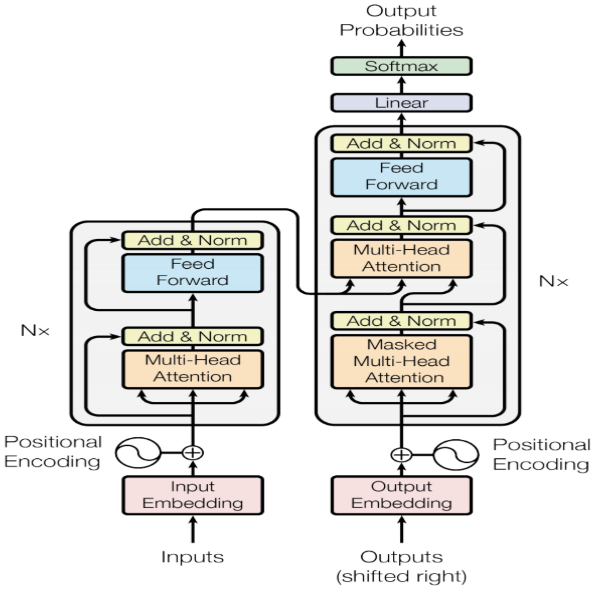


Figure 2.1: Transformer architecture, source : [16]

Encoder: The Encoder block as represented in the left side of figure 2.1 has two layers, namely, multi-head attention layer and feed forward layer. The transformers model consist of a stack of encoders. The bottom encoder takes the vector embeddings of input sequence as input and all other encoders take the output of the below encoders. The vector size of these inputs and output does not change and is the hyperparameter we can set. As the attention layers takes the input it encodes each word of the sentence considering other words of the sentence. The output of this layer is normalized and fed to the feed-forward layer.

Decoder: The decoder takes the input from the encoder. In addition to the layers used by the encoder, it has a multi-head attention layer. The decoder predicts the next word with the help of the encoder output and output the sequence it has predicted so far. The masked attention layers helps in focusing on one position and restricts attention on subsequent positions. This ensures that the prediction for a position depends only on the sequence which lies before it.
BERT [17] was a revolution in NLP and contextual word representation methods. BERT is an autoencoder(AE) language model which uses the transformer model's design to learn the context of a word by scanning the text from both directions. This model was pre-trained on large corpus in unsupervised manner with two goals, i.e., Masked Language Modeling and Next Sentence Prediction. In the pre-training

phase, a token in the input text is replaced with `[MASK]` and the goal of AE model is to reconstruct the original text. The benefit of using this model is that the context is evaluated from both forward and backward direction. The disadvantage of AE model is that `[MASK]` symbols used in pretraining phase. When finetuning, the `[MASK]` are absent in raw data which results in pretrain-finetune discrepancy. Another disadvantage of using `[MASK]` is that the predicted token is independent of other unmasked tokens. These disadvantages are solved by the XLNet model [18] which is the new state-of-the-art model which outperformed BERT model.

2.3.2 XLNet Model

XLNet is an AutoRegressive (AR) language model that achieved state-of-the-art performance in the standard NLP tasks that comprise the GLUE benchmark [19]. Like other language models, the first component is the word embeddings which is a fixed length vector that is fed to the language model. The vector is created by assigning ids to the tokens of the input sequence. The XLNet model uses the SentencePiece tokenizer [20] which tokenizes the input text which is then fed to the language model. The AR language model uses the context of a word to predict the next word. A disadvantage of this model is that, it is constrained to either forward or backward direction. XLNet model proposed a new method in which the AR language model learns from bi-directional context. This is called the **Permutation Language Modeling**. In the following sections, the XLNet tokenizer and model is explained in detail.

2.3.2.1 XLNet Tokenizer

The XLNet tokenizer is based on SentencePiece [20]. SentencePiece is a language independent subword tokenizer which is built specifically for neural network language models and text processing. The size of the vocabulary of the tokenizer is pre-determined in the pre-training phase. XLNet model has a vocabulary of size 30 thousand tokens. SentencePiece is implemented using two algorithms namely, Byte-Pair Encoder and Unigram Language model. It comprises of four components - *Normalizer*, *Encoder*,*Decoder* and *Trainer*. The Encoder executes the normalizer internally which normalizes the semantically equivalent Unicode into Canonical forms. The normalized input text is tokenized into subwords. *Encoder* and *Decoder* are the tokenizing and detokenizing methods which are used to convert the tokens to id mapping and vice versa. *Decoder* is the inverse of *Encoder*.

$$\textit{Decode(Encode(Normalize(Text)))} = \textit{Normalize(text)}$$

The SentencePiece tokenizer follows a lossless tokenization, which is achieved by considering the text information as a sequence of Unicode characters including the white spaces between words. The white space is first escaped using a meta symbol _(U+2581). This method makes the detokenizing easy by using a joining the token and replacing the meta symbol with space. Using python code, the conversion can be done by

```
detok = ‘’.join(tokens).replace(‘_’,‘ ’)
```

2.3.2.2 Permutation Language Modelling

The XLNet language model's learning objective is to learn the conditional distribution for all permutation of tokens in a sequence. For an input sequence, the AR model calculates the probability of a token given the condition of all the previous tokens present before. The authors of XLNet improvised this by using permutation. Using the permutation operations, the context for each token is considered from all the positions and each position learns to utilize contextual information. For a sequence x of length *N*, there are *N!* different order to perform a valid AR factorization which lets the model to gather information from all positions on both sides [18].

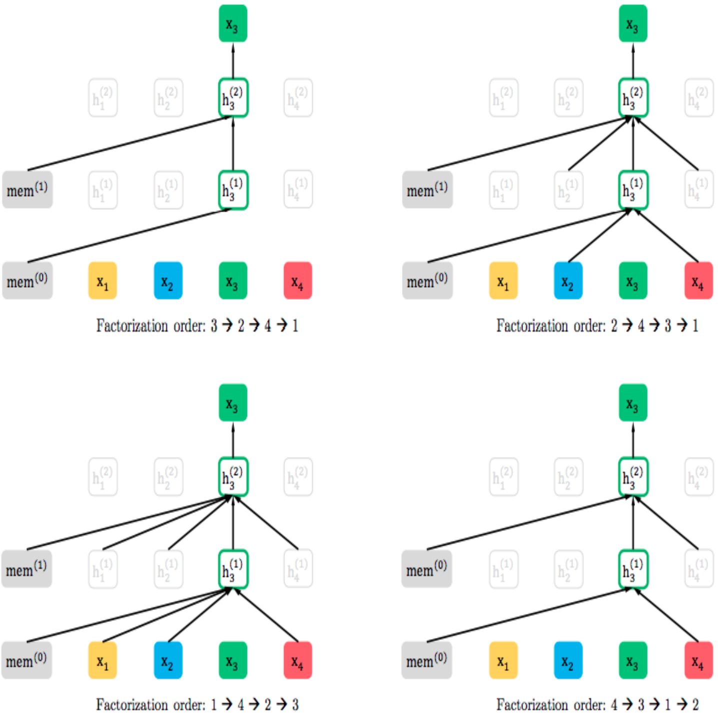

x_3
$h_1^{(2)}$ $h_2^{(2)}$ $h_3^{(2)}$ $h_4^{(2)}$
mem$^{(1)}$ $h_1^{(1)}$ $h_2^{(1)}$ $h_3^{(1)}$ $h_4^{(1)}$
mem$^{(0)}$ x_1 x_2 x_3 x_4
Factorization order: 3 → 2 → 4 → 1
Factorization order: 2 → 4 → 3 → 1
Factorization order: 1 → 4 → 2 → 3
Factorization order: 4 → 3 → 1 → 2

Figure 2.2: source : XLNet: Generalized Autoregressive Pretraining for Language Understanding [18]

In this thesis, the XLNet model is used in calculating the semantic similarity between the occupations of ESCO and O*NET. This is because,

1. It is the current state-of-the-art model which has outperformed BERT model.

2. The BERT model has been used in previous works of matching ESCO and O*NET and based on our knowledge, this is the first attempt to use XLNet model in the process of matching ESCO and O*NET occupations.

2.4 Semantic Similarity and Semantic Similarity Score

Finding the Semantic textual similarity (STS) between two text data is one of the important tasks in Natural Language Processing. Semantics is the meaning behind natural language text, and semantic similarity is the closeness of two texts based on their meaning. To calculate the semantic similarity between two text data, a vector representation of the text is produced using language models. The semantic similarity score is calculated using the cosine distance between these two vectors. Cosine similarity is one of the methods that can be used to find the similarity between two texts. The cosine angle between the two vectors tells us about the similarity which ranges between 0 and 1. *cos (90)* = 0, which says that if the angle is 90, then the vectors are far apart and the texts are dissimilar. If the angle is 0, then *cos(0)* = 1,which says that the vectors are overlapping and the texts are equal. This shows that the greater the angle between the vectors, the more dissimilar the texts are, and vice versa.

2.5 Chapter Summary

In this chapter, we explained the background knowledge related to the thesis. First, an ontology is explained which is how the data that is used in this thesis is represented. Then, ontology matching and the different techniques are briefly explained. For this thesis the language-based technique and resource-based methods are of particular interest to obtain the matching between ontologies. The occupations in the two ontologies are described in natural language texts and comparing these texts determines the similarity between the occupations. The reasons to select these two methods are further discussed in chapter 3. The NLP technique to extract the meaning behind text and find the similarity between different texts is called semantic similarity which is explained in section 2.4. In addition, the semantic relations that should be established between the occupations of the ontologies are also discussed. In the next chapter, the existing works in ontology matching from different authors are discussed. This includes the related works of ontology matching in general and then in the labor market domain specifically.

Chapter 3

Related Work

This chapter contains a review of the literature and related studies. The chapter begins with a quick review of the literature on ontology matching before delving into works relating to the two main types of ontology matching. These utilize “background knowledge” and “semantic similarity”. The following section discusses how the existing occupational ontologies are matched.

3.1 Ontology Matching

Ontology matching is a challenging task. Ramar et al. [21] explored the different mapping techniques that are discussed in section 2.1 and provided a comparative summary which showed that most of the techniques that exist are based on lexical and structural methods. The authors also pointed out that these methods cannot handle contextual difference between the ontologies. The text used in the ontologies can have ambiguous meanings which depends on the context. This makes it difficult to match based on labels and terms. To solve the problem of ambiguity, the context of text can be considered with the help of background knowledge [22]. Recently, Liu et al. [23] summarized and analyzed the existing state-of-the-art methods in ontology matching. These methods are based on the three types of information available in the ontologies namely, lexical information, structural information, and semantic information which correspond to string-based, structure-based, and language-based matching. For each category of information, the current approaches and challenges are mentioned. The authors have pointed out that the semantic information based methods are more effective compared to structure-based and string-based methods as they focus mainly on the surface level similarity and fail to capture the meaning behind the entities. This can be addressed using NLP techniques like word embeddings and using external sources for extending the meaning. These two methods are further discussed in the following sections.

3.2 Ontology Matching using Background Knowledge

Ontology matching with the help of background knowledge is helpful when the elements of two ontologies are related but are not supported by lexical similarity or the structure of ontologies. A background knowledge can be used to provide a path between the source and target ontologies that are to be aligned or it can be used to enrich the concepts of ontologies with more information. Using a larger and detailed ontology of the same domain as background knowledge has improved the matching process and it can be maximized by combining different knowledge from different sources [24]. The background knowledge used is selected manually and it should be processed before it is used in the matching process. To overcome this problem, *Swoogle*, an online ontology searching tool which searches for ontologies in the same domain as the input ontology can be used [25]. The background knowledge can also be used to enrich the entities with more context information. The enriched ontologies can then be used to calculate the similarity between concepts [26] [27].
Husein et al. [28] gives an overview of ontology mapping using background knowledge. The process of using background knowledge consists of 2 steps: anchoring and Inferencing. Anchoring is the process of finding an anchor between the ontologies and the background knowledge while inferencing uses the relations that already exist between the anchored concepts and the ontologies. Synonyms, lexicons can be used as a reference background. Some of the most used general purpose background knowledge are Wikipedia and WordNet. Annane et al. [29] provided a two step method for mapping of ontologies which increases the efficiency by matching only a part of the background knowledge. The 2 steps are to select and build background knowledge using external knowledge. The first step is to select different ontologies from which the instances are selected. In the second step, the instances selected as combined to form a single resource.

3.3 Ontology Matching using Semantic Similarity

Semantic similarity is used to measure the similarity between two concepts of two ontologies based on the text used in the concepts. WordNet is a lexical database of the English language which contains the set of synonyms called synset. Ontology alignment techniques have used WordNet to calculate the semantic similarity by calculating the distance between the word and its synonym. If the words from source and target ontology belong to the same sysnet, then there is a similarity otherwise, there is no similarity [30] [31] [32].

Word embeddings are used to represent the sequence of words as vectors in the semantic space. This technique was introduced in the field of ontology alignment by using the Word2Vec model which was trained on Wikipedia data. The similarity between the vectors of entities is calculated using cosine similarity [33]. Fasttext [13] is another word embedding method where the vector representations of concepts are calculated by averaging the word embedding vectors of all words in the concept. Dhouib et al. [34] used the fasttext word embedding method to align the Silex ontology to other ontologies of the same domain which includes ESCO.

The development in word embeddings method addressed the problem of ambiguity by considering the meaning of a word based on the context, which was the drawback with fasttext and Word2vec models. The Bidirectional Encoder Representation from Transformers (BERT) Model [17] revolutionized word embeddings by using a transformer model that is pre-trained on a large dataset using bidirectional representations. This model creates different vector representations for the same word by considering the context. This is called Contextualized embeddings. Neutel et al. [6] used the BERT model in the ontology alignment between ESCO and O*NET using the occupation label and description. The fasttext model, BERT model and Sentence BERT (SBERT) were compared to get the best results of alignment. SBERT is - an adaptation of BERT model for building semantically meaningful sentence embeddings and supports the use of cosine similarity as compared to BERT. The results showed that SBERT model performed the best. The BERT Model has been trained on general text corpus like Wikipedia and BooksCorpus. In some fields like finance, biomedical etc., there are high quality text data which are not used for training the BERT model. This drawback was addressed by using domain specific BERT models which are trained from scratch using domain-specific data like FinBERT, a BERT model trained on financial data [35], BioBERT, a BERT model trained on biomedical data [36], and ClinicalBERT, a BERT model trained on clinical notes [37].

In the field of labor market, there are many applications which use NLP techniques like the labor market analysis, and recruitment process. The Stanford Digital Economic Lab developed Job2vec, a NLP model to classify job postings. The model was trained on real-world job posting available on the internet. The model was trained on the job descriptions to predict the job titles based on the Standard Occupation Classification(SOC) and the salaries. The description texts were encoded using the BERT embeddings. The model performed with a accuracy of 60% in predicting job titles [38]. These applications and method shows that using domain-specific knowledge in training the models has given better results when compared to the models trained on general data. After BERT, the XLNet model was developed by Google. XLNet is generalized autoregressive model which outperformed BERT on 20 tasks by a large margin. This has been explained in the background section [2]

3.4 Existing crosswalks in Labor Market

The national and international classifications which exist in different countries sometimes develop official crosswalks between other classifications or between different versions of the classifications.In this section the existing crosswalks between different occupational classifications are discussed.

Hoen et al. [39] developed crosswalk between the Norwegian occupations(STYRK) and the O*NET database. It was constructed manually with the help of a series of other crosswalks which lead to the mapping of the STYRK and O*NET.

Hardy et al. [40] used skills, work activities, work context and abilities of O*NET to map with the corresponding occupations of the Polish classification of occupations and specialisation called KZiS (Klasyfikacja zawodów i specjalności). The crosswalk was created using a series of conversions from O*NET-SOC to ISCO- ISCO to KZiS while considering the modified versions of both the classifications. Hardy et al. [41] also mapped the O*NET occupations with European Commission-Labor Force Survey (EU-LFS) using the 3 digit occupation codes. The EU-LFS data are coded using ISCO, so a crosswalk between O*NET and EU-LFS was created by first mapping O*NET to SOC, and then the SOC occupations are linked to ISCO using the official crosswalk created by the International Labor Organization.

There are two existing alignments between ESCO and O*NET occupations developed by Sophie et al. [6] and Kanders et al. [7] which used the then state-of-the-art model BERT to calculate the semantic similarity. Neutel et al. [6] experimented with three embedding techniques (fasttext, BERT, sentence-BERT) to calculate the semantic similarity score of the occupations' *labels* and *descriptions*. The results showed that the sentence-BERT has better alignment results compared to the other systems in terms of coverage and mean reciprocal rank. Kanders et al. [7] used the sentence-BERT model to calculate the semantic similarity score between the occupations data which includes *title*, *description*, *skills*, *work activities* and *work characteristics* of each occupation. This work aligns 1627 occupations out of 2942 occupations of ESCO. The authors did not use the final granular level of ESCO and only looked at occupations that were connected to ISCO, which is the foundation of ESCO. They used a two-step method for the mapping. First, a mapping was created using the existing mappings between O*NET and ISCO. These are called 'constrained' occupations. Then semantic similarity was calculated between each

ESCO and O*NET occupations to find the best match. The evaluation of the mapping was done manually by giving a confidence level to each mapping.
A common observation from these crosswalks ([39], [40], [41], [42]) is that they are manually aligned and/or use a series of crosswalks to finally arrive at the required alignment. The crosswalks developed by [6] and [7] use semantic similarity to derive a relation between the occupations. As told by the authors, the results of [6] was better using context sensitive embeddings and results of [7] had 31% of the matches with a confidence level of 2.0 which are also obtained by calculating the semantic similarity between the occupations. However the two alignment systems lacks a more detailed semantic relation between occupations.

3.5 Ontology Matching with Semantic Relations

While the above sections detailed about the different ontology matching techniques and the ways to improve matching, these matching techniques focus on finding equivalent matches. Raunich et al. [43] showed that establishing a relationship like *is-a*, *part-of* between the concepts of ontologies adds values and support in ontology merging process. S-Match [44]and STROMA [45] used a two step approach to establish a relation. In the frst step, a standard ontology matching technique is used to find the corresponding concepts and in the second step, background knowledge is used to specify the relations. Recently, Chen et al. [46] developed a BERT based ontology alignment system which utilized the structure and logic of the ontology. The first step of the process was to fine-tuning the BERT model on a corpus containing synonyms and non-synonyms and then find equivalent mappings between the ontologies with the help of string similarity and semantic similarity. These mappings are then refined using the *locality principle* [47] which says that if two concepts are related then the parent and child concept of these concepts are likely to be related.

3.6 Chapter Summary

The results of ontology matching approaches surveys were first examined in this chapter, which revealed that the methods lacked domain knowledge and could not solve the problem of ambiguity when lexical information was used. This prompted researchers to investigate the use of background knowledge and NLP approaches such as semantic similarity. The crosswalks developed in the labor market, which is the main theme of this thesis, are then described in depth. These existing crosswalks were mostly created manually, demonstrating the potential for automatic matching techniques. Following the discussion of ontology matching techniques, the next section looked at the various methods for expressing relationships and the relevance of having an informative relationship in situations like ontology merging. In the following chapter, the data used in this thesis is explored. These works showed that the mappings can be further improved and used for ontology integration.

Chapter 4

Data

The information available in ESCO and O*NET is described in this chapter. The data preparation for each of these ontologies, which are then used in the experiments, is also explained.

4.1 ESCO: European Skills, Competence and Occupation

The *European Skills, Competences, Qualifications and Occupations* - ESCO, is the European classification of occupations which defines the occupations and the skills relevant for the European labor market. ESCO is developed by the European Commission and updated constantly. The first version was available in 2017. At the time of this thesis, the latest version was version 1.0.8 which is used for the research. ESCO provides 2942 occupations and 13485 skills information which are related to those occupations and are also translated to all the European languages including Icelandic, Norwegian, and Arabic. The main idea behind the development of ESCO is to help in the exchange of information between employers, education providers, and job seekers. On the other hand, the availability of multiple languages helps in occupational mobility within Europe. The data can also be used for the analysis of skills supply and demand in real-time [48]. The ESCO data model is constructed based on three pillars namely, *Occupation pillar*, *Skills pillar*, and *Qualification pillar* which are interconnected.

The ESCO classification is expressed in the Simple Knowledge Organization System (SKOS) [49], which is a data model for representing knowledge organization systems like classification systems, taxonomies etc. The concepts of ESCO are represented as `esco:Occupation`, `esco:Skills`, and `esco:Qualification`. These are the subclasses of ESCO and have their own instances. The occupations used in this these are instances of `esco:Occupation` and skills information are instances of `esco:Skills`.

Occupation pillar

Occupations are not the same as jobs. Occupations are the set of jobs whose tasks and responsibilities are similar, whereas a job is the set of tasks and responsibilities carried out by one person [48]. ESCO provides information about the occupations at European level, including self-employment, volunteers, subsistence based occupations, arts and craft occupations, and political mandates. Each ESCO occupation contains detailed information regarding the occupation and the relationship that exists between the other pillars. The occupations are organized according to their mapping to the International Standard Classification of Occupations (ISCO-08) [49]. ISCO is a statistical classification that is structured hierarchically with four levels. The occupations are classified into 426 unit groups. These ISCO occupations and groups do not provide skills information related to the occupations. All the ESCO occupations are assigned to one of the ISCO unit groups. The top four levels of thecome from ISCO-08 and all the levels below are defined by ESCO. Any mapping to the ESCO classification has to be equal to more specific than the ISCO unit group but not more general [50]. The ESCO hierarchy is as follows :

Major group ⇒ Sub-major group ⇒ Minor group ⇒ Unit group ⇒ Occupation ⇒ Narrow occupation

The occupation level and narrow occupation level are of the type `esco:Occupation`. The occupations have a set of properties like `esco:prefLabel`, `esco:altLabel`, and `esco:description`. Occupation in different levels are related with the relation `skos:broader` and `skos:narrower`. `skos:broader` relation indicate that one occupation is more general than other and `skos:narrower` indicate that one occupation is more specific than other.

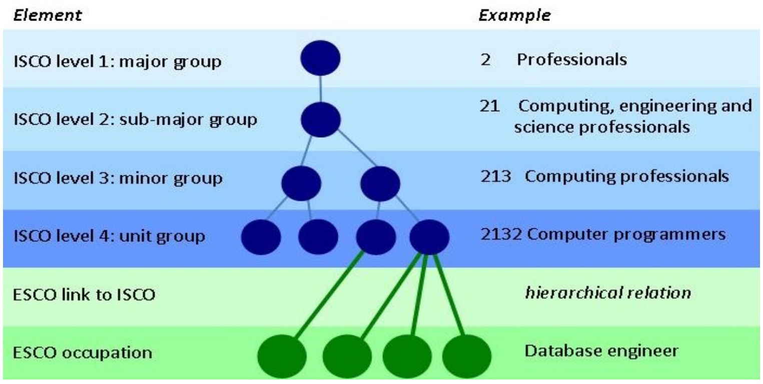

Element
Example
ISCO level 1: major group
2 Professionals
ISCO level 2: sub-major group
21 Computing, engineering and science professionals
ISCO level 3: minor group
213 Computing professionals
ISCO level 4: unit group
2132 Computer programmers
ESCO link to ISCO
hierarchical relation
ESCO occupation
Database engineer

Figure 4.1: ESCO occupation heirarchy, source: ESCOPedia[1]

Knowledge, Skills, and Competences pillar

This is also known as the Skills pillar, which includes the knowledge, skills and competences required for the occupations across the Europe. The skills/competences and knowledge are distinguished by indicating the skill type. This pillar contains 13,485 concepts and each of the concepts comes with a description and a preferred term. The relationship between occupation and skills is categorized as essential and optional. Essential skills are useful for the occupation regardless of the work context, employer or country. Optional skills are those which are dependent on the employer, working context and/or country. Optional skills are critical for occupation matching since they indicate the variety of jobs available within the same occupation [48]. The skills and competences are structured in 4 ways:

- Knowledge
- Skills
- Attitude and values
- Language skills and Knowledge

The Qualification pillar

The qualifications pillar gathers data from two sources: (1) member states qualifications databases (2) qualifications given directly to ESCO by individuals granting these credentials. The qualifications were still being updated by ESCO and were not available for the version of ESCO used in this thesis.

ESCO data preparation

From ESCO the level 5(ESCO occupations) and below occupations are considered. There are 2 reasons for this;

1. The top 4 levels of ESCO are taken from ISCO, which does not provide skills and knowledge information.

2. The level 5 and below are the occupations created by ESCO. An observation made from previous work of Neutel et al. [6] and Kanders et al [7] is that, only either the level 5 occupations are considered or the most granular level occupations which do not have a child node are considered. This led to missing occupations like `Chief Executive Officer` which are in level 5 and also has a child node. The child node is `airport Chief Executive Officer` which is very specific to the airport and there is no alignment to the CEO occupation. To overcome this limitation the level 5 occupations are also included. This thesis considers all the occupations defined by ESCO.

The information available from ESCO is summarized in table 4.1.

ESCO	
Occupation	Title
	Description
	Alternate Labels
Essential/Optional Skills	Title
	Description
	Alternate Labels
Essential/Optional Knowledge	Title
	Description
	Alternate Labels

Table 4.1: Information available in ESCO

4.2 O*NET: Occupational Information Network

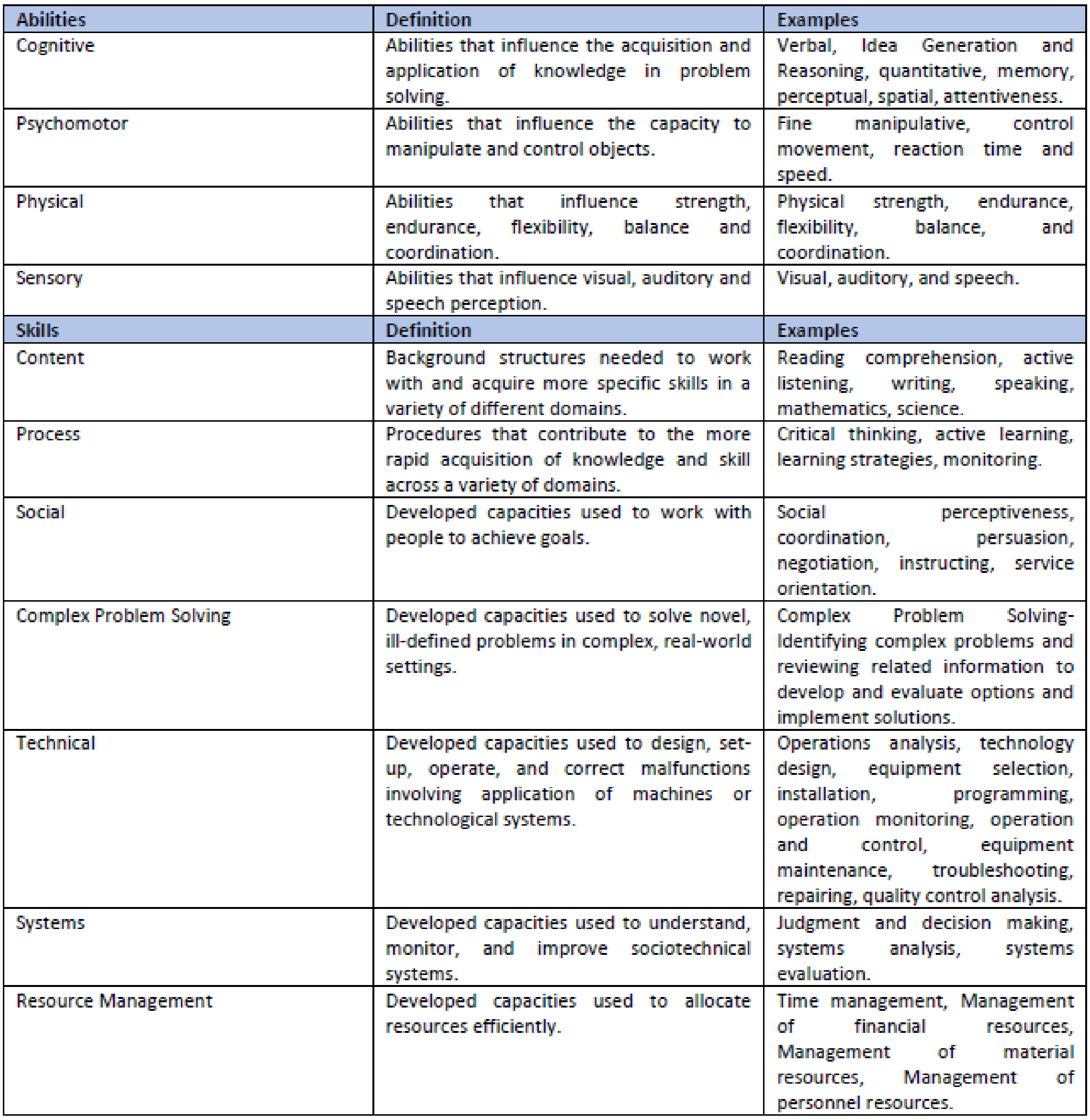
The O*NET is the database which describes the work, worker characteristics and the skills requirement for the occupations of US labor market [51]. The database is supported by the O*NET-SOC taxonomy which is based on the Standard Occupation Classification (SOC-2018). The O*NET model was developed using research on job and organizational analysis. The O*NET taxonomy structure includes 1016 occupations of which 852 has titles and details, and the other occupations have only titles and description [52].

It is built with respect to two views, the job oriented and the worker oriented views. The worker oriented and job oriented descriptors are classified into six domains, namely *worker characteristics*, *worker requirements*, *experience requirements*, *occupational requirement*, *workforce characteristics* and *occupational specific information*.

Worker Characteristics: These are the characteristics that may influence the performance as well as the capacity to acquire knowledge and skills required for the work performance. The worker characteristics include abilities, occupational interests, work values and work styles [53].

Ability is the capability to perform various tasks. The O*NET ability taxonomy includes 52 specific abilities, 15 general abilities, and 4 more general abilities. Work styles are the personal skills required for the occupations. This includes 16 work styles which are nested within 7 more generic work styles.

Worker Requirement: These are the attributes of an individual related to work performance such as the work related knowledge and skills. Worker requirements include basic skills, cross functional skills, knowledge, education [51]. 35 skills are divided into basic and cross-functional skills. Basic skills include content skills and process skills whereas the cross-functional skills refer to competences like social



Abilities	Definition	Examples
Cognitive	Abilities that influence the acquisition and application of knowledge in problem solving.	Verbal, Idea Generation and Reasoning, quantitative, memory, perceptual, spatial, attentiveness.
Psychomotor	Abilities that influence the capacity to manipulate and control objects.	Fine manipulative, control movement, reaction time and speed.
Physical	Abilities that influence strength, endurance, flexibility, balance and coordination.	Physical strength, endurance, flexibility, balance, and coordination.
Sensory	Abilities that influence visual, auditory and speech perception.	Visual, auditory, and speech.
Skills	**Definition**	**Examples**
Content	Background structures needed to work with and acquire more specific skills in a variety of different domains.	Reading comprehension, active listening, writing, speaking, mathematics, science.
Process	Procedures that contribute to the more rapid acquisition of knowledge and skill across a variety of domains.	Critical thinking, active learning, learning strategies, monitoring.
Social	Developed capacities used to work with people to achieve goals.	Social perceptiveness, coordination, persuasion, negotiation, instructing, service orientation.
Complex Problem Solving	Developed capacities used to solve novel, ill-defined problems in complex, real-world settings.	Complex Problem Solving- Identifying complex problems and reviewing related information to develop and evaluate options and implement solutions.
Technical	Developed capacities used to design, set-up, operate, and correct malfunctions involving application of machines or technological systems.	Operations analysis, technology design, equipment selection, installation, programming, operation monitoring, operation and control, equipment maintenance, troubleshooting, repairing, quality control analysis.
Systems	Developed capacities used to understand, monitor, and improve sociotechnical systems.	Judgment and decision making, systems analysis, systems evaluation.
Resource Management	Developed capacities used to allocate resources efficiently.	Time management, Management of financial resources, Management of material resources, Management of personnel resources.

Figure 4.2: O*NET skills and abilities, Source: [1])

skills, problem solving, technical and system skills [53].

Experience Requirements: These are related to the background of workers in an occupation which can include certificates and other training. The experience requirements include Experience and Training, basic skills- entry requirement, cross functional - entry requirement, licensing [51].

Occupational Requirement: This requirement gives information regarding the activities across occupations and are divided into 3 categories based on the specificity of the occupations like generalized work activities, intermediate work activities, and detailed work activities. This also includes the organizational context and work context [51].

Workforce characteristics: These are the variables that describe the characteristics of the occupation that influence the occupational requirements [51].

Occupation-Specific Information: This gives detailed information about the occupation and includes requirements of other domains like knowledge, skills and tasks in addition to the tools and equipment used in the workplace. It includes the title of the occupation, description, alternate titles, tasks, technology skills, and tools [51].

O*NET Data preparation

The O*NET occupation is available as a database and the version 26.1 is used in this thesis. In O*NET, not all the available occupations are detailed occupations, which means that for some occupations only the title and description information is available. There are 852 occupations for which detailed information is available, while the other occupations only have a title and description. Only the detailed O*NET occupations are used in this thesis. The information available from O*NET is summarized in table 4.2.

O*NET	
Attribute	**Information available**
Occupation	Label
	Description
	Sample Report Titles
Skills	Label
	Description
Knowledge	Label
	Description
Tasks	Label
Technology Used	Label
	Software/Application Used
Tools Used	Example
Abilities	Label
	Description
Work Activities	Label
	Description
Detail Work Activities	Label
Work Context	Label
	Description
Education	Label
Work Styles	Label
Interests	Label
	Examples
Work Values	Label
	Description

Table 4.2: Information available in O*NET

Chapter 5

Methodology

The methodologies used in this thesis are detailed in this chapter. The first section describes the matching procedure, which is used to find an O*NET occupation for each ESCO occupation based on semantic similarity. After that, three methods are defined to answer the research questions. Method 1 is to find an O*NET occupation match based on the semantic similarity which is calculated with help of the embeddings created by the generic XLNet model. Method 2 is similar to first method but here a fine-tuned XLNet model is used to create embeddings. Third method uses Wikidata as the background knowledge which acts as an anchor between ESCO and O*NET. It is then followed by the method for identifying the relationship between the ESCO-O*NET occupation pair.

5.1 Matching Process

This section outlines the fundamental matching technique. The ESCO ontology is the target ontology in this thesis. Between the two ontologies, ESCO *occupation title* corresponds to O*NET *occupation Label*, ESCO *occupation description* corresponds to O*NET *occupation description*, ESCO *occupation Alternate labels* corresponds to O*NET *occupation Sample report titles*, and ESCO *skills* and *knowledge* are combined which corresponds to O*NET *skills*, *abilities*, *knowledge*, etc. To identify a match for each ESCO occupation, the semantic similarity score with all O*NET occupations is calculated, and the top five highest-scoring occupations are obtained. In the matching process, label embeddings, description text embeddings, and alternate label text embeddings are compared using cosine similarity to obtain a score. These scores are calculated individually. The skills score is determined by calculating the number of skill pairs that exist in a predetermined set of skill pairs - *look-up skills* corpus, which is also determined by calculating the semantic similarity between skill labels and descriptions. This is explained in section 5.1.2. All the scores are in a scale of 0-1 and the final score is determined by taking the weighted average of all four scores. The weighted average is calculated by giving certain weights to these attributes.

label score

$\cos(ESCO\_label\_embeddings\_vector, ONET\_label\_embeddings\_vector)$

description score

$\cos(ESCO\_description\_embedding\_vector, ONET\_description\_embedding\_vector)$

Alternate label score

$\cos(ESCO\_alt\_labels\_embeddings\_vector, ONET\_alt\_labels\_embeddings\_vector)$

skills score

$$\frac{matching\ skills*}{Number\ of\ skills\ in\ ESCO\ occupation}$$

* calculated in skills matching

Overall Semantic Similarity Score

$$\frac{labelscore * L\_W + descscore * D\_W + AltLabelsscore * AL\_W + skillsscore * S\_W}{10}$$

L_W = weights for label, D_W = weights for description, AL_W = weights for alternate labels, S_W = weights for skills.

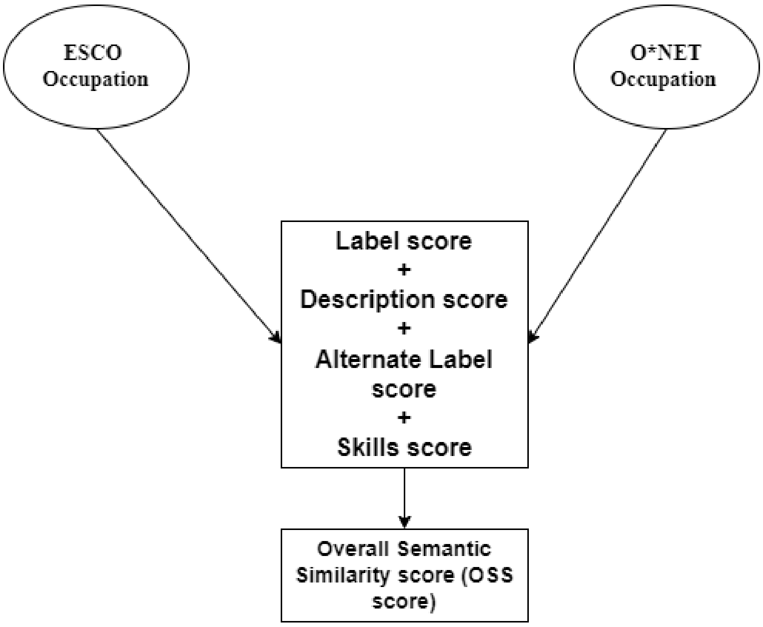


Figure 5.1: Matching Process

5.1.1 Creating Embeddings

The semantic similarity score, derived by the distance between the vectors of embeddings, is used to compare the occupations of ESCO and O*NET. The XLNet model and the fine-tuned XLNet model are used to generate the contextual embeddings. For each input text, this produces a vector of length 512. Embeddings are created for both ontologies' *occupation labels*, *occupation alternate labels*, *occupation description*, *skill label*, and *skill description* individually and then used in the matching process.

5.1.2 Skills Matching

The way the skills are classified in the two ontologies differs. The *Knowledge* indicated in ESCO resembles with the *Knowledge* defined in O*NET; however, the skills/competences of ESCO are not labeled and distinguished as in O*NET. The *Abilities* and other categories of skills from O*NET collectively resemble the skills defined by ESCO. Computing the skills similarity in the matching process was a time-consuming task because of computing the semantic similarity between each pair of skills for each ESCO-O*NET occupation pair. To address this problem in matching process, a skill matching procedure is carried out separately. In this, all ESCO skills are compared to all the skills available in O*NET irrespective of the skill types based on the semantic similarity of the *skill label* embeddings and *skill description* embeddings. For each ESCO skill, an O*NET skill with the highest semantic similarity is discovered. A threshold of 0.75 was used on the semantic similarity score which resulted in a set of 629 skills pairs which are stored as a look-up skills corpus. In the matching process, all unique combination of skills pairs from the ESCO-O*NET occupation is created and stored as a set. The number of pairs which are common in both set created for each ESCO-O*NET occupations and the *look-up skills* corpus, then it is considered as the skill is common between both occupations.

Matching skills

$$set(skillspairsfromESCO - ONEToccupation) \cap set(look - upskillspairs)$$

5.2 METHOD 1: ESCO and O*NET occupations matching

This method answers RQ 1.a and the generic XLNet model is used to create contextualized sentence embeddings for the occupation's data and the matching process explained in section 5.1 is used to find the best match. The relations are derived as described in section 5.5. The results of this system are used as the baseline to compare against the other two systems which are explained in the next sections.

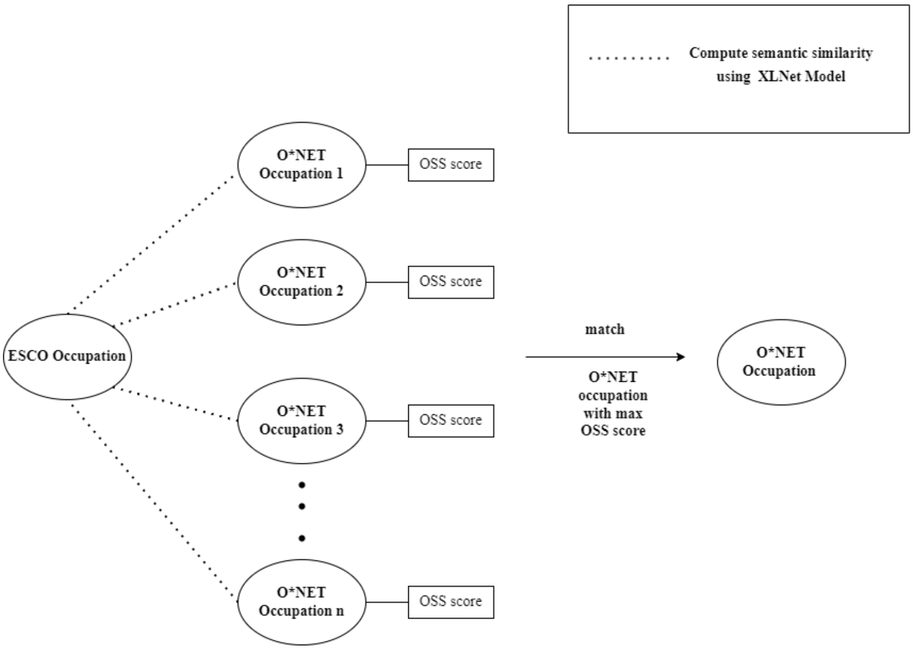


Figure 5.2: ESCO-O*NET ontology matching using XLNet model

5.3 METHOD 2: ESCO and O*NET occupations matching using a domain-specific XLNet model

To answer RQ 1.b, domain-specific knowledge is used to train the XLNet language model. XLNet model and other language models are usually trained on huge amount of general data from English Wikipedia, BookCorpus [54], etc. The aim of this method is to use a language model, in this case the XLNet model, which is familiar with the vocabulary of labor market. Training a domain-specific language model requires a large amount of training data and huge amount of GPUs for the computation. In our case, it is crucial to fine tune the model such that the vocabulary is a mixture of original vocabulary and also has the domain specific vocabulary. In the next section, an analysis of the data and the need for domain-specific knowledge is explained in detail. The method to find a match for ESCO occupations is same as used in section 5.2 but here the XLNet model is fine-tuned on domain-specific data. The relations are derived as given in section 5.5.

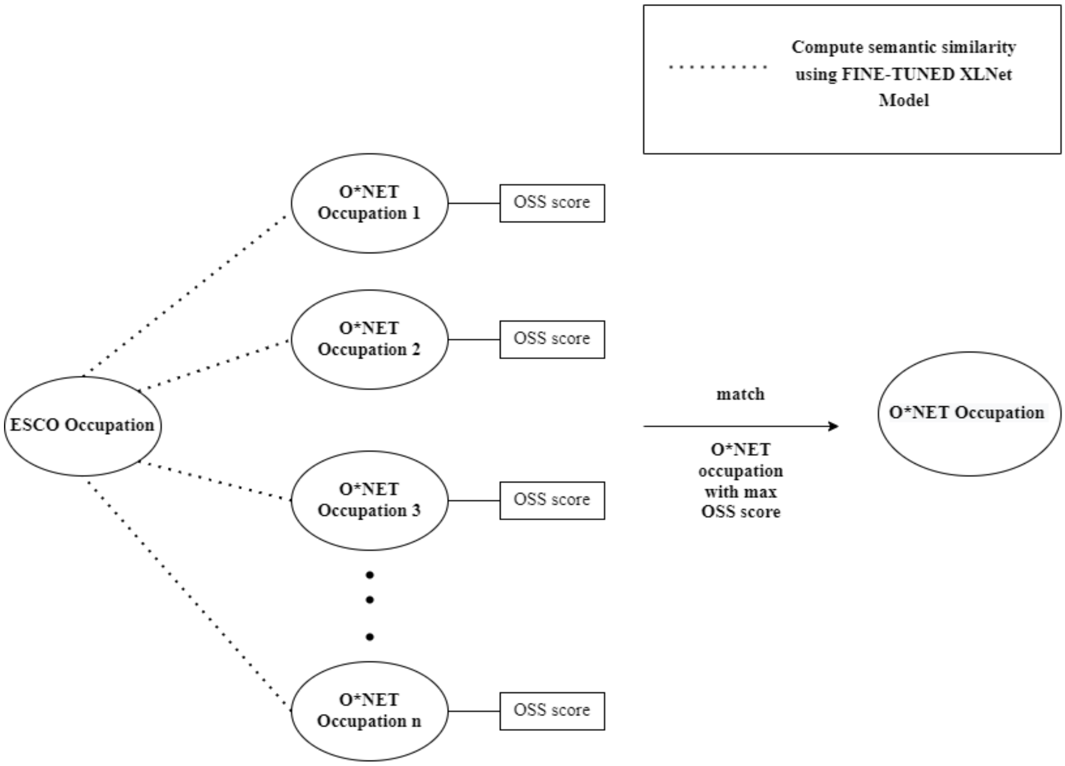
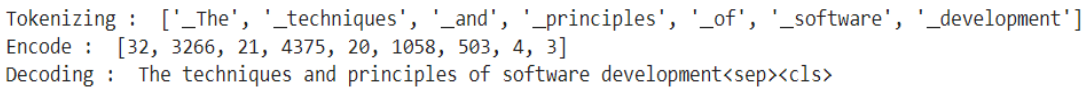


Figure 5.3: ESCO-O*NET ontology matching using domain-specific XLNet model

5.3.1 What is the need to train on domain specific data?

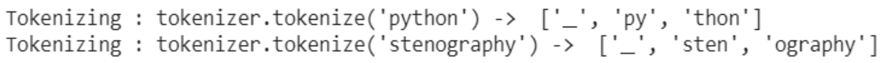
The language and words used in label, description, and skills information of occupations from the ontologies do not seem to be complicated. However, when evaluating the tokenizing procedure and the vocabulary of the XLNet model, several terms were broken down into subwords to construct embeddings. The XLNet vocabulary contains words extracted from the data utilized in the pre-training phase. As mentioned earlier, the XLNet tokenizer is a lossless tokenization process which keeps majority of the words in their original form. The following is an example of tokenizing and detokenizing input text.



```
Tokenizing :  ['_The', '_techniques', '_and', '_principles', '_of', '_software', '_development']
Encode :  [32, 3266, 21, 4375, 20, 1058, 503, 4, 3]
Decoding :  The techniques and principles of software development<sep><cls>
```

Figure 5.4: Tokenizing and Detokenizing

As seen above in figure 5.4, the vocabulary of XLNet tokenizer is rich and did not split words into subwords. But there are words which are split into subwords. An example is given below in figure 5.5.



```
Tokenizing : tokenizer.tokenize('python') ->  ['_', 'py', 'thon']
Tokenizing : tokenizer.tokenize('stenography') ->  ['_', 'sten', 'ography']
```

Figure 5.5: Sub word formation from tokenizing

As seen above, the words like *python* and *stenography* are not retained but are split into [*py*,*thon*] and [*sten*,*ography*] respectively. This is because these sub words can be used in creating other words which include these sub words. While creating an embedding for an input text, the XLNet model first tokenizes the input text into tokens and obtains the token IDs which is referred as '*encode*' in figure 5.4. These ids are then used in creating a vector representation of the input text using the XLNet model. In the cases where the words are split into sub-words, then the number of tokens for that input text increases and the model should find the token ids of all the sub-words which are distributed inside the 30 thousand tokens of XLNet vocabulary.

ExBERT (EXtended BERT) [55] is a BERT model which lies between a general language model and fine-tuning a language model from scratch to a specific domain. It fine-tunes the BERT model to a new domain by extending the vocabulary in the training phase. The same approach is used in this experiment to enhance the vocabulary of XLNet model to the labor market domain. When the vocabulary consists of these words then an improvement in the embeddings is expected as it can reduce the number of tokens. HuggingFace library[1] provides a function called

[1] https://huggingface.co/

tokenizer.add_tokens() which is used to add the list of tokens to the existing XLNet tokenizer vocabulary and ignores a token if it already exists. Once the vocabulary is extended, the model is adjusted to the new vocabulary size to ensure that the token embedding matrix of the model matches with the embedding matrix of the tokenizer. This is done using the *resize_token_embeddings()* method.

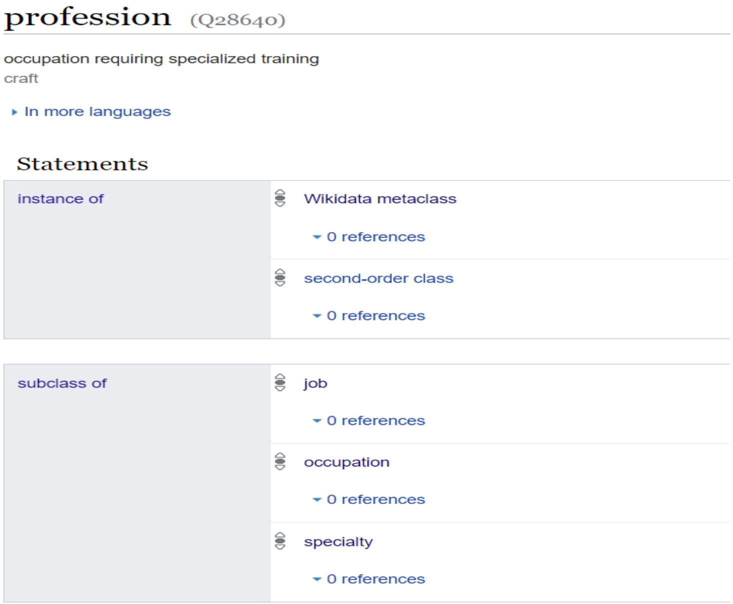
5.3.2 Data Preparation

The list of tokens which were added to the vocabulary list was prepared using the job posts datasets which consist of more than 40,000 job posts extracted from online job portals like LinkedIn, Glassdoor, etc. The term frequency of each word was calculated for the complete dataset and the top 5% of the terms which has the highest frequency were taken as the list of tokens which are then added to the vocabulary.

5.4 METHOD 3: ESCO and O*NET occupations matching using Background Knowledge

To answer the research question RQ 1.c, the two ontologies are matched using a helper background knowledge. Here ESCO and O*NET are both matched to a background knowledge ontology which is WikiData[2] in our case. The WikiData is the target ontology and the ESCO and O*NET occupations are mapped to instances of the ***professions*** item in the Wikidata ontology. Wikidata is structured as a repository of items. Each item has a label and description. An item in Wikidata also gives other information like `instanceOf` and `subclassOf`. The `instanceOf` attribute tells the class to which the item is an instance of and the `subclassOf` attribute gives the next higher class or type. All the instances of these items are instances of the higher classes as well. All the instances are also an item with a label and description. As shown in figure 5.6, the *professions* item is a subclass of *job*, *occupation*, and *speciality*. It has 8000 instances which describe professions like `economist`, `Sommelier` etc.

[2] https://www.wikidata.org/wiki/Wikidata:Main_Page



profession (Q28640)

occupation requiring specialized training
craft

▸ In more languages

Statements

instance of	Wikidata metaclass ▾ 0 references second-order class ▾ 0 references
subclass of	job ▾ 0 references occupation ▾ 0 references specialty ▾ 0 references

Figure 5.6: Wikidata details of Profession

A match is found for ESCO occupations when an O*NET occupation also matches to the same Wikidata profession. The process of matching ESCO and O*NET occupation uses the background knowledge as an anchor. In this process, the ESCO and O*NET occupations are first matched to the instances of *profession* from Wikidata. The semantic similarity score is calculated between the occupations and the profession using contextual embeddings of the *label* and *description* information created by the XLNet model. When the ESCO and O*NET occupations are matched to the same instance of *profession*, then it can be concluded that a there is a match between those occupations.

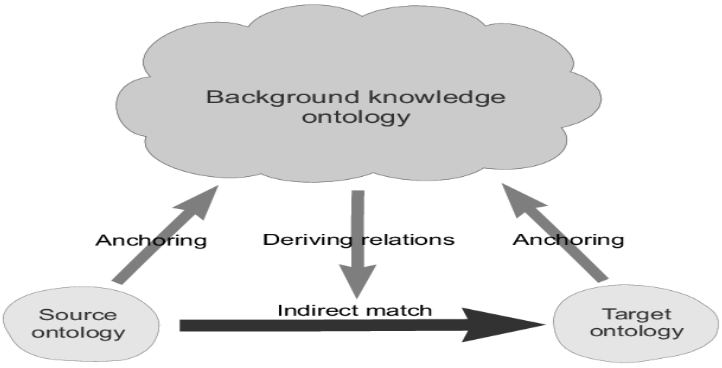


Figure 5.7: ESCO-O*NET ontology matching using domain-specific background knowledge. [56]

Figure 5.7 summarizes the process of mapping ESCO and O*NET with the help of background knowledge ontology which is the Wikidata.

5.5 Method to Derive Relations

Ontology matching is useful for merging data from many sources. The results of ontology matching provide correspondences between two ontologies' concepts and can be beneficial in information retrieval tasks. The methodologies for ontology matching employed by [6], [7] find the correspondence between concepts based on the semantic similarity score above a certain threshold and focus on finding equivalent matches. In this thesis, a rule-based technique is used to identify more expressive semantic relations, such as those described in section 2.2. The rules are based on the taxonomical structure of the target ontology, ESCO and it was not possible to use the structure of O*NET because of the unavailability of information. The overall structure of an ontology is that the child node of a concept contains more information and is more specialized. This arrangement conveys the impression that the concepts at the top level are broader, and as the layers descend, the concepts get smaller. Figure 5.8 shows an example of a concept and its sub concepts in SKOS format.

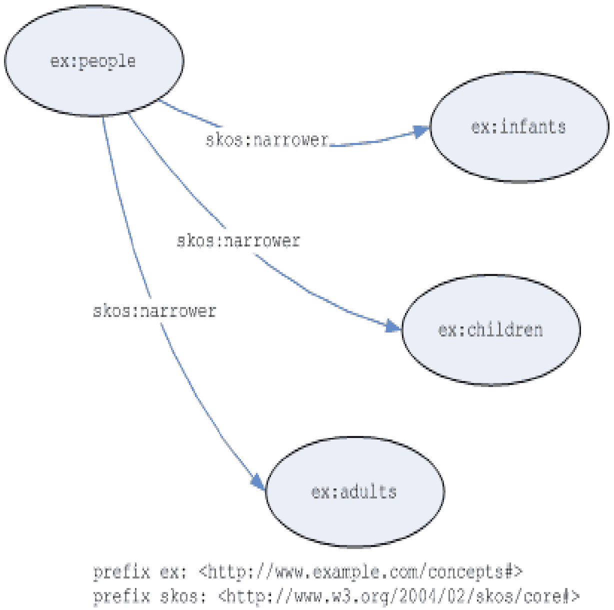

ex:people
skos:narrower
ex:infants
skos:narrower
ex:children
skos:narrower
ex:adults
prefix ex: <http://www.example.com/concepts#>
prefix skos: <http://www.w3.org/2004/02/skos/core#>

Figure 5.8: SKOS Structure

ESCO ontology is also available in SKOS format. To structure the occupations, the occupations are formed by mapping the occupation to ISCO groupings. The top four levels are provided by ISCO-08, and ESCO occupations are at levels 5 and lower [48]. There are 426 ISCO groups to which ESCO occupations are assigned, and they are organized in a hierarchical system with ESCO occupations at levels 5, 6, 7, and 8. The occupations become narrower as the levels decrease. Figure 5.9 depicts an example of the hierarchy of ISCO group 5165. The occupations are

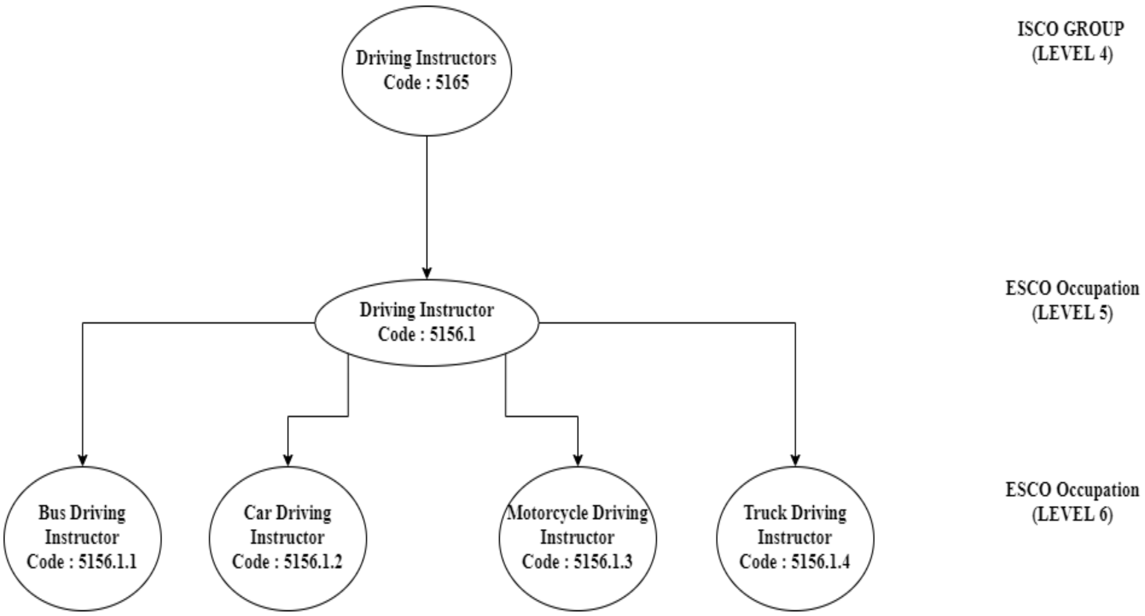


Figure 5.9: ISCO group 5165

in a hierarchical structure and then become narrower down the levels. Some ISCO groups have occupations that are very different within the same level. In figure 5.9, the occupations in level 6 are all similar as they describe an occupation of an instructor of different vehicles like bus, car, motorcycle, and truck. In this case, we can apply the *locality principle*, which states that if the parent node is matched, then so are all of the child nodes. However, there are occupations that are specialized to 'casinos' as shown in figure 5.10, and it is reasonable to establish a relationship between O*NET occupations related to 'casinos' and ESCO occupations of the same topic, which also provides insight into where it can be placed when merged.

In this thesis, an experiment of deriving the relations is carried out based on some rules. Some of the key decisions that were made before establishing a relationship are as follows,

Decision 1: The ESCO occupations were grouped based on their ISCO group which gives a set of 426 ISCO groups. The decision to consider individual groups while establishing a relation was because in some cases, an O*NET occupation was matched to more than 30 ESCO occupations from different ISCO groups. This raised a conflict in deriving the relations based on the level of ESCO occupation that it was matched to. For example, the occupations `Actuarial Consultant` and `Actuarial Assistant` are related with respect to description and the label, but these occupations are placed different ISCO groups **Mathematicians, Actuaries,**

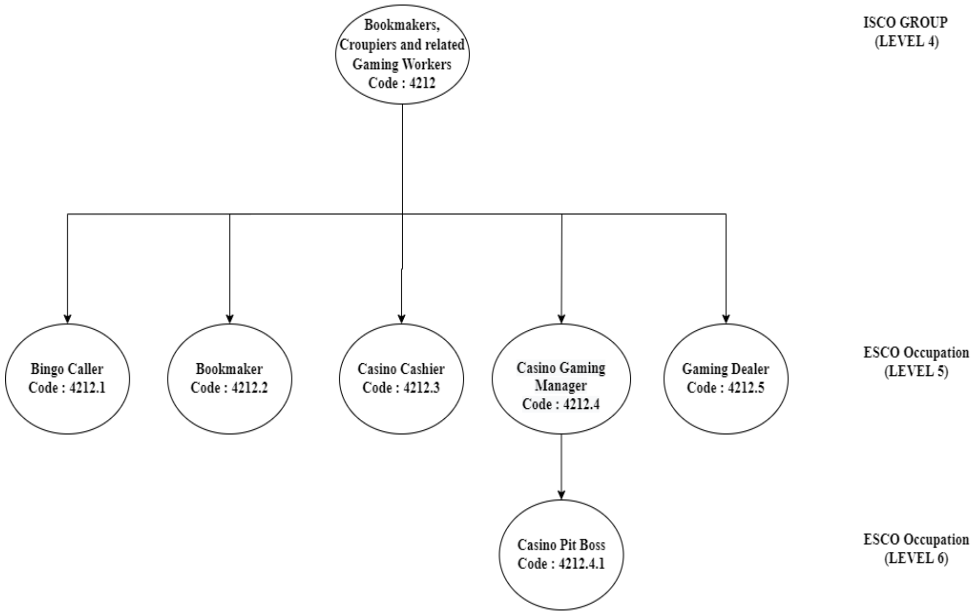

Bookmakers, Croupiers and related Gaming Workers Code : 4212
ISCO GROUP (LEVEL 4)
Bingo Caller Code : 4212.1
Bookmaker Code : 4212.2
Casino Cashier Code : 4212.3
Casino Gaming Manager Code : 4212.4
Gaming Dealer Code : 4212.5
ESCO Occupation (LEVEL 5)
Casino Pit Boss Code : 4212.4.1
ESCO Occupation (LEVEL 6)

Figure 5.10: ISCO group 4212

and Statisticians - 2120 and **Statistical, Mathematical and related Associate professionals - 3314** respectively and belong to different Major groups [57]. When an O*NET occupation was matched to both these occupations it is difficult to establish a relation which can help in placing the O*NET occupation appropriately in the ISCO group. Dividing them based on the groups reduces the complication and give a clearer view as to where it can be placed within the ISCO groups.

Decision 2: The relations `skos:broadMatch`, `skos:narrowMatch`, `skos:exactMatch` and `skos:closeMatch` are established only when two or more ESCO occupations are matched to the same O*NET occupation which is a case of many-to-one match. These are further referred as *shared_matches*. The relation depends on the semantic similarity score and level of the ESCO occupations. This was decided because most of the ISCO groups are similar to the case as desccribed in figure 5.10 and to the fact that we encountered many-to-one matches.

Decision 3: Two new semantic relations are introduced called the `Match` and `noMatch` relation. These are established when the O*NET occupation is matched to only one ESCO occupation within the ISCO group. These relations are not a part of the SKOS semantic relations and depends solely on the semantic similarity score and therefore has a lower confidence when compared to other relations.

Let X be level of an ESCO occupation. Then X-1 is the level above it, and X+1 is the level below the occupation if it exists. The rules to derive a relation is as follows :

Rule 1: A `skos:exactMatch` is established between the ESCO-O*NET match

if the semantic similarity score is the highest when compared to the other *shared_matches*.
Rule 2: A `skos:closeMatch` is established between the ESCO-O*NET match which has a lower semantic similarity score compared to the pair which has `skos:exactMatch` relation and also these ESCO occupations belong to the same level.
Rule 3: A `skos:broadMatch` is established between the ESCO-O*NET match which has a lower semantic similarity score compared to the pair which has `skos:exactMatch` relation and also these ESCO occupations belong to different levels. If ESCO-O*NET match with `skos:exactMatch` belong to level X, then `skos:broadMatch` relation is established to the ESCO-O*NET match at level X-1.
Rule 4 : A `skos:narrowMatch` is established between the ESCO-O*NET match which has a lower semantic similarity score compared to the pair which has `skos:exactMatch` relation and also these ESCO occupations belong to different levels. If ESCO-O*NET match with `skos:exactMatch` belong to level X, then `skos:narrowMatch` relation is established to the ESCO-O*NET match at level X+1.

The steps to find a relation between the ESCO-O*NET occupation pair is as follows :
Step 1: The total Semantic Textual Similarity (STS) score is calculated as per the matching process for all the ESCO occupations and the match with highest semantic similarity score is considered.
Step 2: The ESCO occupations are grouped and sorted in descending order based on the semantic similarity score.
Step 3: The ESCO-O*NET match within each group is compared based on three conditions: one, if the O*NET occupation is shared or not. Second, which ESCO-O*NET match has the semantic similarity score. Third, the ESCO occupation level.
Step 4: When the O*NET occupation is matched to only one ESCO occupation, then the relation `Match` is established when the semantic similarity score above a certain threshold, in this case a threshold of 0.6 was used and `noMatch` is established when the score is lower than 0.6.

5.6 Chapter Summary

In this chapter, we have presented the methods that are used in this thesis starting with the matching process used to find an O*NET occupation match for each ESCO occupation. The method 1 uses generic XLNet model to create embeddings, method two uses fine-tuned XLNet model, and method three uses domain specific knowledge - Wikidata - as an anchor to find a match. After finding a match using these methods a relation has to be established between the matches, this done based on the method to derive relations which is also detailed at the end of the chapter. In the next chapter, the experimental setup and the evaluation method is described.

Chapter 6

Experimentation

This chapter details the experiments conducted during the study to address the research questions listed in section 1.3. The experiments are conducted such that the two research questions concerning the usage of domain-specific knowledge and also using skills and alternate labels information in the matching process. The sections that follow details the experimental design as well as the evaluation metrics that will be utilized in the study.

*

6.1 Experimental Setup

To answer the research questions RQ1 and RQ2 which deals with improving the matching process, five experiments were conducted which are based on the three methodologies explained in section 5.2, section 5.3 and section 5.4. The XLNet model and the fine-tuned XLNet model with an extended vocabulary set are used in these methods to create contextual embeddings.

The five experiments are :

1. ESCO-O*NET occupation matching using only *label* and *description* data with the help of general XLNet Model. Further this method is denoted as **METHOD_1$_{\mathbf{LD}}$**

2. ESCO-O*NET occupation matching using *label, description, skills,* and *alternate labels* data with the help of general XLNet Model. Further this method is denoted as **METHOD_1$_{\mathbf{ALL}}$**

3. ESCO-O*NET occupation matching using only *label* and *description* data with the help of Fine-Tuned XLNet Model. Further this method is denoted as **METHOD_2$_{\mathbf{LD}}$**

4. ESCO-O*NET occupation matching using *label, description, skills,* and *alternate labels* data with the help of Fine-Tuned XLNet Model. Further this method is denoted as **METHOD_2$_{\mathbf{ALL}}$**

5. ESCO-O*NET occupation matching using Wikidata as an anchor. This experiment uses only *label* and *description* from ESCO and O*NET as other information is not available in Wikidata. Further this method is denoted as **METHOD_3$_{\mathbf{LD}}$**

For **METHOD_1$_{\mathbf{LD}}$**, **METHOD_2$_{\mathbf{LD}}$**, and **METHOD_3$_{\mathbf{LD}}$**, to calculate the overall semantic similarity score, a weight of 6,4 is used as *label weight* and *description weight* respectively. For **METHOD_1$_{\mathbf{ALL}}$** and **METHOD_2$_{\mathbf{ALL}}$**, weights of 3, 3, 2, and 2 are used as*label weight*, *description weight*, *skills weight* and *alternate label weight* respectively. The five experiments are performed to get an insight of the following statements which can answer the research questions:

- The effect of using various other metadata
- The effect of using Fine-Tuned XLNet model with domain-specific vocabulary
- The effect of using domain-specific knowledge (Wikidata) as an anchor

The following comparisons were made to answer the research questions defined in 1.3.

- **METHOD_1$_{\mathbf{LD}}$** is compared with **METHOD_2$_{\mathbf{LD}}$** and **METHOD_1$_{\mathbf{ALL}}$** is compared with **METHOD_2$_{\mathbf{ALL}}$** to evaluate the effect of Fine-tuned XLNet model and answers the **RQ1.b**
- **METHOD_1$_{\mathbf{LD}}$** is compared with **METHOD_3$_{\mathbf{LD}}$** and **METHOD_2$_{\mathbf{LD}}$** is compared with **METHOD_3$_{\mathbf{LD}}$** to answer **RQ1.c**
- **METHOD_1$_{\mathbf{LD}}$** is compared with **METHOD_1$_{\mathbf{ALL}}$** and **METHOD_2$_{\mathbf{LD}}$** is compared with **METHOD_2$_{\mathbf{ALL}}$** to know the effect of using more information in the matching process and answers **RQ2**.

6.2 Evaluation

6.2.1 Evaluation Process

The first step in the evaluation was to check if the matches that were found are correct or not. Among the previous studies of the ESCO and O*NET matching, the results of Kanders et al. [7] involved the judgement of two reviewers so, it was used as the ground truth. This result set matched 1627 occupations out of 2942 occupations of ESCO. The authors did not use the final granular level of ESCO and only looked at occupations that were connected to ISCO, which is the foundation of ESCO. They used a two-step method for the mapping. First, a mapping was created using the existing mapping between O*NET and ISCO. These are called ‘constrained’ occupations. Then, semantic similarity was calculated between each ESCO and O*NET occupation to find the best match. Semantic similarity was measured using sentence embeddings generated by sentence-BERT model. The mapping was based on the *skills*, *work activities* and *work characteristics* of each occupation in the two ontologies. The evaluation of the mapping was done manually by two reviewers by giving a confidence level to each matching as follows.

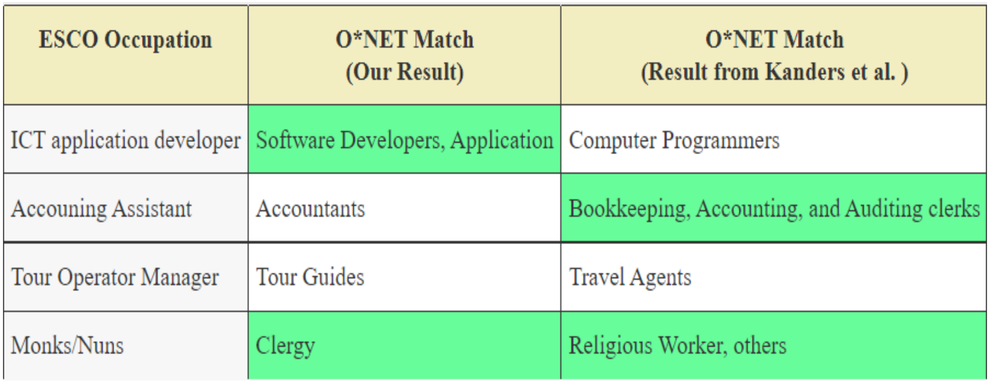
- A score of 0.5 indicates that the reviewers did not agree mutually with mapping and the mapping suggested by the second reviewer is considered.

- A score of 1 indicates that the matches obtained by calculating the semantic similarity was not correct but the two reviewers agreed on the best match found after two rounds of manual review of matches.

- A score of 2 was given to the matching that were the best ‘constrained’ match as well as the most semantically similar.

Considering only the matches with highest confidence level, i.e., 2.0, gave us a set of 480 occupations which was then used as the ground truth set for our experiments.

The results of our studies were compared to those of Kanders et al. [7]. First, the **METHOD_1$_{\textbf{ALL}}$** results were used as the evaluation set, which was likewise obtained by utilizing all of the information. They made use of ESCO version 1.0.5 and O*NET version 24.1. To match this, we used the same versions’ data and assessed their results. A maximum of 62.9 percent accuracy was attained by using various weight combinations for the attributes used in the matching procedure. When we evaluated these results, we discovered some matches that we thought would be a better fit. A domain expert from TNO assessed ten ESCO-O*NET match samples to investigate the conflict of matches between our results and the ground truth set. The study revealed that for five occupations, the match discovered by us was correct, one match was not correct, and for the remaining occupations, it was difficult to pick between the two provided matches. Four example matches from each of these cases is given in table 6.1 where the green cells are the correct matches according to our annotator.



ESCO Occupation	O*NET Match (Our Result)	O*NET Match (Result from Kanders et al.)
ICT application developer	Software Developers, Application	Computer Programmers
Accouning Assistant	Accountants	Bookkeeping, Accounting, and Auditing clerks
Tour Operator Manager	Tour Guides	Travel Agents
Monks/Nuns	Clergy	Religious Worker, others

Figure 6.1: Our Results v/s Results from Kanders et al. [7]

Given the uncertainty regarding which match is correct, as well as the fact that we were using old versions of ESCO and O*NET, we chose to manually analyze a sample of 200 occupations with the assistance of a domain expert from TNO. The domain expert assigned the correct O*NET occupation out of the given options for each of the 200 ESCO occupations. The relations were not determined here.

6.2.2 Sampling

As annotating whether a match is correct or incorrect takes time, 200 ESCO occupation samples were chosen and the matches for these occupations were evaluated. So, a preliminary analysis of the results was carried out, which aided in the selection of the 200 samples. The following sections details the analysis and highlight the key takeaways.

Matchings Found by all the Methods When examining the results of **METHOD_3$_{\mathbf{LD}}$** -the approach using domain-specific background knowledge as an anchor, it was found that the method was unable to match all 2942 ESCO occupations to Wikidata due to the limitation of the source. The approach was able to match 713 ESCO occupations. So, the first decision was to use only these 713 occupations for which an O*NET match was available from all three methodologies and five experiments.

The second step was to check if all the systems were able to find a match given a threshold of 0.55. This threshold was chosen since the experiment's goal was not only to identify a correct match but also to find different associations based on semantic similarity scores, such as broad or narrow. These relations are based on the rule given in section 5.5, which states that relations that are not an `skos:exactMatch` have a lower score. For the 713 ESCO occupations which were obtained as per the previous analysis, METHOD_1$_{\text{LD}}$, METHOD_2$_{\text{LD}}$, and METHOD_3$_{\text{LD}}$ found top five matches within the given threshold. The number of matches found at each rank in METHOD_1$_{\text{ALL}}$ and METHOD_2$_{\text{ALL}}$ are shown in table 6.1. METHOD_1$_{\text{ALL}}$ has a single match for 306 ESCO occupations and has five matches for only 105 occupations. The 306 ESCO occupations had a match from all the experiments thus resulted in selecting the samples from among these ESCO occupations that had at least one matches across all methods.

METHOD_1$_{\text{ALL}}$					**METHOD_2$_{\text{ALL}}$**				
rank 1	rank 2	rank 3	rank 4	rank 5	rank 1	rank 2	rank 3	rank 4	rank 5
306	193	150	123	105	528	444	403	370	344

Table 6.1: Number of matches found at each rank in METHOD_1$_{\text{ALL}}$ and METHOD_2$_{\text{ALL}}$

The domain expert was tasked with analyzing 200 random samples from the 300 ESCO occupations. To find the 200 samples, a simple sampling procedure was employed on the ESCO occupation list. To overcome the problem of losing a match that was found by one experiment but not by others, each of the 200 occupations had the O*NET occupation match result from all five result sets. After accounting for overlapping matches between methodologies and removing duplicate matches for the ESCO occupation, each occupation had a maximum of 10 to 12 matches to analyze. For each ESCO occupation, the annotator chose an O*NET occupation from the possibilities provided. The discovered O*NET occupation was the best match. Further, these matches are referred as *correct* matches.

6.2.3 Evaluation Metrics

The evaluation of the matches found is challenging as there is no ground truth available. As a result, typical evaluation criteria such as precision, recall, and f1-score cannot be used. A human judgement was made to determine if the match between ESCO-O*NET occupation pair is correct or not within the given set of options, this is further explained in section 6.2.2. After finding the correct matches, it is then used as the ground truth to find the accuracy of different methods. Accuracy is further explained in the next section.

6.2.3.1 Accuracy

Accuracy measures the number of correct matches found for a given set of samples. In this experiment, the accuracy was found for each rank, i.e, the number of correct matches found at rank 1 to rank 5 when the matches are placed in descending order based on the semantic similarity score with the highest score being in rank 1.

6.2.3.2 Top N Accuracy

The top n accuracy metric is a measure to find if the correct match is within a certain N predicted values. If the correct match is found within the N predicted matches then it is considered as found and used for the calculation of accuracy. This is the sum of the accuracies at all ranks.

6.3 Chapter Summary

In this chapter we went through the experimental setup of the thesis which help in answering the research questions. After the experiments are performed, the obtained results are evaluated by a human judgement for a sample of 200 occupations. In the next chapter, the results and performance of each method are discussed in detail.

Chapter 7

Results and Discussion

7.1 Results

As mentioned earlier, a sample of 200 ESCO-O*NET occupations were evaluated by the annotator to find the correct matches. As shown in the figure 7.1, out of the 200 matches, 155 of the ESCO occupations found a correct O*NET match and the other 45 ESCO occupations did not have any correct match within the given options.

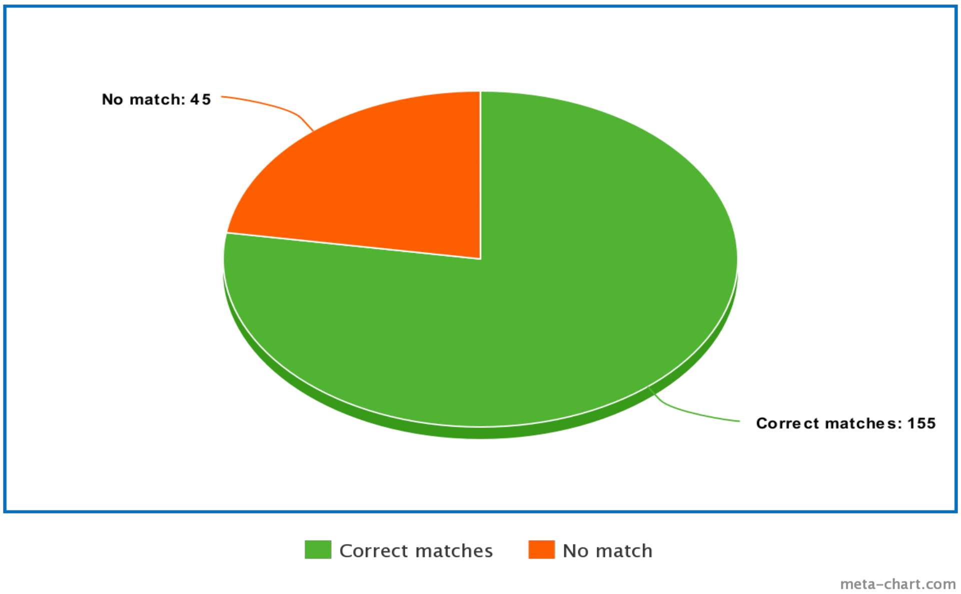


Figure 7.1: Number of Correct matches and no matches in sample

These 155 *ESCO-O*NET occupation matches* were considered as the ground truth and was used for evaluating the five experiments. Figure 7.2 and 7.3 show the performance of each experiment and the accuracy at each rank. On top of the bar, we can find the *Top 5 accuracy* which is calculated as explained in section 6.2.3.2. The generic XLNet model has performed well when compared to the methods using domain-specific knowledge. When a threshold of 0.55 was used on the semantic similarity score to determine top five matches, there were occupations in *METHOD_1$_{ALL}$* and *METHOD_2$_{ALL}$* which did not find five matches. So, a lower threshold of 0.45 was used to check if any occupation could find a match. An increase of 4% and 1% in accuracy was found in the *METHOD_1$_{ALL}$* and *METHOD_2$_{ALL}$* respectively. From the results we can also see that the accuracy at *rank one* and also the *top 5 accuracy* is higher in experiments that use more information in the matching process.

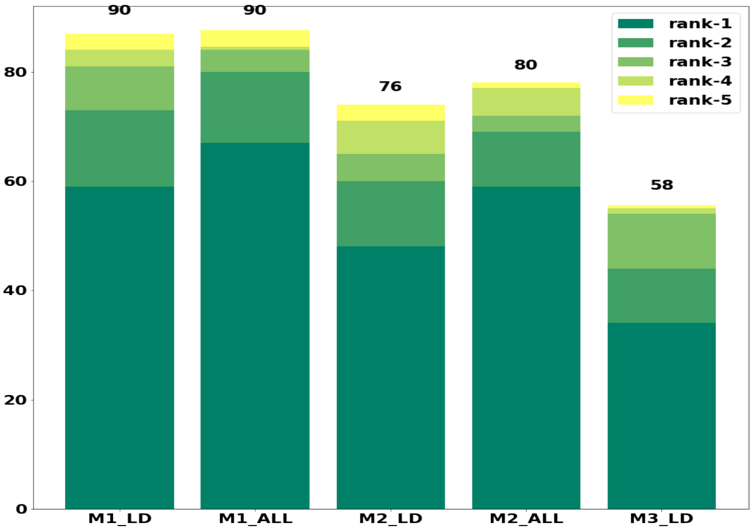


Figure 7.2: Accuracy of Different Methods with a threshold = 0.55

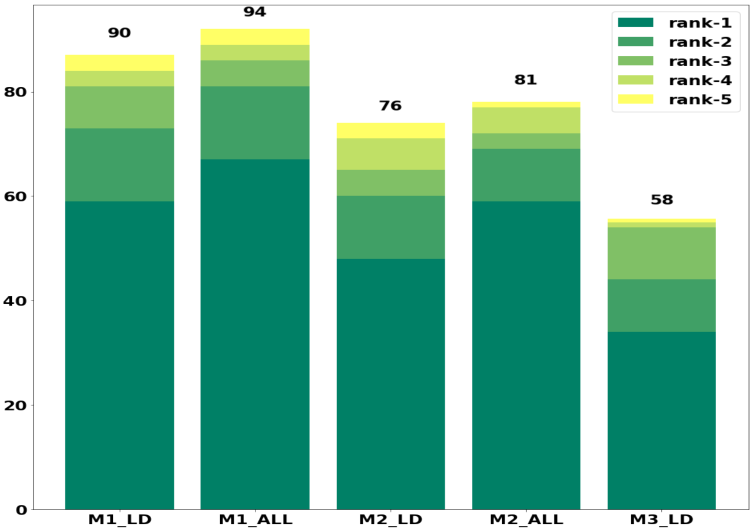


Figure 7.3: Accuracy of Different Methods with a threshold = 0.45

Evaluation of the semantic relations was another challenging task as it requires more time and a deeper knowledge in the structure of ESCO. To overcome this situation, a different approach was taken to evaluate the correctness of semantic relations. Two cases were established to makes the analysis. First, when the *correct match* was same as the *predicted match*, these are referred as *correct-predicted matches*. Here the *predicted match* means the O*NET occupation found at rank one because the method of deriving relations only considers the O*NET occupation with the highest overall semantic similarity score. *Correct matches* are the matches found by the annotator. *correct-predicted matches* means that the *Correct matches* was found at rank one. An intuition is that in case of *correct-predicted matches*, the relationship between the ESCO and O*NET occupations should belong to one of the following relations: `skos:broadMatch`, `skos:narrowMatch`, `skos:closeMatch`, `skos:skos:exactMatch` or `Match` and it should not belong to `noMatch` as the match is already annotated as correct and must have a relation between them.

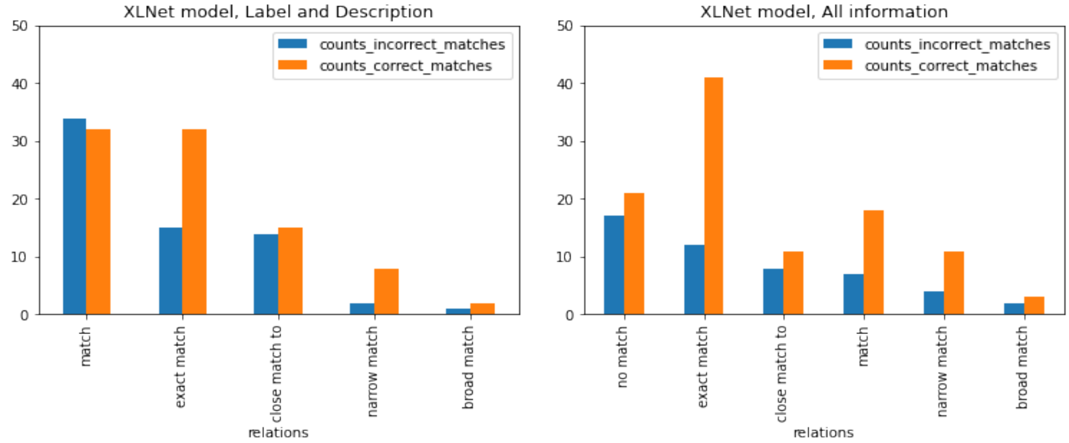


Figure 7.4: Number of relations in $METHOD\_1_{LD}$ and $METHOD\_1_{ALL}$

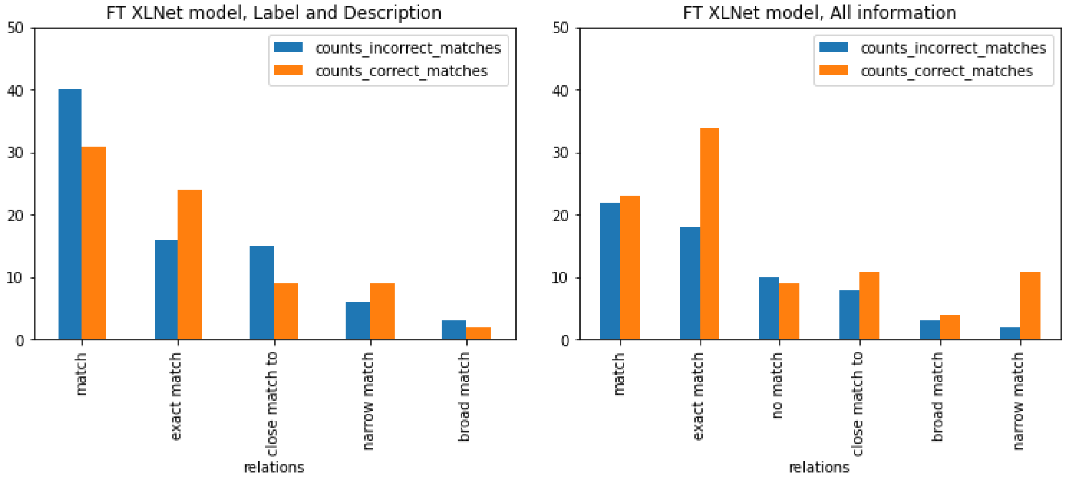


Figure 7.5: Number of relations in $METHOD\_2_{LD}$ and $METHOD\_2_{ALL}$

From figure 7.4 and 7.5,we can see that `noMatch` relation does not exist in the experiments which use only *label* and *description* information in the matching process. We can also observe that *METHOD_1*$_{ALL}$ has more `noMatch` relations in *correct-predicted matches* case which should not occur as they are already annotated as a correct match which means that a relationship exist between the occupations. In *METHOD_1*$_{ALL}$, `noMatch` relations are more in *incorrect-predicted matches* case which is logically true as they are not the correct match. In the next section, the results are discussed in detail.

7.2 Discussion

In this section, a quantitative analysis is carried out to look at the results which are discussed in detail. The section is divided into two sub-sections namely, shallow Analysis, which discusses the results irrespective of if the match found is correct or not, i.e., all the matches including and outside of the 200 samples. It also provides information on the O*NET occupations that were not matched to any ESCO occupations, as well as overlapping matches between experiments. It is then followed by the next section which gives an in-depth analysis of the results and discusses the relations established with the support of examples. Finally, the chapter is concluded with the limitations and the lessons learnt from the analysis.

7.2.1 Shallow Analysis of Matches

This section thoroughly examines the matches and the O*NET occupations that are matched to the ESCO occupations. The analysis does not assess whether the found match is a correct match or not; nonetheless, this is covered in section 7.2.2. Here, the 306 ESCO occupations which found at least one *NET match are considered for the analysis.

The initial step in the analysis was to see if all of the approaches could find five matches with a threshold of 0.55 on the overall semantic similarity score. This threshold was chosen since the experiment's next stage is to build relationships other than an exact match, and these matches with broad/narrow relationship will have a lower similarity score. As a result, a lower threshold was chosen.

7.2.1.1 Analysis of Number of Unique O*NET Occupations

In this section, the number of unique O*NET occupations that were matched to the ESCO occupations by each experiment is observed. This gives an insight into how the methods and the data used in the matching process helps in finding a variety of O*NET occupations. Two ways of comparison was performed.

1. Comparing within the methods: In method 1, the results from $\textbf{METHOD\_1}_{\textbf{LD}}$ had 30% more unique O*NET occupation matches compared to the result set from $\textbf{METHOD\_1}_{\textbf{ALL}}$. In method 2, there was an increase of 10% in results from $\textbf{METHOD\_2}_{\textbf{LD}}$ when compared with the result set from $\textbf{METHOD\_2}_{\textbf{ALL}}$ experiment. From this we can say that using only *Labels* and *Descriptions* information in the matching process matches with more unique O*NET occupations compared to using more metadata related to the occupations in the process.
2. Comparing between the methods: The percentage of unique O*NET occupations in $\textbf{METHOD\_1}_{\textbf{LD}}$, $\textbf{METHOD\_2}_{\textbf{LD}}$ and $\textbf{METHOD\_3}_{\textbf{LD}}$, which are based on *Labels* and *Descriptions* are all within the same range, with $\textbf{METHOD\_1}_{\textbf{LD}}$ showing a little rise over the other results.
When compared between $\textbf{METHOD\_1}_{\textbf{ALL}}$ and $\textbf{METHOD\_2}_{\textbf{ALL}}$, $\textbf{METHOD\_1}_{\textbf{ALL}}$ has 60% unique matches while $\textbf{METHOD\_2}_{\textbf{ALL}}$ has 40% unique matches. This tells that in method 2, in which the XLNet model was supported with domain-specific vocabulary, the ESCO occupations were matched to the same O*NET occupation in more numbers. The analysis of how finding a variety of O*NET matches depends on finding a correct match is discussed in section 7.2.2.3.

7.2.1.2 Unmatched O*NET Occupations

187 out of the 852 O*NET occupations were not matched to any of the ESCO occupations in any of the experiments. When examined, most of these O*NET occupations resembles to an ISCO unit group rather than an ESCO occupation which are defined under the ISCO groups. A few examples of these occupations are `‘Painting, Coating, and Decorating Workers’` and `‘Molders, Shapers, and Casters, Except Metal and Plastic’`
which doesn’t define one specific occupation but a group of occupations in ESCO. For example the table 7.1 shows the ESCO occupations that can be a potential match to the O*NET occupation `‘Painting, Coating, and Decorating Workers’`. We can also see that these ESCO occupations are from different ISCO groups and refer to only ‘Painter’ topic.

ESCO Occupations	ESCO Occupation COde	O*NET Occupation
Construction Painter	7131.1	Painting, Coating, and Decorating Workers
Scenic Painter	3432.4.1	
Ceramic Painter	7316.1.1	
Glass Painter	7316.1.2	
Artistic Painter	2651.1	

Table 7.1: Potential matches for the Unmatched O*NET occupation: `'Painting, Coating, and Decorating Workers'`

7.2.1.3 Analysis of Overlapping Results

In this section, an analysis of the overlapping of matches between the results of different experiments is discussed. Since there were five set of results obtained from the experiments, there was high chance of overlapping. Overlapping of results means that the same match has been found by more than one experiment. This gives an insight into how the methods impact in finding a match and how different they are. If there is less overlapping, then the results are different which means that the experiments are unique and it is worth making a comparison. Two ways of overlapping analysis was made:
1. Within the method: In this, the number of overlapping was checked to see how adding more information affected the results. This gives an insight if adding more information would find a different match or if it is the same.
The analysis showed that in method 1 (using general XLNet model) only 30% of the results were overlapping and within method 2 (using Fine-tuned XLNet model), 36% of the matches are overlapping. It can be concluded that adding more information in the matching process have an effect in finding different matches. 2. Between the methods: This analysis is carried out to find the effect of using domain-specific knowledge. This is again performed in three parts,

- comparing between **METHOD_1**$_{\textbf{LD}}$, **METHOD_2**$_{\textbf{LD}}$ and **METHOD_3**$_{\textbf{LD}}$ results which consider only the label and descriptions information.
- Results of **METHOD_1**$_{\textbf{ALL}}$ and **METHOD_2**$_{\textbf{ALL}}$ which used more information.

When compared between the **METHOD_1**$_{\textbf{LD}}$ and **METHOD_2**$_{\textbf{LD}}$ which use only *label* and *description*, there was approximately 50% overlap and between the results of **METHOD_1**$_{\textbf{ALL}}$ and **METHOD_2**$_{\textbf{ALL}}$ using all the available information, there was

26% overlap. The very low overlap between **METHOD_1$_{\textbf{ALL}}$** and **METHOD_2$_{\textbf{ALL}}$** shows that the extended vocabulary helped in finding new matches for the occupations and also reveals that the vocabulary had an impact majorly in the skills and alternate labels information text. When compared to **METHOD_3$_{\textbf{LD}}$**, there was the least overlap of 1% with **METHOD_1$_{\textbf{LD}}$** and **METHOD_2$_{\textbf{LD}}$** which tells that this method gives very different results compared to direct semantic similarity between ESCO and O*NET occupations.

7.2.2 In-depth Analysis of Matches

An in-depth study of the matches identified by each experiment is undertaken in this section. The 200 samples of ESCO-O*NET occupations matches which were annotated by the domain expert are considered.

7.2.2.1 Analysis of Unmatched ESCO-O*NET pairs

Out of the 200 samples that were given to the annotator, 155 ESCO occupations had a *correct match* within the given options. The remaining 45 ESCO occupations did not find a correct match, these occupations are referred as *unmatched* ESCO occupations. The *unmatched* ESCO occupations were analyzed again manually to check if there was an O*NET occupation out of all the available O*NET occupations which is similar to the ESCO occupation but could not be found by any of the experiments within the top five matches of the experiments. The result of the analysis is as follows.

1. **Occupations which did not find a match in O*NET:**
`‘Astronaut’`, `‘Kinesiologist’`, `‘Media Scientist’`, `‘Senator’`, `‘cosmologist’`and other ESCO occupations did not find a match in O*NET with a keyword search in all the fields like label, description and alternate labels which can mean that they are not defined by O*NET.
2. **Occupations which found a match manually:**
Out of the 45 occupation which did not find an O*NET occupation match, only the ESCO occupation `derrickhand` was able to find a match in O*NET which is similar. The O*NET occupation `Derrick Operators, Oil and Gas` is very similar but was not found in the top five matches of any experiment.

7.2.2.2 Analysis of Unique and Overlapping Matches

Here, the overlapping of *correct ESCO-O*NET occupation matches* is analyzed to know how the methods are different compared to each other. There are 155 *correct*

*ESCO-O*NET occupation matches* out of the 200 samples. Overlapping occurs when the *correct ESCO-O*NET occupation matches* is found in more than one experiment. Some interesting observations regarding the overlapping are as follows:

- 73 out of 155 *correct ESCO-O*NET occupation matches* were found in all the experiments. All these matches have high string similarity between the occupation's *label*. Example: `Mechanical Engineer` (ESCO) →`Mechanical Engineers` (O*NET). We can see in this example that *labels* of the ESCO and O*NET occupations are same except for the additional 's' at the end of O*NET occupation label. These occupations shows that irrespective of using more information in the matching and using domain-specific knowledge, the match can be found.

- 30 out of 155 ESCO-O*NET occupation matches are found in all the experiments except METHOD_3$_{\mathrm{LD}}$.

- 6 out of 155 *correct ESCO-O*NET occupation matches* were found only in METHOD_1$_{\mathrm{ALL}}$. For example the match `Geophysicist` (ESCO) →`Geoscientists, Except Hydrologists and Geographers` (O*NET) is not found only in this method due to the fact that one of the alternate labels of O*NET occupation is `Geophysicists`. This was not found in METHOD_2$_{\mathrm{ALL}}$ which shows that the extended vocabulary has not performed well and has declined the performance of XLNet model.

7.2.2.3 Analysis of using Domain Specific Knowledge

To answer the research question RQ1 which investigates the effect of using domain-specific knowledge in the matching process, an analysis is performed to know the occupations for which no match was found within top five rank without knowing the domain and understanding the vocabulary of the texts. The ESCO-O*NET occupation pairs `Embedded System Designer` (ESCO) →`Computer Programmers` (O*NET) was found only in **METHOD_2$_{\mathrm{LD}}$** and **METHOD_2$_{\mathrm{ALL}}$**. The match `Dressmaker` (ESCO) →`Fashion Designers` (O*NET) was found only in **METHOD_3$_{\mathrm{LD}}$**. The match `Landscape Architect` (ESCO) →`Landscape Architects` (O*NET) is in **METHOD_2$_{\mathrm{LD}}$**, **METHOD_2$_{\mathrm{ALL}}$**, and **METHOD_3$_{\mathrm{LD}}$**. Nonetheless, majority of the ESCO occupations found a match without the usage of domain-specific knowledge within the given set of samples. With such a small number of samples, reaching a conclusion is quite difficult.

7.2.2.4 Analysis of Using More Information

To answer the research question RQ2 which investigates the effect of using *skills* and *alternate labels* information in the matching process, an analysis is performed to find the matches which were found only in **METHOD_1$_{\textbf{ALL}}$** and **METHOD_2$_{\textbf{ALL}}$**.
From figure 7.2 and 7.3, we can see that there are more number of correct matches found using more information compared to using only the label and description of occupations in both methods. An example of this is `Guide` (ESCO) →`Tour Guides and Escorts` (O*NET) and `Librarian` (ESCO) →`Librarians and Media Collections Specialists` (O*NET). This was found only in **METHOD_1$_{\textbf{ALL}}$** and **METHOD_2$_{\textbf{ALL}}$**. This shows that using various metadata in the matching process has a prominent effect in finding matches.

7.2.3 Analysis of Semantic Relations

Another focus of the thesis was to establish a relationship between the *ESCO-O*NET occupation matches* which gives more information other than being a correct or incorrect match. These relations are based on the semantic similarity score and the taxonomical structure of ESCO i.e., the levels of the ESCO ontology. The method to establish a relation used in this thesis only considers the matches which have the highest semantic similarity score (Rank 1 matches). So even if the ESCO occupation had found a match at a different rank, it is not considered in this analysis. Out of all the experiments performed, in method one (using generic XLNet model), **METHOD_1$_{\textbf{ALL}}$** has the best performance with an accuracy of 69% at rank one and from method two (using fine tuned XLNet model) **METHOD_2$_{\textbf{ALL}}$** has the best performance with an accuracy of 59% at rank one. Only these two results are used for further analysis of the semantic relations. The results of **METHOD_3$_{\textbf{LD}}$** (using background knowledge as anchor) is not considered here because it was not able to find a match for all the ESCO occupations. The method of deriving relations used in thesis depends on the structure of ESCO and since not all the ESCO occupations have an O*NET occupation match, the structure remains incomplete for method three. This is also a limitation of method three.
Evaluating semantic relations is a challenging task since it requires an in-depth understanding of the occupations and the structure of ESCO. It is significantly easier to declare that a relationship that has been established is incorrect than it is to say that a relationship has been built correctly. Based on this approach, at the end of each analysis, a table is produced that shows the number of correct relations as well as the number of incorrect relations that can be transformed into another relation and can be considered as correct. These relations are analyzed manually by us and not by the annotator.

Before going into the details of the discussion, a recap of the terminologies are given below.

- *ESCO-O*NET occupation matches*: This refers to the ESCO occupation and corresponding O*NET occupation match found in the matching process.
- *correct match*: This refers to the *ESCO-O*NET occupation matches* which are declared as correct by the annotator.
- *predicted match*: This refers to the *ESCO-O*NET occupation matches* where the O*NET occupation was found at rank one i.e., the occupation with the highest overall semantic similarity score.
- *correct-predicted matches*: This refers to the *correct matches* that was also found at rank one.
- *incorrect-predicted matches*: This refers to the *ESCO-O*NET occupation matches* where the *correct match* was not found at rank one. This can mean that the *correct match* was found at a different rank or not at all.

In the following sections, the different relations are discussed in detail combined with the methods in which the relations are established.

7.2.3.1 Analysis of `skos:exactMatch` and `skos:closeMatch` Relations

In this section, the `skos:exactMatch` and `skos:closeMatch` relations established between the matches found by **METHOD_1$_{\textbf{ALL}}$** and **METHOD_2$_{\textbf{ALL}}$** are discussed in detail. For each method, the results are divided into two parts, one, *correct-predicted matches* and second, *incorrect-predicted matches*. These are then used for further analysis.

Analysis of `skos:exactMatch` and `skos:closeMatch` Relations of METHOD_1$_{\textbf{ALL}}$

Case 1: *correct-predicted matches* There are 41 *correct-predicted matches* that have the `skos:exactMatch` relation. An interesting observation is that 29 out of 41 of these *correct-predicted matches* exhibit significant string similarity in terms of labels. For example `Security Guard` (ESCO) →`Security Guards` (O*NET). Two of the *ESCO-O*NET occupation matches* were not an `skos:exactMatch` relation. For example, `Legal Consultant` (ESCO) →`Lawyers` (O*NET) and `Dog Trainer` (ESCO) →`Animal Trainers` (O*NET) are incorrect `skos:exactMatch` relation but can be considered as a different relation like a `skos:closeMatch` or `skos:narrowMatch`. These are the kind of matches which have the wrong relation but can be converted to a different relation.

11 of the

textitcorrect-predicted matches have `skos:closeMatch` relationship. All these *ESCO-O*NET occupation matches*' relationships are appropriate. This is determined by taking into accoount the `skos:exactMatch` relation that is correlated to the `skos:closeMatch` relation. As the ESCO occupations are matched to the same O*NET occupation and belong the the same level in ESCO, these relationships are correlated.

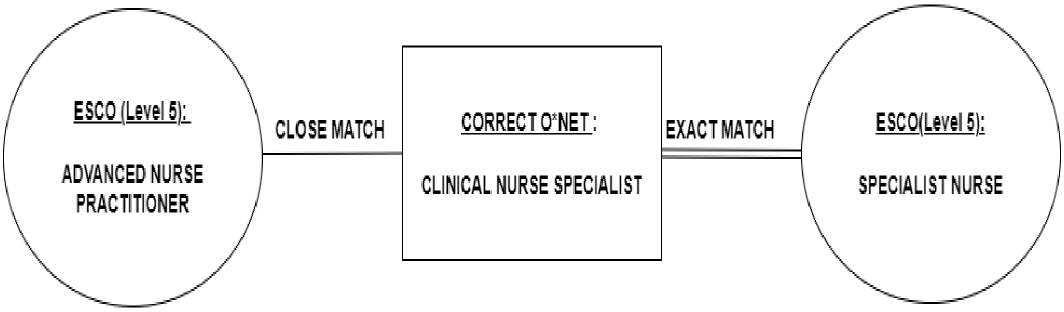


Figure 7.6: Exact and Close Match relation for correct matches found

An example of the relationship between the ESCO and O*NET occupations is depicted in Figure 7.6. Two ESCO occupations, `Advanced Nurse Practitioner` and `Specialist Nurse`, are *correct match* to the same O*NET occupation, `Clinical Nurse Specialists` in this example. Both of these ESCO occupations belong to the same level and since the match `Specialist Nurse` (ESCO) →`Clinical Nurse Specialists` (O*NET) has highest overall semantic similarity score among the matches, the `skos:exactMatch` relation is established here.

Case 2: *incorrect-predicted matches* There were 12 `skos:exactMatch` relationships and 8 `skos:closeMatch` relationships. The majority of the 12 *ESCO-O*NET occupation matches* with the `skos:exactMatch` relations were completely erroneous, such as `Dance Therapist` (ESCO) →`Nannies` (O*NET), which are completely unrelated. However, others were about the same topic or line of work for example, `Software Developer` (ESCO) →`Web Developers` (O*NET), the *correct match* for this ESCO occupation is `Computer Programmers`, as indicated in figure 7.7. This cannot be considered a valid `skos:exactMatch` relationship but can be a different relationship.
Three of the eight `skos:closeMatch` relations can be considered as `skos:closeMatch` relation, for example, `Dressmaker` (ESCO) →`Sewer, Hand` (O*NET), the correct match for this ESCO occupation is `Fashion Designer` as per the annotator. In this case, the description of `Dressmaker` (ESCO) is very similar to that of `Sewer, Hand` (O*NET). This relation is dependent on the relationship between `Tailor` (ESCO) and `Sewer, Hand` (O*NET), which is `skos:exactMatch`

relation because this pair has a higher semantic similarity score.

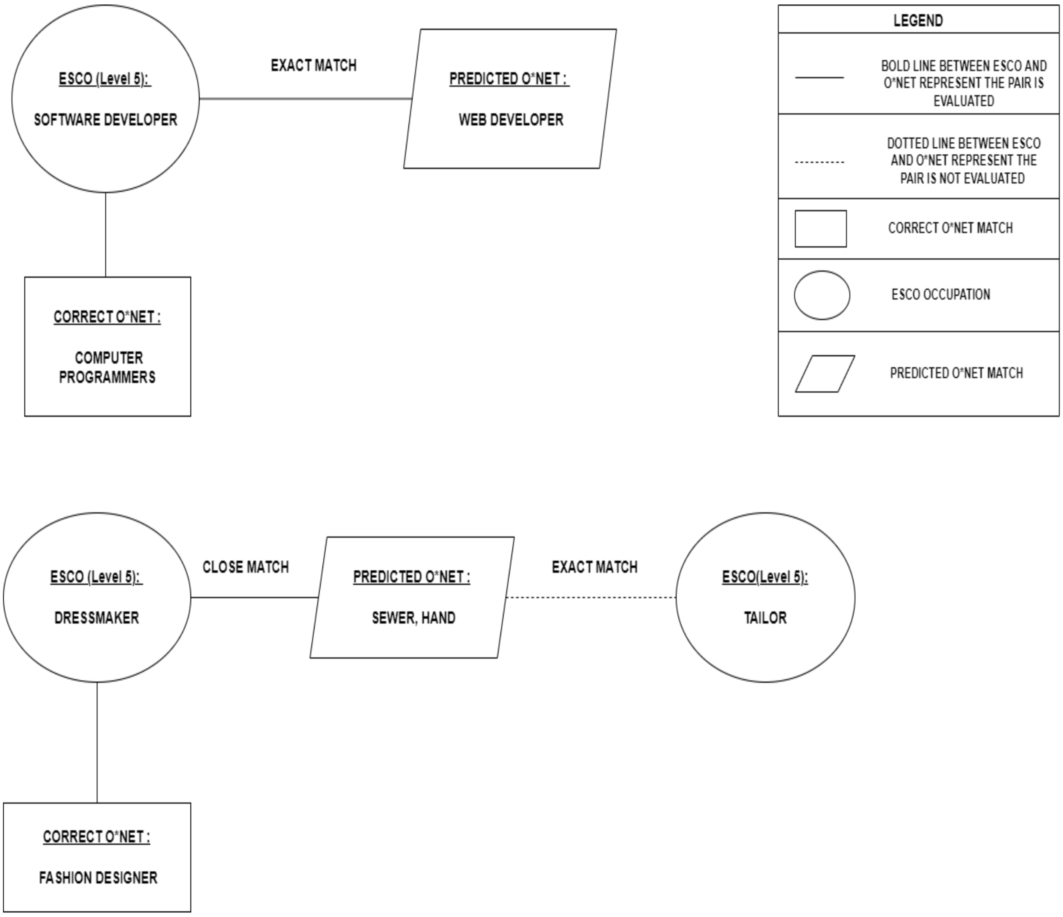


Figure 7.7: Exact and Close Match relation for incorrect matches found

In figure 7.7, the matches for the ESCO occupation `Dressmaker` are evaluated but the matches for `Tailor` were not evaluated because it was not present in the sample and hence the line between `Tailor` and `Sewer, Hand` is *dotted line*. All other *ESCO-O*NET occupation matches* in this case were wrong.

Analysis of `skos:exactMatch` and `skos:closeMatch` Relations of METHOD_2$_{\text{ALL}}$

Case 1: *correct-predicted matches* There are 34 *correct-predicted matches* with skos:exactMatch relation and 11 *correct-predicted matches* with skos:closeMatch relation.

The figure 7.8 shows the relation between the ESCO and O*NET occupations. In this example, two ESCO occupations - Babysitter and Nanny are matched correctly to the same O*NET occupation Nannies. As the semantic similarity score between Babysitter (ESCO) and Nannies (O*NET) is higher, it has the skos:exactMatch relation and skos:closeMatch relation between Nanny (ESCO) and Nannies (O*NET).

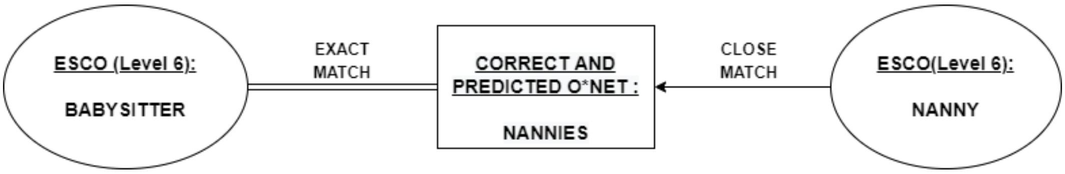


Figure 7.8: Exact and Close relations for correct matches

Case 2: *incorrect-predicted matches* Out of 63 incorrect-predicted matches there are 18 skos:exactMatch relations and 8 skos:closeMatch relations. In the 18 matches with skos:exactMatch relation, there were 4 matches that were completely wrong and other matches were difficult to judge. For example, the match Data Analyst (ESCO) →Operations Research Analysts (O*NET) corresponds with the definition which could be regarded as skos:closeMatch. But, considering the relation that was established by our method, it cannot be an skos:exactMatch. The correct match that was found for this ESCO occupation was Business Intelligence Analysts (O*NET). Out of the 8 skos:closeMatch relations that were found, four of them are completely wrong, for example: Dressmaker (ESCO) →Carpenters (O*NET), the correct match for this occupation is Fashion Designers (O*NET) as per the annotator. The other four pairs can be considered skos:closeMatch relations, one example of these matches are Engine Designer (ESCO) →Mechanical Engineers (O*NET), the correct match for this occupation is Automotive Engineers (O*NET), the description of Engine Designer corresponds closely with the description Mechanical Engineers.

Section Summary

METHOD_1$_{ALL}$ *correct-predicted matches* = 105, *incorrect-predicted matches* = 50					
	Case 1 *correct-predicted matches*			Case 2 *incorrect-predicted matches*	
	skos:exactMatch	skos:closeMatch		skos:exactMatch	skos:closeMatch
Total Number of Relations	41	11	Total Number of Relations	12	8
Number of Correct Relations	37	11	Number of Relations Completely Wrong	4	3
Number of Relations which can be Different Relation	4	0	Number of Relations which can be Different Relation	8	5

Table 7.2: Analysis of `skos:skos:exactMatch` and `closeMatch` Relations of METHOD_1$_{ALL}$

METHOD_2$_{ALL}$ *correct-predicted matches* = 92, *incorrect-predicted matches* = 63					
	Case 1 *correct-predicted matches*			Case 2 *incorrect-predicted matches*	
	skos:exactMatch	skos:closeMatch		skos:exactMatch	skos:closeMatch
Total Number of Relations	34	11	Total Number of Relations	18	8
Number of Correct Relations	29	8	Number of Relations Completely Wrong	6	5
Number of Relations which can be Different Relation	5	3	Number of Relations which can be Different Relation	12	3

Table 7.3: Analysis of `skos:exactMatch` and `skos:closeMatch` Relations of METHOD_2$_{ALL}$

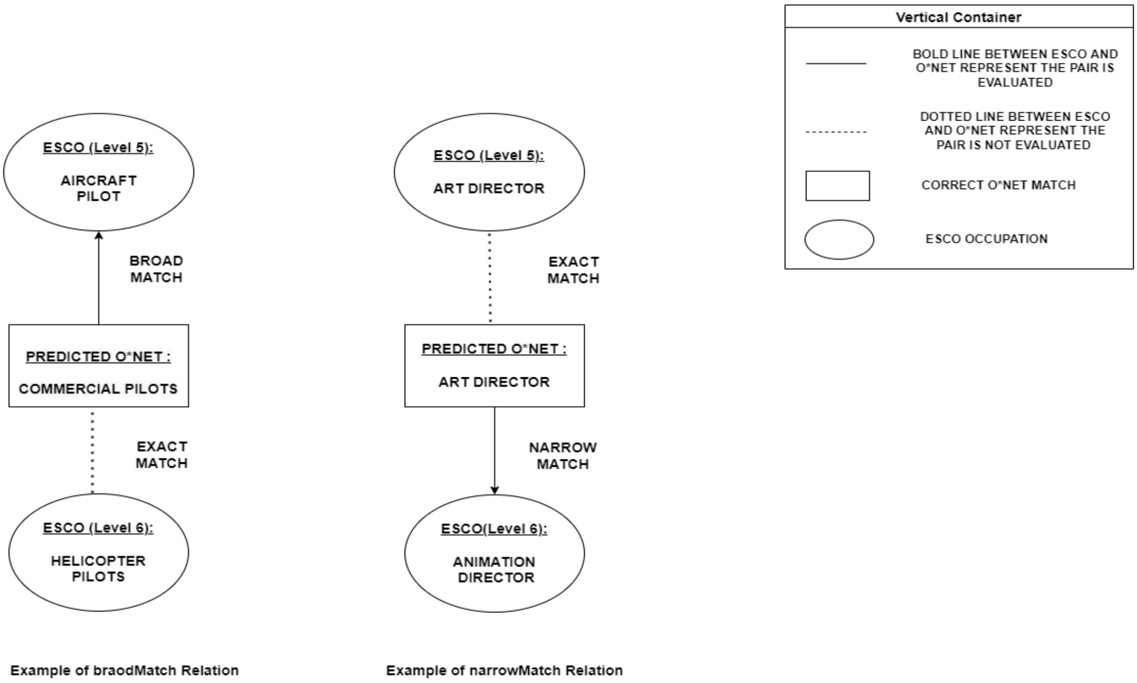
From the table 7.2 and 7.3 and from the discussion given in previous sections, we can say that the METHOD_1$_{ALL}$ and METHOD_2$_{ALL}$ has established appropriate relationships for majority of the matches in the case of *correct-predicted matches*. The interesting part of this analysis is that there is a chance of establishing appropriate relations even in the case of *incorrect-predicted matches*. This is greater for METHOD_1$_{ALL}$ compared to METHOD_2$_{ALL}$ which means that METHOD_1$_{ALL}$ has found close or appropriate matches for the ESCO occupations.

7.2.3.2 Analysis of `skos:broadMatch` and `skos:narrowMatch` Relations

Analysis of `skos:broadMatch` and `skos:narrowMatch` Relations of METHOD_1$_{ALL}$

Case 1: *correct-predicted matches* Three *correct-predicted matches* have a `skos:broadMatch` relationship and eleven *correct-predicted matches* have a `skos:narrowMatch` relationship.

Figure 7.9 shows an example of `skos:broadMatch`: `Aircraft Pilot` (ESCO) →`Commercial Pilots` (O*NET) with the relation `skos:broadMatch`. Because `Aircraft Pilot` (ESCO) and `Helicopter Pilot` (ESCO) are both matched to `Commercial Pilots` (O*NET), and `Helicopter Pilot` (ESCO) has the most semantic similarity with the O*NET occupation and is similar considering the description and other metadata, the `skos:broadMatch` relation was established.


Vertical Container
BOLD LINE BETWEEN ESCO AND O*NET REPRESENT THE PAIR IS EVALUATED
DOTTED LINE BETWEEN ESCO AND O*NET REPRESENT THE PAIR IS NOT EVALUATED
CORRECT O*NET MATCH
ESCO OCCUPATION
ESCO (Level 5): AIRCRAFT PILOT
BROAD MATCH
PREDICTED O*NET : COMMERCIAL PILOTS
EXACT MATCH
ESCO (Level 6): HELICOPTER PILOTS
ESCO (Level 5): ART DIRECTOR
EXACT MATCH
PREDICTED O*NET : ART DIRECTOR
NARROW MATCH
ESCO(Level 6): ANIMATION DIRECTOR

Example of braodMatch Relation

Example of narrowMatch Relation

Figure 7.9: Broad and Narrow Match relations for correct matches found

Example of `skos:narrowMatch`: `Animation Director` (ESCO) →`Art Directors` (O*NET) with the relation `skos:narrowMatch`. This relation was established because there exists an `Art Director` (ESCO) occupation which was also matched to `Art Directors` (O*NET) with higher semantic similarity score and was given the `skos:exactMatch` relation. This demonstrates that the use of hierarchy in the establishment of relationships was successful.

Case 2: ***incorrect-predicted matches*** Out of 50 *incorrect-predicted matches*, 2 matches have `skos:broadMatch` relationship and 4 matches have `skos:narrowMatch` relationship.

Example of `skos:broadMatch` : `credit adviser` (ESCO) →`Credit Analysts` (O*NET) with the relation `skos:broadMatch`. The correct match for this ESCO occupation is `Credit Counselors`. The `skos:broadMatch` relation was established because `credit adviser` (ESCO) and `credit analyst` (ESCO) are matched to `Credit Analysts` (O*NET) and `credit analyst` (ESCO) has the highest semantic similarity with the O*NET occupation and lies in a level below `credit adviser`.

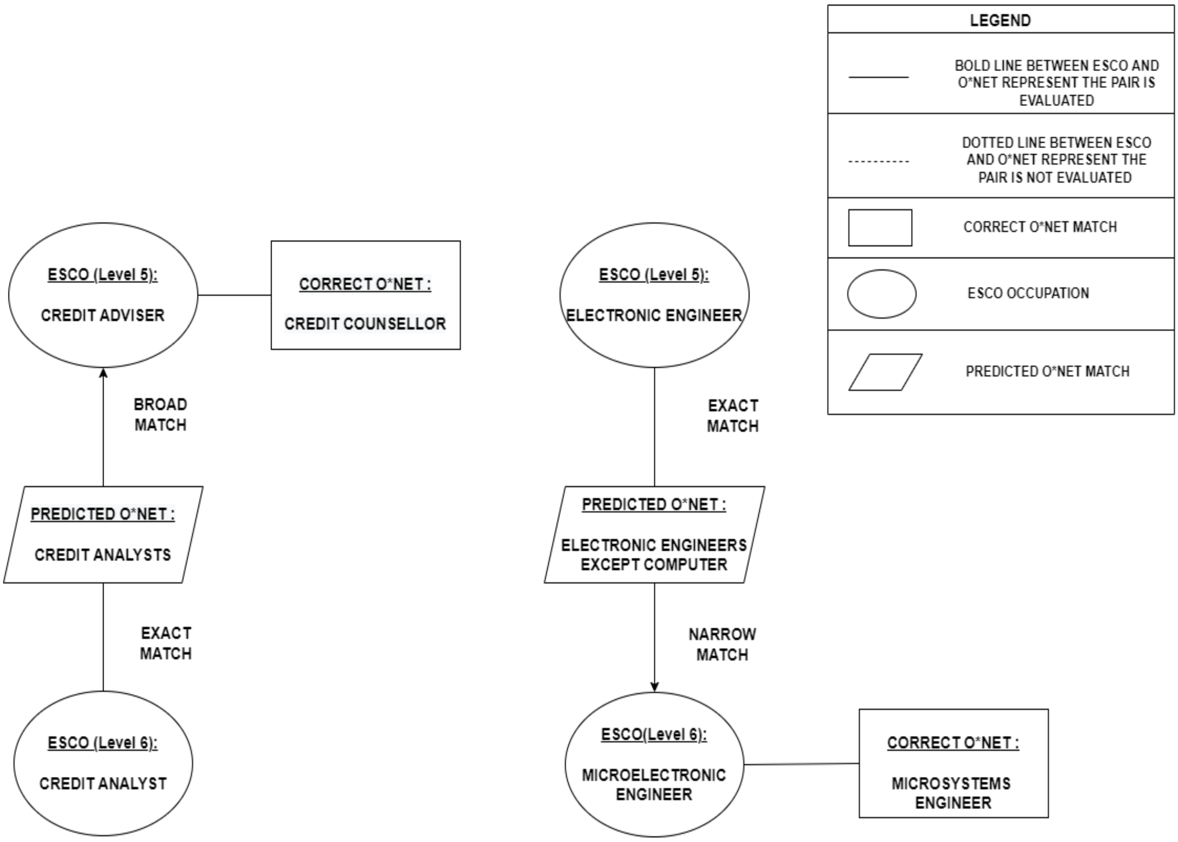

LEGEND
BOLD LINE BETWEEN ESCO AND O*NET REPRESENT THE PAIR IS EVALUATED
DOTTED LINE BETWEEN ESCO AND O*NET REPRESENT THE PAIR IS NOT EVALUATED
CORRECT O*NET MATCH
ESCO OCCUPATION
PREDICTED O*NET MATCH
ESCO (Level 5): CREDIT ADVISER
CORRECT O*NET : CREDIT COUNSELLOR
BROAD MATCH
PREDICTED O*NET : CREDIT ANALYSTS
EXACT MATCH
ESCO (Level 6): CREDIT ANALYST
ESCO (Level 5): ELECTRONIC ENGINEER
EXACT MATCH
PREDICTED O*NET : ELECTRONIC ENGINEERS EXCEPT COMPUTER
NARROW MATCH
ESCO(Level 6): MICROELECTRONIC ENGINEER
CORRECT O*NET : MICROSYSTEMS ENGINEER

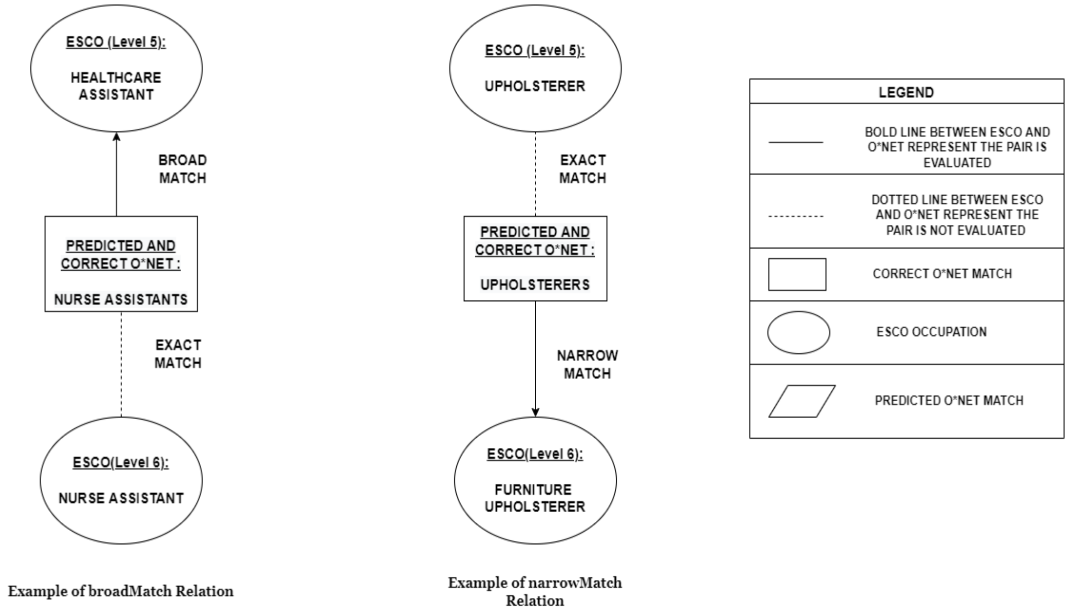
Figure 7.10: Broad and Narrow Match relations for incorrect matches found

Example of `narrowMatch` : `microelectronics engineer` (ESCO) →`Electronics Engineers, Except Computer` (O*NET); relation : `narrowMatch`. The correct match for this ESCO occupation is `Microsystems Engineers`. Considering the information corresponding between these occupations, the relationship established is appropriate. for example, the skill - “design integrated circuits” is common in both the ESCO and O*NET occupations.

Analysis of `skos:broadMatch` and `skos:narrowMatch` Relations of METHOD_2$_{\text{ALL}}$

Case 1: *correct-predicted matches* Out of 92 *correct-predicted matches* in this result set, 4 matches have `skos:broadMatch` relationship and 11 matches have `skos:narrowMatch` relationship.

An example of `skos:broadMatch` relation is shown in figure 7.11. Out of the 4 `skos:broadMatch` relations established, one of the relation was wrong and this was judged based on the `skos:exactMatch` relation established with the help of this match. This was `Civil Engineer` (ESCO) → `Civil Engineers` (O*NET) and the `skos:broadMatch` relation was established for this match. This O*NET occupation is also matched with ESCO’s `Water Engineer`, for which the `skos:exactMatch` relation is established since this match has a higher semantic similarity score than `Civil Engineer` (ESCO). Checking the label and description it was evident that this match was wrong which implies that the `skos:broadMatch` is also wrong. This example shows that even though the match is correct, the relation is wrong because of the decision to establish relationship only when two or more ESCO occupations are matched to same O*NET occupation. This is one of the limitations of the method of deriving relations.



Example of broadMatch Relation

Example of narrowMatch Relation

Figure 7.11: Broad and Narrow relations for correct matches

All of the 11 `skos:narrowMatch` relations found by this method are appropriate. All the ESCO occupations in this set are narrowed down from more general ESCO occupations. An example is given in figure 7.11. Other examples include `Corporate Lawyer` (ESCO) → `Lawyers` (O*NET) with the relation `skos:narrowMatch`, `Lawyer` (ESCO) → `Lawyers` (O*NET) with the relation `skos:exactMatch`. In this example the O*NET occupation is matched to two ESCO occupations and as the semantic similarity score between `Lawyer` (ESCO) → `Lawyers` is higher it has the `skos:exactMatch` relation. In this case, the method of deriving relations has established relations correctly.

Case 2: ***incorrect-predicted matches*** Out of 63 *incorrect-predicted matches*, 3 matches have the `skos:broadMatch` relationship and 2 matches have `skos:narrowMatch` relationship. Figure 7.12 depicts the examples of `skos:broadMatch` and `skos:narrowMatch`. All the matches in case have the appropriate relationship even if they are not the *correct* matches. This is based on the descriptions and skills of the occupations.

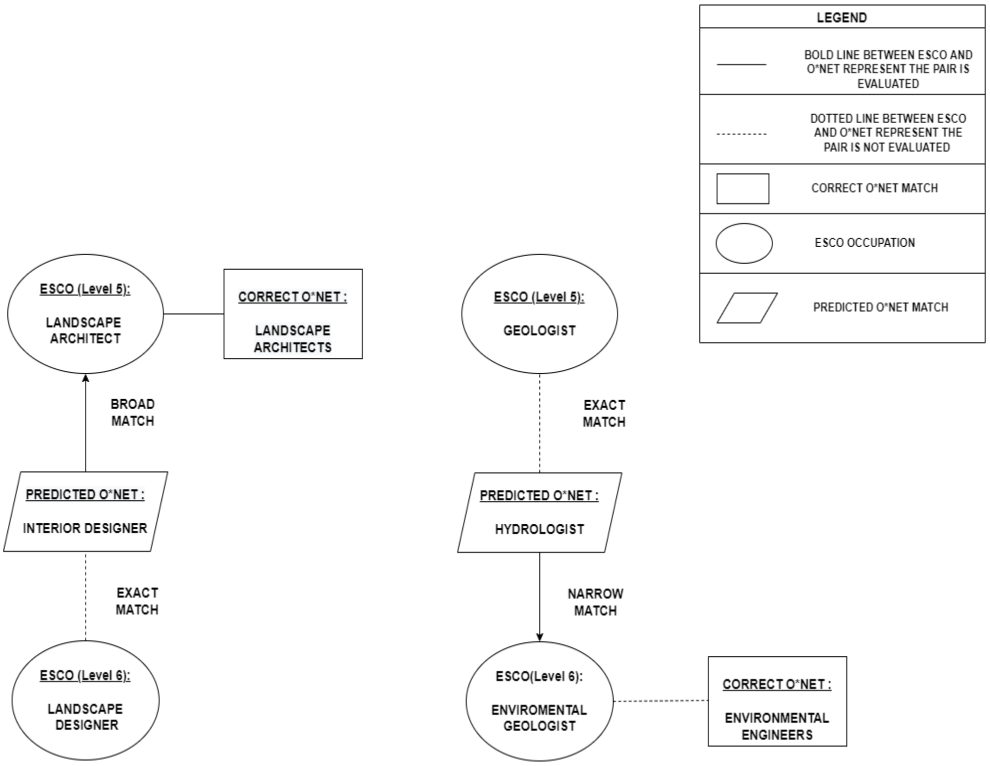

LEGEND
BOLD LINE BETWEEN ESCO AND O*NET REPRESENT THE PAIR IS EVALUATED
DOTTED LINE BETWEEN ESCO AND O*NET REPRESENT THE PAIR IS NOT EVALUATED
CORRECT O*NET MATCH
ESCO OCCUPATION
PREDICTED O*NET MATCH
ESCO (Level 5): LANDSCAPE ARCHITECT
CORRECT O*NET : LANDSCAPE ARCHITECTS
BROAD MATCH
PREDICTED O*NET : INTERIOR DESIGNER
EXACT MATCH
ESCO (Level 6): LANDSCAPE DESIGNER
ESCO (Level 5): GEOLOGIST
EXACT MATCH
PREDICTED O*NET : HYDROLOGIST
NARROW MATCH
ESCO(Level 6): ENVIROMENTAL GEOLOGIST
CORRECT O*NET : ENVIRONMENTAL ENGINEERS

Figure 7.12: Broad and Narrow relations for incorrect matches

Section Summary

METHOD_1$_{ALL}$: *correct-predicted matches* = 105, *incorrect-predicted matches* = 50					
	Case 1 *correct-predicted matches*			Case 2 *incorrect-predicted matches*	
	`skos:broadMatch`	`skos:narrowMatch`		`skos:broadMatch`	`skos:narrowMatch`
Total Number of Relations	3	11	Total Number of Relations	2	4
Number of Correct Relations	1	10	Number of Relations Completely Wrong	1	2
Number of Relations which can be Different Relation	2	1	Number of Relations which can be Correct/ Different Relation	1	2

Table 7.4: Analysis of `skos:broadMatch` and `skos:narrowMatch` relations of METHOD_1$_{ALL}$

METHOD_2$_{ALL}$: *correct-predicted matches* = 92, *incorrect-predicted matches* = 63					
	Case 1 *correct-predicted matches*			Case 2 *incorrect-predicted matches*	
	`skos:broadMatch`	`skos:skos:narrowMatch`		`skos:broadMatch`	`skos:narrowMatch`
Total Number of Relations	4	11	Total Number of Relations	3	2
Number of Correct Relations	1	11	Number of Relations Completely Wrong	0	1
Number of Relations which can be Different Relation	3	0	Number of Relations which can be Correct/ Different Relation	3	1

Table 7.5: Analysis of `skos:broadMatch` and `skos:narrowMatch` relations of METHOD_2$_{ALL}$

From the table 7.4 and 7.5, we can see that METHOD_1$_{\text{ALL}}$ and METHOD_2$_{\text{ALL}}$ has established `skos:narrowMatch` relationships appropriately in both cases while the `skos:broadMatch` relationships are mostly inappropriate and could be a different relation.

7.2.3.3 Analysis of `Match` and `NoMatch` Relations

The `Match` and `NoMatch` relations are established when the O*NET match is not shared with any other ESCO occupations within the same ISCO unit group. These relations depend entirely on the semantic similarity score.

Analysis of `Match` and `NoMatch` relations of METHOD_1$_{\text{ALL}}$

Case 1: *correct-predicted matches* As seen in figure 7.13, the `Match` relation is is correct for the first match. This relation is established because the overall semantic similarity score is above the threshold. The `NoMatch` relation is incorrect for the second relation even though it is the correct match because the overall semantic similarity score is below the threshold. As these relations depend only the score, it is crucial to select a threshold which can balance between finding the correct and incorrect relation.

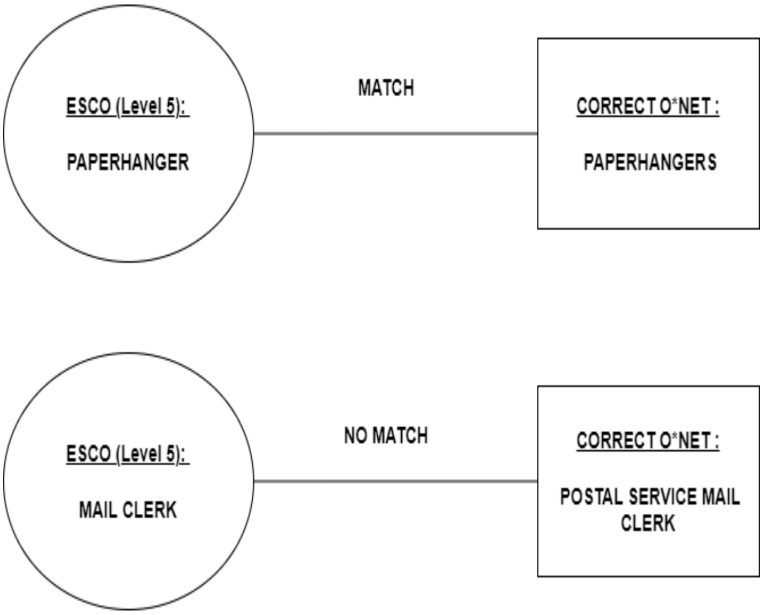


Figure 7.13: `Match` and `NoMatch`relations for correct-predicted matches

Case 2: *incorrect-predicted matches* As we can see in the figure 7.14, in some cases, the match found are related, but some matches found are totally unrelated.

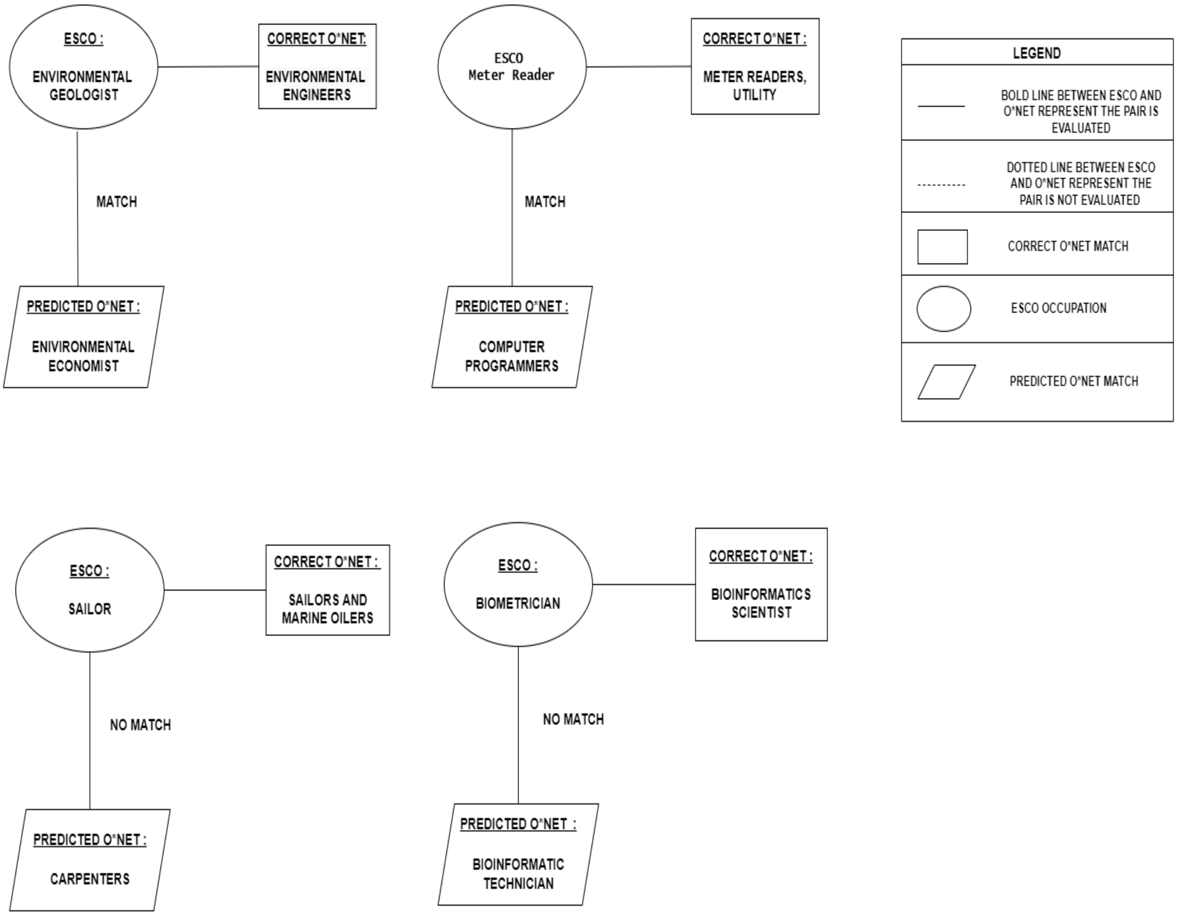


Figure 7.14: `Match` and `NoMatch`relations for incorrect-predicted matches

Analysis of `Match` and `NoMatch` Relations of METHOD_2$_{\mathrm{ALL}}$

Case 1: *correct-predicted matches* Out of the 92 correct-predicted matches, 23 of the ESCO-O*NET occupation pairs have the `Match` relation and 9 of the ESCO O*NET occupation pairs have `NoMatch` relation. An example of each of these relations are given in figure 7.15

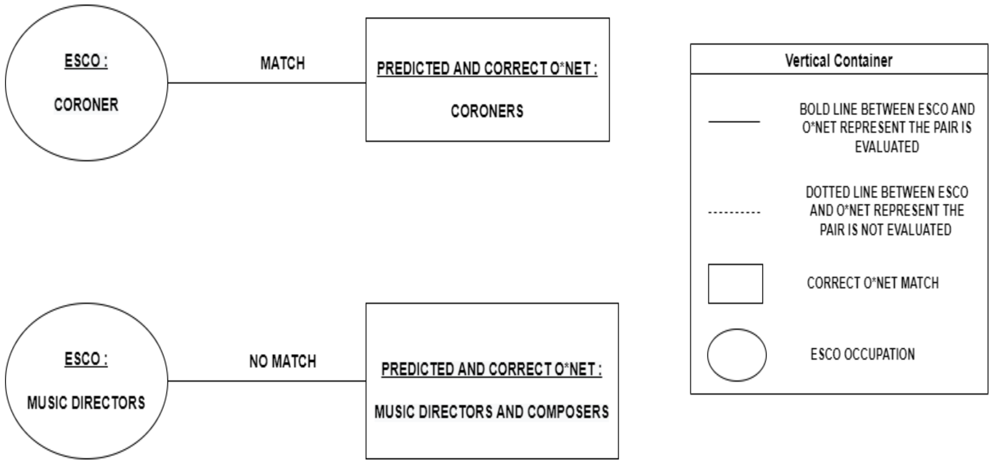

ESCO :
CORONER
MATCH
PREDICTED AND CORRECT O*NET :
CORONERS
ESCO :
MUSIC DIRECTORS
NO MATCH
PREDICTED AND CORRECT O*NET :
MUSIC DIRECTORS AND COMPOSERS
Vertical Container
BOLD LINE BETWEEN ESCO AND O*NET REPRESENT THE PAIR IS EVALUATED
DOTTED LINE BETWEEN ESCO AND O*NET REPRESENT THE PAIR IS NOT EVALUATED
CORRECT O*NET MATCH
ESCO OCCUPATION

Figure 7.15: `Match` and `NoMatch`relations

Case 2: *incorrect-predicted matches* Out of the 63 incorrect-predicted matches, 22 of them have the `Match` relation and 10 are `NoMatch` relation. To examine these relations, the O*NET occupation which was found and the correct match was compared. This gives an idea about how different or close the match was compared to the correct match. Out of the 22 `Match` relations, 11 of the pairs were completely wrong and the relation should have been a `NoMatch`. The other 11 ESCO-O*NET occupation pairs were close to the correct match and an appropriate match for these pairs would have been the `skos:closeMatch` relation. On the other hand, all the `NoMatch` relations established by this method were indeed wrong and the relation established was correct. Examples of the `Match` and `NoMatch` relations are given in figure 7.17 AND 7.16

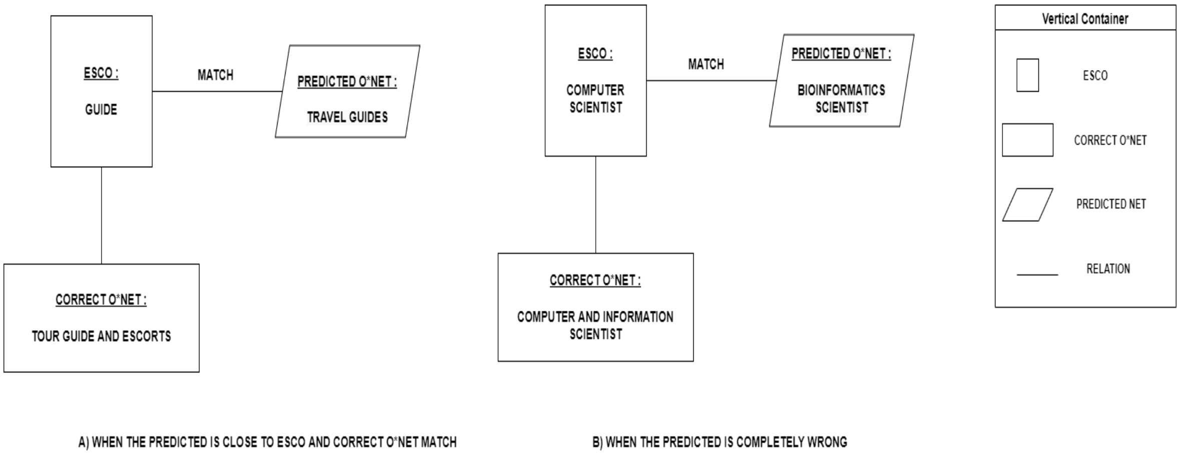


Figure 7.16: `Match` relations of incorrect-predicted matches

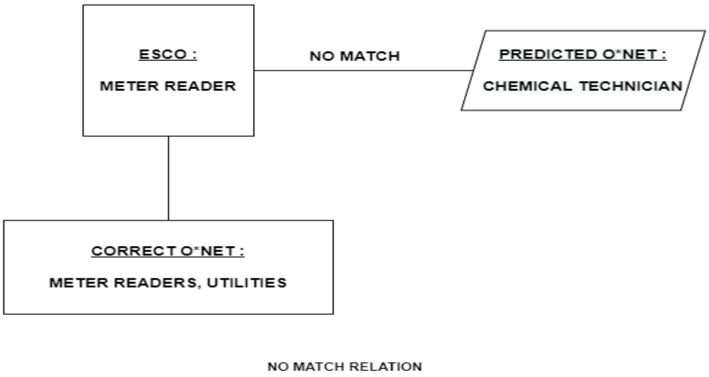


Figure 7.17: `NoMatch` relations relation incorrect-predicted matches

Section Summary

METHOD_2$_{ALL}$ *correct-predicted matches* = 105, *incorrect-predicted matches* = 50					
	Case 1 *correct-predicted matches*			Case 2 *incorrect-predicted matches*	
	Match	NoMatch		Match	NoMatch
Total Number of Relations	18	21	Total Number of Relations	7	17
-	-	-	Number of Relations Completely Wrong/*correct	6	12*
Number of Relations which can be Different Relation	0	21	Number of Relations which can be Different Relation	1	5

Table 7.6: Analysis of `Match` and `NoMatch` relations of METHOD_1$_{ALL}$

METHOD_2$_{ALL}$: *correct-predicted matches* = 92, *incorrect-predicted matches* = 63					
	Case 1 *correct-predicted matches*			Case 2 *correct-predicted matches*	
	Match	noMatch		Match	noMatch
Total Number of Relations	23	9	Total Number of Relations	22	10
-	-	-	Number of Relations Completely Wrong/*correct	14	10*
Number of Relations which can be Different Relation	23	9	Number of Relations which can be Different Relation	8	0

Table 7.7: Analysis of `Match` and `NoMatch` relations of METHOD_2$_{ALL}$

The tables 7.6 and 7.7 give the number of relations that can be a different relation in METHOD_1$_{ALL}$ and METHOD_2$_{ALL}$. In the first case- *correct-predicted matches* of both methods, when all the matches are already evaluated it is difficult to establish a relation without in-depth knowledge of occupations and support of structure. So all the *correct-predicted matches* are considered as the match which can have any of the relations.

The interesting case is *incorrect-predicted matches* because here we can analyze the matches and find if there could be a chance for different relation. For the relation `Match` in METHOD_1$_{ALL}$, there are 6 matches which are completely wrong, meaning that they cannot be a match as they differs completely. There was one match - `Librarian` (ESCO) →`Library Technician` (O*NET) which can have a relation as these occupations belong to the same line of work, which is library. Whereas

in METHOD_2$_{ALL}$, there are 14 matches which are wrong and cannot have a relation and 8 matches which can have any relationship. An example of this is, Forest Worker (ESCO) →Forest Conservation Technician (O*NET); the *correct match* found by the annotator for this ESCO occupation is *Forest and Conservation Workers*. For the NoMatch relation in METHOD_1$_{ALL}$, there were 12 *incorrect-predicted matches* with NoMatch relation which means that they have the correct relation (this is represented by the asterisk*). Five of the *incorrect-predicted matches* can have a different match which means that they are related in some way. An example of this is, Fire Commissioner (ESCO) →Firefighters (O*NET); the *correct match* found by the annotator for this ESCO occupation is *Fire Inspectors and Investigators*. In the METHOD_1$_{ALL}$, none of the matches could be related with a different relation. This means that the relation - NoMatch was correct.
From this analysis we can see that METHOD_1$_{ALL}$ has more predicted matches which can have different relations even if the predicted match is incorrect whereas, METHOD_2$_{ALL}$ has the right NoMatch relation established but not did not perform well with Match relations.

7.3 General Discussion

The above sections discussed ESCO-O*NET occupation matches and also the relations that are established between them. Here are some of the observations that were not mentioned explicitly before.

1. Most of the ESCO occupations were matched with the O*NET occupation Nannies and most of the managerial occupations of ESCO are matched to Marketing Managers of O*NET. This could be because of the more elaborate description given compared to the length of the descriptions of other occupations.

2. **METHOD_3$_{LD}$** did not find a O*NET occupation match for all the ESCO occupation as the anchor ontology was limited. This also restricted in finding a relationship between the occupations.

3. When looking at the semantic relations that were established, although the correct match was not found, the analysis revealed that there is a chance that different relationship can be established which can be helpful. For example:

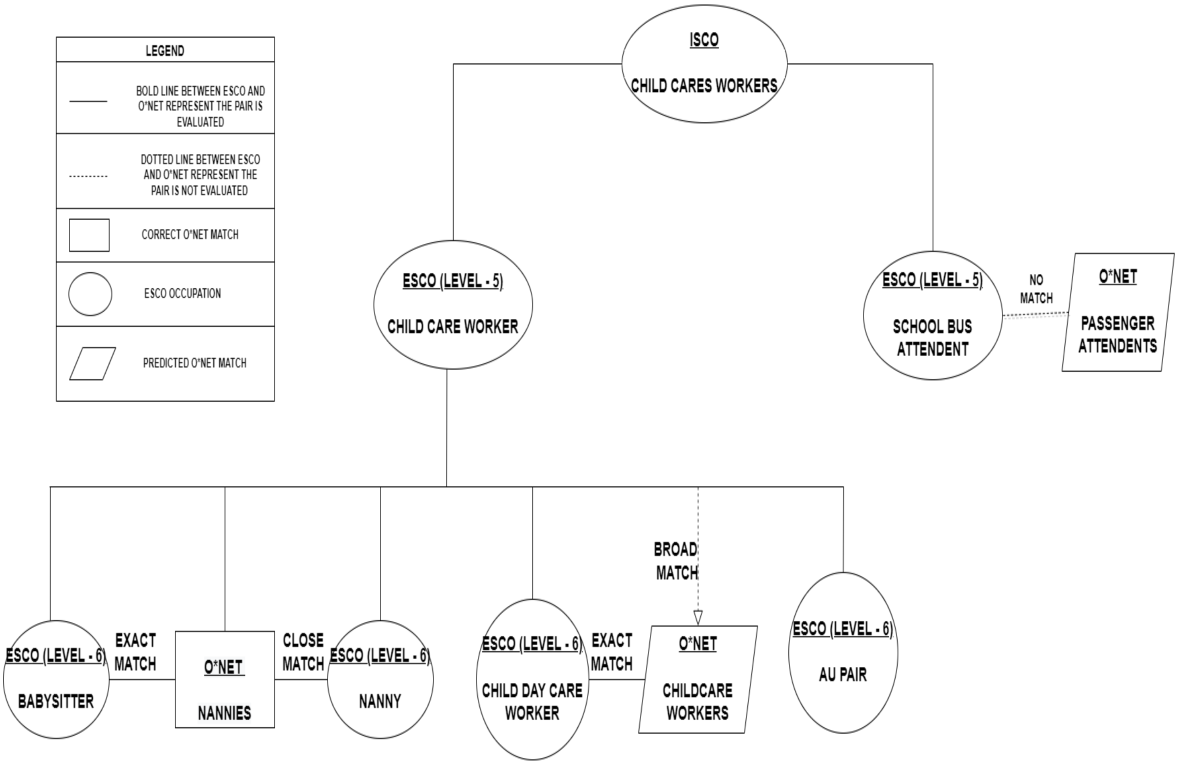


Figure 7.18: 5311 ISCO group when O*NET with corresponding O*NET matches

From figure 7.18 when looking at the relations, we can see where the O*NET occupations can be placed within an ISCO unit group. The usage of structure and similarity score helped in placing the O*NET occupations in the right place. If the *locality principle* - which says that if the Parent node is matched then all the child nodes are also matched - was followed in finding the relations then all the occupations below a level would have a relation of broad/narrow. As we can see from the figure above, even within the group, there are occupations like `Nannies` and `Child Day Care Worker` which are different are in the same level. The relation established with the O*NET occupation `Childcare Workers` (O*NET) and `Child Day Care Worker` (ESCO) and `Childcare Worker` (ESCO) is much more informative and appropriate than deriving relation between texttt**Babysitter** (ESCO) →`Childcare Workers` (O*NET) because the parent node - `Childcare Worker` (ESCO) is matched to `Childcare Workers` (O*NET).

7.3.1 Limitations

In this section the limitation found in the methods of matching process and deriving relations are detailed.

- Evidently from the results of the experiments, the methods which used domain-specific knowledge to did not perform well based on the evalaution of the samples. **METHOD_2$_{\text{LD}}$** and **METHOD_2$_{\text{ALL}}$** used the XLNet model with extended domain-specific vocabulary. It is very difficult to conclude if this did not performed well for all the ESCO occupations as only limited number of samples were evaluated. A limitations that was found in using information from job description on online job portals is that these information are concentrated more on Information Technology and Engineering jobs rather than for other professions like daily wage workers or teaching jobs. These occupations are not posted online and therefore the information related to these occupations are missing. Some other potential reasons that the model did not perform well are,

  - XLNet model is not suitable and adapting to extended vocabulary.
  - The number of tokens added to the vocabulary is excessive which increases the distribution of token representation.

- The anchor ontology - Wikidata - used in **METHOD_3$_{\text{LD}}$** was very limited and did not help in finding a match to all the ESCO occupations.

- In the method of deriving relations, considering only the many-to-one match cases has helped in all the cases but considering the semantic similarity score in this process has it's own advantage and disadvantage. In most cases the relations are correct and in some cases the relations could be interchanged. Fro example, `Dog trainer` (ESCO) →`Animal Trainer` (O*NET) and `Animal Trainers` (ESCO) →`Animal Trainers` (O*NET) has the `skos:exactMatch` relation because both these matches have the same overall semantic similarity score. But, the correct relationship should be `narrowMatch` and `skos:exactMatch`. This can be achieved by setting the correct weights for the attributes while calculating the overall semantic similarity score. It is very crucial to find the correct weights which can perform well.

- The `Match` and `NoMatch` relations depend only on the overall semantic similarity score and therefore, it was difficult to find a threshold to find the correct matches. The current threshold of 0.6 used on the score has given more correct relations as compared to the previous thresholds which established more `NoMatch` relations if the threshold was high or more `Match` relation with lower threshold and resulted with more wrong relations.

- There were many `Match` relations which could potentially be a `skos:exactMatch` relation considering how the occupations correspond with regards to the label, description an other information. Since, `skos:exactMatch` relations are established only in the case of many-to-one matching, it was not established independently. This is one of the disadvantages of making the relations dependent on each other.

Chapter 8

Conclusion and Future Work

8.1 Conclusion

In this thesis, we computed the semantic similarity between the natural language text information available for the ESCO and O*NET occupations in order to improve the matching process. According to the findings of the literature review, using domain-specific knowledge should improve the matching process by bridging the uncertainty between the occupational information text and hence improve the matching process. A relationship was established after a match was found, providing more information regarding the type of relationship between the occupations. Three primary research questions were used to accomplish this and a domain expert examined a sample of 200 match results. In the following section we answer the research questions.

RQ1: *How can we improve matching between ESCO and O*NET occupation using domain-specific background knowledge?* This question was further divided into sub-questions. The answer to the RQ1.a sub-question which used the generic XLNet model in the matching process was used as the baseline and the other two sub-questions investigate the effect of using domain-specific knowledge in two ways, *(a) extending the vocabulary of XLNet model with domain-specific information [RQ1.b]* and *(b) by using Wikidata as helper domain-specific ontology - [RQ1.c]*. A sample of 200 ESCO occupations and its corresponding top 5 O*NET occupation matches were annotated by domain expert. 155 out of the 200 ESCO occupations had a correct match within the top 5 and the other 45 ESCO matches did not find any correct match. The 155 ESCO-O*NET occupation matches were used as the ground truth to evaluate the three methods. 139 out of 155 correct ESCO-O*NET matches were correct in the results obtained by using generic XLNet model and 125 correct matches were found in the results obtained using XLNet model with extended vocabulary. The third method of using a domain-specific ontology as an anchor found correct match for 91 out of 155 correct ESCO -O*NET matches. As 50% of the correct ESCO-O*NET occupation matches were found in the results of using generic XLNet model and also in the results of using domain-specific knowledge.Moreover, the small sample size made it difficult to investigate the effect in depth. Considering the results of the samples, it is best to use the generic XLNet model to create the embeddings and finally calculate the semantic similarity between the occupations.

RQ2: *How does using various metadata like skills and alternate label related to the occupations improve matching between ESCO and O*NET occupation?* The impact of using other available metadata, such as skills and alternate label information for occupations is investigated in this question. Looking at the results, it was clear that incorporating these information had a significant impact, and it can be concluded that using skills and alternate labels in the matching process is important in finding correct matches.

RQ3: *How can we establish different types of relations between the ESCO and O*NET occupations using semantic similarity and taxonomic structure of ontology?* When looking at the relations established between the occupations based on the structure of ESCO and the semantic similarity, we can understand how they are related to each other. The usage of structure does not seem to have any limitations with regards to the relations but the usage of semantic similarity score has established wrong `Match` and `NoMatch` relations in most cases. The other relations are correct most of the time and in some cases the relations established can be a different relation which can be achieved by adjusting the weights while calculating the semantic similarity score. One of the main rules in establishing relations was to consider only the many-to-one matching cases and this seems to have given good results for all the relations but limiting the `skos:exactMatch` relation has some disadvantages. Overall, the relations give a clear view into where the O*NET occupation can fit within the ESCO ontology and help in merging the ESCO and O*NET ontologies.

Overall, it can be concluded that ontology matching of ESCO and O*NET using the semantic similarity between the occupations of these ontologies and establishing a relationship between the occupations has given promising results which has improved automatic matching compared to the manual crosswalk which were created previously. The XLNet model has given the best result in finding a correct match and then expressing the match with semantic relations has added more weight to the matching. Evaluation of these matching was a challenging task due to the lack of ground truth. With the help of domain experts from ESCO or O*NET, the matching process can be improved much more and used in applications like job matching for job seekers.

8.2 Future Work

There were two components that made up this thesis. One, the matching process and second, establishing a relationship between the matched occupations. Some of the aspects that can be enhanced in the future are.

8.2.1 Matching Process

- The matching process with the generic XLNet model outperformed other methods. This approach can be improved further by considering the *padding* operation, which can aid in dealing with input text of varying lengths.
- As indicated in section 7, many ESCO and O*NET occupations have a high string similarity between the occupation labels. A combination of string similarity and semantic similarity could help in matching.
- The domain-specific XLNet model used in this thesis did not perform well due to limited source and less number of samples in the evaluation. Further analysis can help in knowing the in-depth effect of using domain-specific knowledge.

8.2.2 Semantic Relations

The established semantic relations have offered a clear understanding of how O*NET occupations relate to ESCO occupations. It still has room for improvements. For instance,

- In the process of establishing a relationship, it would be interesting to leverage the rank at which the correct match was discovered. The rank at which the correct match was found can be used to determine a relation.
- The current method establishes a `skos:exactMatch` relation only when the O*NET occupation is shared by more than one ESCO occupation. It would be beneficial if a rule was defined to find this relation even if the O*NET occupation is matched to only one ESCO occupation. This is because there were several ESCO-O*NET occupation matches with `Match` and `NoMatch` relations which could potentially be a `skos:exactMatch` relation.

Bibliography

[1] C. Ospino, “Occupations: Labor market classifications, taxonomies, and ontologies in the 21st century,” *Inter-American Development Bank*, Sep 2018.

[2] T. Gruber, *Ontology*. New York, NY: Springer New York, 2016, pp. 1–3. [Online]. Available: https://doi.org/10.1007/978-1-4899-7993-3_1318-2

[3] R. Rentzsch and M. Staneva, “”skills matching” and ”skills intelligence” through curated and data-driven ontologies,” in *Proceedings of DELFI Workshops 2020*. Bonn: Gesellschaft für Informatik eVz, 2020.

[4] C. Kuptsch and P. Martin, “Actors and factors in the internationalization of labour markets,” *The internationalization of labour markets*, p. 115–134, 2010.

[5] A. Emmel and T. Cosca, “Occupational classification systems: Analyzing the 2010 standard occupational classification (soc) revision,” 2010.

[6] S. Neutel and M. H. T. de Boer, “Towards automatic ontology alignment using BERT,” in *Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021), Stanford University, Palo Alto, California, USA, March 22-24, 2021*, ser. CEUR Workshop Proceedings, A. Martin, K. Hinkelmann, H. Fill, A. Gerber, D. Lenat, R. Stolle, and F. van Harmelen, Eds., vol. 2846. CEUR-WS.org, 2021. [Online]. Available: http://ceur-ws.org/Vol-2846/paper28.pdf

[7] K. Kanders, J. Djumalieva, C. Sleeman, and J. Orlik, *Mapping Career Causeways: Supporting workers at risk*. [Online]. Available: https://www.nesta.org.uk/report/mapping-career-causeways-supporting-workers-risk/

[8] B. Chandrasekaran, J. Josephson, and V. R. Benjamins, “What are ontologies, and why do we need them?” *Intelligent Systems and their Applications, IEEE*, vol. 14, pp. 20 – 26, 02 1999.

[9] J. Euzenat and P. Shvaiko, *Ontology Matching*, 2nd ed. Springer Publishing Company, Incorporated, 2013.

[10] A. Miles and S. Bechhofer, "Skos simple knowledge organization system reference," Aug 2009. [Online]. Available: http://www.w3.org/TR/2009/REC-skos-reference-20090818/

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1301.html#abs-1301-3781

[12] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [Online]. Available: https://aclanthology.org/D14-1162

[13] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017. [Online]. Available: https://aclanthology.org/Q17-1010

[14] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018, cite arxiv:1802.05365Comment: NAACL 2018. Originally posted to openreview 27 Oct 2017. v2 updated for NAACL camera ready. [Online]. Available: http://arxiv.org/abs/1802.05365

[15] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," 2018. [Online]. Available: https://arxiv.org/abs/1801.06146

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[18] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[19] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 353–355. [Online]. Available: https://aclanthology.org/W18-5446

[20] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 66–71. [Online]. Available: https://aclanthology.org/D18-2012

[21] K. Ramar and G. Gurunathan, "Technical review on ontology mapping techniques," vol. 15, pp. 676–688, 01 2016.

[22] M. Kamdar, T. Tudorache, and M. Musen, "A systematic analysis of term reuse and term overlap across biomedical ontologies," *Semantic Web – Interoperability, Usability, Applicability*, vol. 1, pp. 1–5, 02 2016.

[23] X. Liu, Q. Tong, X. Liu, and Z. Qin, "Ontology matching: State of the art, future challenges, and thinking based on utilized information," *IEEE Access*, vol. 9, pp. 91 235–91 243, 2021.

[24] Z. Aleksovski, W. ten Kate, and F. van Harmelen, "Ontology matching using comprehensive ontology as background knowledge," in *Proceedings of the International Workshop on Ontology Matching at ISWC 2006*, P. S. et al., Ed. CEUR, 2006, pp. 13–24.

[25] M. Sabou, M. d'Aquin, and E. Motta, "Using the semantic web as background knowledge for ontology mapping," in *The 1st International Workshop on Ontology Matching (OM-2006)*, 2006, oM-2006 Ontology Matching Proceedings of the 1st International Workshop on Ontology Matching (OM-2006) Collocated with the 5th International Semantic Web Conference (ISWC-2006) Athens, Georgia, USA, November 5, 2006. Edited by Pavel Shvaiko, Jérôme Euzenat, Natalya Noy, Heiner Stuckenschmidt, Richard Benjamins, Michael Uschold. [Online]. Available: http://oro.open.ac.uk/23566/

[26] X. Su and J. A. Gulla, “Semantic enrichment for ontology mapping,” in *Natural Language Processing and Information Systems*, F. Meziane and E. Métais, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 217–228.

[27] L. L. Wang, C. Bhagavatula, M. Neumann, K. Lo, C. Wilhelm, and W. Ammar, “Ontology alignment in the biomedical domain using entity definitions and context.” in *BioNLP*, D. Demner-Fushman, K. B. Cohen, S. Ananiadou, and J. Tsujii, Eds. Association for Computational Linguistics, 2018, pp. 47–55. [Online]. Available: https://aclweb.org/anthology/W18-2306

[28] I. G. Husein, S. Akbar, B. Sitohang, and F. N. Azizah, “Review of ontology matching with background knowledge,” *2016 International Conference on Data and Software Engineering (ICoDSE)*, pp. 1–6, 2016.

[29] A. Annane, Z. Bellahsene, F. Azouaou, and C. Jonquet, “Building an effective and efficient background knowledge resource to enhance ontology matching,” *Journal of Web Semantics*, vol. 51, pp. 51–68, Aug. 2018. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01809627

[30] P. Lambrix, H. Tan, and Q. Liu, “Sambo and sambodtf results for the ontology alignment evaluation initiative 2008,” in *Proceedings of the 3rd International Conference on Ontology Matching - Volume 431*, ser. OM’08. Aachen, DEU: CEUR-WS.org, 2008, p. 190–198.

[31] Y. Jean-Mary and M. Kabuka, “Asmov: Results for oaei 2008,” *CEUR Workshop Proceedings*, vol. 431, pp. 132–139, Dec. 2008, 3rd International Workshop on Ontology Matching, OM-2008 - Collocated with the 7th International Semantic Web Conference, ISWC-2008 ; Conference date: 26-10-2008 Through 26-10-2008.

[32] J. Euzenat and P. Valtchev, “An integrative proximity measure for ontology alignment,” in *Proc. ISWC-2003 workshop on semantic information integration*, ser. Proc. ISWC-2003 workshop on semantic information integration. Sanibel Island, United States: No commercial editor., Oct. 2003, pp. 33–38, euzenat2003h. [Online]. Available: https://hal.inria.fr/hal-00922318

[33] Y. Zhang, X. Wang, S. Lai, S. He, K. Liu, J. Zhao, and X. Lv, “Ontology matching with word embeddings,” in *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, M. Sun, Y. Liu, and J. Zhao, Eds. Cham: Springer International Publishing, 2014, pp. 34–45.

[34] M. Tounsi Dhouib, C. Faron Zucker, and A. G. B. Tettamanzi, “An ontology alignment approach combining word embedding and the radius measure,”

in *Semantic Systems. The Power of AI and Knowledge Graphs*, M. Acosta, P. Cudré-Mauroux, M. Maleshkova, T. Pellegrini, H. Sack, and Y. Sure-Vetter, Eds. Cham: Springer International Publishing, 2019, pp. 191–197.

[35] Z. Liu, D. Huang, K. Huang, Z. Li, and J. Zhao, "Finbert: A pre-trained financial language representation model for financial text mining," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, ser. IJCAI'20, 2021.

[36] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, pp. 1234 – 1240, 2020.

[37] K. Huang, J. Altosaar, and R. Ranganath, "Clinicalbert: Modeling clinical notes and predicting hospital readmission," *ArXiv*, vol. abs/1904.05342, 2019.

[38] S. Bana, E. Brynjolfsson, D. Rock, and S. Steffen, "job2vec: Learning a representation of jobs," *Stanford Digital Economy Lab*. [Online]. Available: https://conference.iza.org/conference_files/DATA_2021/bana_s26582.pdf

[39] M. Hoen, "Occupational crosswalk, o*net characteristics, and occupation data," 01 2016.

[40] W. Hardy, R. Keister, and P. Lewandowski, "Do entrants take it all? The evolution of task content of jobs in Poland," *Ekonomia journal*, vol. 47, 2016. [Online]. Available: https://ideas.repec.org/a/eko/ekoeko/47_23.html

[41] P. L. Wojciech Hardy, Roma Keister, "Technology or Upskilling? Trends in the Task Composition of Jobs in Central and Eastern Europe," HKUST Institute for Emerging Market Studies, HKUST IEMS Working Paper Series 2016-40, Dec. 2016. [Online]. Available: https://ideas.repec.org/p/hku/wpaper/201640.html

[42] M. Somers, "The changing demand for skills in the netherlands," May 2019, 16th Belgian Day for Labour Economists, BDLE ; Conference date: 24-05-2019 Through 24-05-2019. [Online]. Available: http://roa.sbe.maastrichtuniversity.nl/?p=15747

[43] S. Raunich and E. Rahm, "Atom: Automatic target-driven ontology merging," 05 2011, pp. 1276 – 1279.

[44] F. Giunchiglia, P. Shvaiko, and M. Yatskevich, "S-match: an algorithm and an implementation of semantic matching," in *The Semantic Web: Research and Applications*, C. J. Bussler, J. Davies, D. Fensel, and R. Studer, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 61–75.

[45] P. Arnold and E. Rahm, “Enriching ontology mappings with semantic relations,” *Data Knowledge Engineering*, vol. 93, pp. 1–18, 2014, selected Papers from the 17th East-¬-European Conference on Advances in Databases and Information Systems. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169023X14000603

[46] Y. He, J. Chen, D. Antonyrajah, and I. Horrocks, “Bertmap: A bert-based ontology alignment system,” *CoRR*, vol. abs/2112.02682, 2021. [Online]. Available: https://arxiv.org/abs/2112.02682

[47] B. C. Grau, I. Horrocks, Y. Kazakov, and U. Sattler, “A logical framework for modularity of ontologies,” in *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, ser. IJCAI’07. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2007, p. 298–303.

[48] *ESCO handbook*. European Commission. [Online]. Available: https://ec.europa.eu/esco/portal/document/en/0a89839c-098d-4e34-846c-54cbd5684d24

[49] *ESCO ServicePlatform: Data model*. European Commission. [Online]. Available: https://ec.europa.eu/esco/portal/document/en/87a9f66a-1830-4c93-94f0-5daa5e00507e

[50] *ESCO Implementation manual*. European Commission. [Online]. Available: https://ec.europa.eu/esco/resources/escopedia/20181213_145926/cdd888b6-73d9-47f6-813b-3f29dfc0919c05_A_Annex_Draft_ESCO_Implementation_manual.pdf

[51] “The o*net® content model.” [Online]. Available: https://www.onetcenter.org/content.html

[52] “The o*net-soc taxonomy.” [Online]. Available: https://www.onetcenter.org/taxonomy.html

[53] J. Burrus, T. Jackson, N. Xi, and J. Steinberg, “Identifying the most important 21st century workforce competencies: An analysis of the occupational information network (o*net),” *ETS Research Report Series*, vol. 2013, pp. i–55, 12 2013.

[54] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, “Aligning books and movies: Towards story-like visual explanations by watching movies and reading books,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 19–27.

[55] W. Tai, H. T. Kung, X. Dong, M. Comiter, and C.-F. Kuo, "exBERT: Extending pre-trained models with domain-specific vocabulary under constrained training resources," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1433–1439. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.129

[56] Z. Aleksovski and F. Harmelen, "Using multiple ontologies as background knowledge in ontology matching," *CEUR Workshop Proceedings*, vol. 351, 01 2008.

[57] "Esco-mathematicians, actuaries and statisticians." [Online]. Available: http://data.europa.eu/esco/isco/C2120