Matching ontologies with distributed word embeddings

Harmen Prins

July 6, 2016

[1]

[1] Credit: Sanna Dinh

Acknowledgements

Arjen de Vries, professor at the Radboud University of Nijmegen and my mentor - Kennis enthusiasme en als het tegenzat kon hij me toch motiveren. Sanna Dinh, my partner - Nachtenlang opgebleven om het schrijfwerk en plaatjes Jeroen Vuurens, Delft University of Technology - Sparren over word2vec Xander Wilcke - Inspireren over het Semantisch Web Marianne - op de gekste tijdstippen skypen en proof-readen Suzan Verberne, professor at the Radboud University of Nijmegen - Nakijken Flynn, my pug - For the fresh air when I needed a minute to relax

Abstract

McCallum proposed using distributed word embeddings for knowledge base completion, the problem of finding relations *within* a knowledge base. In this study distributed word embeddings were used for ontology matching, the problem of finding relations *between* knowledge bases. To test if distributed word embeddings also work on ontology matching, multiple algorithms that adapt these embeddings to ontology graphs were implemented.

The conclusion is that distributed word embeddings can be applied to ontology matching. The performance is not as good as the state-of-the-art methods yet. However, distributed word embeddings have potential to improve with more data as opposed to current methods which are hand-crafted.

Summary

Contents

1 Introduction

It is a Friday afternoon and you realise you want to watch a movie with some friends this weekend. So, what do you do? To achieve this goal you must take three steps. These steps are also shown in Figure 1.

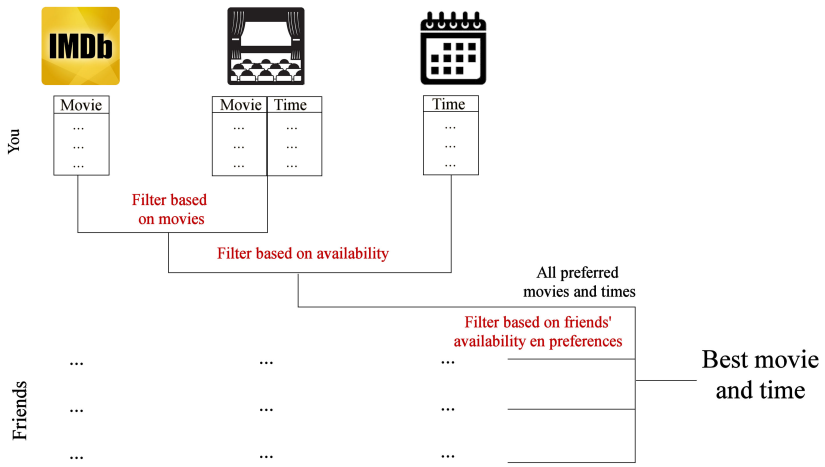


Figure 1: To find the best movie, the shown movies need to be filtered by the ones you like and the showtimes need to be filtered by the times you are available. All remaining showtimes need to be cross-references with the preferences of your friends.

Firstly, search for movies that you like that are also playing this weekend. To do this you must either manually go through the list of movies playing this weekend and check if any movies on that list match your preferences. In Figure 1, this is called "Filter based on movies".

Once you have found one or more movies that you like, you have to select a time at which you can see the candidate movies, so you check for every movie every time slot at every cinema near you. The number of time slots is equal to the number of movies you chose in the previous step, times the number of cinemas near you times the number of times every movie is shown at one cinema. You also have to filter the out time slots based on your availability, so you cross-reference the available movie showtimes with your agenda. This step is called "Filter based on availability" in Figure 1

Then onto the third and final step, getting friends to go with you. In this step you do not want to burden your friends too much by having them pick both the movie and the time, so you will have to decide which movie to watch or when, even though you do not know if your friends will like this movie or are available at that time. Once you contact your friends, most of them do not respond as they are not online at that time or they do not want to fill out an

availability poll or movie choices. This is called "Filter based on friends" in Figure 1.

Most of the time people will limit themselves to suboptimal options by avoiding these steps, considering only a few movies, cinemas, time slots and friends. But limitations like this are not necessary. What would happen if the above process was automated?

The first step, finding movies that match two criteria would be considered trivial with modern technology. One finds the two lists of movies that are recommended for you and movies that are playing this weekend and intersect them. A ranking can be added based on how expensive the tickets are, how certain it is that you may like the movie, and other factors.

The second step, cross-referencing the list of movies with cinema play times, would already be harder. Most of the time the cinema play times are written in a human-readable format, something that is often hard to interpret for a computer. Either the play times have to be written in or transformed into a computer-readable format manually or a computer has to interpret the format and convert it into a computer-readable one. However, errors might occur because the local cinema labelled a sequel as "Movie II" whereas the list we have has it labelled as "Movie 2", for example.

Checking an agenda to find an appropriate time would be easy if the times from the cinemas can be transformed to the same format as your agenda. If we are automating the process, we may be able to add more features, like changing standing appointments (of course incurring a penalty to the score of that time spot) and such.

The last step, contacting friends can be done in one click. Let us call the program that automated the first three steps a *softbot* and assume they are universal, i.e. everyone has one. Rather than bothering your friends with scheduling, your softbot can contact the softbots of your friends and find the movie and time slot that optimize a cost function, taking into account which friends you like the most, which movie everybody likes the most and at which time the fewest dinners with wives need to be moved. Once this best time slot is chosen, everyone gets an invite to see the movie and once it is known who will go, a car pooling route is calculated and you can sit back an enjoy your Friday afternoon with no planning or scheduling required.

The most amazing thing about this story, is that it is already possible with the current technology. The algorithms to convert for example cinema website text to a computer-readable format exist. Many open databases already contain a lot of structured data that can be used to reason about data. And algorithms to combine these pieces of information to make decisions already exist and work.

Two things are required before this web of computers that communicate and understand, this Semantic Web, will come into existence. Firstly, the Semantic

Web needs to be adopted by humans. They must see the value of these personal assistants that can schedule things for you, that take into account your personal context when searching the web, that understand questions and can ask you for more information that can improve the search. People have to realise that the current search engines are not good enough, and that having to read through many papers to find the one nugget of information or the one connection that you are looking for is not acceptable or necessary. But also, the different knowledge bases and unstructured texts need to be *aligned*. Aligning means that softbots should be able to combine information that is contained in different domains, documents and databases. This last problem is the one that I will work on in this thesis. Once this problem has been solved, the Semantic Web is technically possible and only needs to be socially accepted before it will become the norm[14].

1.1 Motivation

In this section I will explain how the study came into existence.

At the 14th International Semantic Web Conference McCallum presented his keynote on *Representation and Reasoning with Universal Schema embeddings*[23].

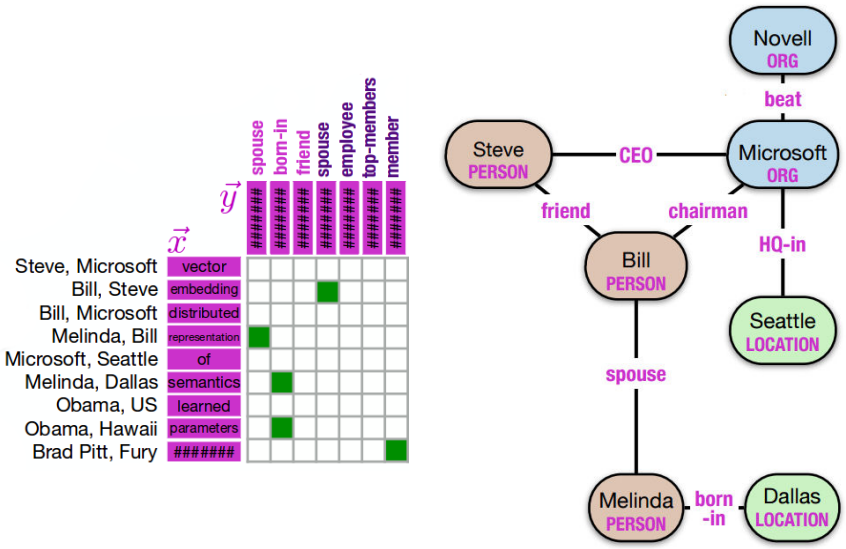


Figure 2: The Universal Schema. Every row and every column is represented by an embedding (in purple). Green cells mean a row and column are in the context of the other. For example, the pair Bill and Microsoft are connected to *chairman*. The embeddings are then trained so that the context can be predicted. That means the embeddings of Bill & Microsoft and *chairman* will become more similar.

The Universal Schema uses two steps. Firstly, the system learns embeddings of entities and relations found in text and knowledge bases[33], see Figure 2. Secondly, these embeddings are then used to predict new relations for the knowledge bases. The system finds embeddings such that the embeddings of entities and relations that co-occur are highly similar, see Figure 2.

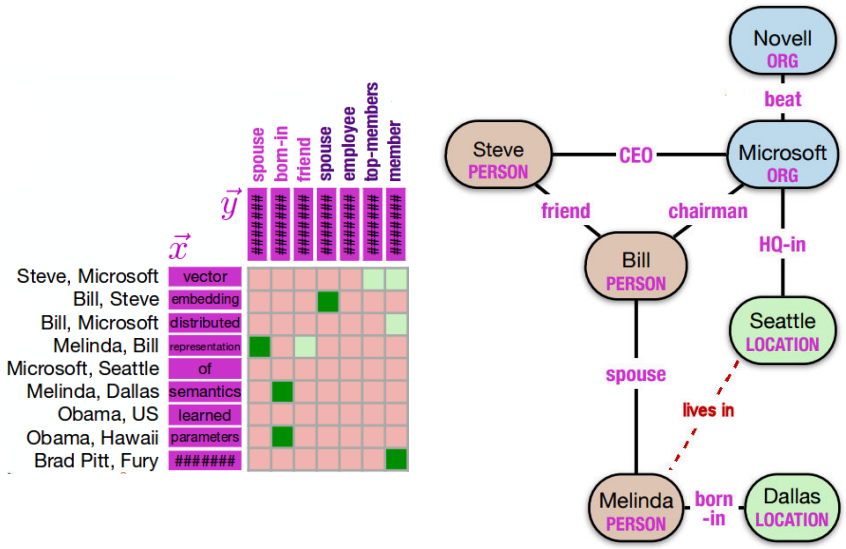


Figure 3: After the embeddings are learned, rows and columns with similar embeddings are connected (in light green). For example, the embeddings of Melinda & Seattle and *lives in* may have become very similar. In that case, the system connects the pair with that relation.

If some entities and relations have similar embeddings but are not known to co-occur, they are flagged as being similar. Similarly, if two entities or two relations have similar embeddings, they are likely referring to the same or similar things, see Figure 3. This can be used to infer even more similarities by collaborative filtering. The technique had very promising results in knowledge kase completion tasks.

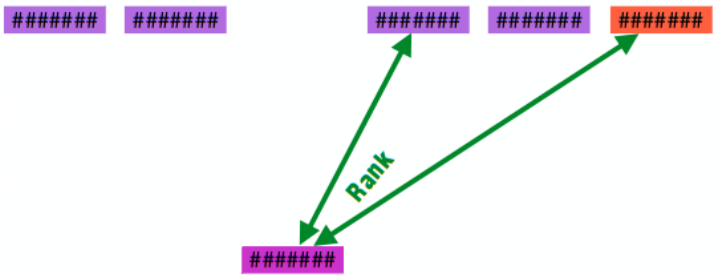


Figure 4: The embedding of a row (in pink) is optimized to be close to its context embeddings (in purple) but far away from other embeddings (in red).

The embeddings are optimized by a model called the skip-gram word embedding model[26], a neural network model that learns to predict the context of a word, see Figure 4. McCallum discussed a new extension to the algorithm that allowed for homonyms and as such can deal with more ambiguous data[23].

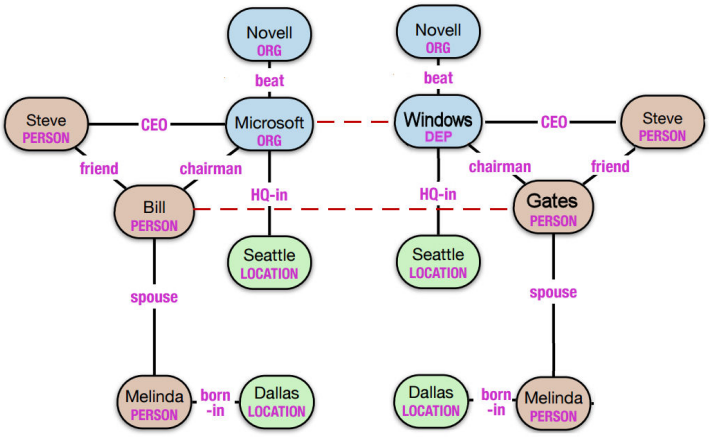


Figure 5: Similar to how the embeddings are used to predict relations within an ontology, ontology matching can be done by predicting relations between two ontologies. For example, since Gates and Bill will have very similar sets of embeddings (as their contexts are very similar) the system will recognise they are the same entity.

McCallum plans on using these embeddings to predict new intra-knowledge base relations. However, from the keynote it is clear that this technique should work with multiple knowledge bases as the Universal Schema is applied to a set of knowledge bases including Freebase, knowledge extracted from text and the TAC knowledge base population track dataset. This means that inter-knowledge base relations can also be predicted using the word embedding tensor factorisation technique.

Applying the embedding tensor factorisation to inter-knowledge base relation finding (i.e. ontology aligning) makes it objective. This is because for ontology aligning there is a ground truth possible and available. There are also many projects on ontology aligning so it is possible to compare the results with the best methods available. This means ontology aligning is a good test for this algorithm.

An example of an inter-knowledge base relation is shown in Figure 5. If these relations can be predicted, multiple knowledge bases can be aligned. Testing if this word embedding algorithm is versatile enough to work in the field of ontology aligning is the goal of this work.

1.2 Justification

This study is important to the future of the World Wide Web. The next goal of the Web is the Semantic Web, where computers can interact with web pages as humans do. It is necessary to build the Semantic Web as soon as possible, as the amount of information on the Web is growing so fast it will be infeasible for humans to handle[30].

The core of the Semantic Web are knowledge bases with the corresponding ontologies. Ontology matching is essential for the Semantic Web to work, as the information contained in the knowledge bases has to be combined. Since the Semantic Web will have many small knowledge bases, they will be combined at the moment information from them is required. To combine these knowledge bases, their ontologies first need to be aligned. This is an act that will happen very often on the Semantic Web, since every interaction with the Semantic Web will usually require information from multiple domains. Large domains may be stored on multiple devices, which can be viewed as different knowledge bases which also need to be aligned. Therefore, it is save to say that efficient ontology matching will be very valuable.

The most important goal of this study is to find an ontology matcher that is able to deal with ambiguous concepts. Since most of the knowledge on the Semantic Web is automatically gathered, it may be unreliable, ambiguous and even wrong. Therefore, ontology matching that is capable of handling this uncertain knowledge is required. Word embeddings and especially multi-sense word embeddings should be able to handle a high amount of ambiguity, since multi-sense word embeddings are made to differentiate between different meanings of the same word.

Normally, the raw text representation of words is used for matching, for example to calculate the string distance. However, word embeddings models use a vector representation, which should be more efficient. Calculating the distance between vectors can be done very efficiently, therefore ontology matching with distributed concept representations fits that efficiency criteria.

I believe the algorithms proposed in this study will work on new data without the need for redesigning. This is because distributed representation algorithms can be trained unsupervised on new domains, so they can be applied to those new domains without effort from humans. As opposed to for example the string edit distance. This is also an improvement for the semantic web.

This study should be done now, as ontology matching is already used in areas other than the semantic web. It is, for example, used by companies for merging knowledge from different knowledge bases when merging with other companies. If the proposed matching algorithm is better than the current algorithms, or

at least improves the performance when used in ensemble with the current algorithms, it will be useful immediately in these situations, not just in the future, on the Semantic Web.

1.3 Context

This section will enable the reader to put the problem in perspective as well as understand the reasoning behind the solution to the problem. Everything that is mentioned in this section will be explained more in-depth further on. Firstly, I will give a short summary of the context by explaining why the Semantic Web requires ontology matching. Then I will illustrate the problem of ambiguity. Lastly, I will describe the algorithm that this study is based on.

On the Semantic Web, ontologies are used to define knowledge[11]. The term ontology will be properly defined later on, but for now it suffices to know that an ontology is a set of rules that determine what relations are allowed and required in a given domain of discourse. As a consequence, if two agents want to communicate in a given domain, they need to agree on the rules of that domain, or in other words, share the ontology of that domain.

However, it is unlikely that a conversation only covers one domain. To solve this, there are two potential solutions. Either you ensure that all possible combinations of domains are covered by an ontology, which means one ontology that covers all domains is required, or, alternatively, ontologies are combined on-the-fly.

The first option, using one single ontology, is infeasible. This is due to the fact that inference and search time scales with ontology size[5], and the number of concepts that are needed on the Semantic Web is enormous. Every single concept that is present in the billions of terabytes of text, images and sounds that are on the web needs to be represented, for every professional domain and science, in every language. And all these concepts are related. Searching through this ontology just to find one concept would be incredibly costly. And since every interaction on the Semantic Web requires lookups in an ontology, these lookups need to be fast.

The second option, the one I address in this thesis, is the problem of combining ontologies from multiple domains. Combining in this sense means connecting the related concepts from the different ontologies. Some research has focused on ontology merging, with promising results. However, recently research has been slowing to the point where there is little improvement.

One big problem in ontology matching is the fact that in different domains, words can have different meanings. Common examples are synonyms like author and writer, homonyms like bank (for money or by a river) and ambiguity.

Ambiguity means that similar words mean slightly different things, or a word might have multiple related meanings.

A recently developed model for word embedding may be used to deal with ambiguity. It maps words to a vector space that has semantic properties, and thus can be used to calculate the relatedness of two words. There are a number of extensions that can deal with homonyms as well as data structured other than the continuous bag of words representation that the initial model was developed for. These facts show that it may be possible to make an embedding model for ontologies and use the word relatedness to combine ontologies[28].

Homonyms are two concepts that share a label, or in other words, a homonym is one word with multiple distinct meanings. As such, homonyms are inherently ambiguous, since the meaning is uncertain without context. This ambiguity is very hard for computers to deal with, as computers are used to dealing with absolute certainties.

1.4 Problem

We have seen that when information from multiple domains needs to be integrated so that it is possible to reason over them together, the ontologies of those domains need to be aligned or merged. If we want to merge two ontologies, we need to find the concepts from the different ontologies that are related by an 'is-same-as' or other rule. However, since the ontologies are made for different domains, the same concept can be represented by different labels in the two ontologies. This problem is called the synonymy problem, as synonyms are two words that refer to the same concepts.

The opposite problem is called the homonymy problem, in which two concepts are represented by the same label. Since they have the same label, naive algorithms might align the concepts as being the same, when they are not.

Then there is the problem of ambiguity, which encompasses many other problems. It encompasses the fact that some concepts are used wrongly by humans, but also the fact that some concepts are extremely similar, and may occur in the same context, but are still slightly different.

The problem is:
Matching ontologies and dealing with the ambiguity and homonymy contained in them.

I will use distributed word representations to answer the above question, as McCallum suggested in his keynote. Distributed word representations, specifically multi-sense *Word2vec*, should be used for this problem since multi-sense *Word2vec* is specifically designed to combat homonymy while the neural networks in distributed word representations use the context to deal with ambiguity.

1.5 Research question

The main question that I will answer with this work is the following research question:

> *Can distributed representations of concepts be used to find relationships between concepts in ontologies better than existing matchers?*

To answer the research question, I must investigate the following points:

- The algorithm uses distributed representations of concepts.
- The algorithm finds relationships between concepts in ontologies.
- The algorithm performs well.

The answer is positive for distributed representations matching in general if it is positive for one distributed representation matcher.

1.6 Overview

This report is structured as follows: firstly, I will give all the background information required to understand the context and the algorithms I will use. This includes the history and future of the Semantic Web, the roles of ontologies in the Semantic Web, ontology aligning and the recent developments in word embeddings.

Secondly, I will describe the method, which includes the design of the algorithm, the whole system and the evaluation method. Multiple algorithms will be described that all use word embeddings differently.

Thirdly, the results. These show how the algorithms perform when applied to ontology aligning. The results will be compared to a baseline and the best performing system so far.

Lastly, the conclusion and discussion, which interpret the results and give recommendations for future research.

2 Background

In this section I will discuss all concepts that are required to understand the study. Firstly, the Semantic Web, the key technology that depends very strongly on ontology matching, will be discussed. Secondly, the concept of the ontology itself will be addressed. Thirdly, current alignment strategies and their their problems are discussed. Lastly, other related research and technologies will be discussed. Later sections will refer back to the concepts discussed in this section.

2.1 The Semantic Web

The first concept I will discuss is the Semantic Web, also called Web 3.0, a vision of a Web that enhances the user experience in ways that the current Web can not provide.

2.1.1 Progression of the Web

The original Web connected authors to readers. It allowed authors to create content, mostly static websites. Readers, however, could not change anything, merely read the sites and click on links. In this iteration of the Web there was no interaction[7]. See Figure 6.

Figure 6: The first iteration of the Web consisted of static content, written by authors and read by readers.

Web 2.0, the current Web, models the internet as *hubs* (or sites) with a specific purpose and interaction between users, who can also add content, as in Figure 7. However, there is only limited interconnectivity between those hubs[2][32]. It is possible to link to content from other hubs using URLs, but this moves the user to another hub, rather than connecting the content present in the hubs.

[2]For example, Facebook is a hub for sharing content with people you consider friends, but if you want to share that content with more people you will have to go to another hub, like Imgur or Pinterest.

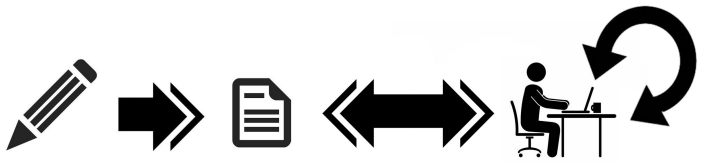


Figure 7: In the second iteration of the Web, readers have become users that can alter and add to the content found on the interactive hubs. This also allows users to interact with each other.

We are currently progressing towards Web 3.0, where not only users interact, but also virtual agents, see Figure 8. The so-called Semantic Web contains sources of *information* which can be combined whenever the user needs it. If someone wants to find information on a certain topic, the virtual agent can collect all documents that are relevant to that topic, summarize each document in a way that is relevant to the search request and provide sources for every fact it finds. It would even be possible to have a question answering session on that specific topic, where the agent finds the answer to every question posed in the material that is linked to the topic.

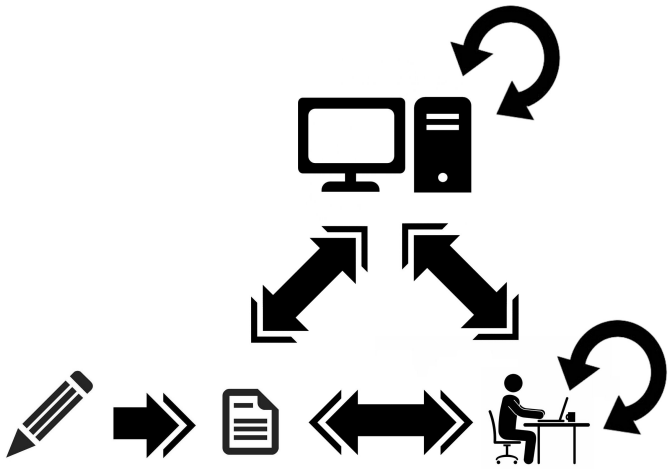


Figure 8: On the third iteration of the Web, virtual agents are also able to read and write content and interact with humans (through question answering as an example). Virtual agents may also work together.

Another possibility on the Semantic Web is the scenario from the introduction (Section 1), where a user wanted to see a movie, and ordered his softbot, to find him a suitable time and movie to see. This was shown to reduce a

lot of planning and allow for much more favourable decisions since many more variables can be taken into account.

All of this is possible if the softbots can access all of the information reliably. In the current Web, this is not possible, as all information is written in an ambiguous, unstructured format called natural language. To allow softbots to access the same information that we can, the Semantic Web proposes to add a layer to the internet where all information is saved in a unified format that is unambiguous, robot-interpretable and can be used to store any type of knowledge.

The Semantic Web also improves user interactivity by connecting sources of information. To stick with the social media example, a person can create a piece of content, say a picture of a tree, and share it with his friends. One of his friends then can share this piece of content with a group of people that likes nature pictures. However, everyone can access all references someone makes. So if a person from the nature group notices the tree is ill and comments on this, the original poster can see this comment and act accordingly. On the current web the friend who shared the picture with the group has to be contacted directly by the commenter and then has to relay the message manually.

2.1.2 Goal

The goal of the Semantic Web is very pragmatic: help people in everyday activities by leveraging all information available on the internet. The fact that it is pragmatic goes a long way of making it a reality: if even one application comes into existence that uses online information in a structured way and thus makes everyday tasks a little easier, the goal is accomplished. Of course, then the goal is stretched and we should make more applications leveraging more information for more tasks.

Ultimately, the goal is to leverage all information that is available to an entity on the internet to improve the lives of humans. This information is not limited to the information that is online *now*, but can include sensor data from the Internet of Things, data from robots, facts deduced or statistically inferred from existing data and so on. All this information can be used to accommodate the wants and needs of humans, when asked for it *and* before the users are aware of their needs. The personal agent should be able to predict the users need and provide the tools that can satisfy the need.

To provide these tools to the user, an agent uses many different sources. For example, if the need of a researcher is knowledge on a certain topic, the agent must chart relevant topics, map relations between different domains and connect these new ideas to topics that the researcher is already familiar with. In fact, the largest part of the job an agent has to carry out on the Semantic Web is

finding relationships between pieces of information on the Web. Based on this observation, we can state that ontology and knowledge base alignment is the cornerstone of the Semantic Web, since without this vital system the Web is merely a collection of separate pieces of information. Only when those pieces are aligned can we talk of knowledge, of semantics.

To summarize, the goal of the Semantic Web is to connect all online information to help humans. Only when all information can be connected will it be possible to aid humans in their needs. But humans will only accept the Semantic Web if it is useful to them, so the Semantic Web must show its applications and usefulness before it will be used in daily life.

2.1.3 Workings

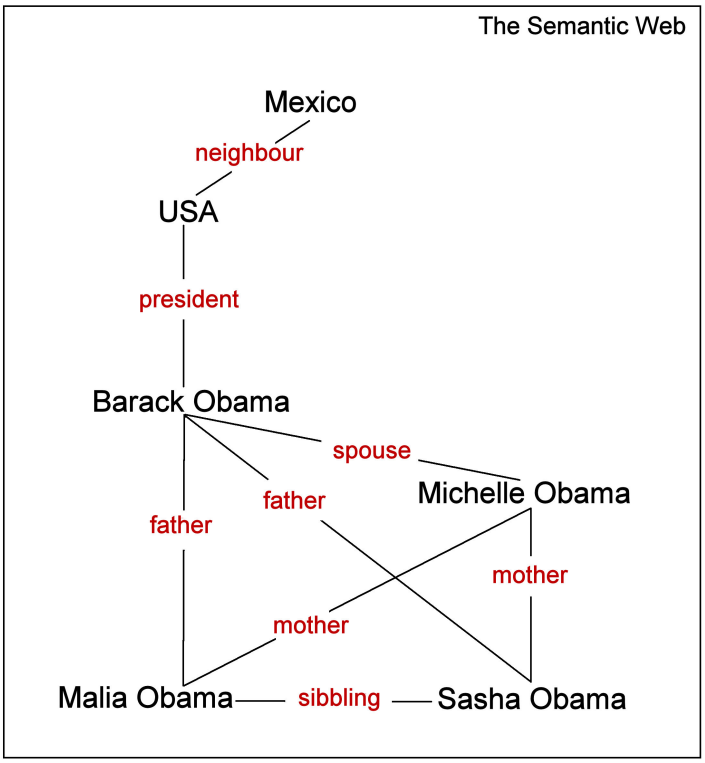


Figure 9: An example of the Semantic Web represented as a graph of entities (in black) and relationships (in red). It contains many different kinds of entities, including countries and people.

The Semantic Web is one big graph of entities and relationships between those entities, containing information about, amongst others, people and countries as in Figure 9.

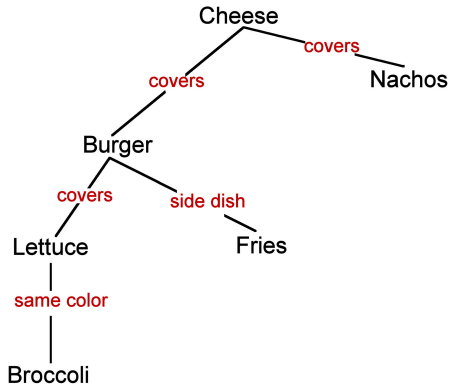


Figure 10: New information to be added to the Semantic Web.

If new information is added to the Semantic Web, a new concept or connection is added to this graph. In Figure 10 a whole set of related entities and their connections are shown. These entities are also connected to the rest of the Semantic Web graph so that it stays one connected web, as in Figure 11.

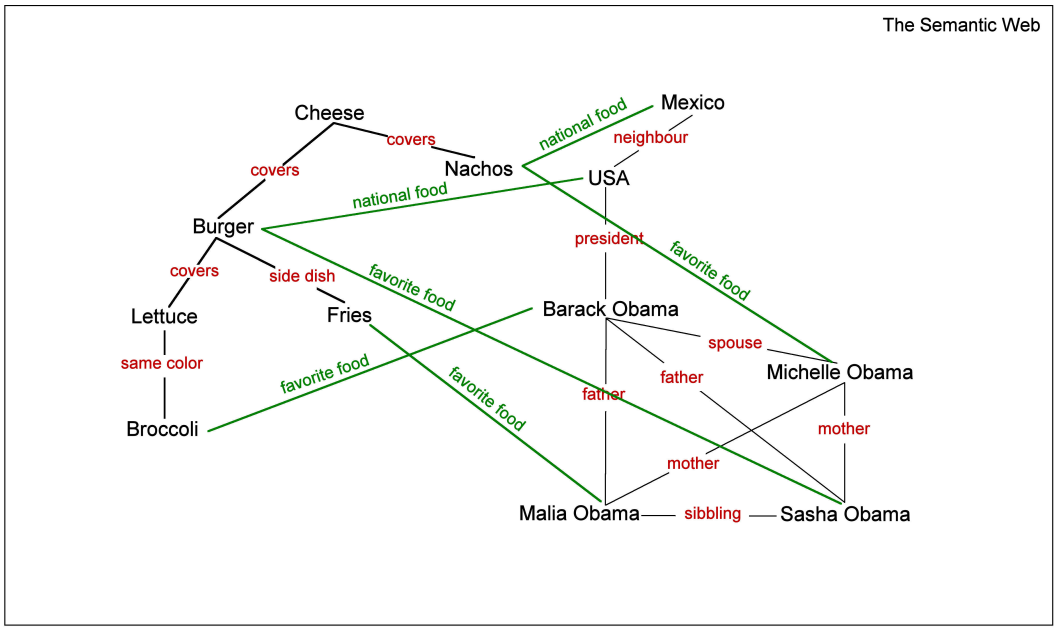


Figure 11: The new information is connected to the rest of the Semantic Web by connecting the new entities with entities in the Semantic Web.

This big graph covers many different domains, and every domain is covered by a part of the graph, a sub-graph. For example, all people are in the same sub-graph, as are countries and the recently added foods. Every such sub-graph is one knowledge base that covers that specific domain. See Figure 12 for the knowledge bases in the example.

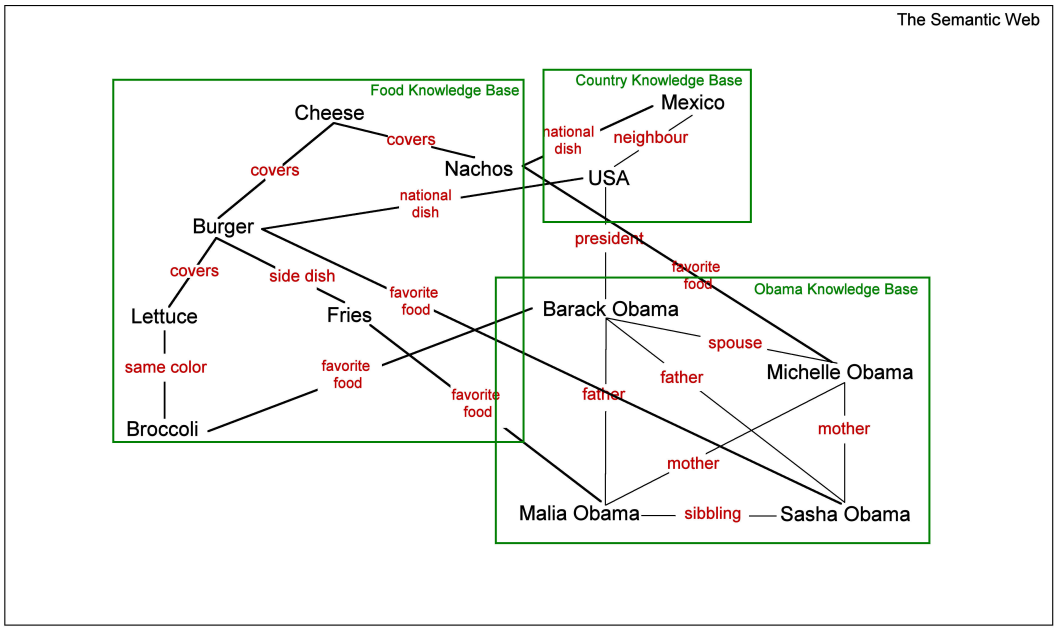


Figure 12: The different sub-graphs in the Semantic Web graph. They are contained in different knowledge bases and connected by URIs.

To ensure that the Semantic Web is one big connected Web, the knowledge bases are connected by inter-knowledge base relations. The relations are made by referring to the URI of that entity. The URI consists of a link to the knowledge base and then the identifier of the entity. For example, the people are connected to their country of origin, so in the people knowledge base there is a reference from Obama to *countries#USA*.

In this paragraph I will explain the use of ontologies on the Semantic Web. Ontologies describe the rules by which the knowledge bases must play. For example, every food must have a country of origin, and those countries of origin need to be of the Country *Class*. This rule enables search algorithms to limit search to the countries knowledge base, as it is the only knowledge base that contains countries[34]. This is just one example of a constraint that improves the work of one system. There are many more reasons, for example adding the rule "all mothers are parents" allows inference systems to reason about richer information without having to explicitly write down every relationship.

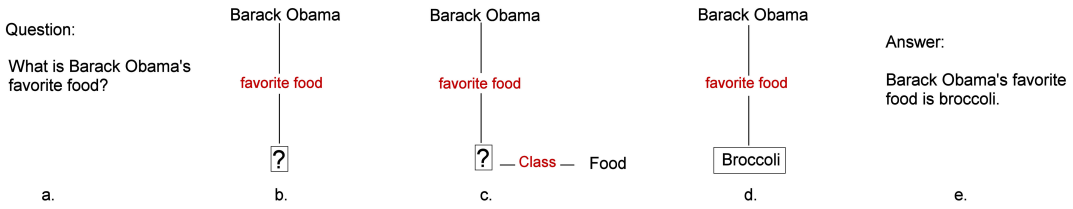


Figure 13: The question posed in a) is converted to a query pattern b). Ontology information is added in c). The result is found after pattern matching in d) and then converted to a natural language answer in e).

There are many ways of interacting with the Semantic Web, but they all interact with the graph in some way. For example, a user can ask a question "What is Obama's favorite food?" Firstly, this question is converted into a format that can be used by a search system. Then, information from the ontologies is added so that the search can be performed more efficiently. Then the pattern is matched in the Semantic Web graph, listing all results. The results are then combined or filtered such that one answer pattern remains, and that pattern is converted to natural text. This process is depicted in Figure 13.

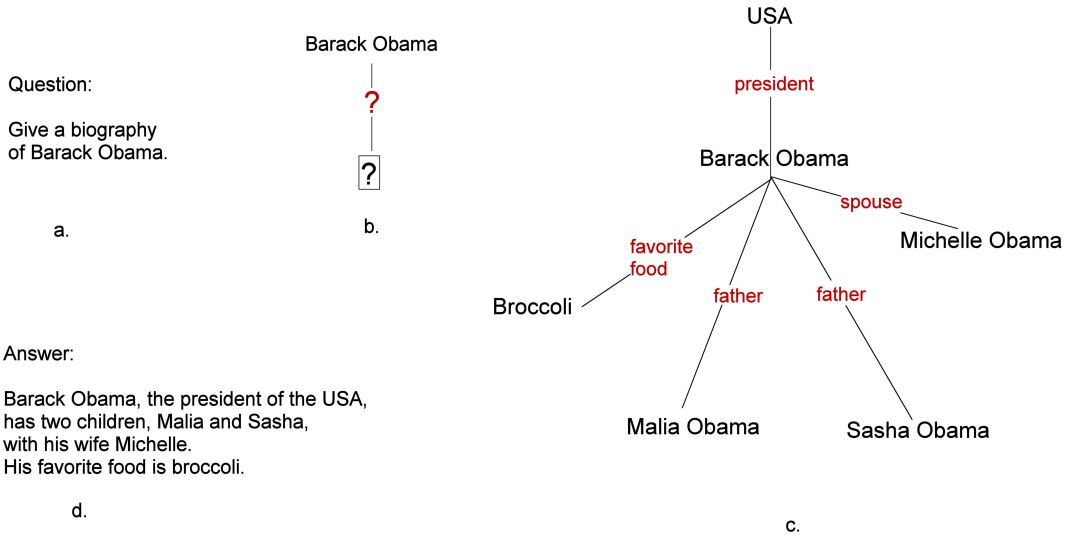


Figure 14: Similarly to a question with a single answer, the question in a) is converted to the pattern in b). The answers that match b) are combined in c) and converted to natural language in d).

More complex question can be posed, although they do not need to be longer. If the question can be mapped to more parts of the Semantic Web, the answer

will be longer. In this case all knowledge of Barack Obama is extracted, combined and converted into an answer.

2.1.4 Current state

Currently, people are not aware of the Semantic Web. Either websites and applications do not use any semantic information, or it is used on the back-end, hidden with other complex programs that users will not understand. So what is the Semantic Web used for nowadays, if at all? In this section, I will list a number of general ways people use Semantic Web technologies as well as specific applications. Note that these are Semantic Web technologies, i.e. technologies that came into existence through Semantic Web research and can eventually be used in the Semantic Web, but currently are not used for the Semantic Web since it does not exist yet.

The main languages W3C endorses are RDF, OWL and SKOS[22, 24, 27]. They are important for the Semantic Web and are already in use. Most of this use, however, is for hand-crafted knowledge bases, created by experts, containing only facts that are accurate, verified and relevant[25]. This method of creating a knowledge base does not scale very well and usually consists of fewer than ten thousand facts. Because of this, the knowledge bases are restricted to small domains, including some parts of medicine and cultural heritage[3, 15]. As such it does not help regular uses of the Web but only people who are willing to become adept with the tools used in the knowledge bases.

The Linked Open Data project, on the other hand, is building the foundation of the Semantic Web mostly automated. These automated methods work fast, parsing thousands of websites a minute, but are inclined to make mistakes. Often, these methods use crowd-sourcing to improve accuracy, using experts or the general public to suggest or check facts[37].

The largest knowledge base of the LOD project, Freebase, boasts almost two billion triples. The three hundred combined Large Open Data databases contain over thirty billion[16, 2]. The LOD is simply a single alignment over many different knowledge bases.

Applications include improved web searching, question answering, product comparison, context merging, data integration, decision support, translation, all the way up to the intelligent softbot mentioned earlier[1, 35].

An example: since the Hummingbird update, when someone Googles "Who is Barack Obama" they will find a part of the Knowledge Graph, see Figure **??**.. This system improves web searching by finding answers rather than pages.

The many uses of ontology matching and its role in the Semantic Web show that it is important to research all avenues that might improve ontology matching, as it will have a direct result in many fields.

2.2 Ontologies

Earlier, ontologies were referred to as 'a set of rules that determine what relations are allowed and required.' Now, the term will be defined more extensively, the standard format will be described and the uses of ontologies will be discussed.

An ontology describes concepts as their relationships to other concepts. The concepts that are described determine the domain the ontology covers, and the relationships determine the rules the ontology imposes on the domain. For example, a hierarchy is an ontology that describes concepts that are subclasses of other concepts. Therefore the relationships are 'is-a' relationships: a bird *is an* animal. Concepts can be instantiated, which is done in a knowledge base (KB). KBs and ontologies are often confused, and are indeed very similar in structure. The knowledge base in our example can contain actually existing birds or fictional birds, and refers to the ontology concept of bird to embed the instantiations with meaning[12].

2.2.1 Example

In this section I will give an example of a pair of anatomical ontologies with their alignment, which will be used throughout this work.

A mouse consists of the following body parts:

- Head
- Torso
- Whiskers
- Tail
- Front paws
- Hind paws

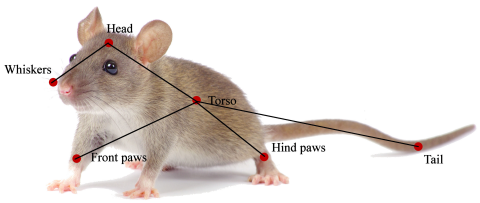


Figure 15: The different body parts of the mouse are connected as in this image.

Humans consist of:

- Head
- Torso
- Moustache
- Arms
- Legs

Figure 16: The different body parts of the man are connected as in this image.

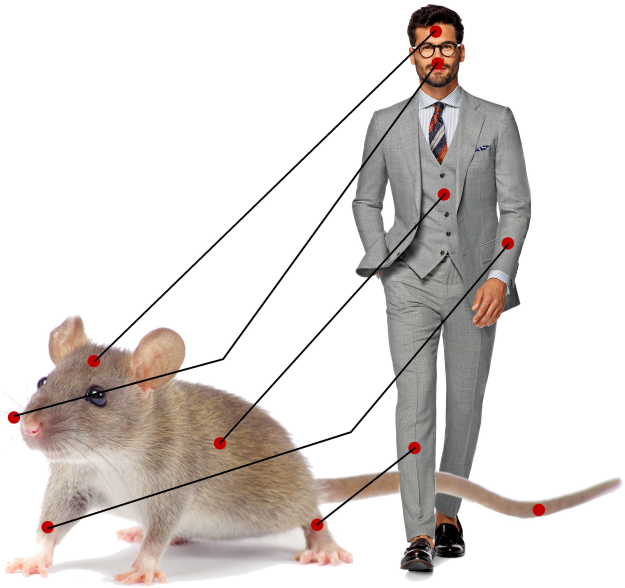


Figure 17: Humans and mice are connected as shown in this image.

2.2.2 OWL

OWL is a family of languages and syntaxes that can be used to create an ontology. The different languages are designed around requirements of possible relationships and concept definitions, and thus may differ a lot between them. The W3C has defined three variants which are the bases for all other adaptations. The variants trade off levels of expressiveness versus computability.

A relationship in OWL is written as follows:

```
<owl:Class rdf:about="http://mouse.owl#Whiskers">
  <rdfs:connectedTo rdf:resource="http://mouse.owl#Head"/>
</owl:Class>
```

2.2.3 Uses of ontologies

There are four main reasons ontologies are used. Firstly, it allows querying languages to optimize the search process. Secondly, it allows expansion of information on a node. Thirdly, it allows for consistency checking a knowledge base. Lastly, it allows merging of knowledge bases.

Search can be optimized by taking into account constraints the ontology provides. For example, no professor is a student, therefore no professor can be connected to a course with 'follows' predicate. This allows a search algorithm to skip all professors when looking for people that might follow a certain course.

Expanding the available information of a certain node can be done by taking into account *positive* constraints of a concept. For example, since all birds have beaks, it would be inefficient to store this fact for every instance of bird in the knowledge base. However, if this is a constraint given in the ontology, it can be accessed for all instantiations of the bird concept with less space required. Superclassing enables ontologies to store these types of information even more efficiently, allowing for enormous amounts of information to be extracted for every node without the need to store it all explicitly for that node.

One important aspect of data bases is consistency. Manipulations on a data base such as a knowledge base must not result in a knowledge base that does not adhere to the data base constraints. In a knowledge base, these constraints are defined in the ontology.

When two knowledge bases need to be merged, aligned or matched, instantiations that refer to the same thing need to be merged. For example, if two person databases both contain references to the same person, all information on that person needs to be linked to the same object, such that information of that person from both knowledge bases can be combined. Merging the ontologies of the two knowledge bases vastly improves the alignment between the

two knowledge bases. However, ontology and knowledge base merging is not a trivial problem and actually the problem that this thesis addresses. Therefore it will be explained more in-depth later on[13].

2.2.4 Aristotle and Plato

All ontologies lie on a spectrum of formality. On the one end of the spectrum are the Platonic ontologies, which should only contain concepts that are perfectly defined and constraints which are completely binding. The name refers to the Platonic philosophy that all concepts are existing entities and can thus be perfectly captured in a definition.

On the other end of the spectrum are the Aristotelian ontologies, which do not necessarily contain perfectly defined concepts, but rather concepts as we observe them. Aristotle disagreed with Plato's theory that concepts are existing things, and thus these ontologies are named after him[21].

Aristotelian ontologies have the advantage of being easier to create. One can use statistical methods given a sample of all possible observations or crowd sourcing to create concepts and constraints. This enables enormous ontologies and knowledge bases, especially with the big amount of data currently available and cheap mental labour with services like Amazon's Mechanical Turk.

The advantage of a Platonic ontology would be the fact that every query would result in a fact, since the ontology itself is perfect. The disadvantage is that creating such an ontology is much more costly than an Aristotelian ontology, since every concept and constraint needs to be correct in all cases.

These ontologies are often used in areas where precision is important, like medicine, where lives depend on the information contained in the ontology and its knowledge base. Aristotelian ontologies are more common on the Semantic Web, where massive amounts of data need to be represented and queried.

Currently, alignment strategies are focussed more on Platonic ontologies. In the Ontology Alignment Evaluation Challenge the datasets are relatively small Platonic ontologies. Since most of the ontology matching systems are produced to compete in the OAEI, almost all techniques focus on small, perfect knowledge, perfect ground truth ontologies.

The proposed method should scale to much larger datasets without effort, allowing for ontology matching that is suitable for the Semantic Web, rather than just for individual small knowledge bases.

2.3 Current alignment strategies

In this section I will give a more in-depth description of the ontology matchers that compete in the OAEI.

Many different ontology matchers have already been developed. All matchers can be divided into two groups: those that require just the labels of the concepts and those that require an initial correspondence set and the structure of the ontologies. The first group is always used first to generate the correspondence set of the second group. The first step is called a terminological matcher since it only uses the terms of the concepts. If it uses a vector space model it is also called an extensional matcher.

The second step is referred to as structural matching. If the matcher uses logic, it is also called an inference matcher[35].

Popular terminological strategies are WordNet comparison and edit distance. The former uses the popular WordNet hierarchy as a distance measure between two concepts, e.g. the number of edges between the concepts and their least general common superconcept[18]. For example, the distance between moustache and whiskers is 3, since their least general common superconcept is hair the distance between moustache and hair is 2 (with facial hair between them) and the distance between whisker and hair is 1. Edit distance is a purely string-based similarity measure. The similarity is a (weighted) count of edits required to transform one word into another, where common edits are insertion, deletion and replacement. Similar methods include substring matching and n-gram matching, which compare parts of the strings to find similarities[36, 17]. For example, arm and front paw have an edit distance of 8, since they have 1 character in sequence in common (either the a or the r) and thus 8 characters that are different, see Figure 18.

		f	r	o	n	t		p	a	w
	0	1	2	3	4	5	6	7	8	9
a	1	1	2	3	4	5	6	7	7	8
r	2	2	1	2	3	4	5	6	7	8
m	3	3	2	2	3	4	5	6	7	8

Figure 18: The edit distance between *arm* and *front paw*.

An example of a structural matcher is the children matcher. It matches all children of a pair in the correspondence set with some low confidence. When two nodes are matched more often, their confidence is combined until it reaches some threshold. Then it is added to the correspondence set.

2.3.1 Problems in ontology alignment

How should the wildly differing demands the applications place on the ontology matching systems be satisfied by a single or just a few systems? When merging ontologies on one end of the Aristotle-Plato spectrum, often a completely different approach is required than an ontology on the other side. Sometimes the matching needs to be very fast, for example when the ontologies are used in a search query which the user expects to be done in milliseconds.

Based on the previous point, it would also be useful to have different benchmarks to test those different types of challenges and matching systems. This would allow for better comparison of methods and better tracking of the improvements in the field.

Since the Linked Open Data keeps growing, it, and other resources, should be used in matching. This is already done to some degree, as discussed in the WordNet matching strategy section 2.3. The challenge is to all available ontologies and other sources rather than just WordNet.

Word embeddings should be good at using other ontologies, since they can very quickly learn embeddings for all concepts in those ontologies. These embeddings can then be used during matching as an a priori believe of the locations of the word embeddings in semantic space.

When these and other problems are solved, ontology matching has matured enough to be used in Semantic Web applications. Then it is merely the challenge to have the public adopt these technologies.

2.4 Word representation

An important area of research in Natural Language Processing is the representation of words. The most obvious representation is the collection of characters humans are used to. However, this is not useful for computers as these characters do not say anything about the meaning of the words, nor is it a representation that computers can quickly do calculations on.

Many different representations have been proposed. The main purpose of word representations in NLP is to represent the meaning of the word. One way to represent a word is the term frequency. This representation counts the number of occurrences of a word per document in a set of documents. The resulting vector of occurrence counts is then used to represent the words. Since the documents have a subject, there is some meaning embedded in the vector representation. However, the representation is very sparse and the semantics or meaning of the word are not captured very well. Also, the method requires many discrete documents to train on, which are not available in ontology matching.

The rest of this subsection will describe the latest word representation model which has very interesting results in NLP. It will cover the history and the most recent advances that are relevant to this study.

2.4.1 Skip-gram models

Skip-gram models have been in use for some time. Like the document frequency representation, it represents words as vectors. However, the vectors are much more dense and can be trained on sentences rather than documents. The vector distance of skip-gram models has been shown to be related to the semantic similarity between words. This semantic similarity property is very useful for ontology alignment, as parts of one ontology have to be aligned with parts of another ontology depending on their semantic relation.

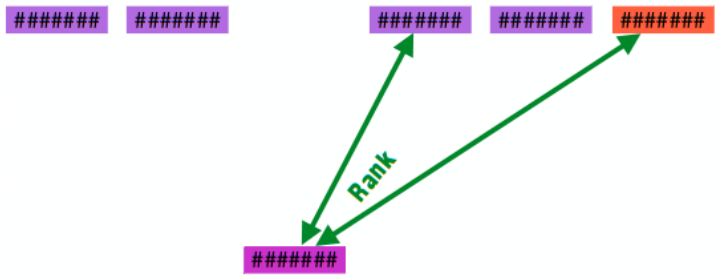


Figure 19: The embedding of a word (in pink) is optimized to be close to its context embeddings (in purple) but far away from other embeddings (in red).

Skip-gram models work by teaching a 3-layer neural network to predict the context of a word. In NLP the context is a number of words before and after the word. Firstly, all words a represented by 1-hot encoding, which is a vector with all zeroes and a single one on the index of the word. This vector has length V equal to the number of words in the corpus. This vector represents the first layer, the input layer of the neural network. The second layer, the hidden layer is obtained by multiplying the first weight matrix with the first layer. Since the first layer is one-hot encoded, this is equal to the row of the matrix corresponding to the index of the word. This representation will later be used as the representation of the word.

Then the representation is multiplied with another matrix to create the third layer, the output layer. The word representations and second matrix are optimized in such a way that the output layer is as close as possible to the one-hot encodings of the context.

As a consequence words that occur in similar contexts will be represented similarly. This again causes contexts that contain similarly represented words to also become more similar. This positive feedback loop causes co-occurring words to be represented more and more similarly. On convergence, words with similar meanings are represented similarly.

2.4.2 The state of the art

Recently, an effective and efficient skip-gram model has been developed, called *Word2vec*, which uses a number of extensions over previous methods that enable it to efficiently learn an effective representation[26]. One example of such an extension is negative sampling, which is a step that teaches the network that randomly selected words should have non-similar representations.

These extensions enable *Word2vec* to be trained on large corpora and represent words in a way that is very useful in NLP. It is used as a language model in part-of-speech tagging and machine translation.

Interestingly, the representations have been shown to contain semantic information too. A common example is the following formula: $vec(\text{"queen"}) = vec(\text{"}king\text{"}) - vec(\text{"}man\text{"}) + vec(\text{"}woman\text{"})$. This means that gender is encoded in the vectors and other types of meaning may also be encoded.

After the authors showed the usefulness of *Word2vec*, many researchers developed new uses for the algorithm. Some extensions that are relevant to this study will be discussed. Firstly, I will describe a model that represents syntax as opposed to semantics. Secondly, I will describe a model that is capable of representing multiple meanings per words.

2.4.3 Word order

Word2vec has been expanded in a number of different ways. For example, word order can be preserved, leading to a similarity measure that is closer to syntax, as syntax defines positional properties of words within a sentence[19]. To enable this alternative representation, the model had to be changed structurally. Rather than a single output layer generated by a single matrix, one layer per word in the context was used. As such, the model learns to predict the context words based on their position relative to the source word. More important than the fact that these models can be used for syntax is the fact that the research shows that the input does not have to be a continuous bag-of-words representation. This sparked the idea that it may also be possible to have a graph as the input to the network. This assumption is the basis of this study and will be tested thoroughly.

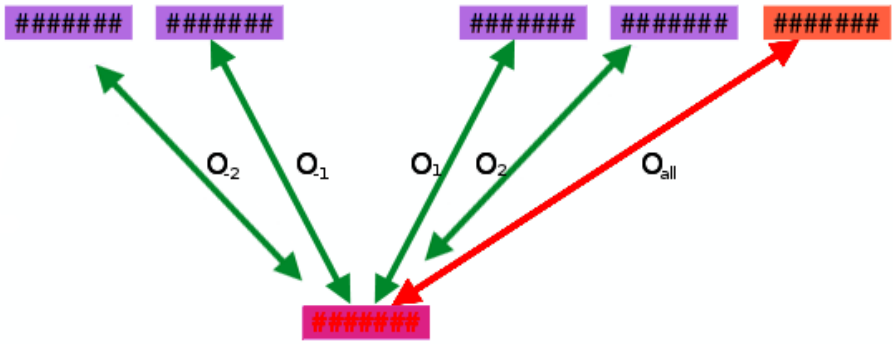


Figure 20: Rather than ranking the output of the network with respect to the whole context, one output per item in the context is computed and ranked.

2.4.4 Multi-sense

Another interesting development is the multi-sense *Word2vec*, which allows for multiple vector representations per word, depending on the number of different definitions a word has. For example, the word bank represents both the monetary institute and the riverside, which would both have a different representation in this extension. It can differentiate between two different meanings based on the context. The importance of allowing for multiple meanings is that between multiple ontologies, the same word can have different meanings and as such should not be matched. On the other hand, a concept in a single ontology can have multiple meanings and should therefore be matched with more than one concept from another ontology[28]. This idea sparked the second reason why I started investigating semantic word representations for ontology matching, since current methods do not allow for disambiguation.

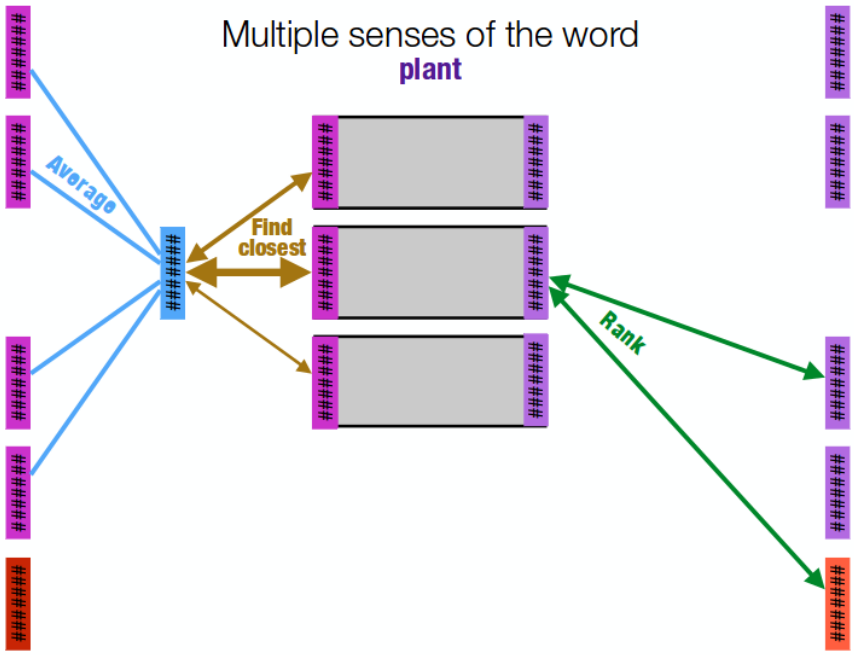


Figure 21: The context is first used to find the right sense, which is then optimized to match the context.

String matchers only look at the phrases representing the concepts, and as such do not care about semantics. Structural matchers use the context, but do not use the meaning of the concept since they only look at the location of the concept within the graph. Lastly, lexical matchers do use synonyms and antonyms to see which different words could refer to the same concept, but if two concepts can be represented by one word. As such, initially the multi-sense representation seemed like a very strong candidate for improving matching results.

2.5 Used technology

2.5.1 Factorie

Factorie is a toolkit that contains a number of tools which includes word-embedding models [9, 26].

Factorie provides a number of utilities for writing an algorithm in the shape of abstract classes with existing functionality, namely IO, parallelisation, quick parsing and command line argument parsing. It also provides example implementations and programs. Factorie is written in scala, a language that has been gaining popularity recently and interfaces with java. See [31].

The automated IO allows for quick reading in of the data set and secure saving of the trained model. The file reading is quicker than a naive CSV parser would allow for, and thus enabled quicker training of the models. This meant less time is used and more time can be used for experimenting and improving the algorithm. Saving the different models allows for model comparison after all models have been trained, which means new investigations can be performed after conclusions have been drawn from earlier investigations, without having to retrain all models. This also allows multiple models to be trained during a period where the experimenter is absent, after which he can still manually investigate the models.

Parallelisation can massively improve training speed. It allows an algorithm to update its model for multiple learning instances at the same time, thus reducing the time needed by the number of instances it can process simultaneously. The number of parallel processes differs per device but modern personal computers already boast 8 parallel processors.

Since I investigate many different implementations of the same algorithm, it should be easy to switch between these implementations at run-time without much effort. Command line parsing allows for this to happen, and thus improves experimentation ease.

2.5.2 Jena

Jena is the most used library for storing, manipulating, querying and reasoning on ontologies[4]. It has an interface that allows for easy access to an ontology. For example, if one would like to obtain a list of all neighbours of a certain node, one can simply use the following Jena command:

```
ontology.listStatements(givenNode, null, null)
```

It automatically gives all statements that have the given node as subject. In terms of graphs this is equal to obtaining all edges that go out of the given node.

In Jena it is possible to obtain the label of any node, if it has it, by calling:

```
label = node.isBlank() ?
                    node.getBlankNodeLabel() :
                    node.getLiteral()
```

The ternary is to distinguish between blank and non-blank nodes. Blank nodes are a feature of ontologies and thus need to be taken into account. Blank nodes can be used to represent concepts with multiple representations, for example a label, a description and a known synonym. More advanced algorithms can take into account all these representations. Another use of a blank node is

representing a complex relationship that cannot be represented by an object-predicate-subject triple. For example, if the relationship has a certain confidence or more than two concepts are involved.

2.5.3 AgreementMaker

AgreementMaker is an ontology matching system that obtained the highest F-measure in 6 of the 7 ontology matching tracks of OAEI 2015 [6]. The system has also been updated the most recent at the time of writing[8].

The system will be used as a benchmark to compare the results of this study. It uses many matchers in ensemble and therefore it is more realistic to compare my single matcher with a single matcher from the system, so I will do that too.

Ideally it would be possible to extend the Agreementmaker system with the matcher developed in this study. However, their claim that the system is easily extendible does not ring true, as it was too hard for the scope of this project to accomplish.

3 Method

This section contains the description of the matching system and experimental setup. It consists of four parts. Firstly, it describes the idea and challenges of implementing the system. Secondly, it describes the data used to test the system. The third part describes the whole system, including all different implementations of the algorithm and a system check to ensure valid results. The fourth and last part describes the evaluation criteria.

3.1 Idea

The goal of any ontology matcher is to find an alignment between two ontologies that is as similar as possible to the true alignment. Such an alignment consists of node pairs where the two nodes in each pair come from the two different ontologies. Sometimes a label is added to each pair, indicating the relationship the two nodes have. For example, 'part-of', 'similar-to', and such. However, we only consider equality relationships and thus the label can be omitted.

Then the question becomes: *how do we measure similarity of concepts between ontologies?* As shown in 2.3, attempts have been made to model similarity of nodes as the string similarity of their labels. Also, WordNet is used to find similar nodes. The string similarity assumes that similar labels refer to similar concepts. While this is true to a certain degree, homonyms, synonyms, and other phenomena make this assumption very weak. WordNet-based similarity measures are better since WordNet links concepts semantically, and thus similar concepts will be close in WordNet. The major problem with WordNet is that it is constructed by hand, and thus does not scale very well to new domains.

The idea behind the algorithm introduced in this study is to find an alternative to WordNet that is at least as good in modelling semantic distance but scales to new domains. It should be trainable on this new domain through existing ontologies and texts in an unsupervised way so that it does not require the interference of experts. This will allow it to overcome the weakness of WordNet. This is only relevant, though, if the performance is equal to WordNet-based matchers.

Word2vec is an unsupervised method that implicitly maps words to a vector space that has been shown to contain semantic properties. Explicitly, it predicts words from the context of that word[26]. It can be trained on large corpora quickly and the resulting vectors can easily be used to measure the distance between words. All these properties make it an excellent candidate for a concept distance measure to use for ontology matching. As an added benefit, it can also take into account the context of a concept, which may be able to improve its performance over a matcher that only uses the label of a concept.

3.2 Challenges

The semantic representation will be used for finding nodes in the two ontology graphs that are similar, i.e. close in the vector space. This step is relatively straightforward if proper vector representations are found. However, this may be hard due to a number of problems which I have listed below.

- The problem I am trying to solve is the problem of differently labelled nodes referring to the same concept. For example *writer* and *author* will not be matched by a string matcher, but should be matched. The two concepts may also have different labels in their context even though their contexts refer to the same concepts.
- The size of the training corpus should be large enough for proper representation to be learned for every concept. This is a problem since the ontologies may be relatively small.
- The representations should also be relevant to the labels that are in the ontologies. For example, a model pre-trained on text will likely have very different embeddings than trained on an ontology, and may as such not be very useful.
- Some labels may be ambiguous, for example homonyms, which are concepts that have the same label but represent a different context.

To solve these problems I will adapt *Word2vec* to graphs to create an algorithm that converts nodes to vectors, in other words *Node2vec*. It can be extended with multi-sense embeddings, different ways of considering the neighbourhood and hot-starting.

3.3 Data

The data used comes from a list of data sets that the OAEI provides[10]. Although the data is not as big as the dataset McCallum will use, there are a few reasons to use these data sets.

Firstly, it allows comparison to the state-of-the-art matchers as they all competed in the OAEI. Secondly, DeepWalk, an alternative way to transform graphs into vector representations, works on small graphs. Consequently, it is likely that other embedding models should work on small graphs as well.

Every data set contains two ontologies that need to be aligned and a ground truth alignment. The goal is straightforward: find the alignment given only the ontologies and whichever outside resources are needed.

3.3.1 Mice and men

The first data set considered consists of two anatomical ontologies. The first ontology describes the anatomy of mice and the second ontology the anatomy of

men. Since they are similar creatures, there are many correspondences between the anatomy. The ontologies contain 1838 and 3298 nodes, and 1807 and 3761 edges, respectively. There are 1516 correspondences between the two ontologies.

The ontology is structured as a hierarchical tree with the only label being *subClassOf*. This poses a few problems for some of the implementations as they take into account the edge labels. Since there are no edge labels, these implementations will not be different from their edge-less counterparts.

3.3.2 Medical

The second data set is larger, boasting 23692 and 16923 medical concepts with 31289 and 18932 relations. There are 18476 correspondences between the two ontologies.

3.4 Algorithms

In this section, I will discuss the different implementations that were made of the algorithm. All implementations use the same basic algorithm in different ways by taking into account the context in different ways. All of these implementations were tested and the results can be found in Section 4. Some implementations extend others, but all implementations are covered to ensure completeness. This way, all improvements are recorded precisely. All algorithms use word embeddings, though the way they use word embeddings is different. The explanations refer back to the example in Section 2.2.1.

3.4.1 Pre-trained models

The simplest way to use node embeddings for ontology matching is to obtain a model already trained on normal text. This model should already have learned some semantic context, although neighbourhood in text is different from that in an ontology. Also, since the words it is trained on might not match the labels from the ontology, two things need to be done. Firstly, if a word is not in the model, it is ignored. Secondly, the labels are split up into separate words and the resulting vectors are averaged. This model was chosen because of its ease of implementation and to see if the graph-based context improves ontology matching over the text-based context.

3.4.2 DeepWalk

DeepWalk generates sentences by randomly walking through the ontology and processes those sentences using a normal word embedding model. This means it is possible to have a wide neighbourhood, as the neighbours of neighbours are also used as context, or even further neighbours depending on the scope of the algorithm. It also allows for weighting of edges to change the sampling rate of the random walk, which can be used to prioritize important nodes or highlight

informative neighbours. One adaptation over normal Deepwalk that was considered is switching between ontologies when on a concept from the correspondence set. For example, if both ontologies contain *eye*, when the algorithm lands on the corresponding node, it will have a chance to switch to the other ontology. As such the sentence will contain concepts from both ontologies and they will move closer together. See 3.5.5 for an example of the sentences generated by DeepWalk.

3.4.3 Node vectorisation

The most basic *Node2vec* model converts a node to a vector purely based on its own label. This method has one advantage over simple string matching: if no node is found in the other ontology that matches exactly, we can still find a node that is similar since it is close in the vector space. Therefore, this method should already be an improvement over the most basic string matching algorithms. It may even be competitive with more advanced string matchers (that look at substrings of labels). For example, *whisker* and *moustache* may be matched by a substring matcher, but are also semantically similar since they will often be mentioned in the same context. This method was chosen because it is the foundation of all other models that use the graph-context. It is required to see if the extensions improve results.

3.4.4 Bidirectional connections

More connections means bigger context. For neural networks more data is always a good things. Therefore it is likely that this will improve results somewhat.

3.4.5 Select from senses

One way to take into account the influence of neighbouring nodes is to train a multi-sense node embedding model as described in Section 2.4.4, and selecting from the different node senses based on which is closest to the context average.

3.4.6 Hot start

It is hard to classify the proposed algorithm in terms of the conventional ontology alignment method classes. It is not purely string matching, since it takes into account information from the context. However, it is also not structural, as it can work without a seed alignment and uses information other than the structure as well. It is not semantic (or logical) since it does not use inference. Nor is it terminological, which looks at dictionaries or other ontologies to find matches. However, it can use dictionaries and other ontologies to improve performance. The algorithm can find matches based purely on the given node and edge labels. However the algorithm might benefit from a hot start. Such a seed alignment might allow the algorithm to train with certain words that are known synonyms as if they were the same word, thus increasing the number of training

samples per word (on known relevant words) and decreasing sparsity. This may help the model, improving performance. However, since some information from other algorithms is now used, it would not be fair to compare the results to the cold start algorithm results. For example, it may just add the results from the generic string matcher to its own results, improving performance, without learning any new relationships. A new testing method had to be designed for this algorithm. This method would have to compare it to the string matchers, to see if it improved over their results, rather than compare it to the cold start algorithm. However, if we compare the cold and hot algorithms with the generic string matchers, we may be able to compare them indirectly. An example of a hot start is that *mouse torso* is already matched with *human torso*. Then it is likely that their neighbours are also related. Since the *front paws* and *arms* are next to the *torso*, they are more likely to be matched now. In the case of *Node2vec* they are more likely to be matched, since their neighbourhoods are the same (the only neighbour of both *arms* and *paws* is *connectedTo torso*).

3.5 System check

To ensure the results are valid, I ran a system check. The check is designed to detect the following conditions.

1. The data breaks the considered algorithms.
2. There are bugs in the developed system unrelated to the algorithms.
3. The considered context extraction does not work.

I will investigate all three to ensure they cannot be the cause of bad results and as a sanity check to prevent confirmation bias. If potential cause one is the case, the study is invalid and should be repeated on one or more different data sets. If either potential cause two or three is the actual cause, the study is invalid and the results should not be published until the system is fixed and the results do not change. If it is the case that none of these three potential causes is the possible, the results are valid, the results reflect the performance of the algorithm and valid conclusions can be drawn from this study.

3.5.1 The data

If the data set is broken, switching to another data set will substantially change performance. Of course, different data sets are inherently unequal in their difficulty. Therefore I will check the performance relative to the baseline and the best known algorithm. If it performs similarly relative to those references, it can be concluded that the data works as intended.

Possible reasons for the data causing bad results include the data set being too small, labels that are not useful or the data set having a structure that does not work with the algorithm. For example, the pre-trained *Word2vec* model is expected to work better on a data set that has commonly used words as labels.

The algorithm has been run on multiple data sets which are different. Despite that there is no significant difference in performance. For example, compare the F-measures for the medical dataset and the anatomy dataset. From Section 4 we can see that both F-measures improve over the baselines, therefore the datasets used do not seem to influence the algorithm by much.

3.5.2 The system

Since the system parts are connected in sequence, if any of the steps in the software system contains an error, the whole system fails. This could cause all algorithms to learn badly or the results to be misinterpreted. The system consists of a data loader, preprocessor, context extractor, the algorithm, the matcher and multiple visualizers. If all of these systems work properly, the system is not the reason for the underperformance. I will analyse every part except for the context extractor, which will be analysed in the next section.

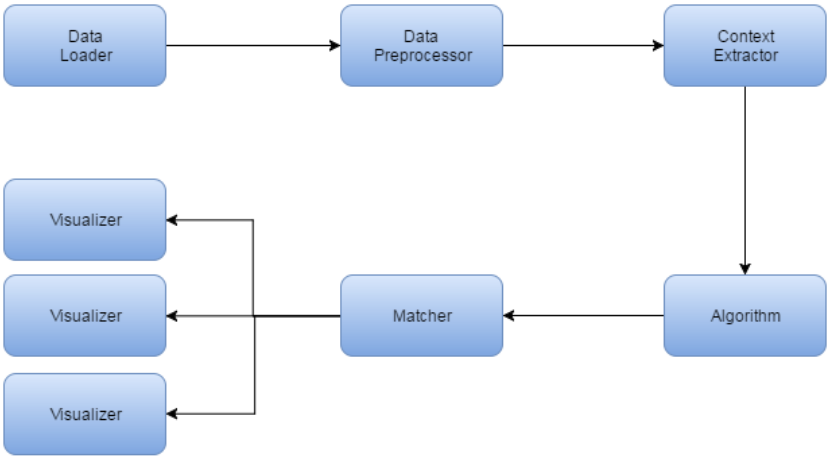


Figure 22: The program is structured in sequence, except the visualisers which are structured in parallel.

Preprocessing The preprocessing transforms the OWL graph into a graph with just the nodes and edges. Since OWL uses blank nodes to structure the ontology and connects everything to *Thing* and *Nothing*, two OWL concepts, those need to be resolved and removed, respectively.

We can conclude that this works if the context extraction works. Please refer to Section 3.5.3 for those results.

Matcher The matcher gets as input the embeddings generated obtained by the algorithm. It uses these to calculate the distance between nodes and output the most likely alignments. As such it completes the system so that it produces a set of alignments from the two ontologies.

The matcher is a simple KD-tree that finds the nearest neighbours of every node in the opposing ontology. The KD-tree has been tested by its creator and as such does not contain errors.

Data loading Since the preprocessing and context extraction works, we can conclude that the data is also loaded properly.

Statistics and visualisation Since the visualizers work in parallel, errors in them would not cascade to the others, so they are all faulty or none of them are.

Since every part works, the system is not faulty.

Secondly, as an extra measure, the following steps have been replicated by another piece of software: the data loader, preprocessor, context extractor and the algorithm. The results of that system are titled "alternative DeepWalk" in the results Section. As this system has been used in other projects and is written by someone else, if the results match the ones from this study, the system from this study works.

3.5.3 Context extraction

There are different context extractors for some of the different algorithms. I will research every one: the sentence generator of DeepWalk, the context extractor for the graph-based models and the context extractor for the pre-trained model. If all three work well, the context extractors are not the cause of the underperformance.

The context extractor for DeepWalk works as generates sentences as described in Section 3.5.5. Since the sentences are generated properly and DeepWalk only requires sentences, the context extraction for DeepWalk works.

For the Node2vec algorithms, the context extraction is more advanced. All edges are extracted from the ontology using Jena. Then, for every node the context is decided as every node it has an edge with. In the case of the bidirectional algorithm, both outgoing and ingoing edges count as being connected, whereas in the monodirectional algorithm only outgoing edges count as the connectedness. For the algorithms that also use the edge labels for context, double the context size is expected.

The context extraction of the pre-trained Word2vec consists of splitting up every label into multiple words which can be looked up. For example "Adrenal vein" can be split up into "Adrenal" and "vein". The pre-trained model does not use any other form of context integration.

First, I obtained the number of edges from VOWLview[20], an ontology graphing program, represented in Figure 23. This was necessary since the creators of the data did not share how many edges were in the database. With the actual number of edges, obtained, it is possible to compare them with the number of edges found. If these match, the context extraction works.

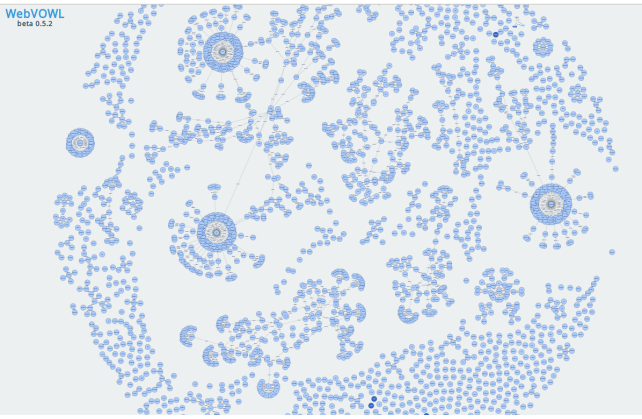


Figure 23: The mice dataset visualized in VOWLview. It shows the hierarchical nature of the data, as subclasses circle their superclass. It also shows that many individual classes are not connected, and as such structure cannot be used in the matching process.

Table 1: The number of edges compared to show that edge parsing was successful.

	#edges mice		#edges humans	
Algorithm	expected	actual	expected	actual
Truth*	**1810**		**3780**	
Monodirectional	1810	1807	3780	3761
Bidirectional	3620	3614	7560	7522

Since the expected number of edges and actual number of edges are very similar, we can conclude that the context extraction works. The difference can be explained by oddities in the VOWLview representation, as it contains edges that are not relevant to the ontology such as in Figure 24.

3.5.4 The algorithm

At the beginning of this section, I reasoned that there are three possible causes of system failure, namely, bad data, bad code and bad context extraction. I have shown that all three are unlikely to cause problems. Since those three possibilities have been excluded, it is certain that the results of this study are actual results.

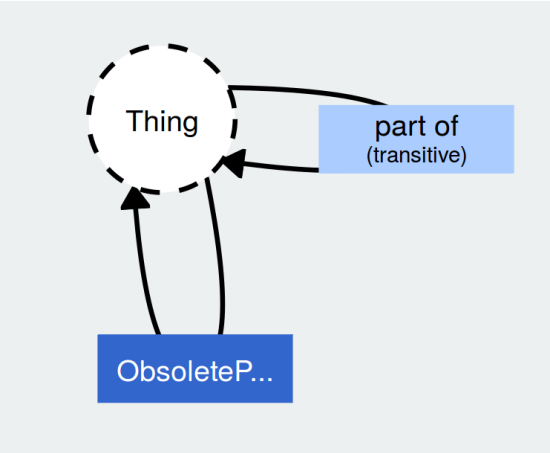


Figure 24: An example of two edges that VOWLview renders that are not relevant to ontology matching.

3.5.5 Example: DeepWalk sentences

Sentence lengths vary from 1 to about a hundred. The lower limit is due to the fact that some nodes do not have neighbours and as such the sentence is cut short. The upper limit not precise since the sentence end is random.

Below are some examples of sentences.

1) incus auditory_bone incus auditory_ossicle ear_part cymba_conchae
2) brain_nucleus brainstem_nucleus red_nucleus brainstem_nucleus raphe_pallidus_nucleus brainstem_nucleus trochlear_iv_nucleus brainstem_nucleus nucleus_of_trapezoid_body
3) hip_joint hindlimb_joint ankle_joint joint_by_site cricoarytenoid_joint

Figure 25: Example of a sentences that run over both ontologies. The words are color coded according to which ontology they occur in. Red is the mice ontology, blue the human ontology and purple words occur in both ontologies. Underlined words indicate a point at which the walker switches between the ontologies.

The first and third sentence both contain words that are in both ontologies. However, in the first sentence the switch does not occur immediately. The second sentence contains words from just one ontology. This will improve the embeddings for that ontology, but will not teach the algorithm about the mapping from one ontology to the other.

thyroid_gland_left_lobe

Figure 26: Some sentences contain only one word. Often these words refer to nodes with no neighbours.

3.6 Evaluation

The goal of this study is to see if the developed ontology matching algorithm performs better than the best existing matchers. In this section I will define better.

3.6.1 Evaluation methods

Since for the used data the ground truth is available, comparison is straightforward. Namely, the algorithm that finds the most items that are also in the ground truth in the lowest number of attempts. This definition, however, has two parts: the number of correct attempts and the total number of attempts. This makes it hard to compare two algorithms, as one may be better with few attempts allowed whereas the other may be better with more attempts allowed.

Precision represents the number of attempts: it is the ratio of the correct attempts divided by the total number of attempts. However, if you make only one attempt and it is correct, you cannot improve your precision by making more attempts. This solution is to balance precision with recall. Recall is the ratio of the correct attempts divided by the number of items in the ground truth. Often a higher recall results in lower precision as items from the ground truth become increasingly harder to recognise. For example, concepts with the same label are easy to link, but synonyms are much harder to link.

To balance precision and recall we use the F-measure. It uses a balance factor to favour either precision or recall. The F_1-measure is the harmonic mean of the two. The F_1-measure is used in this study. This is in correspondence to studies in the same area of research and thus allows for good comparison.

However there is still one problem remaining. The algorithm outputs a confidence measure for every possible alignment. This means that the F-measure depends on the confidence threshold to select which alignments become attempts. The optimal threshold can be decided by using a validation set to check the most likely thresholds.

We want to compare to a baseline to ensure proper interpretation of the results. For example, it is possible that a very simple baseline algorithm is already able to obtain a good score. In that case the results of the algorithm should be corrected. A higher score is still better, but an improvement from 0.1 to 0.2 is not as impactful as an improvement from 0.8 to 0.9.

To test the hypothesis it is also required to compare the results to the best scoring algorithm on the same data set. However, as the best scoring ontology alignment *system* uses many algorithms in an ensemble, it would be an unfair comparison. There are two solutions.

Firstly, if the ensemble is extended with the algorithm of this study, it is possible to see if the algorithm improves the ensemble and thus can find alignments the other algorithms cannot find. However, as the it is not possible to change the ensemble used in the best alignment system, this is not possible.

The second option is to compare the algorithm to every single algorithm in the ensemble. This is possible as the whole system is available and single

algorithm can be separately used. This method was used in this study.

4 Results

In this section I will show the objective results of the study. Firstly, I will show the precision-recall trade-off curve of all algorithms. Then I will go into detail of the results of the most promising algorithm.

4.1 Precision versus recall

4.1.1 Mice and men data

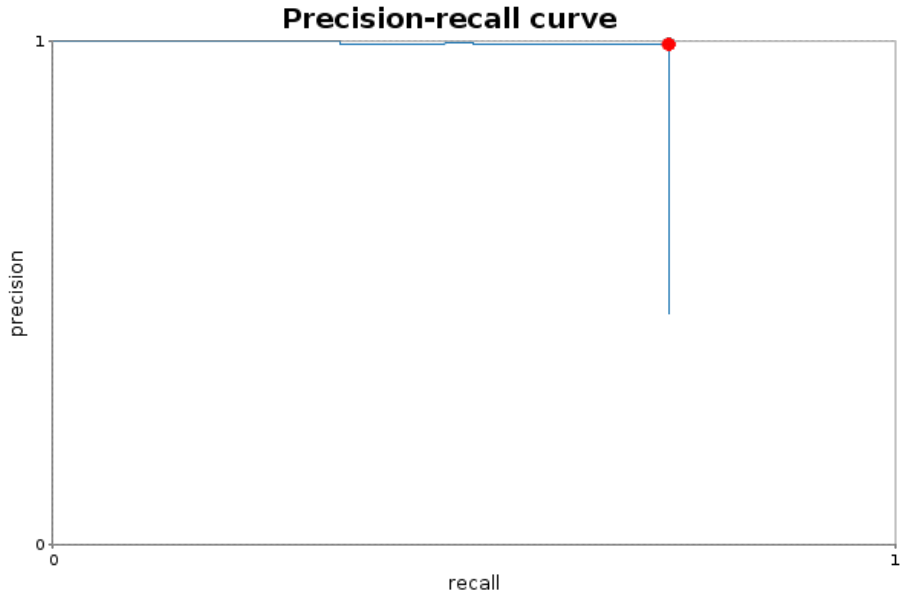
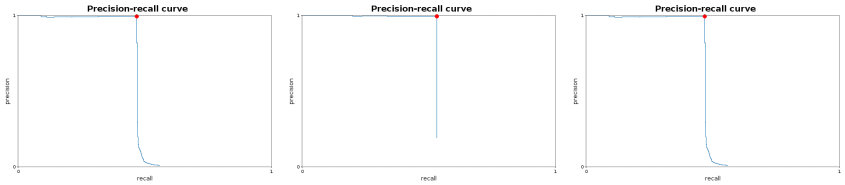


Figure 27: Precision-recall trade-off curve for the results of bidirectional Node2vec. Red indicates the maximum F_1-score.

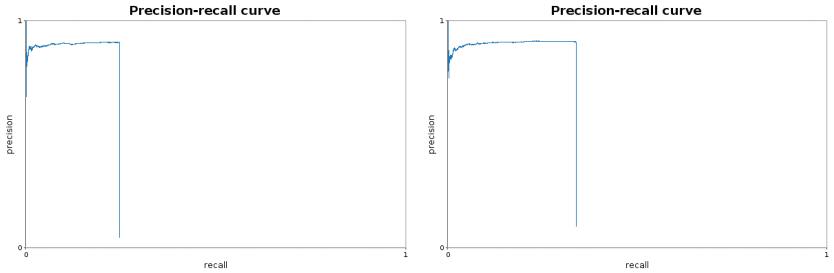


(a) Pretrained Word2vec algorithm.

(b) Random for comparison.

(c) Alternative DeepWalk algorithm.

Figure 28: Precision-recall trade-off curve for the results of three more algorithms.



(a) Pre-trained (b) Bidirectional

Figure 29: Precision-recall trade-off curve for the results of three more algorithms.

4.1.2 Medical data

4.2 F-score

Table 2: The F_1-scores for the different matchers for the mice and medical datasets compared to the result of AML at OAEI 2015 [6]

Matcher	F_1**-score**	
	Anatomy	Medical
Label Match	0.6926	0.4905
Random	0.6926	0.4905
Pre-trained	0.6926	0.3874
Bidirectional	0.8432	0.4946
Multi-sense	0.6926	0.4905
Deepwalk	0.6926	0.4906
Alternative DeepWalk	0.6926	0.4905
AML ensemble	**0.94**	**0.81**

4.3 Interpretation

4.3.1 Unexpected results

4.3.2 Expected results

5 Conclusion and discussion

In this section I will draw conclusions from the results, answer the research question and give a verdict on the hypothesis. In the second part of this section I will discuss these conclusions. In the third and last part of the section, I will describe a few avenues that might be interesting to explore further, including a number of implementations I did not consider and why.

5.1 Conclusions

The goal of this study was to answer the research question from Section 1.5:

Can distributed representations of concepts be used to find relationships between concepts in ontologies?

To answer this question I have examined the following points from Section **??** for every algorithm (Pre-trained Word2vec, DeepWalk, Node2vec):

The algorithm uses distributed representations of concepts From Section 3.4 it is clear that every single algorithm uses distributed representations. To summarize:

pre-trained Word2vec Uses existing distributed representations on every word in a label.

Node2vec All implementations of the Node2vec algorithms, monodirection, bidirectional and multi-sense, convert nodes into vectors.

DeepWalk Trains a normal distributed representation model on labels with the context generated from a random walk.

The algorithm finds relationships between concepts in ontologies This is correct as the inputs are always two ontologies and the output is always a set relationships between the ontologies.

The algorithm performs well One of the algorithms finds new relations. Therefore we can conclude that the algorithm can be used for ontology matching.

All algorithms perform worse than the full AML ensemble. However, as detailed further in Section 5.2, there is potential for embedding methods.

5.2 Discussion

This study shows that the algorithm of McCallum et al as shown in [23] does not help improve the results of ontology alignment more than current methods already do.

The results show that one assumption that was made by the developers of the Universal Schema was incorrect, namely the assumption that when combining

knowledge bases matching labels refer to the same entity and vice versa. For example, the assumption that Bill Gates is called Bill Gates in FreeBase as well as in all unstructured text does not hold up. In reality, there may be multiple entities named Bill Gates and the same entity may be labelled differently.

This assumption become even weaker as an increasingly growing part of the Semantic Web is automatically generated and as such different labels can be assigned to the same entity. One of the goals of ontology matching is to link those differently labelled entities.

In the following sections I will describe a number of extensions that might improve the algorithm further such that it may one day rival hand-crafted methods in their domain and be unparalleled in the big data domain.

5.3 Discarded algorithms

5.3.1 Combine with neighbours

To improve the model, context information from the neighbours of a node can be added. The most basic context adaptation would be to create a vector as the average of the context vectors. The resulting vector should be combined with the node vector of the node that is being investigated. This can be done by averaging or weighted averaging where the context is weighed more if there are more neighbours, though not necessarily linearly. For example, since the *front paws* and *arms* are both attached to the *torso*, their context vectors will be similar as well.

5.3.2 Adding edge labels

To add more information to the model, edge labels can also be added to the context of a node. This method effectively doubles the training data and context size when creating the context average. This should make the model more robust, although edge labels may be duplicate (one has many unique family members) and less informative in general, so a lower weight may be appropriate. A different possible problem with this method is that the relation between a neighbour node and its corresponding edge may be lost, since they are just treated as independent contexts. In the case of the mouse ontology, the *head* is above the *torso* so *isAbove* will be added to the context of *head*. The same goes for the *human head*, which will make them more similar.

5.3.3 Combining edges with nodes

As can be read in Section 2.4.3, it is possible to drop the bag-of-words assumption that the context is sequence invariant. This means we can for example have the context be (previous word, next word) and those words will be treated differently. Similarly we can separate the edge and node labels and treat them differently. This ensures that the model will find any relationship between the edge and the node if it exists and will take this into account. For example, *isBelow* and *torso* will combine into a vector that is similar to *tail* whereas

isAbove and *torso* combine into a vector that is similar to *head* since the edge and node labels are combined in the model, rather than considered separately. As the datasets used in this study did not provide any edge labels, the method could not be studied.

5.3.4 Training corpus

Word embedding models need to be trained on a large corpus. These corpora need to cover the concepts that are contained in the ontology, but also need to be large enough to build good embeddings. Since the ontologies themselves do not necessarily contain enough examples to embed the concepts properly, other ontologies that contain the same concepts can be useful. In the case of the mice and men ontologies, anatomical or medical ontologies may be used, since they may list more anatomical relationships or common relationships (for example *whisker* and *moustache* are related to *hair*). Since no such ontologies exist for the OAEI datasets, this extension could not be investigated.

6 Lists of figures and tables

List of Figures

List of Tables

7 References

[1] A Amin, MFJ van Assem, V de Boer, L Hardman, M Hildebrand, L Hollink, Z Huang, J van Kersen, M de Niet, B Omelayenko, et al. Multimedian e-culture demonstrator. 2006.

[2] Chris Bizer, Anja Jentzsch, and Richard Cyganiak. State of the lod cloud. *Version 0.3 (September 2011)*, 1803, 2011.

[3] Kate Byrne. Populating the semantic web: combining text and relational databases as rdf graphs. 2009.

[4] Jeremy J Carroll, Ian Dickinson, Chris Dollin, Dave Reynolds, Andy Seaborne, and Kevin Wilkinson. Jena: implementing the semantic web recommendations. In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, pages 74–83. ACM, 2004.

[5] Venkat Chandrasekaran, Nathan Srebro, and Prahladh Harsha. Complexity of inference in graphical models.

[6] Michelle Cheatham, Zlatan Dragisic, Jérôme Euzenat, Daniel Faria, Alfio Ferrara, Giorgos Flouris, Irini Fundulaki, Roger Granada, Valentina Ivanova, Ernesto Jiménez-Ruiz, et al. Results of the ontology alignment evaluation initiative 2015. In *10th ISWC workshop on ontology matching (OM)*, pages 60–115. No commercial editor., 2015.

[7] Graham Cormode and Balachander Krishnamurthy. Key differences between web 1.0 and web 2.0. *First Monday*, 13(6), 2008.

[8] Isabel F Cruz, Flavio Palandri Antonelli, and Cosmin Stroe. Agreementmaker: efficient matching for large real-world schemas and ontologies. *Proceedings of the VLDB Endowment*, 2(2):1586–1589, 2009.

[9] Andrew McCallum et al. Factorie github page, 2009.

[10] Jérôme Euzenat, Maria-Elena Roşoiu, and Cássia Trojahn. Ontology matching benchmarks: generation, stability, and discriminability. *Web Semantics: Science, Services and Agents on the World Wide Web*, 21:30–48, 2013.

[11] Dieter Fensel. Ontologies. In *Ontologies*, pages 11–18. Springer, 2001.

[12] Thomas R Gruber. A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2):199–220, 1993.

[13] Nicola Guarino, Daniel Oberle, and Steffen Staab. What is an ontology? In *Handbook on ontologies*, pages 1–17. Springer, 2009.

[14] Rohit Khare and Tantek Çelik. Microformats: a pragmatic path to the semantic web. In *Proceedings of the 15th international conference on World Wide Web*, pages 865–866. ACM, 2006.

[15] Sophie Le Moigno, Jean Charlet, Didier Bourigault, Patrice Degoulet, and Marie-Christine Jaulent. Terminology extraction from text to build an ontology in surgical intensive care. In *Proceedings of the AMIA Symposium*, page 430. American Medical Informatics Association, 2002.

[16] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195, 2015.

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

[18] Feiyu Lin and Kurt Sandkuhl. A survey of exploiting wordnet in ontology matching. In *Artificial Intelligence in Theory and Practice II*, pages 341–350. Springer, 2008.

[19] Wang Ling, Chris Dyer, Alan Black, and Isabel Trancoso. Two/too simple adaptations of word2vec for syntax problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1299–1304, 2015.

[20] Steffen Lohmann, Stefan Negru, Florian Haag, and Thomas Ertl. Vowl 2: user-oriented visualization of ontologies. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 266–281. Springer, 2014.

[21] Alexander Maedche. *Ontology learning for the semantic web*, volume 665. Springer Science & Business Media, 2012.

[22] Frank Manola, Eric Miller, Brian McBride, et al. Rdf primer. *W3C recommendation*, 10(1-107):6, 2004.

[23] Andrew McCallum. Representation and reasoning with universal schema embeddings presentation. `http://videolectures.net/iswc2015_mccallum_universal_schema/`, 2016.

[24] Deborah L McGuinness, Frank Van Harmelen, et al. Owl web ontology language overview. *W3C recommendation*, 10(10):2004, 2004.

[25] Peter Mika. On schema. org and why it matters for the web. *IEEE Internet Computing*, 19(4):52–55, 2015.

[26] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[27] Alistair Miles and Sean Bechhofer. Skos simple knowledge organization system reference. *W3C recommendation*, 18:W3C, 2009.

[28] Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. Efficient non-parametric estimation of multiple embeddings per word in vector space.

[29] Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. Multi-sense skipgram implementation, 2014.

[30] Sibangiso Ngwenya and Khesani Richard Chilumani. A semantic web solution for information overload, 2011.

[31] Martin Odersky, Philippe Altherr, Vincent Cremet, Burak Emir, Sebastian Maneth, Stéphane Micheloud, Nikolay Mihaylov, Michel Schinz, Erik Stenman, and Matthias Zenger. An overview of the scala programming language. Technical report, 2004.

[32] Tim O'reilly. What is web 2.0: Design patterns and business models for the next generation of software. *Communications & strategies*, (1):17, 2007.

[33] Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. Relation extraction with matrix factorization and universal schemas. 2013.

[34] Mario Schlosser, Michael Sintek, Stefan Decker, and Wolfgang Nejdl. Hypercup—hypercubes, ontologies, and efficient search on peer-to-peer networks. In *International Workshop on Agents and P2P Computing*, pages 112–124. Springer, 2002.

[35] Pavel Shvaiko and Jérôme Euzenat. Ontology matching: state of the art and future challenges. *Knowledge and Data Engineering, IEEE Transactions on*, 25(1):158–176, 2013.

[36] Vikram Singh, Pradeep Joshi, and Shakti Mandhan. Concept integration using edit distance and n-gram match. *International Journal of Database Management Systems*, 6(6):1, 2014.

[37] Lina Zhou. Ontology learning: state of the art and open issues. *Information Technology and Management*, 8(3):241–252, 2007.

8 Appendices

Appendix 1

Appendix 2