Ontology Matching with word2vec

Harmen Prins

April 16, 2016

Acknowledgements

Abstract

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 four steps.

Firstly, you must open up an internet Movie Database, 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 or movies the Database recommends for you and check if any movies on that list match your full criteria.

Once you have found one or more movies that match your search criteria, you have to select a time at which you can see the candidate movies, so you check for every movie every available 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.

The next step is to filter the number of time slots based on your agenda, so you cross-reference the available movie times with the times you are available during the weekend.

Then onto the last 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 feel like filling in a poll about availability or movie choices.

Most of the time people will limit their options to avoid these steps, considering only a few movies, cinemas, time slots and friends. But this is 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 piece 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.

1.1 Short Context

In this section, I will give a short summary of the context for the problem. This 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.

On the Semantic Web, Ontologies are used to define and share knowledge. 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. Without ontologies, agents on the Semantic Web would be unable to communicate, as all concepts are defined in an Ontology.

However, using one single ontology is infeasible. This is due to the fact that inference and search time scales with Ontology size, 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 or corresponding Knowledge Base, these lookups need to be fast.

The solution to this problem is to split this Ontology into many smaller Ontologies. However, as often is the case in Computer Science, this creates two new, albeit smaller, problems.

The first problem is the problem of splitting the Ontology. How do we find the distribution of all concepts over the different Ontologies to maximize their usefulness. Usefulness in this case should take into account co-occurrence of concepts. Naturally, connected concepts should be more likely to occur in the same Ontology as concepts that are not connected, since it is likely that agents that want to know about a concept also want to know things about its neighbours.

The second problem, the one I address in this thesis, is the problem of conversations that cover multiple domains. These conversations will occur more often if there are more Ontologies, but they can always occur if the number of Ontologies is larger than one. When a conversation requires information from multiple domains, the Ontologies corresponding to those domains need to be merged. This means connecting the related concepts from the different Ontologies. Some research has focused on ontology merging, with promising but slowing results.

The problem of Ontologies on the Semantic Web is to find those Ontologies that are as small as possible to improve search speed, but need to be merged as little as possible, to ensure as little time is used to merge the Ontologies before they can be searched. If it is possible to quickly and reliably merge Ontologies, Ontologies can be made small and communication on the Semantic Web can be done quickly.

Recently, a model that embeds words in a vector space that appears to contains semantics has become popular. Since it is unsupervised and can be used to calculate a semantic distance between words, it may very well be applicable to ontology matching.

1.2 Overview

This thesis is structured as follows...

2 Context

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. Then the concept of the Ontology itself will be addressed. Lastly, current alignment strategies, their problems and other research and technologies will be discussed.

2.1 The Semantic Web

The Semantic Web, also called Web 3.0, is a concept of a Web that enhances the User Experience in ways that the current Web can not provide. Web 2.0, the current Web, models the internet as *hubs* (or sites) with a specific purpose and limited interconnectivity. 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. 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. The Semantic Web, on the other hand, contains sources of *information* that can be combined whenever the user needs it. To stick with the social media example, a person can create a piece of content, say a picture of a tree, en shares 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.

The linking of every piece of content to another allows not only humans to browse the internet, but also virtual agents, which can use this information to aid their users in ways that are simply impossible nowadays. If someone wants to find information on a certain topic, the 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.

Another possibility is the scenario from the introduction **??**, 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 human 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.

2.1.1 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.

2.1.2 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 Semantic Web is currently divided into two groups: the people who need high accuracy and the people who need high recall. The first group consist of scientists that work on specific fields that require exact knowledge, whereas the second group are data scientists that want big amounts of data available. In the past it was not possible to extract information from texts and other sources automatically, so

The group that requires high recall use hand-crafted knowledge bases, created by experts, containing only facts that are accurate, verified and relevant. 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. [2, 6]

The group that requires a high recall, i.e. requires a huge amount of data, automated methods are required. These 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. [17]

Beside these knowledge bases, many other pieces of the Semantic Web are already in use. For example, the W3C [11, 12]

Large ontologies already exist, with the largest, DBPedia, boasting almost two billion triples and the three hundred combined Large Open Data databases contain over thirty billion. [7, 1] The LOD is simply one huge alignment of hundreds of 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.

Overall it is clear that

2.1.3 Technologies

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 'part-of' relationships: birds are part-of the animals concept class. 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.

2.2.1 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.

2.2.2 SPARQL

SPARQL is the query language for knowledge bases and ontologies. It allows a user to select nodes whose context match the query, similar to how rows are selected in SQL based on the row contents.

2.2.3 Role

The reason ontologies are used in databases consists of multiple parts. 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.

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.

Aristotelian ontologies have the advantage of being easier to create. One could use statistical methods given a sample of all possible observations or crowd sourcing to create concepts and constraints. This would enable enormous ontologies and knowledge bases with the 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.

Another way to represent this spectrum is the precision-recall trade-off. A Platonic ontology has maximum precision but its recall is limited since it can only represent very little if creation budget is limited, or paradoxes exist. The Aristotelian ontology on the other hand, can theoretically reach perfect recall given enough storage size and observations for its statistical methods. Again, in reality this is limited by a budget. All ontologies fall between these two extremes. Often, a combination of statistical methods with expert confirmation is used.

Platonic 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.

2.3 Current alignment strategies

Many different alignment strategies have already been developed. All strategies follow the same two-step approach. The two steps are independent and as such different methods can be used interchangeably. The first step is to generate an initial correspondence set, where correspondences between nodes are found based on just their labels and meta-data. The second step is to use this initial set and the structure of the ontology to find more correspondences. These steps are called the terminological and the structural steps. Some strategies also

use extra steps like the extensional and semantic steps, which use vector space models and inference, respectively. https://hal.inria.fr/hal-00917910/document

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. [9] 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.[16, 8]

2.4 Problems in ontology alignment

There are a number of different problems ...

How should the wildly differing demands the applications place on the ontology matching systems be satisfied by a single of 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.

Online matching would allow for updates in a knowledge base to be represented in the alignments of that knowledge base. This would be required on any ontology that tracks changing knowledge, like stores, movie theatres and social media, for example.

Based on the previous two points, we can state that 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.

Matchers should be able to explain the results it produces, i.e. making those results more interpretable. This would allow a number of other improvements to occur. Firstly, it would allow users of a system to have more confidence in the results and make better decisions based on the source of the result. Secondly, it would allow people manually aiding the alignment to check if the matcher did a good job, and thus improve the interaction between the system and the domain expert. Thirdly, it would allow researchers to improve the matching system because they are able to detect the source of errors.

When these 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.

2.5 Other related research

2.5.1 Word sense representation

Skip-gram models have been widely used to represent words as vectors. This is useful as it is easy to calculate the distance between two vectors, which is a useful property for ontology matching. This vector distance has been shown to be related to the semantic similarity between words, which 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.

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. /citew2v Word2vec learns this representation by trying to predict the context of a word. First the vocabulary is one-hot encoded. Then the model learns to predict a context vector that minimizes the distance to the context one-hot encoded vectors. The model is a neural network with one hidden layer. Since the input is one-hot encoded, every word in the vocabulary is represented by one row in the weight matrix between the input layer and hidden layer. This row is the vector representation of the word.

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 which words go where in a sentence. This preservation was done by having multiple output layers, one for every word in the context. The research shows that the input does not have to be a bag-of-words representation, which will be useful later on. [10] 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. This will help with resolving ambiguity, a very important factor in ontology alignment.[15]

2.6 Used technology

2.6.1 AgreementMaker

AgreementMaker is an ontology matching system that obtained the highest F-measure in 6 of the 7 ontology matching tracks of OAEI 2015. [5] The system has also been updated the most recent at the time of writing[3].

The system uses an extensible framework which allows its users to add new modules to the system. This is very useful for research as it allows the researcher to compare different modules by swapping in just those modules in the framework. The module-swapping technique will also be used in this work, but more of that will be discussed in 5.1.

2.6.2 Factorie

"FACTORIE [is a] a toolkit for probabilistic modeling based on imperatively-defined factor graphs" [4]. As a toolkit it contains many different parts that can be used together. One example is the NLP package in the toolkit, which provides many algorithms to process text. One of those algorithms is the distributed representation of words algorithm by [13].

Factorie is written in scala, a relatively new language that interfaces with java. See **??**. This allows for interoperability between factorie and AgreementMaker. Any algorithm that is implemented in factorie or on top of factorie can therefore be used as a matching algorithm in AgreementMaker.

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.

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.

When model training is extremely fast, which is the case for parallelised neural network training, parsing the data from string format to vectors in a proper data structure can become a bottleneck. Recently, factorie added JFlex to parse its data. This improved its parsing by %5000, removing the bottleneck.

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.

3 Problem

3.1 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. Those ontologies contain rules about concepts and concept classes in the shape of triplets, with two concepts and a relation. For example, the rule `Professors are humans` contains the concepts *Professor* and *human* and relationship *being*.

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.

Lastly, there is the problem of different types of alignment relations. For example, one concept can be a *part of* another concept, or an example of one. These different relations all require a different approach and come with their own sets of problems.

In summary, ontology aligning is required for multi-domain communication but poses a number of problems that make it difficult for concepts, the building blocks of ontologies, to be matched.

3.2 Research questions

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 terminological matchers?

I will also investigate the following two related but different questions:

Can distributed representations use the context of a concept to find relationships between concepts in ontologies better than existing structural matchers?

Can distributed representations perform well in a domain it was not designed for?

3.3 Hypotheses

Since multiple implementations of the algorithm are considered, every hypothesis that follows is considered for every implementation.

This specific matcher that uses a distributed representation of concepts can find relationships between concepts in ontologies better than existing terminological matchers.

To prove this hypothesis, I must proof the following points:

- The algorithm uses distributed representations of concepts.
- The algorithm finds relationships between concepts in ontologies.
- The algorithm has been compared to the currently best matchers.
- The algorithm performs better.

The first three points will be shown in **??**, whereas the last point is shown in **??**.

The distributed representations matcher can use the context of a concept to find relationships between concepts in ontologies better than existing structural matchers.

The distributed representations matcher can perform well in a domain it was not designed for.

The following hypotheses are true if for at least one implementation the corresponding hypothesis is true.

A distributed representations of concepts matcher can find relationships between concepts in ontologies better than existing terminological matchers.

A distributed representations matcher can use the context of a concept to find relationships between concepts in ontologies better than existing structural matchers.

A distributed representations matcher can perform well in a domain it was not designed for.

3.4 Justification

On the Semantic Web, many relatively small knowledge bases will be combined temporarily so that it is possible to reason over the combined knowledge of those knowledge bases. 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 over 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.

Since most of the Semantic Web will consist of Aristotelian ontologies, some of the information it contains can be unreliable and even wrong. Therefore, ontology matching that is capable of handling this uncertain data is required.

Calculating the distance between vectors can be done very efficiently, therefore ontology matching with distributed concept representations fits that efficiency criteria. Since these representations can use the context of a concept, they are also most robust than methods that do not take the context into account.

Since distributed representation algorithms can be trained unsupervised on new domains, they can be applied to those new domains without effort from humans, unlike current matching algorithms. This is also an improvement for the semantic web.

However, ontology matching is also used outside of the semantic web. It is already in use for companies that want to merge knowledge from other companies, for example. 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 in any situation, not just on the Semantic Web.

4 Method

In this section, I will describe the method that I developed. Firstly, I will describe the general idea. In the second section, I will give a theoretical justification for why this method should work and how it works conceptually. After that I will describe the different ways that the input graphs can be used to train the word2vec model and create the alignment, some of which I have discarded on theoretical grounds as I will explain in the relevant section. Lastly, I will talk about the implementation process.

4.1 Idea

As explained earlier, word2vec is a model that learns two things: implicitly it learns the semantic vector representation of words, and explicitly it learns to predict words from their context or vice versa. I will use both of these pieces of information in our alignment algorithm. 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. It should also be relevant to the labels that are in the ontologies. This is a problem since the ontologies may be relatively small, so they need to be extended while keeping relevance.
- 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 and extend it with multi-sense embeddings. Also hot-starting is considered as an improvement.

4.2 Considered implementations

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 7. Some implementations are extensions of other implementations, but all implementations are mentioned for completeness and to enable explanation of the source of improvements.

4.2.1 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, I need to use ontologies that are likely to contain the concepts I want to align. As a benchmark, I will use the WordNet and NELL ontologies. To check if adding the ontologies that are to be aligned helps, I will also train a model on an ontology consisting of WordNet, NELL and the two alignment candidates.

4.2.2 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, *ear lobe* and *ear* will be matched by a substring matcher, but are also semantically similar since they will often be mentioned in the same context.

4.2.3 Combine with neighbours

To improve the model, we can add context information from the neighbours of a node. 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. These possibilities were also investigated.

4.2.4 Select from senses

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

4.2.5 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.

Another problem with this method is that the relation between a neighbour node and its corresponding edge is lost, since they are just treated as independent contexts.

4.2.6 Select from senses

This method is exactly the same as the selecting from node senses with just the neighbour nodes except that you now also take into account the edge labels.

4.2.7 Combining edges with nodes

As can be read in appropriate section, 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, the system might use just the information 'isMarriedTo' to infer that someone is a human, but may infer that 'isMarriedTo' in combination with a man usually refers to a woman.

4.2.8 Hot start

It is hard to classify the algorithm in terms of the normal ontology alignment methods. 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 logical since it does not use inference. Nor is it terminological, which looks at dictionaries or other ontologies to find matches. However, it does use other ontologies (or text corpi) implicitly, so if it must be grouped it would be terminological. 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.

4.3 Discarded implementations

4.3.1 Training on text

Rather than training on ontologies, it would be possible to train the model on text. Since there is more text available than ontologies when counting the number of training samples, it would make sense to use a purely text-trained model. However, this method has downsides which made me decide not to use it. Namely, the context of a word in a text corpus is very different from a node in an ontology. A word is surrounded by the words directly around it, or the sentence it is in, or the paragraph or entire document. However all these contexts are just words, usually represented as a bag of words to ensure enough data for the model. In an ontology, the context is a combination of a node with an edge, which have a very different relation than merely co-occurring. The relation may be inferred from one sentence, or many different documents.

4.3.2 Just context nodes

4.3.3 Just edges

4.3.4 Context matrix representation

4.4 Implementation

The two foundations the algorithm has been build and tested on are Factorie, described in 2.6.2, and AgreementMaker, described in 2.6.1, respectively. Earlier work on a forked branch of factorie was done by [14] to implement multi-sense word2vec.

5 Evaluation

The goal of the research is to find an ontology alignment algorithm A and proof that it performs better than current techniques. However, as current techniques may combine many different algorithms and sources of information, maybe such a comparison is not fair, nor is a comparison with individual algorithms as they may be optimised for being combined. So, instead we may want to see if performance is improved when adding A to an ensemble of algorithms $E = (A_1, .., A_n)$.

5.1 Evaluation methods

The evaluation should be done using a performance measure $P(Al, T)$ that takes the alignment an algorithm produces given an ontology and the ground truth. If both the alignment and truth are considered sets of alignments, i.e. $Al = \{a_1, .., a_n\}$ and $T = \{a'_1, .., a'_m\}$, then the True Positives $TP = \{a | a \in Al, a \in T\}$, False Positives $FP = \{a | a \in Al, a \notin T\}$ and False Negatives $FN = \{a | a \in T, a \notin Al\}$ determine the performance of the alignment. The True Negatives do not matter, as there are extremely many, and the goal of ontology matching is to find True Positives. Therefore the Jaccard similarity should be used as a performance measure.

The Jaccard similarity is calculated as follows

$$J = \frac{TP}{TP + FP + FN}$$

i.e. the fraction of the true positives over the sum of the true positives and mistakes.

The Jaccard similarity is a number between 0 and 1, with a higher number being better. Therefore if we compare two alignments, the one which is most similar to the truth, i.e. has a higher Jaccard similarity, is the better one.

However, if we want to find out if an algorithm improves an ensemble, we calculate the improvement as follows:

$$I = J_{E \cup \{A\}} - J_E$$

where J_E is the Jaccard coefficient of an ensemble E. To see if an algorithm is useful to an ensemble, the improvement score I has to be *significantly* higher than zero. To check if a score is significant, a statistical significance test will have to be carried out. The test works as follows: generate N solutions randomly and score them. Then score the method you want to investigate. If the score is in the top p%, it is significant, otherwise it is not. In this case the solutions are $E \cup \{R\}$, where R is an algorithm that generates random alignments.

It is in principle possible to weigh False Positives and False Negatives differently, but this is not done in practice. To make the results of this study easily comparable to others, equal weights will be used.

5.2 Data

The data used comes from a list of data sets that the OAEI provides. 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.

A reference alignment can be made by experts or be created by bipartitioning an existing ontology.

5.2.1 Parts of OWL ignored

6 Algorithm

6.1 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 maps words to a vector space that has been shown to contain semantic properties. 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.

6.2 Theory

6.3 Product

7 Results

8 Conclusions

8.1 Recap research questions

8.2 Recap hypotheses

8.3 Conclusion per hypothesis

8.4 Overview of results

8.5 Final conclusion

9 Discussion

9.1 Interpretation

9.2 Unexpected results

9.3 Expected results

9.4 Future research

References

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

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

[3] 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.

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

[5] Daniel Faria, Catarina Martins, Amruta Nanavaty, Daniela Oliveira, Booma S Balasubramani, Aynaz Taheri, Catia Pesquita, Francisco M Couto, and Isabel F Cruz. Aml results for oaei 2015. In *ISWC International Workshop on Ontology Matching (OM), CEUR Workshop Proceedings*, 2015.

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

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

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

[13] 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.

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

[15] Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. Efficient non-parametric estimation of multiple embeddings per word in vector space. *arXiv preprint arXiv:1504.06654*, 2015.

[16] 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.

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

A Appendix 1

B Appendix 2