Radboud University Nijmegen

Master's thesis in Computing Science

Word translation with Wikipedia

Author:
Bas van Berkel

Supervisor:
Dr. Suzan Verberne

July 15, 2016

Abstract

Wikipedia contains a large collection of articles which are connected across languages. In this research two models are proposed which use this structure to obtain word translations. The first model solely uses the link structure within and between two language projects. The second model uses cosine similarity between word embeddings and uses the first model as seed model. The second model uses two monolingual corpora and is based on the assumption that the most similar words in the source and goal language of a word are the same. The first model's performance lies between 30-55% correct translations, the second model has a performance between 1-22% correct translations. The poor performance of the second model shows that the assumption that translation is possible based solely on similar words might be too naive.

Contents

1 Introduction

Wikipedia is an online encyclopedia written by volunteers. Currently Wikipedia holds millions of articles in multiple languages. These articles are connected across languages by inter-language links. Within each language articles also link to each other from relevant words in their texts. The large amount of text in Wikipedia and the way in which it is connected could possibly be used for word translation.

Word embeddings are numerical vectors which represent words in a corpus. Each word is represented by a word embedding. With the arrival of word2vec, word embeddings can be calculated faster than before [5]. This allows one to train word embeddings on very large corpora. Word2vec is trained using a skip-gram model. The skip-gram model optimises the word embeddings by maximising the prediction of a word by another word in the same sentence [5].

Using word embeddings for translation could result in a new way of translating, where one can use two monolingual corpora, and thus use all the available text in two languages, to generate a translation corpus from.

In this thesis I propose two models for word translation using Wikipedia. The first model uses the inter-language connections between Wikipedia articles in different languages. This first model is also used to generate seed translations for the second model. The second model is a more principle attempt to derive word translations using two monolingual corpora. The second model uses similarity measurements between word embeddings and combines these with the inter-language connections to obtain translations.

This thesis addresses two research questions:

1. To what extent is it possible to obtain good word translation performance using two unconnected monolingual corpora in the source and goal language?

2. How can the inter-language structure of Wikipedia be used in word translation?

Both models are evaluated using mean reciprocal rank (MRR) and the percentage of correctly translated words (hit rate). Besides looking at performance I will also investigate which types of words are the easiest and the most difficult to translate and which issues arise with both translation methods. Finally I will investigate the possibilities and issues of using Wikipedia as a corpus.

2 Background

2.1 Machine translation

Most high performing machine translation methods, such as the statistical machine translation method by Och and Ney [8], are trained on parallel corpora, sometimes helped by hand generated bilingual lexicons. In parallel corpora texts in the source and goal language are aligned on a sentence level. A downside to parallel corpora is that the amount of available text is limited. Depending on the language a sufficient amount of text might, or might not, be available.

Another option of corpora for translation are comparable corpora [10]. Comparable corpora consist of a set of texts which are aligned across two languages based on topic. There are more of these texts available, thus there is more text to generate a comparable corpus from [11]. When comparable corpora are used for machine translation mostly a bilingual seed lexicon is used to connect the two languages [11].

An even further step would be to use all available text in both the source and goal language to generate two unconnected monolingual corpora. This allows for all the available text in a pair of languages to be used. Separate models could be created for each of the two language corpora. Similar as with comparable corpora a seed lexicon can be used to connect the two language corpora. These seed words can be generated from a parallel corpus or bilingual lexicon.

2.2 Word embeddings

Word embeddings are vectors of real numbers which represent each words in a corpus. Each word is represented by its own word embedding. Word embeddings can be used in natural language processing tasks such as finding similar words. In 2013 Mikolov et al. introduced a skip-gram model [5], called word2vec, which can learn word embeddings from huge amounts of texts. The skip-gram model optimises the word embeddings by maximising the prediction of a word by another word in the same sentence [5]. Words within a certain range around the current word are predicted by classifying the current word using a log-linear classifier. Words are weighted according to their distance. Especially the efficiency in calculations of word2vec makes it usable on a large scale [6].

Using word embeddings it is possible to perform linear calculations on words. For example adding the vector of “king” to the vector of “woman” and subtracting the vector of “man” results in a value which has the vector of “queen” as the closest vector.

2.3 Wikipedia

Wikipedia is a crowd sourced encyclopedia where everybody can contribute by writing articles on relevant topics. There are Wikipedia versions in 293 languages, 58 of those languages contain more than 100,000 articles [12]. The articles in Wikipedia range from short bot-generated stubs to very large and thoroughly reviewed articles. Articles on the same topic are connected across languages via inter-language links, which are called interwikis. A majority of the articles is about a named entity, but there are also articles on regular words. In the articles some words, mainly those which are considered relevant further reading, are linked to other articles in the same language Wikipedia. The visible text, or anchor text, and the article it links to, are often the same. In (almost) all cases the text links to the most relevant article describing the concept.

3 Methods

First both proposed models are explained. In the second subsection the data which is used is described and in the final subsection the approach for the experiments is described.

3.1 Models

3.1.1 Translation using inter-language links

The inter-language links from Wikipedia can be used as a bilingual lexicon. As these inter-language links connect the same topic in different languages they likely contain useful information on possible translations for words. There is however not an article for all words in all languages. And often there are multiple interpretations of one word or term which are described in multiple articles. To account for this variety of possibilities the links between articles within one language version and their respective anchor texts are used to determine which article is the most relevant given a certain a term.

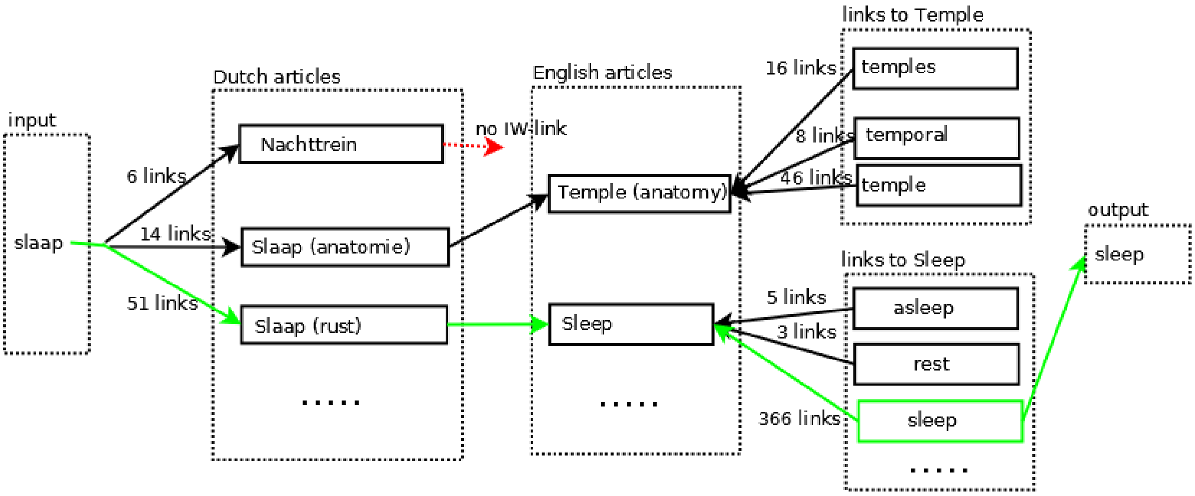

Figure 1: An example of how the word "*slaap*" is translated to "*sleep*" using the inter-language links of Wikipedia. In a first dictionary all usages of *slaap* as an anchor text are listed, with the articles they link to. Using this dictionary the most frequently linked article is found, in this case *Slaap (rust)*. After that a table of inter-language connections is used to determine which article in the English Wikipedia is the equivalent of *Slaap (rust)*, that is *Sleep*. In the next step a dictionary containing all anchor texts, which link to the article *Sleep*, is used to determine which anchor text links to *Sleep* most often. This is *sleep*, the most linked word is also the output or proposed translation. If *Slaap (rust)* would not have an English article which is connected, the second most frequent article (*Slaap (anatomie)*) would be used, and so forth.

The model can best be described by describing the different steps which are taken to translate a word. In figure 1 an example of how the word "*slaap*" ("sleep") is translated to "*sleep*" using inter-language links is given.

For the first step a dictionary containing, for each unique anchor text, the articles it links to in the source language is generated. Using this, for each word which is used as an anchor text at least once, the article which is most often connected to that word can be found. The article which is linked the most from the anchor text is seen as the best possible article to use in the translation. If a word is never used as an anchor text this step cannot be performed, and thus the word cannot be translated.

The second step is to find the article in the goal language which is linked, via an inter-language link, to the article in the source language. If the most frequently linked to article from the previous step has no inter-language link than the second most frequent article is used, and so forth. If none of the articles has an inter-language link then this step cannot be performed, and thus the word cannot be translated.

The third step is the inverse from the first step. For each article there is a list of all the anchor texts which link to the article, with the respective numbers of links from each anchor text. The anchor text which links most frequent to the article is the text that will be proposed as translation. If the article is never linked to then this step cannot be performed, and thus the word cannot be translated.

3.1.2 Word embedding similarity

The word embeddings model uses similarity between words both in the source and the goal language to translate words. As the similarity between words the cosine similarity between their word embeddings is used. The word embeddings are generated by training word2vec [5] on all articles of Wikipedia. This results in two separate word embedding networks, one in the source language for the translation and one in the goal language.

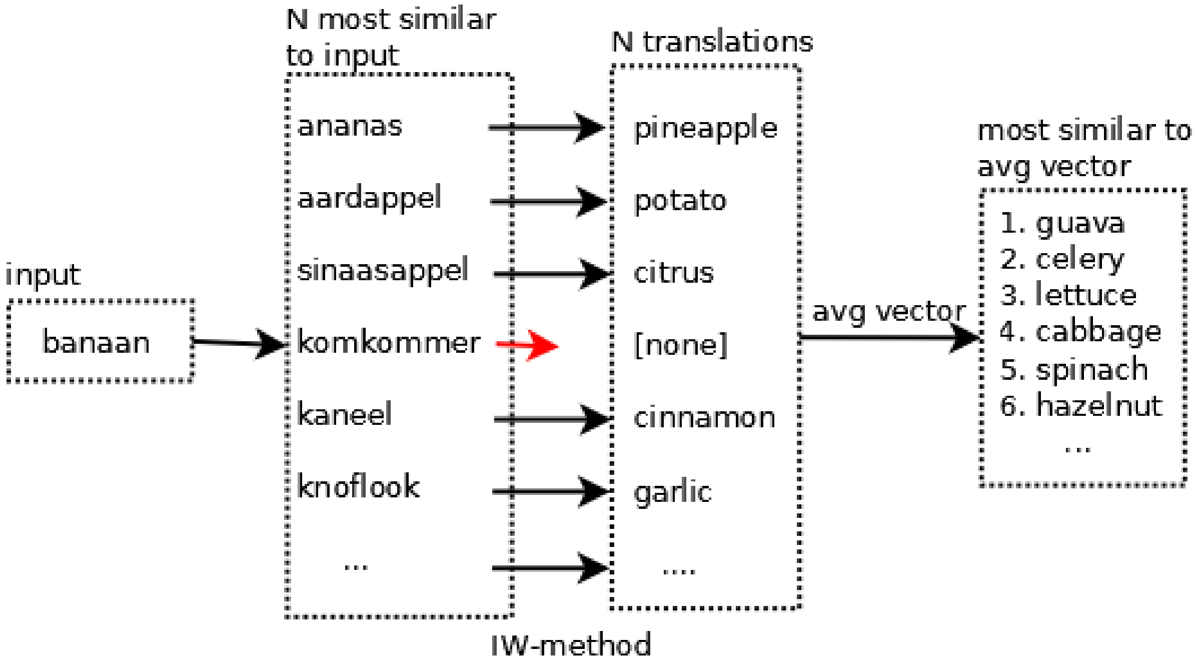

Figure 2: An example of how the word "*banaan*" ("banana") is incorrectly translated to "*guava*" using the similarity between word embeddings. In the first step the N vectors with the biggest cosine similarity to the vector of "*banaan*" are found. All these words are then translated using the inter-language method (described above). Some words cannot be translated using this method. In the final step the average vector of the translated words is used to find the words with the biggest cosine similarity to the translated most similar words. This results in a ranked list of proposed translations. In the case of this example, at least, the first 6 suggestions are incorrect.

I will again describe the method by describing the different steps of the method. To translate a word the first step is to determine the N words which are most similar to the source word. After this step these N most similar words have to be translated. To translate these words a bilingual lexicon is necessary. As a bilingual lexicon the inter-language method, described in the previous subsection, is used. This causes two issues. Firstly the inter-language method does not always result in a translation, secondly the translations are not always correct. This means that the word embeddings method has to be robust enough to handle the fact that not all similar words can be (correctly) translated. All the similar words which can be translated are used to determine the word which is the most similar to these translated words in the goal language. That word, the most similar in the goal language, is proposed as the translation of the input word in the source language. In figure 2 you can see how the Dutch word for "banana" ("*banaan*") is (incorrectly) translated using the word embeddings model.

For both the source and goal language the plain article texts from Wikipedia are used to train the word embeddings on. Anchor texts are considered as one word. When multiple words are present in an anchor text they are connected using underscores. This way concepts that are described by a single word in Dutch can be translated to their equivalent English concepts described by multiple words. For example "slaapzak" which translates to "sleeping bag". Named entities described by multiple terms in both languages, for example "Zuid Afrika"/"South Africa", can also be translated because of this.

By using similarity, in both the source and goal language, some implicit assumptions are made. An important assumption is that a word and its translation have exactly the same function in the two languages. If the word means something different in the two languages than one would expect the most similar words to be different. When words are ambiguous this could cause issues because the word will have multiple functions. If the translations of the multiple concepts of the ambiguous term are different then one would expect the most similar words to be different in the two languages.

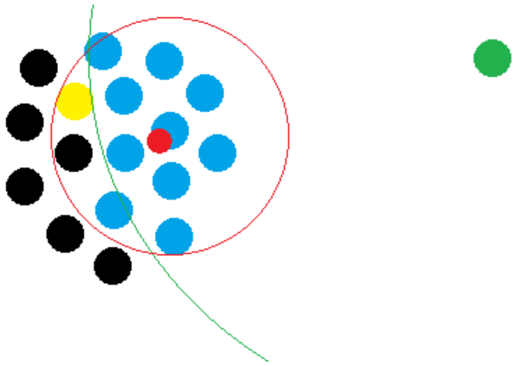

Figure 3: An example where the assumption that a word is closest to the average vector of its most similar words does not hold. The green dot represents the source word, the 10 blue dots are the 10 words which are most similar to the source word (cut-off line in green). The red dot is the average of the blue dots and represents the average vector. The red circle shows the closest dots to the red dots. The yellow dot is closest. If the vectors in the source and goal language would be perfectly aligned the word represented by the yellow dot would be the proposed translation.

Another assumption is that the most similar words to the word to translate are the same in the source and goal language, and more importantly that their average is closest to the translation in the goal language. This could go wrong when the word to translate is near, but outside, a cluster of words. The most similar words are then found in this cluster. If one takes the average of these words this would most likely be closest to another word inside of the cluster and not closest to the goal word, which is outside of the cluster, and thus result in an incorrect translation. A 2D-example of how this can go wrong can be found in figure 3.

3.2 Data

Both models only use data derived from Wikipedia. The data consists of all articles on the Dutch Wikipedia, all articles on the English Wikipedia and the inter-language links between the Dutch and English Wikipedia articles.

The articles were first parsed from wiki-markup to plain text and a list of links. For this the Wikipedia Extractor [1] was modified and used. The Wikipedia Extractor removes templates, tables, categories, enumerations, images, headers and all wiki-markup. A list of links is preserved separately from the plain text by the Wikipedia Extractor. An example of how the article Baker Bridge looks before and after parsing can be found in figures 4 and 5.

```
{{Infobox NRHP
  ... }}
‘‘‘Baker Bridge’’’, also known as Huntingdon County Bridge No. 14,
is a historic [[reinforced concrete]] closed spandrel [[arch bridge]]
spanning the [[Great Trough Creek]] and located at [[Todd Township,
Huntingdon County, Pennsylvania|Todd Township]], [[Huntingdon County,
Pennsylvania]].  It was built in 1917, and measures
{{convert|114|ft|m|adj=mid|-long}} and has a
{{convert|17|ft|m|adj=mid|-wide}} bridge deck.
It has two arch spans.<ref name="arch">{{cite web| url =
https://www.dot7.state.pa.us/ce/SelectWelcome.asp| title =...}}</ref>

It was added to the [[National Register of Historic Places]] in 1990.
<ref name="nris"/>

==References==
{{reflist}}
...
[[Category:Bridges completed in 1917]]
...
```

Figure 4: Baker Bridge before wiki-markup has been removed (some lengthy parts replaced with "..." by me) (article is CC-BY-SA-3.0 by Pubdog).

```
Baker Bridge.\nBaker Bridge, also known as Huntingdon County Bridge
No. 14, is a historic reinforced concrete closed spandrel arch
bridge spanning the Great Trough Creek and located at Todd Township,
Huntingdon County, Pennsylvania. It was built in 1917, and measures
and has a bridge deck. It has two arch spans.\nIt was added to the
National Register of Historic Places in 1990.
```

Figure 5: Baker Bridge after wiki-markup has been removed by the Wikipedia Extractor. A list of links and anchor texts is also provided by the Wikipedia Extractor.

The Wikipedia Extractor provides the cleanly parsed article text and a list of links and anchor texts. The first step after the Wikipedia Extractor was to reintroduce the anchor texts, using underscores, as one word. The next step was to split the articles into sentences using a sentence splitter and tokenizer. An example of how these steps affect the article Baker Bridge can be seen in figure 6. Finally each array representing a sentence was used to train the word embeddings with.

The word embeddings have 70 dimensions in the vector space. The default of gensim word2vec implementation is 100 dimensions, but due to long training times I decided to use a slightly lower value. For the Dutch word embeddings, words have to occur at least 10 times to be used. For the English word embeddings, words have to occur at least 20 times to be used. I chose to use different values because of the different sizes of the corpora. The two times higher minimum for the English Wikipedia still resulted in more words being included (there are 1,071,977 English word embeddings and 465,125 Dutch word embeddings). The dimensions and the minimum amount of times a word has to occur results in reasonable training times (circa 12 and 36 hours for the Dutch and English word embeddings respectively). The settings also result in reasonable performance on a number of test queries.

```
[’baker’, ’bridge’]
[’baker’, ’bridge’, ’also’, ’known’, ’as’, ’huntingdon’, ’county’
, ’bridge’, ’no’]
[’14’, ’is’, ’a’, ’historic’, ’reinforced_concrete’, ’closed’, ’spandrel’
, ’arch_bridge’, ’spanning’, ’the’, ’great_trough_creek’, ’and’
, ’located’, ’at’, ’todd_township’, ’huntingdon_county_pennsylvania’]
[’it’, ’was’, ’built’, ’in’, ’1917’, ’and’, ’measures’, ’and’
, ’has’, ’a’, ’bridge’, ’deck’]
[’it’, ’has’, ’two’, ’arch’, ’spans’]
[’it’, ’was’, ’added’, ’to’, ’the’, ’national_register_of_historic_places’
, ’in’, ’1990’]
```

Figure 6: Baker Bridge as word arrays after sentence splitting and tokenization. These arrays are used as input to train the word embeddings on.

To substantiate the cut-off value of 10 words for the Dutch corpus and 20 words for the English corpus a list of randomly selected words at the cut-off value can be seen in table 1. In the English corpus quite some terms are named entities which are used as multi-word anchor texts. The Dutch examples are mainly rather specific words in a specific form and named entities or parts of named entities.

Dutch	English
witkopmaki	consciousness-based
contractonderzoek	zta
alcalá-zamora	film_capacitors
boé	homer_township
bloederziekte	rock_action
ontologieën	geoffrey_wheatcroft
ipsilateraal	reiling
reinartz	navy_army_an_air_force_institutes
kosjere	cían
toesi	andreatta
ineu	lothar_von_arnauld_de_la_perière
gelijkgekleurde	matsuyama_castle
oreias	65000km

Table 1: 13 examples of words at the cut-off value. This means for the Dutch words that they occur 10 times in the corpus and for the English words that they occur 20 times.

3.3 Experiments

To determine the performance of both models they are tested by translating words from Dutch to English. For the experiments I took 4 sets of 100 random words within 4 different frequency ranges in the collection of Dutch words. 100 words from the 1,000 most used words; 100 words from the 1,000-10,000 most used; 100 words from 10,000-100,000 most used and 100 words from anything outside the 100,000 most used words. As explained in the data section only words with more than 10 occurrences are included in the word embeddings in Dutch, therefore only words with more than 10 occurrences are considered during sampling.

Both models propose translations from Dutch to English for all of these 400 words. The inter-language model only proposes one translation per word. The word embeddings model provides a ranked list of proposed translations. I chose to limit the number of proposed words to 10. Because of the different amount of proposed translations, different methods of evaluation are better for the different approaches. Therefore I decided to use both the mean reciprocal rank (MRR) and the percentage of correctly translated words or hit rate at 1. The MRR is the average of the reciprocal ranks of the individual results. The MRR=$\frac{1}{R}\sum_{i=1}^{R}\frac{1}{rank_i}$, where R is the number of results and $rank_i$ is the position of the correct translation for translation i. For all proposed translations I verified whether they were correct or incorrect. Translations to a different form, but with the same lemma, are considered to be correct. For example translating the word "boeken" ("books") to "book" would be considered a correct translation.

The word embeddings from the Dutch Wikipedia are trained on words which are used at least 10 times. The English word embeddings are trained on words which are used at least 20 times. These settings result in a reasonable training time (circa 12 and 36 hours). The settings also result in reasonable performance on a number of test queries.

Besides discussing the performance I will also discuss some individual examples and the assumptions made for the word embeddings model.

4 Results and discussion

method	above 1,000	1,000- 10,000	10,000- 100,000	below 100,000
inter-language	0.55	0.55	0.50	0.30
word embeddings N=3	0.20	0.19	0.027	0.025
word embeddings N=10	0.26	0.17	0.062	0.021
word embeddings N=100	0.17	0.15	0.038	0.014

Table 2: MRR of the different methods. "N" indicates the number of similar words (translated via the inter-language method) which is used to obtain the translation.

method	above 1,000	1,000- 10,000	10,000- 100,000	below 100,000
inter-language	55%	55%	50%	30%
word embeddings N=3	16%	15%	1%	2%
word embeddings N=10	22%	13%	4%	1%
word embeddings N=100	14%	11%	3%	1%

Table 3: hit rate@1 of the different methods. "N" indicates the number of similar words (translated via the inter-language method) which is used to obtain the translation.

In table 2 and 3 the performance of the inter-language and the word embeddings method can be seen. Table 2 shows the MRR, this performance score is not particularly interesting for the inter-language method, as the inter-language method only results in 1 suggested translation. To compare the inter-language method with the word embeddings method the hit rate at 1 was also calculated. For the word embeddings different amounts of most similar words ("N") were used to propose translations.

The performance of the inter-language method ranges from 30% to 55% correct translations. The performance of the word embeddings method ranges from 1% to 22% correct translations, with MRR scores between 0.014 and 0.26. For all methods the performance on the most frequent words (above 1,000) is the best. However for all methods, especially for the inter-language method, the performance on the test set of 1,000-10,000 most frequent word is almost as good. From 10,000 downwards the performance of the word embeddings method decreases quickly. The word embeddings method performs better when using 10 similar words than when using 3 or 100 similar words.

The performance of both algorithms is rather poor when compared to other methods. For example Google Translate, which is based on statistical machine translation [8], and uses parallel UN corpora [7]. While Google Translate still has issues on a sentence level [7, 9], on a word level its performance is almost perfect, at least when its performance is measured on proposing a, not the, correct translation. Similar performance of up to 96-99% correct word translations were already reported by Gale & Church and Kay & Röscheisen in 1993 using parallel corpora [2, 3]. In a 1999 paper by Rapp 72% correct word translations have been reported using comparable corpora [10].

It is important to note that the sample of 400 words is randomly sampled, over all terms which occur more than 10 times, within the corpus per frequency range. The effect of this is that quite some words are named entities. There also some abbreviations (eg. “nb” and “tpa”) and unexpected terms (eg. “200910”) such as some German (eg. “illustrierte”) and English terms (eg. “october”) in the Dutch corpus, mainly in the lower frequency ranges. In the higher frequency ranges the number of years is quite high. There are 14 years (eg. 1980 and 1945), from these 14 years all 14 are correctly translated using the inter-language method. More important however is that with 11 correct translations of years using the word embeddings method, with N=10, the impact of the years is quite high on the total of 40 correct translations for the word embeddings method.

4.1 Inter-language method

frequency range	no Dutch article	no interwiki	no English links	proposed translation	correct
above 1,000	20%	5%	5%	70%	55%
1,000-10,000	12%	8%	2%	78 %	55%
10,000-100,000	22%	15%	9%	54%	50%
below 100,000	42%	21%	4 %	33%	30%
average	24%	12%	5%	59%	48%

Table 4: For the sample of 400 words it is shown in which step of the inter-language method the translation fails or whether a (correct) translation is proposed.

Given the rather poor performance of both the inter-language method and the word embeddings method it is interesting to investigate where the issues lie with both methods. As the inter-language method consists of three steps I looked at the percentage of translations which failed in each step. As can be seen in table 4 24% of the words is not linked to any article and thus fails in the first step. For 12% of the words none of the Dutch articles the word links to has an inter-language link to an English article. Finally for 5% of the words the English article does not have any links going to it with a frequency of above 20 (and thus in the corpus). What remains is 59% of the words for which a translation is proposed which results in a final score of 48% correct translations.

Notable is that the difference between the number of proposed translations and the number of correct translations is much higher for both the top 1,000 and 10,000 frequent word compared with the words outside of the 10,000 most frequent words. For the top 10,000 words the percentage of correct translations, given a translation is proposed, is 75%, for the words outside of the 10,000 most frequent words the percentage of correct translations is 91%. Thus, while the performance of the inter-language method is worse for less frequent words, the accuracy of the proposed translations by the inter-language method is higher for the less frequent words. This could be caused by the fact that the less frequent terms are often non-ambiguous and more specific. While the less frequent terms might not always have an article related to them, non-ambiguous and specific words are likely more often linked to a directly related article. More general and terms are used more often, and even if there is no article on the general term, it might be incorrectly linked to an article which does not specifically handle the term. Which causes incorrect translations to be proposed.

An example of the above issue is the translation of the word “enige” (“only”) which is used 59,895 times in the Dutch Wikipedia. The word “enige” is used twice in a link, once as a link to “Monopoly” (“Monopoly”) and once as a link to “Monetheïsme” (“Monotheism”). Both of these articles do not describe the term “only”, however, they are related (describing an only god and only competitor). The term is finally incorrectly translated to "monotheism". Another example is the word "officieel" ("officially") with 19,254 occurrences, which is incorrectly translated to the related term "official language".

4.2 word embeddings method

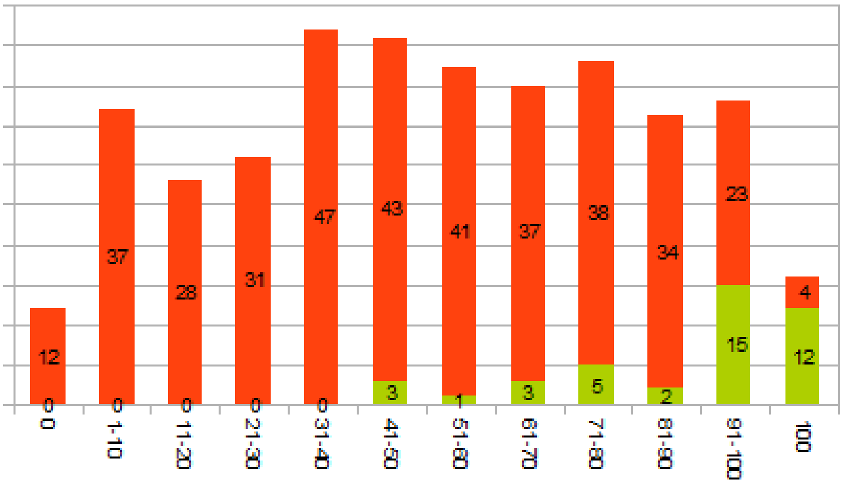

Figure 7: Effect of the number of translatable similar words on the number of correct translations for the word embeddings method. On the x-axis the number of similar words which could be translated using the inter-language method is shown (in ranges with an exception for 0 and 100). For each amount of translatable similar words the number of correct (green) and incorrect (orange) translations for the word embeddings method with N=100 is shown.

For the word embeddings method two assumptions about similarity across languages were made. The first assumption was that a word and its translation have exactly the same function in the two languages. The fact that a lot of words are semantically ambiguous is opposing this assumption [4]. An example which illustrates the issues with ambiguous terms is "soort" which translates to "kind" and "species". Because of this the closest words which are: "wetenschappelijke" ("scientific"), "geldig" ("valid"), "gepubliceerd" ("published"), "naam" ("name"), "eerst" ("first") and "operastuk" ("opera piece") seem quite unrelated, especially "opera piece". Due to the very low performance (when years are not included) I suspect that there might also be other causes which harms performance.

The word embeddings method uses the inter-language method to translate the most similar words. As the inter-language method only translates 59% of the words and only 48% correctly a lot of the translations in the intermediate step are not present or incorrect. The second assumption for the word embeddings method was that the closest words in both the source and goal language are the same. The word embeddings method can only use 48-59% of the most similar words. This could be one of the main causes of the poor performance.

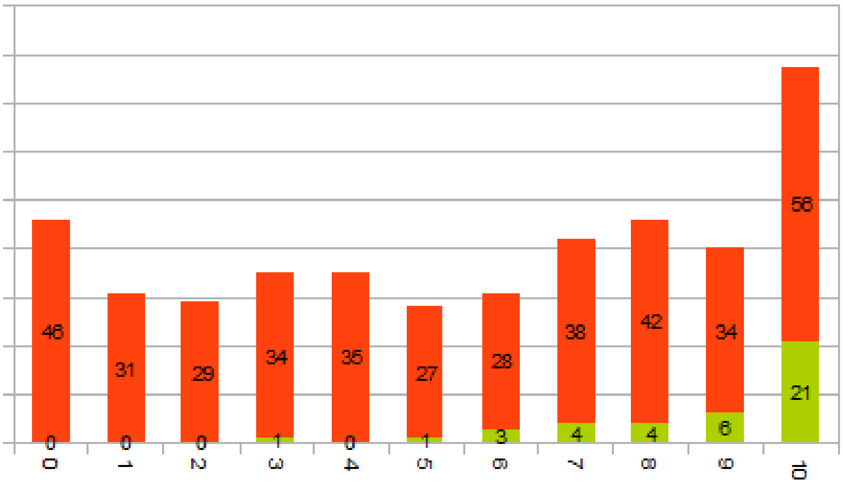

Figure 8: Effect of the number of translatable similar words on the number of correct translations for the word embeddings method. On the x-axis the number of similar words which could be translated using the inter-language method is shown. For each amount of translatable similar words the number of correct (green) and incorrect (orange) translations for the word embeddings method with N=10 is shown.

To investigate this I looked into the number of translatable similar words, by the inter-language method, and how this effects correct translations. The results can be seen in figure 7, 8 and 9. As can be seen the amount of similar terms which is translatable is pretty evenly distributed. In both figure 7 and 8 we can see that the number of successful translations when less than half of the similar words (first 6 bars, 0-5 and 0-50) can be translated is rather poor. With only 2 and 3 words successfully translated out of 200 and 198 words respectively. For all three cases it can clearly be seen that when all similar words can be translated the performance gets quite a boost with 75%, 38% and 19% words correctly translated using 100, 10 and 3 similar words respectively. This might however be slightly deceiving as for the 16 terms for which 100 of the most similar terms could be translated, 12 were years, of which 11 correctly translated and one incorrectly. The only non-year which got correctly translated, and had 100% of the related terms translated, is a named entity, “Alexander III”.

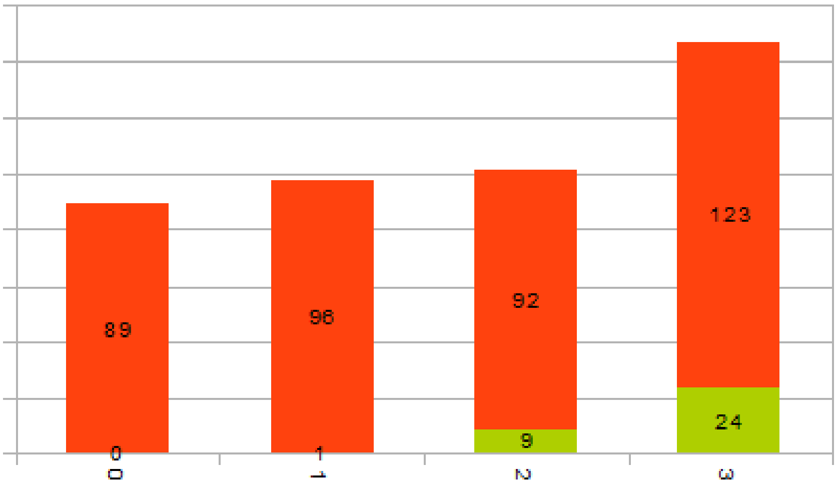

Figure 9: Effect of the number of translatable similar words on the number of correct translations for the word embeddings method. On the x-axis the number of similar words which could be translated using the inter-language method is shown. For each amount of translatable similar words the number of correct (green) and incorrect (orange) translations for the word embeddings method with N=3 is shown.

5 Conclusion

One of the main goals of this thesis was to investigate whether the inter-language structures on Wikipedia can be used for translation. The results of the inter-language method show that using this method it is possible to propose a translation for 59% of the terms. And that those proposed translations are correct 81% of the time. Compared to other methods, which propose correct translations for more than 96% of the terms, this performance is rather poor. It does however show that a simple method, purely using the inter-language and link structures on Wikipedia, can obtain reasonable performance.

In this thesis the inter-language method also served a secondary goal in providing the word embeddings method with seed translations, in the form of translations for the most similar words.

The second main goal of this thesis was to investigate whether good word translation performance can be achieved using two monolingual corpora. To investigate this the word embeddings method was designed and tested. The results of the word embeddings method show that only 10% of the words can be translated successfully. With a large part of the correct translations being years, which are quite easy to translate, the method does not seem to be particularly robust for untranslatable or incorrectly translated similar words. It is also doubtful whether the assumption that taking the average vector of the most similar words which are closest to the source word results in a vector closest to the source word holds. A final issue with the current method is that it requires all similar words to be translatable. This means that a large amount of correct translations is already required to be able to translate new words.

5.1 Future research

Using unconnected monolingual corpora in combination with word embeddings is the main new idea proposed in this thesis. As such most opportunities I see for related future research relate to the word embeddings method.

One of the main advantages of using two monolingual corpora is that one can use huge amounts of text in both the source and goal language. In this thesis this was not fully taken advantage of, as only texts on Wikipedia were used. Using more texts, and different or more natural texts, would be a good idea when further developing translation methods based on unconnected monolingual corpora. Another improvement could be to use another seed lexicon, one which has almost perfect word translation performance.

One of the main issues with the word embeddings method is that the assumptions of the similarity of the two languages are quite important. A more robust similarity measure is likely the main thing which has to be improved upon. Another option is to make use of the one-to-one assumption, which holds that if one translates a word to the goal language and then tries to translate this word back to the source language it should hold the original source word. Using this to test proposed translations would likely result in improved results. This notion could also be used for the similarity measure. Where in the source language one could test if based on the similar words the source word is found, if this is not the case in the source language it would likely also not be the case in the goal language.

6 Acknowledgements

I would like to thank my supervisor, Dr. Suzan Verberne for her valuable advice and suggestions during my thesis project.

References

[1] Wikipedia Extractor. `http://medialab.di.unipi.it/wiki/Wikipedia_Extractor`. Accessed: 2016-04.

[2] William A Gale and Kenneth W Church. A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):75–102, 1993.

[3] Martin Kay and Martin Röscheisen. Text-translation alignment. *Computational linguistics*, 19(1):121–142, 1993.

[4] Robert Krovetz and W Bruce Croft. Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems (TOIS)*, 10(2):115–141, 1992.

[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[6] 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.

[7] Franz Josef Och. Statistical machine translation: Foundations and recent advances. *Tutorial at MT Summit*, 2005.

[8] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51, 2003.

[9] Sumant Patil and Patrick Davies. Use of google translate in medical communication: evaluation of accuracy. *BMJ*, 349:g7392, 2014.

[10] Reinhard Rapp. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519–526. Association for Computational Linguistics, 1999.

[11] Ivan Vulić and Marie-Francine Moens. Detecting highly confident word translations from comparable corpora without any prior knowledge. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 449–459. Association for Computational Linguistics, 2012.

[12] List of Wikipedias. `https://meta.wikimedia.org/wiki/List_of_Wikipedias`. Accessed: 2016-06-17.