Connecting Multilingual Wordnets: Strategies for Improving ILI Classification in OdeNet

Johann Bergh
Lingolutions
Munich, Germany.
johann@lingolutions.com

Melanie Siegel
Darmstadt University
of Applied Science
melanie.siegel@h-da.de

Abstract

The Open Multilingual Wordnet (OMW) is an open source project that was launched with the goal to make it easy to use wordnets in multiple languages without having to pay expensive proprietary licensing costs. As OMW evolved, the interlingual indicator (ILI)[1] was used to allow semantically equivalent synsets in different languages to be linked to each other. OdeNet[2] is the German language wordnet which forms part of the OMW project. This paper analyses the shortcomings of the initial ILI classification in OdeNet and the consequent methods used to improve this classification.

1 Introduction

A wordnet is a lexical database of semantic relationships between words in a specific language. The first wordnet was created for the English language at Princeton University (also known as the Princeton WordNet, (Fellbaum, 1998)). As the usefulness of wordnets as lexical resources became apparent, the Princeton WordNet (PWN) was expanded and some wordnets were constructed from scratch in other languages.

The Princeton WordNet is distributed in electronic format as part of NLTK (Natural Language Processing Toolkit) and can be accessed with a corresponding Python library[3]. NLTK offers translations for synsets (groupings of synonyms) in various languages, although these translations are incomplete; meaning that not every synset in English has an equivalent translation in another language. There are also wordnets in other languages which were developed completely independent of the PWN, such as GermaNet (Hamp et al., 1997). Many of these wordnets contain high quality data which were constructed manually in a resource-intensive and time-consuming manner. Therefore, these wordnets are commercially licensed and not free to use, except for research and teaching. An example of a large Wordnet built independently from PWN and available on open-source licence is plWordNet (Piasecki et al., 2009; Dziob et al., 2019).

OMW is an open source project that was launched with the goal to make it easy to use wordnets in multiple languages with cc-by-sa-4.0 open-source licenses that include commercial and private use (Bond and Foster, 2013). OMW has the added benefit of connecting equivalent synsets in different languages by means of the ILI (Fellbaum and Vossen, 2008; Bond et al., 2016). The English version of OMW called EWN (McCrae et al., 2020) is basically a copy of the PWN with some enhancements and additions, most notably the addition of an ILI for each synset. Many of the OMW wordnets in other languages were developed by using the already existing translations in NLTK. These translations were extracted and packaged into new wordnets. Consequently, the equivalent synsets in the resulting wordnets were linked to each other via the ILI. Goodman and Bond (2021) developed the WN Python library that can be used to access the wordnets that form part of the OMW project. In Listing 1 we see how the translated lemmas of a synset in PWN can be accessed with NLTK Python library. Listing 2 on the other hand shows how to access these same synsets through the ILI or by searching directly for it in the other language.

Listing 1: Get French translation for EWN synset in NLTK

```
from nltk.corpus import wordnet as wn

s = wn.synsets('dog')
s[0].lemma_names()
['dog', 'domestic_dog', 'Canis_familiaris']

s[0].lemma_names('fra')
['chien', 'canis_familiaris']
```

[1] The next version was called CILI (Collaborative Interlingual Index), https://www.luismc.com/omw/ili

[2] https://github.com/hdaSprachtechnologie/odenet

[3] https://www.nltk.org

Listing 2: Get French synset via ILI or directly in WN

```
import wn
s = wn.synsets('dog')
s[0].lemmas()
['dog', 'Canis familiaris', 'domestic dog']

#Get equivalent French synset via ILI
ili = s[0].ili.id
s = wn.synsets(ili=ili, lang='fr')
s[0].lemmas()
['chien', 'canis familiaris']

#Search for French synset directly
s = wn.synsets('chien', lang='fr')
s[8].lemmas()
['chien', 'canis familiaris']
```

Though NLTK offers translations in many languages, German is so far not included. This means that a German wordnet for OMW could not easily be constructed with the exiting NLTK translations as a base, as was the case with many of the other languages. Therefore, an initiative was launched to create an open source German wordnet (OdeNet) which could form part of the OMW project. OdeNet was constructed from open source linguistic resources in combination with some manual and semi-manual corrections. Since OdeNet was constructed independently of existing resources in NLTK, it was not as easy to connect equivalent synsets in OMW via ILI. As an initial implementation, Google Translate[4] was used in combination with statistical methods as described by Siegel and Bond (2021). However, this implementation has some shortcomings, including:

- incorrect ILI classification for some synsets from a semantic perspective
- duplicate assignment of ILI's to multiple synsets
- Part of Speech (POS) for some ILI's is inconsistent between EWN and OdeNet

This paper describes these shortcomings and proposes solutions for improved ILI and POS classification in OdeNet.

2 Problem Description

A significant problem in using machine translation to connect equivalent synsets in different languages occurs, when translating homographs (words with similar spelling but different meanings) and polysemes. This is particularly noticeable when a word translated from a source language is a homograph or polyseme in the target language. As an example, we take the German word `Unterlegscheibe` from OdeNet. The corresponding English translation is `washer`. Searching for `washer` in EWN, we find three synsets containing the word:

- Name: `washer`
  EWN ID: `ewn-10788571-n`
  ILI: `i94042`
  Definition: `someone who washes things for a living`
- Name: `washer`
  EWN ID: `ewn-04562157-n`
  ILI: `i60971`
  Definition: `seal consisting of a flat disk placed to prevent leakage`
- Name: `washer`
  EWN ID: `ewn-04561970-n`
  ILI: `i60970`
  Definition: `a home appliance for washing clothes and linens automatically`

Our aim is to select the correct synset in EWN so that we can take the corresponding ILI and assign it to the synset in German. For somebody with knowledge of German, it is evident that the second synset in the list is the correct corresponding synset in EWN (i.e. we want to take the ILI from this synset and also use it in the corresponding OdeNet synset). It is difficult to do this assignment automatically, because of the missing context.

The usage of machine translation with Google Translate together with some statistical methods in the current OdeNet implementation (Siegel and Bond, 2021) also resulted in many of the synsets having duplicate ILIs, because the assignment of ILIs to synsets in OdeNet was not restricted to one ILI per synset. An example: The synsets `odenet-4330-n (Anzahl, Zahl)` and `odenet-688-n (Summe, Gesamtmenge)` both referred to `i35594 (measure, amount, quantity)`. Furthermore, Siegel and Bond (2021) used automatic methods for assigning the correct POS to synsets. However, they were only able to assign the correct POS to synsets in 93% of the cases. Often, multi-word lexemes were involved in problematic cases, as for example `postmortal, nach dem Tod, post mortem` was categorized as `pos "n"`, although it is `pos "a"`.

[4] https://translate.google.de

3 Proposed Solution

3.1 Basic Approach

Figure 1 depicts the complete algorithm for correcting ILI classification in OdeNet.

All synsets in EWN have a short, concise definition in the `Definition` field. We propose to use this definition to get more context for the disambiguation. First, we combine the word in the synset and the definition with a semicolon and do machine translations with DeepL[5]. Then, we extract the translated word from the machine translation and look for a corresponding match in OdeNet. These are the results for the `washer` example:

- EWN ID: `ewn-10788571-n`
  ILI: `i94042`
  Word-Definition combination:
  `washer: someone who washes things for a living`
  Machine translation:
  `Wäscher: jemand, der beruflich Dinge wäscht`

- EWN ID: `ewn-04562157-n`
  ILI: `i60971`
  Word-Definition combination:
  `washer: seal consisting of a flat disk placed to prevent leakage`
  Machine translation:
  `Unterlegscheibe: Dichtung, die aus einer flachen Scheibe besteht, um ein Auslaufen zu verhindern`

- EWN ID: `ewn-04561970-n`
  ILI: `i60970`
  Word-Definition combination:
  `washer: a home appliance for washing clothes and linens automatically`
  Machine translation:
  `Waschmaschine: ein Haushaltsgerät zum automatischen Waschen von Kleidung und Wäsche`

As is clearly evident, the machine translation of the second item now enables us to make the correct ILI classification (i60971) for the corresponding OdeNet synset.

3.2 Dealing with Ambiguity: ILI Classification Weight

Although we have obtained success with this simplified example, our aim is to construct a system whereby ILI classification for all synsets in OdeNet is possible. In order to achieve this, there are additional scenarios of ambiguity that we have to take into consideration:

Even with the context-based machine translation as described above, we could still find more than one possible candidate in OdeNet for the ILI of the synset we are evaluating in EWN. For example, consider the EWN synset with ILI `i66412`:

- ILI: `i66412`
  Word-Definition combination: `depth: the intellectual ability to penetrate deeply into ideas`
  Machine translation:
  `Tiefe: die intellektuelle Fähigkeit, tief in Ideen einzudringen`

If we now search for the translated lemma `Tiefe` in OdeNet, we will find three synsets (`odenet-847-n: ['Tiefe', 'Tiefsinn']; odenet-6615-n ['Abgrund', 'Tiefe', 'Schlund', 'Hölle']`, `odenet-16328-n ['Tiefe', 'Teufe']`). Which OdeNet synset do we assign the ILI to? Intellectually, this should be `odenet-847-n`, but this cannot be automatically decided.

More than one EWN synset can match a single OdeNet synset. For example, consider the Word-Definition combinations and translations of the EWN synsets with ILIs `i6124` and `i68929` below:

- ILI: `i6124`
  Word-Definition combination:
  `ethic: the principles of right and wrong that are accepted by an individual or a social group`
  Machine translation:
  `Ethik: die Grundsätze des Richtigen und Falschen, die von einem Individuum oder einer sozialen Gruppe akzeptiert werden`

- ILI: `i68929`
  Word-Definition combination:
  `ethics: the philosophical study of moral values and rules`
  Machine translation:
  `Ethik: das philosophische Studium der moralischen Werte und Regeln`

For both of the lemmas in the respective EWN synsets, the translated lemma in German is `Ethik` which is found in the OdeNet synsets `odenet-10-n ['Sittlichkeit', 'Wertvorstellungen',`

[5] https://www.deepl.com (After manual translation quality assessment, we chose DeepL for our implementation as it performed better on context-based translations than Google Translate)

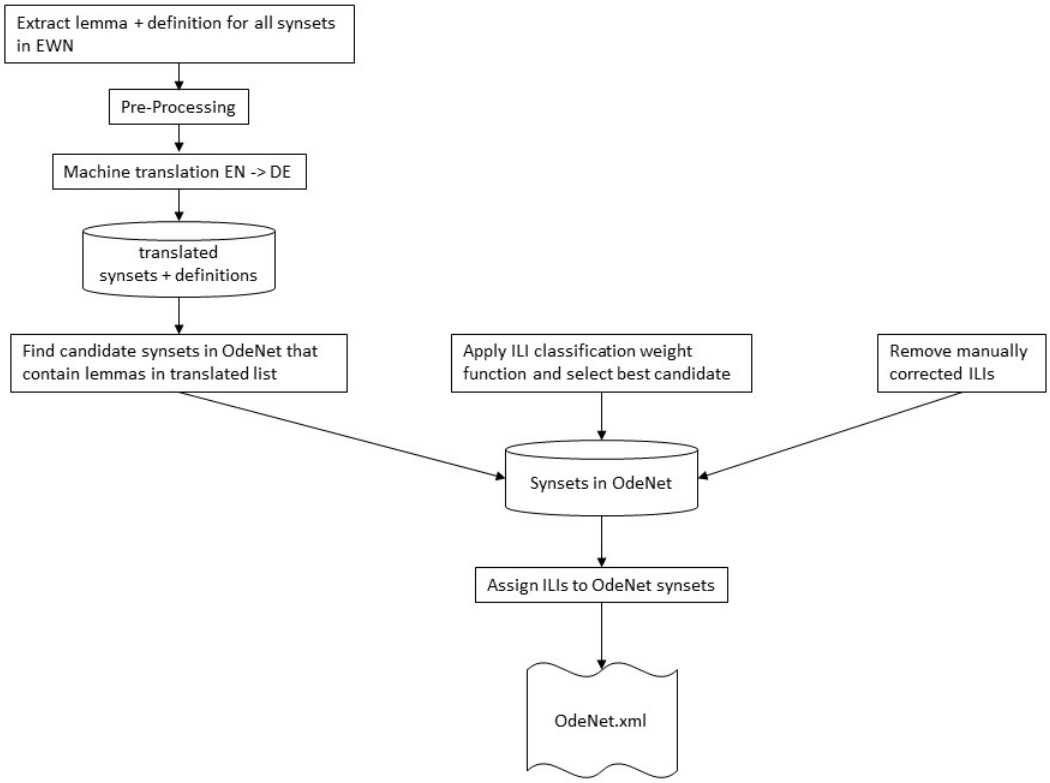

Figure 1: ILI classification

`'Wertmaßstäbe', 'Wertesystem', 'Moral', 'Moralvorstellungen', 'Ethik', 'sittliche Werte', 'moralische Werte']` and `odenet-4879-n ['Ethik', 'Morallehre', 'Sittenlehre', 'Tugendlehre']`. Which one of the EWN synsets' ILI do we assign to which one of the OdeNet synsets?

Since there could be multiple candidates in OdeNet synsets for ILI's in EWN synsets, it is necessary to write a classification function to assign weights to each of the candidates, so that the most optimal assignment can be made. Fortunately, OdeNet is very synonym rich (much more so than other wordnets), and we can use these synonyms in combination with a German spaCy[6] Word2Vec model to do the classification.

$$f(v_1, v_2) = \frac{\sum_i \sum_j dist(v_{1i}, v_{2j})}{|v_1| \times |v_2|} \quad (1)$$

First, we extract the Definition part of the translated Lemma and Definition translation. The content words in this translation are added to a vector (v_1). Only adjectives, adverbs, nouns and verbs are used. Function words, such as prepositions and articles, are discarded. Similarly, all the synonyms (lemmas in the candidate synset) are added to a vector (v_2). For each value in v_1 and v_2 a similarity value is computed. These values are summed and normalised to a value between 0 and 1, which is the weighted value for the candidate synset in OdeNet *competing* for the ILI in a specific EWN synset.

3.3 Optimising Machine Translation for POS by Pre-Processing

In English, there are many nouns and verbs that have the same spelling, such as `search`. Our idea is to use preprocessing in order to obtain better results from machine translation.

Experiments with DeepL machine translation indicated that translation results from English to German for verbs improve when adding `to` in front of the verb. In cases, where we have English nouns and verbs with the same spelling, it also helps the machine translation to distinguish the POS correctly. An example is the word `search`. In the case, where the synset refers to the verb

[6] https://spacy.io/

search, the machine translation performs better when adjusting the word to its infinitive form to search, and is also more likely to translate it as a verb in the target language. The EWN synset with ILI i28263 refers to the verb search. The Word-Definition combination is:

search: try to locate or discover, or try to establish the existence of

Pre-processing changes this to:

to search: try to locate or discover, or try to establish the existence of

Post-processing adjustments were also necessary in some instances for the machine-translated German text. For verbs, the machine translation added the word zu in front of the verb in some cases, as a result of the addition of to in front of the English verbs. Consequently, we removed zu from the translated text as a post-processing cleanup task, if the POS was a verb.

3.4 Correct POS classification in OdeNet

Siegel and Bond (2021) reported that the POS classification for the initial implementation of OdeNet was at 93.3%, with errors occurring mostly in cases where the lemma was a multi-word lexeme, which made correct POS classification difficult by automatic means. The data gathered in the table of translations can be leveraged to address this issue.

For each synset in OdeNet, we extract the first lemma of the synset. We then retrieve all records in the table of translations, where the first lemma from the synset is equal to the translated target lemma. If the POS of the lemma's synset is not equal to any POS's of the relevant records retrieved in the table, then there could be a POS misclassification in the OdeNet synset, since it would be reasonable to assume that the POS of the EWN synset translated to German should also have the same POS in the target language.

4 Results

Table 1 depicts the state of OdeNet, before and after the algorithm has been applied. It can be seen that there were 13,818 synsets with unique ILIs. Further, there were 5,965 synsets with duplicate ILIs; meaning that one unique ILI is assigned to more than one synset in OdeNet. The total number of unique duplicate ILIs were 3,703; meaning that on average, a duplicate ILI was assigned to 1.61 synsets.

The most noticeable difference after applying the algorithm is the complete elimination of duplicate ILIs. The number of synsets with unique ILIs has increased to 19,547 and all duplicate ILIs have been removed. The algorithm identified 361 synsets with possible POS errors. After manual evaluation, 325 of these synsets indeed ended up having a wrong POS. This means a successful identification of 90% of synsets with the wrong POS. Of the 36 false positives, most proposed an adjective for a noun or a verb and in many cases colloquial language was involved, such as in the case of odenet-19938-n (Tüftelei, Getüftel).

5 Concluding remarks

OdeNet is an open-source wordnet that was automatically compiled from an open thesaurus and connected to the multilingual wordnets in the OMW initiative by machine-translating synsets. The result of the machine translation was partly incorrect because the translation context was missing. Further, duplicate interlingual indicators (ILIs) were assigned in OdeNet. Additionally, there was a need to correct the automatically assigned POS.

In this paper, we described a solution for these problems by matching ILIs to OdeNet synsets, taking the English definitions into account. The results have shown that the algorithm is very effective in reducing duplication and improving the correctness of ILI classification.

The algorithm can potentially be improved by providing the ILI classification weight function, as described in section 3.2, with more context information. At the moment, we use synonyms to provide context, and these synonyms could be augmented with the hypernyms of the *candidate* synsets under evaluation. This should lead to higher classification accuracy, but is left for future research.

Although this algorithm was applied to improve the ILI classification for OdeNet, it can be used for any other language in theory. The success of the resulting classification will be dependent on factors such as how synonym-rich the language is and also how good the machine translation support is.

With some minor adjustments to the algorithm, we propose that it will also be possible to connect other lexical resources using the proposed method. For example, two thesauri, developed in two differ-

	EWN	OdeNet (before)	OdeNet (after)
Synsets	120053	36159	36159
Synsets with unique ILIs	117480	13818	19547
Synsets without ILIs	2573	16376	16612
Synsets with duplicate ILIs	0	5965	0
Duplicate ILIs	0	3703	0

Table 1: OdeNet after applying proposed algorithm

ent languages independently of each other, could be merged into a bilingual resource.

Since languages evolve independently of each other, it often happens that not all words in one language have a perfect equivalent in another language. It can happen that some semantic meaning is lost or added in the translation process. Even though you will mostly get the best possible match by applying an algorithm such as described in this paper, there can still be an extent of *fuzziness* or loss/addition of meaning. Currently, the OMW framework is modelled in such as way that a synset in one language can map to a single synset in another language via the ILI. This structure makes it difficult to model fuzzy matching or loss/addition of semantic meaning. This topic may be of interest for future research.

References

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1352–1362, Sofia.

Francis Bond, Piek Vossen, John P McCrae, and Christiane Fellbaum. 2016. CILI: The Collaborative Interlingual Index. In *Proceedings of the Global WordNet Conference*, volume 2016.

Agnieszka Dziob, Maciej Piasecki, and Ewa Rudnicka. 2019. plwordnet 4.1-a linguistically motivated, corpus-based bilingual resource. In *Proceedings of the 10th Global Wordnet Conference*, pages 353–362.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Christiane Fellbaum and Piek Vossen. 2008. Challenges for a global Wordnet. In *Online Proceedings of the First International Workshop on Global Interoperability for Language Resources*, pages 75–82.

Michael Wayne Goodman and Francis Bond. 2021. Intrinsically interlingual: The Wn Python library for wordnets. In *11th International Global Wordnet Conference (GWC2021)*.

Birgit Hamp, Helmut Feldweg, et al. 1997. GermaNet-a lexical-semantic net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.

John Philip McCrae, Alexandre Rademaker, Ewa Rudnicka, and Francis Bond. 2020. English wordnet 2020: Improving and extending a wordnet for english using an open-source methodology. In *proceedings of the LREC 2020 workshop on multimodal WordNets (MMW2020)*, pages 14–19.

Maciej Piasecki, Bernd Broda, and Stanislaw Szpakowicz. 2009. *A wordnet from the ground up*. Oficyna Wydawnicza Politechniki Wrocławskiej Wrocław.

Melanie Siegel and Francis Bond. 2021. Compiling a German wordnet from other resources. In *11th International Global Wordnet Conference (GWC2021)*.