Enriching a terminology for under-resourced languages using knowledge graphs

John P. McCrae[1], Atul Kr. Ojha[1], Bharathi Raja Chakravarthi[1], Ian Kelly[2], Patricia Buffini[2], Grace Tang[3], Eric Paquin[3], Manuel Locria[3]

[1]ADAPT Centre, Data Science Institute, NUI Galway, Ireland
[2] ADAPT Centre, Dublin City University, Ireland
[3] Translators without Borders
E-mail: john@mccr.ae, {atulkumar.ojha,bharathiraja.asokachakravarthi}@nuigalway.ie,
ian.anthony.kelly@gmail.com, patricia.buffini@adaptcentre.ie,
{grace,ericpaquin,manuel}@translatorswithoutborders.org

Abstract

Translated terminology for severely under-resourced languages is a vital tool for aid workers working in humanitarian crises. However there are generally no lexical resources that can be used for this purpose. Translators without Borders (TWB) is a non-profit whose goal is to help get vital information, including developing lexical resources for aid workers. In order to help with the resource construction, TWB has worked with the ADAPT Centre to develop tools to help with the development of their resources for crisis response. In particular, we have enriched these resources by linking with open lexical resources such as WordNet and Wikidata as well as the derivation of a novel extended corpus. In particular, this work has focused on the development of resources for languages useful for aid workers working with Rohingya refugees, namely, Rohingya, Chittagonian, Bengali and Burmese. These languages are all under-resourced and for Rohingya and Chittagonian there are only very limited major lexical resources available. For these languages, we have constructed some of the first corpora resources that will allow automatic construction of lexical resources. We have also used the Naisc tool for monolingual dictionary linking in order to connect the existing English parts of the lexical resources with information from WordNet and Wikidata and this has provided a wealth of extra information including images, alternative definitions, translations (in Bengali, Burmese and other languages) as well as many related terms that may guide TWB linguists and terminologists in the process of extending their resources. We have presented these results in an interface allowing the lexicographers to browse through the results extracted from the external resources and select those that they wish to include in their resource. We present results on the quality of the linking inferred by the Naisc system as well as qualitative analysis of the effectiveness of the tool in the development of the TWB glossaries.

Keywords: under-resourced languages; terminology; linking; natural language processing; knowledge graphs

1. Introduction

Terminology is a vital tool for aid workers in a wide range of crisis situations and the availability of a good-quality terminology in local languages is of vital importance. However, often these are severely under-resourced and so the development of language resources for these languages is significantly complicated. For example, after a devastating earthquake in Haiti in 2012, the natural language processing community rapidly developed tools and resources for the main language of Haiti, Haitian Creole, to help with the aid effort (Lewis, 2010). As such, the development of language resources for under-resourced languages is of critical importance and this is one of the main goals of the non-profit organisation, Translators without Borders (TWB).

The use of natural language processing technologies and existing open resources is a potentially huge benefit for the development of lexical resources for under-resourced languages, and, with this objective, we created a collaboration between the ADAPT Centre and TWB to develop tools to enrich the existing terminologies. For this collaboration, we focused on the work related to the Rohingya refugee crisis and as such the languages of relevance to this population, namely, Bengali, Burmese, Rohingya and Chittagonian. These languages vary in the availability of resources to being under-resourced languages but have significant online presence, namely Bengali

and Burmese, which have large resources such as Wikipedia and support from language technologies such as Google Translate, to Rohingya and Chittagonian, which have nearly no resources or language tool support. Our strategy for expanding these resources was first to increase the corpus resources available for these languages so that we can train natural language processing tools on them. Secondly, we looked at linking them with open resources including WordNet (McCrae et al., 2020; Miller, 1995) and Wikidata so that extra information such as semantic relations, images and translations can easily be added into the glossaries. We examined some techniques for automatically finding candidates from these open resources using the Naisc (McCrae & Buitelaar, 2018) framework. We then have built this into a tool that allows terminologists to validate the data coming into the resources from external sources and thus semi-automatically extend this resource.

The rest of this paper is structured as follows, in Section 2 we lay out some related work and then we present the use case from Translators without Borders in Section 3. We then look at how we constructed the extended corpus in Section 4 and how we linked the existing glossaries with terms from open resources in Section 5. Finally, we show how we built a prototype for semi-automatic enrichment of the glossaries in Section 6 and finish with a conclusion in Section 7.

2. Related Work

As discussed in Section 1, unlike low-resourced languages, such as Bengali and Burmese, high-resourced languages, such as English and French are endowed with ample lexical and other linguistic resources such as WordNet, translated terminologies, corpora, and crowd-sourced resources such as Wikipedia or Wikidata.

Princeton WordNet (Miller, 1995) was the first WordNet which also formed the base for versions in all the other languages. Non-English languages gained focus in 1996 when EuroWordNet (Vossen, 1997) was founded to develop WordNets for several European languages giving way to a multilingual database.

When it comes to Asian WordNets the efforts started late, but significant milestones have been reached. In Asia, Indo-WordNet (Bhattacharyya, 2010) is a huge effort that was built in India to incorporate the major official Indian languages used in the Indian sub-continent, including Bengali. These languages were taken from three language families Indo-Aryan, Dravidian and Sino-Tibetan (Chakravarthi et al., 2018; Bhattacharyya, 2010). A few years ago, the University of Bangladesh (Rahit et al., 2018) also built the Bengali WordNet. Burmese WordNet[1] was developed on Open Multilingual WordNet (Bond & Paik, 2012; Bond & Foster, 2013). EuroWordNet, Indo-WordNet, Burmese and the recent Bangladeshi Bengali WordNet were built using an expand approach. However, Rohingya and Chittagonian do not have a WordNet or any lexical resources. While some effort has been made in the direct translation of WordNets into under-resourced languages (Chakravarthi et al., 2019), the results are still of poor overall quality. Similarly, some work has been done on the automatic development of terminologies for under-resourced languages (Pinnis et al., 2012; McCrae & Doyle, 2019).

Out of the various WordNets, Bengali and Burmese have large text corpora which can be scraped from Wikipedia, CURL (collecting Web Pages for Under-Resourced

[1] https://wordnet.burmese.sg/

Language) (Goldhahn et al., 2016) and An Crúbadán (Scannell, 2007). To the best of our knowledge, Rohingya and Chittagonian do not have any other existing corpora.

3. Use Case

The effectiveness of any aid program depends on delivering the correct information in the correct language. Historically, humanitarian agencies and aid workers have focused on maintaining capacity in major or "world" languages such as English, Spanish, and French. While these may constitute the "official" language of an affected country, they are often not used or well-comprehended by the affected populations. Furthermore, in humanitarian response, field workers must communicate important, sometimes life-saving information to those in need. In many cases, the critical link to ensuring affected people understand is the interpreter. However, too often, that link is broken, either because concepts do not translate well into the target language or because the interpreter does not have the tools to understand the concepts clearly.

TWB is addressing this problem by focusing on under-resourced local languages commonly used by marginalised populations in humanitarian crises. TWB's Glossaries, a critical real-time translation tool, assists front-line aid workers with an online repository of vetted, translated, simplified, and localised emergency-related terminology. It enables interpreters, cultural mediators, and any other field workers to access key concepts, terms, and phrases commonly used in crisis response. Themes include protection; housing, land, and property rights; and water, sanitation and hygiene (WASH). They were developed in collaboration with technical specialists and language partners.

TWB partnered with ADAPT to strengthen and expand TWB Glossaries, specifically the Bangladesh use case through a semantic uplift of the tools. ADAPT is a national research centre in Ireland focused on the digital media technology hosted at Trinity College Dublin and including seven other partner universities in Ireland. The main goal of the partnership was to increase the number of terms available in our glossaries and the discoverability of associated terms. We also used the collaboration to enhance the user experience and explorablity of the glossary content and the functionality e.g., keyword search, linked term review and approval, and search.

4. Corpus Building

We collected corpora from various sources for the target languages. Our target languages are from Bangladesh namely, Bangla (Bengali) (ISO 639-3 ben), Burmese (ISO 639-3 mya), Chittagonian (ISO 639-3 ctg) and Rohingya (ISO 639-3 rhg). All these languages are low-resourced languages.

Bengali is an Indo-European language spoken in Bangladesh and the West Bengal state of India and other places. Bengali is an agglutinative language and there are more than 150 different inflected forms of single verb root in Bengali. Presently, there are several dialects of Bengali that vary mainly in terms of the verb inflections and intonation. For this project, we downloaded the data for the Bangladesh version of Bengali language.

The Burmese language belongs to the Sino-Tibetan language family, it is the largest non-Chinese language from that Sino-Tibetan language family. It is the official language

of the Republic Union of Myanmar and the native language of the Bamar people. The Myanmar script is an abugida system. It consists of 33 characters of standalone consonants, four dependent consonants, and tens of diacritic marks that represent vowels and tones. The orthography of Myanmar is generally syllable - based, although syllables may be merged in special writing forms. One word can be composed of multiple syllables and one syllable can be composed of multiple characters.

Chittagonian is an Indo-European language mainly spoken in the Chittagong Division in Bangladesh Country. Its sister languages include Sylheti, Rohingya, Chakma, Assamese, and Bengali. It is derived through an Eastern Middle Indo-Aryan from Old Indo-Aryan, and ultimately from Proto-Indo-European. Historically Arabic script was used for writing systems. The Bengali script is the most common script used nowadays.

Rohingya is also an Indo-European language spoken by Rohingya people of Rakhine State. The Hanifi Rohingya script is a unified script for the Rohingya language. Rohingya was first written in the 19th century with a version of the Perso-Arabic script.

We downloaded the data from CURL (Collecting Web Pages for Under-Resourced Languages) and WikiDump for Bengali and Burmese languages. For Chittagonia and Rohingya, there were no corpora available in CURL and WikiDump. However, we managed to collect the corpus for Rohingya from Rohingya Poems in Rohingyalish (Basu, 2014), Qur'an Foóila Síarah (Quran translation in Rohingya) and Rohingya Language (Mohammed & Ahmed, n.d.) books. After gathering the data, we cleaned the collected corpora following these steps:

- Removed HTML/file tags, metadata information, non-UTF/illegal characters, etc.
- Split it into one sentence per line
- Removed extra spaces and blank lines
- Removed duplicate sentences

We were able to collect 1,207,285 and 1,883 sentences for the Bengali and Burmese languages, respectively, from CURL. From WikiDump, 1,243,811 and 710,122 sentences were collected for Bengali and Burmese, respectively. From OPUS (http://opus.nlpl.eu/), we collected 681,789 sentences for Bengali and 962,654 sentences for Burmese. A total of 7,177 Rohingya sentences were extracted from books, while 5,100 Chittagonian sentences were extracted from Bible.is, Facebook and YouTube. Details of the of the corpus statistics are presented below:

Language	**Total sentences**	**Total words**
Bengali/Bangla	3,139,915	36,340,082
Burmese	1,674,659	21,568,615
Rohingya	7,177	206,089
Chittagonian	5,100	28,313

5. Linking

5.1 Objectives

The main goal of this project is to enrich the terminologies developed by TWB with the data found in resources such as WordNet and Wikidata. In order to do this, we need to

establish which of the entities in these resources correspond to the terms found in these resources. This is not a trivial task as there are a large number of potential matches in general domain resources such as WordNet and Wiktionary, so it is not clear which of these resources would be a suitable match for which term. For example, the TWB glossary has highly generic terms such as 'cut', which is defined as "to injure a part of your body with something sharp that cuts the skin." Similarly, WordNet has 41 verb senses for the word 'cut' and Wikidata has eight pages whose main label in 'cut'. For WordNet, the most appropriate sense for 'cut' has the definition of "penetrate injuriously", which is quite distinct from the definition give in the TWB terminologies. For Wikidata, none of the main definitions labelled as 'cut' are appropriate and the best match would actually be the page for 'laceration', it should also be noted that a complexity here is that as Wikidata is an encyclopaedic resource, all the concepts are nominal and so any link between these senses necessarily crosses part-of-speech boundaries. However, establishing a linking in a fully manual matter is likely very time-consuming and could be further helped by means of automatic linking tools.

In order to support automatic linking of tools, we have developed a toolkit called Naisc (McCrae & Buitelaar, 2018)[2], which acts as a toolkit for linking resources. This toolkit is designed for general purpose linking of datasets and is highly configurable, such that it can be used for a wide variety of linking tasks. In particular, we have focused a lot of work on the development of this tool for dictionary linking in the context of ELEXIS (Krek et al., 2018) infrastructure, which is developing a new infrastructure for electronic lexicography. As part of this infrastructure we envisage the development of a single large, interlinked matrix of dictionaries, which we refer to as the Dictionary Matrix. A key enabling technology for this is obviously automatic dictionary linking technology, and this is where the contribution of Naisc plays a key contribution to the ELEXIS infrastructure. As such, we have developed specialised modules for dictionary linking in Naisc, that we can also take advantage of for linking the TWB glossaries with WordNet and Wikidata.

5.2 Methodology

Naisc is a pipeline of processes which analyse two input datasets and outputs the set of links between them. This is done in a series of steps that analyse the datasets and find the best link between the elements of these datasets. The first step in this process is the **blocking step**, in which we find all potential matches between the two datasets, and as such the output of this step is superset of the final output, i.e., we can only output links that are identified at this step. As with all steps in Naisc, there are a number of different implementations that can be applied here, however in this case we restricted ourselves to only finding the elements in the target dataset (WordNet or Wikidata) for which we have a matching label. This means that we cannot find links such as 'cut' to 'laceration' described above. More exhaustive blocking strategies could be applied to find such links, however this can be computationally very expensive and lead to a large number of false positive results, so we did not attempt this here. The second step is called the **lens** step, where we analyse the input data in order to find text from each of the datasets that can be compared. In the case of this linking task, this step is fairly trivial as we only

[2] 'Naisc' is pronounced 'nashk' and means 'links' in Irish, the software is open source and available at https://github.com/insight-centre/naisc

extract the definitions from both datasets, but it is easy to see how further information from a dictionary, such as examples or etymological information, could also be extracted and compared. The next step is then the **text feature** step, where we apply natural language processing techniques in order to estimate the similarity of the two pieces of texts. We have several methods implemented for this within the Naisc framework, but in the context of this paper we experimented with two sets of features, firstly a set of text similarity metrics based on surface characteristics, that is referred to as the 'basic string' features of Naisc. These are defined as follows:

Longest common subsequence This measures the largest number of consecutive characters in both strings.

Longest common prefix/suffix The number of characters that these two string share from the start/end of the strings.

Jaccard/Dice/Containment We measure the n-grams in each string in terms of both word n-grams and character n-grams and compare them using the standard methods of Jaccard, Dice and containment as defined below:
$\text{Jaccard} = \frac{|A \cap B|}{|A \cup B|}, \text{Dice} = \frac{2|AB|}{|A|+|B|}, \text{Containment} = \frac{|A \cap B|}{\min(|A|,|B|)}$

Sentence Length Ratio The relative length in words of the two inputs. This is symmetrised using the following formula:
$\text{SLR}(s,t) = 1 - \frac{min(|s|,|t|)}{max(|s|,|t|)}$

Average Word Length Ratio A comparison of the length of the words, symmetrized as above.

Negation A Boolean feature checking for the presence of negation keywords (such as 'not') in both or neither description.

Number Another Boolean feature comparing if all mentioned numbers match.

Jaro-Winkler, Levenshtein String similarity measures based on the edit distance between the strings as implemented by Apache Commons Text.

Monge-Elkan This metric (Monge & Elkan, 1997) uses Jaro-Winkler or Levenshtein as the base similarity function *sim* and is defined as:
$\text{ME}(s,t) = \frac{1}{|s|} \sum_{i=1}^{|s|} \max_{j=1,\ldots t} sim(s_i, t_j)$

In addition, we use the Sentence-BERT model introduced by Reimers & Gurevych (2019), which produces a single vector to represent each of the definitions, we simply take the cosine of these vectors in order to estimate the similarity of the two sentences and this is used as a single feature.

The next step in the Naisc processing extracts features in parallel with the previous two steps and is referred to as the **graph feature** step. Both Wikidata and WordNet are complex graphs with many relations between the elements so we can take advantage of this to ensure that we are linking semantically similar terms. The TWB dataset did not have any links between its terms, however it did group these terms into domains and we created a graph over the TWB dataset by means of linking each term to a pseudo-node for the domain. In this way, we constructed a graph over the TWB dataset and this allows us to compare the graphs using a link prediction methodology. In particular, we take advantage of non-ambiguous nodes within the graph, that is terms which have a single sense in WordNet or in Wikidata, and use these to link the two graphs together. This creates a single graph over both the TWB data as well as the target dataset. We then applied the node proximity metric called *personalised page rank (PPR)* (Page et al., 1999)

to score the likelihood of two terms being linked and in particular we used the FastPPR implementation (Lofgren et al., 2014).

The penultimate step of the algorithm starts with the prediction of the probability of a particular link by means of a **scorer**, which combines the features extracted from both the textual and graph analysis and converts them into a single score. We have explored two methods for this in the current work. Firstly an unsupervised methodology that works by means of micro-ranking the features. In particular, this feature works as follows: the values for each of the features are extracted and these are all ranked. Then we translate each feature value to its relative rank, such that, for example if a feature is the 100th highest value out of 1,000 values returned we would normalise its score to $1 - \frac{100}{1000} = 0.9$. Then we output the final score for each pair as the average of its normalised features. In addition, we used a supervised method, which is a support vector machine (Vapnik, 2000) as implemented by LibSVM (Chang & Lin, 2011).

The final step of the process, **matching**, is to find the most likely link between the terms in TWB and the target datasets. In this case, this is as simple as finding the highest scoring result for each element, however this would be substantially more complex if we also attempted to find multiple non-exact links, such as broader/narrower links. This is an active area of research, but our results in this task are not yet of a high enough quality to be reliable.

5.3 Evaluation

TWB Dataset	Target Dataset	Method	Precision	Recall	F-Measure
COVID	WordNet	Basic Unsupervised	82.30%	85.32%	83.78%
COVID	WordNet	Basic Supervised	79.65%	82.57%	81.08%
COVID	WordNet	BERT Unsupervised	87.61%	90.83%	89.19%
COVID	WordNet	BERT Supervised	87.61%	90.83%	89.19%
COVID	Wikidata	Basic Unsupervised	68.22%	84.88%	75.64%
COVID	Wikidata	Basic Supervised	68.22%	84.88%	75.64%
COVID	Wikidata	BERT Unsupervised	71.70%	88.37%	79.17%
COVID	Wikidata	BERT Supervised	71.70%	88.37%	79.17%
Bangladesh	WordNet	Basic Unsupervised	90.50%	90.53%	90.51%
Bangladesh	WordNet	Basic Supervised	81.05%	81.05%	81.05%
Bangladesh	WordNet	BERT Unsupervised	75.76%	75.79%	75.77%
Bangladesh	WordNet	BERT Supervised	75.76%	75.59%	75.67%
Bangladesh	Wikidata	Basic Unsupervised	84.79%	85.90%	85.34%
Bangladesh	Wikidata	Basic Supervised	70.01%	83.30%	76.08%
Bangladesh	Wikidata	BERT Unsupervised	76.42%	91.03%	83.09%
Bangladesh	Wikidata	BERT Supervised	76.42%	91.03%	83.09%
Average	Average	Unsupervised	81.45%	86.66%	83.82%
Average	Average	Supervised	74.73%	82.95%	78.46%
Average	Average	BERT	77.87%	86.51%	81.80%
Average	Average	BERT + Supervised	77.87%	86.46%	81.78%

Table 1: The results of the linking quality between the two datasets

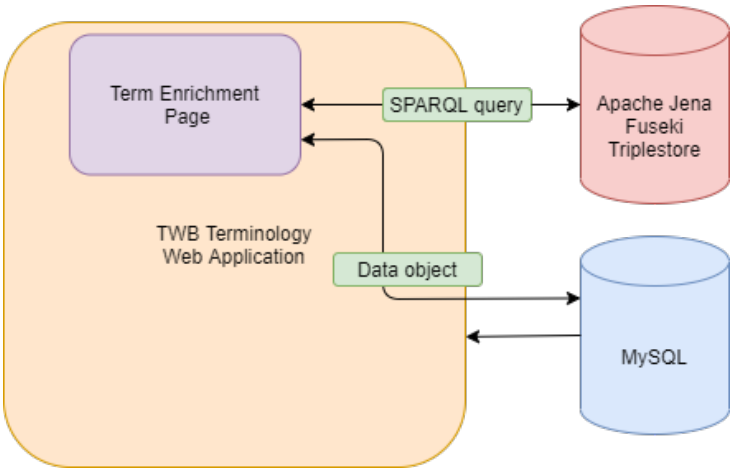

Figure 1: An overview of the enrichment system of the TWB terminology application

In order to evaluate the results of the linking we manually corrected some of the results of the Naisc linking in order to establish a partial gold standard linking. We then applied this to the four linking tasks which were based on the combination of TWB glossaries (on COVID and on Bangladesh) with the two target datasets (WordNet and Wikidata). We also tried four settings, based on whether we were using the 'basic' textual features of the BERT analysis and whether we were using the 'unsupervised' micro-ranking methodology or the 'supervised' SVM methodology; the results are presented in Table 1. Overall the results with all settings are quite strong with nearly four fifths of the links being correct automatically. Perhaps surprisingly the strongest overall system is the 'basic unsupervised' method. This is actually in line with our previous experience, where we have found that the supervised methodology does not fit well with the matching maximisation step, as it tends to predict probabilities that are close to zero or one, whereas the unsupervised method gives a good overall score to each element. Secondly, the use of Sentence-BERT while effective was not fine-tuned to the task and would have had challenges handling the short (and highly variable) nature of textual definitions. It is likely that further experiments could improve these results.

6. Terminology Enrichment

The Naisc linking output is a collection of Resource Description Framework (RDF) data in Turtle or N-Triple format files. These files were uploaded to a Jena Fuseki triplestore. The terminology enrichment (see Figure 1) was implemented into the existing Translators without Borders (TWB) terminology web application. An enrichment page is created for each term in the glossary and is constructed using the dynamic SPARQL Protocol and RDF Query Language (SPARQL) queries to the triplestore based on the ID of each term (see Figure 2).

Due to the open source nature of both Wikidata and WordNet, the results for each term may differ and as such the enrichment page needs to facilitate dynamic results. The SPARQL results are parsed and a page element is built for each returned data object. The Wikidata and WordNet results are separated and broken into sections based on result categories for visual clarity and ease of search. As an example of some of the extra information that would be available through this linking we take the example of the term 'vaccination'. The extra information is as follows:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT * WHERE {
  <http://www.wikidata.org/entity/Q38933> rdfs:label ?label .
}
LIMIT 10
```

QUERY RESULTS

Table | Raw Response

Showing 1 to 10 of 10 entries

	label
1	"fever"@en
2	"pyrexia"@en
3	"febrile response"@en
4	"חום"@he
5	"Bezgak"@uz
6	"Sukar"@eu
7	"gorączka"@pl
8	"Rupharíy"@qu
9	"fiebre"@es

Figure 2: An example SPARQL database query for labels for a certain wikidata term.

- **From Corpora**
  - Examples found from the corpora developed in Section 4.
- **From WordNet**

  Alternative Definition : "taking a vaccine as a precaution against contracting a disease"

  Alternative Terms : inoculation

  Related Terms : immunization, immunisation, immunize, immunise, inoculate, vaccinate
- **From Wikidata**

  Alternative Definition : "administration of a vaccine to protect against disease"

  Alternative Terms : *(none for 'vaccination', example for 'treatment')* medical treatment, therapeutics, treating, intervention, therapy

  Related Terms : treatment, active immunotherapy, active immunity, antibody injection, vaccine, injection

  Translations : 'Impfung' *(German)*, 'vaccination' *(French)*, 'vacsaíniú' *(Irish)*, ... (about 100 languages)

  Images :  [3]

  Wikipedia Link : https://en.wikipedia.org/wiki/Vaccination (and other languages)

The processed data is then stored in a MySQL database for page load persistence and for use in the TWB glossaries. Each term can be included or excluded from the database using an accompanying slider, and includes additional sliders to allow for bulk inclusion and exclusion, indicating that a term has been reviewed, or that the data is mismatched. In the event of the linking generating incorrect term results, the matched slider allows

[3] Public domain image from https://en.wikipedia.org/wiki/File:Typhoid_inoculation2.jpg

for flagging of incorrect terms on the terminology term list page. Similarly, the reviewed slider allows for flagging that the term has been manually reviewed. A screenshot of the application is shown in Figure 3.

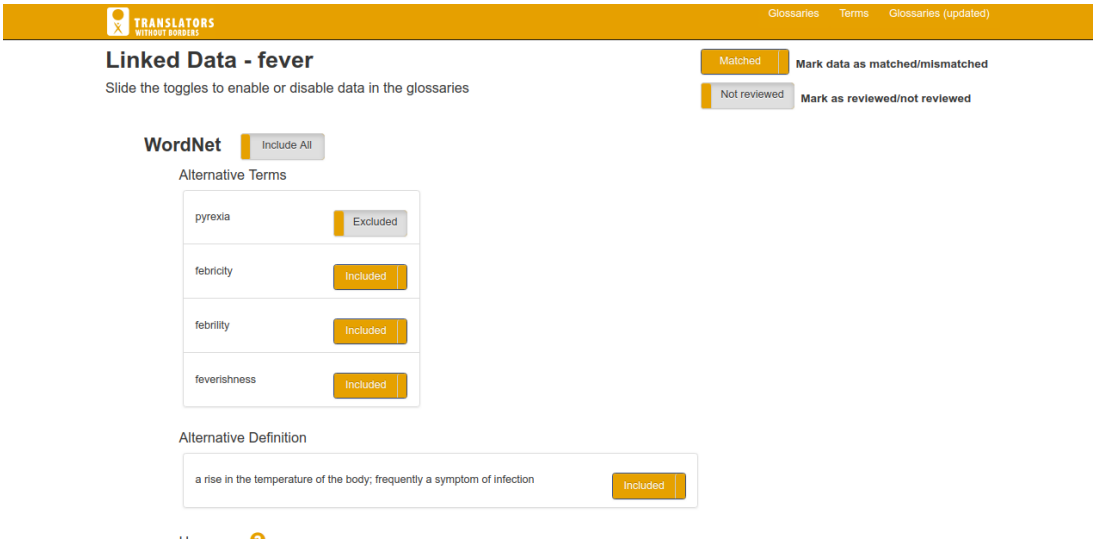

Figure 3: An enrichment page for the term fever showing alternative terms as well as all sliders.

7. Conclusion

We have looked at how we can use terminological resources in order to extend a glossary of terms that are used by front-line aid workers. We examined the use case and saw how we could use open resources in order to improve the data that is available in the glossaries. We first looked at how we can compile a corpus to support these terms and found methods of finding corpus information from social media and other sources that were effective even though the languages were not well-documented. Then, we showed how we could link to Wikidata and WordNet and how to apply the Naisc framework to develop high-quality linking. We experimented with the use of machine learning and deep learning techniques here, but found that the main issues were related to finding suitable candidates in the open resources. We then developed this into a glossary tool that can be used to enrich the terminology and examined some of the extra kinds of data that can be added as the result of this analysis.

Acknowledgements

This work was supported by Science Foundation Ireland and co-funded by the European Regional Development Fund through the ADAPT Centre for Digital Content Technology [grant number 13/RC/2106].

8. References

Basu, E.M.S. (2014). *Rohingya Poems In Rohingyalish.* n.p.

Bhattacharyya, P. (2010). IndoWordNet. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (eds.) *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23*

May 2010, Valletta, Malta. European Language Resources Association. URL http://www.lrec-conf.org/proceedings/lrec2010/summaries/939.html.

Bond, F. & Foster, R. (2013). Linking and Extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*. The Association for Computer Linguistics, pp. 1352–1362. URL https://www.aclweb.org/anthology/P13-1133/.

Bond, F. & Paik, K. (2012). A Survey of WordNets and their Licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*. Matsue. 64–71.

Chakravarthi, B.R., Arcan, M. & McCrae, J.P. (2018). Improving Wordnets for Under-Resourced Languages Using Machine Translation. In *Proceedings of the 9th Global Wordnet Conference*. Nanyang Technological University (NTU), Singapore: Global Wordnet Association, pp. 77–86. URL https://www.aclweb.org/anthology/2018.gwc-1.10.

Chakravarthi, B.R., Arcan, M. & McCrae, J.P. (2019). WordNet Gloss Translation for Under-resourced Languages using Multilingual Neural Machine Translation. In *Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation*. Dublin, Ireland: European Association for Machine Translation, pp. 1–7. URL https://www.aclweb.org/anthology/W19-7101.

Chang, C. & Lin, C. (2011). LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3), pp. 27:1–27:27. URL https://doi.org/10.1145/1961189.1961199.

Goldhahn, D., Sumalvico, M. & Quasthoff, U. (2016). Corpus collection for under-resourced languages with more than one million speakers. *Proc. of Collaboration and Computing for UnderResourced Languages: Towards an Alliance for Digital Language Diversity (CCURL)*, pp. 67–73.

Krek, S., McCrae, J., Kosem, I., Wissek, T., Tiberius, C., Navigli, R. & Pedersen, B.S. (2018). European Lexicographic Infrastructure (ELEXIS). In *Proceedings of the XVIII EURALEX International Congress on Lexicography in Global Contexts*. pp. 881–892. URL http://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202018/118-4-2986-1-10-20180820.pdf.

Lewis, W. (2010). Haitian Creole: How to Build and Ship an MT Engine from Scratch in 4 days, 17 hours, & 30 minutes. In F. Yvon & V. Hansen (eds.) *Proceedings of the 14th Annual conference of the European Association for Machine Translation, EAMT 2010, Saint Raphaël, France, May 27-28, 2010*. European Association for Machine Translation. URL https://www.aclweb.org/anthology/2010.eamt-1.37/.

Lofgren, P., Banerjee, S., Goel, A. & Comandur, S. (2014). FAST-PPR: scaling personalized pagerank estimation for large graphs. In S.A. Macskassy, C. Perlich, J. Leskovec, W. Wang & R. Ghani (eds.) *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*. ACM, pp. 1436–1445. URL https://doi.org/10.1145/2623330.2623745.

McCrae, J.P. & Buitelaar, P. (2018). Linking Datasets Using Semantic Textual Similarity. *Cybernetics and Information Technologies*, 18(1), pp. 109–123. URL http://www.cit.iit.bas.bg/CIT_2018/v-18-1/10_paper.pdf.

McCrae, J.P. & Doyle, A. (2019). Adapting Term Recognition to an Under-Resourced Language: the Case of Irish. In *Proceedings of the Celtic Language Technology Workshop*. Dublin, Ireland: European Association for Machine Translation, pp. 48–57. URL https://www.aclweb.org/anthology/W19-6907.

McCrae, J.P., Rademaker, A., Rudnicka, E. & Bond, F. (2020). English WordNet 2020: Improving and Extending a WordNet for English using an Open-Source Methodology. In *Proceedings of the Multimodal Wordnets Workshop at LREC 2020*. pp. 14–19. URL https://lrec2020.lrec-conf.org/media/proceedings/Workshops/Books/MMW2020book.pdf#page=20.

Miller, G.A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11), pp. 39–41.

Mohammed, M. & Ahmed, R.M. (n.d.). *Rohingya Language Text Book 3*. n.p.

Monge, A.E. & Elkan, C. (1997). An Efficient Domain-Independent Algorithm for Detecting Approximately Duplicate Database Records. In *Workshop on Research Issues on Data Mining and Knowledge Discovery, DMKD 1997 in cooperation with ACM SIGMOD'97, Tucson, Arizona, USA, May 11, 1997*.

Page, L., Brin, S., Motwani, R. & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Pinnis, M., Ljubešic, N., Stefanescu, D., Skadina, I., Tadic, M. & Gornostay, T. (2012). Term extraction, tagging, and mapping tools for under-resourced languages. In *Proceedings of the 10th Conference on Terminology and Knowledge Engineering (TKE 2012), June*. pp. 20–21.

Rahit, K.T.H., Hasan, K.T., Al-Amin, M. & Ahmed, Z. (2018). BanglaNet: Towards a WordNet for Bengali Language. In *Proceedings of the 9th Global Wordnet Conference*. pp. 1–9.

Reimers, N. & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In K. Inui, J. Jiang, V. Ng & X. Wan (eds.) *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. Association for Computational Linguistics, pp. 3980–3990. URL https://doi.org/10.18653/v1/D19-1410.

Scannell, K.P. (2007). The Crúbadán Project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, volume 4. pp. 5–15.

Vapnik, V.N. (2000). *The Nature of Statistical Learning Theory, Second Edition*. Statistics for Engineering and Information Science. Springer.

Vossen, P. (1997). EuroWordNet: a multilingual database for information retrieval. In *Proceedings of the DELOS workshop on Cross-language Information Retrieval, March 5-7, 1997 Zurich*. Vrije Universiteit.